University of Salford MANCHESTER

Towards Forecasting and Prediction of Faults in Electricity Distribution Network: A Novel Data Mining & Machine Learning Approach

By

Charith Silva

School of Science, Engineering & Environment

The University of Salford

Supervised By: Prof. Mo Saraee

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

7th September 2020

Acknowledgements

The work in this thesis is based on research carried out within the School of Science, Engineering & Environment at the University of Salford, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification, and it is all my own work unless referenced to the contrary in the text. Some reference is made to the industry that is confidential and commercially restricted.

There are many people to whom I am indebted for support and assistance in various ways. Without them, this would never have been completed, and it is appropriate that they should share in it. First and foremost, I would like to express my sincere gratitude to Prof. Mo Saraee for his supervision and invaluable suggestions. His invaluable comments on concepts, structures and organisation have greatly enhanced any value the thesis may have. It was an honour to work with him. In addition, I would like to thank Prof.David Parsons who supported me in the final part of my PhD with very helpful suggestions.

Also, I would like to thank Electricity North West Limited for providing support and resources for this research work. Last but not least, my sincere thanks must go to my wife, Sammy, for her love, encouragement and understanding. This thesis would not have been possible without her support.

Abstract

The electricity supply system includes a large-scale power generation installation and a convoluted network of electrical circuits that work together to efficiently and reliably supply electricity to consumers. Faults in the electricity distribution network have a direct effect on its stability, availability and maintenance. Consequently, quick elimination, prevention and avoidance of faults and the causes that generated them, is of special interest.

The possible opportunity to both analyse the distribution of faults and predict future failures that may arise can significantly help electricity distribution operators who are accountable for the detection and repair of such problems. Such information is also crucial for any future planning and design of electricity distribution networks as it would significantly help to prevent problematic areas or and identify any additional measures necessary for the protection of underground and overground cables and equipment. The derived information would also be very useful to avoid any potential penalties associated with future network faults imposed by the regulators.

Any network component faults result in an outage of power not only in the area fed by them but also in the neighbouring area. Fault prediction in distribution systems has always been of immense importance to utilities to ensure reliable power supply. This research aim is to develop data mining, and machine learning models to accurately predict and forecast Electricity Distribution Network Faults. The specific research objectives are to gain a deeper understanding of Electricity Distribution Network faults and to accurately predict network faults using the National Fault and Interruption Reporting Scheme (NAFIRS) database. Furthermore, this research not only proposes solutions but also provides an in-depth discussion of the associated technical, data gathering and data processing challenges.

This research employed multiple case research design, as this allows more opportunities for multiple experiments and cross observation. This research has proposed a new method that analyses historical fault data and seeks to understand the impact of faults with other factors such as the Main Equipment Involved, Component and Direct Cause. This proposed data mining model may be used to safeguard the electrical power distribution system’s key equipment which can be severely damaged by some upcoming faults.

The author of this thesis has proposed a new fault segmentation framework which distributed network operators can use to perform fault segmentation. This approach gives the option of performing multidimensional segmentation using various fault characteristics such as a number of faults, a number of minutes lost, and a number of customers affected. Multidimensional segmentation is a powerful conceptual model for the analysis of large and complex datasets.

This study provides an in-depth discussion of equipment failure related network faults and compares the performance of a range of forecasting methods with a variety of accuracy measures. The study also provides an in-depth analysis of visual data mining concepts and discusses how using 2D and 3D calendar heat map methods can help provide a relatively new perspective in evaluating temporal patterns in electricity distribution network faults.

Finally, the research discusses how external factors, such as local population density, affects electricity distribution network faults. Various classification algorithms were used to build prediction models. Those models were validated and compared for accuracy. The author has also sought to accurately understand the behaviour of Customer Minutes Lost (CML) performance indicators and sought to predict the annual CML figure using other annual financial and network performance indicators such as a number of customers affected, Totex, and Network load.

It is anticipated that the work presented within this thesis will to lead to several original contributions to the scientific community who are working with data mining, machine learning and electricity distribution networks.

Impact Statement

Throughout this study, the author has sought to develop data mining and machine learning models that can be used to understand, predict and forecast electricity distribution network faults, drawing upon research in various case studies. The outcomes of this research will support policy formulation, and network design not just within the electricity distribution industry, but across other utility distribution industries such as gas and water.

The research contributions presented in the thesis could be put to a beneficial use both inside and outside academia. Inside academia, the research could benefit the domains of fault prediction, fault segmentation, visual data mining and machine learning. Firstly, fault prediction and fault segmentation could use the proposed adaptations of machine learning methods to any kind of fault-related data in a utility distribution network such as Water or Gas. Secondly, the thesis develops a novel data analytic framework for data science projects that have already received interest from researchers in the data science domain. Having a good process for data mining and machine learning and clear guidelines is always a plus point for any data science project. It also helps to focus the required time and resources required early in the process to identify a clear definition of the business problem to be solved. Hence, the framework is proposed to aid the data science project lifecycle and bridge the gap with business needs and technical realities.

Outside academia, the thesis could benefit electricity, gas and water distribution network, operators. The thesis has been developed in close collaboration with one of the UK electricity distribution companies, who have provided data for this research. This new fault, prediction and segmentation model will provide opportunities to utilise the engineering staff resource in a more controlled manner and reduce call outs and overtime charges. Ultimately, this research will assist in improving the level of power system availability by helping to reduce network faults.

List of Abbreviations

2D - Two Dimensional

3D - Three Dimensional

ARIMA - Autor Regressive Integrated Moving Average

CAPEX - Capital Expenditure

CI - Customer Interruptions

CML - Customer Minutes Lost

DBSCAN - Density-Based Spatial Clustering of Application with Noise

DNO - Distribution Network Operators

HV - High Voltage

IEEE - Institute of Electrical and Electronics Engineers

IIS - Interruption Incentive scheme

IoT - Internet of Things

KDD - Knowledge Discovery

KPI - Key Performance Indicators

LHS - Left Hand Side

LV - Low Voltage

MAE - Mean Absolute Error

MAPE - Mean Absolute Percentage Error

MASE - Mean Absolute Scaled Error

ME - Mean Error

MEI - Main Equipment Involved

MPE - Mean Percentage Error

NA - Energy Networks Association

NaFIRS - National Fault and Interruption Reporting Scheme

OFGEM - Office of Gas and Electricity Markets

OPEX - Operating Expenditure

R2 - Coefficient of determination

RHS - Right Hand Side

RMSE - Root Mean Absolute Error

SARIMA - Seasonal Autor Regressive Integrated Moving Average

SQL - Structured Query Language

UK - United Kingdom

VDM - Visual Data Mining

WSS - Within the Sum of Squares

Declaration

I declare that I am the sole author of this thesis and no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university. Therefore, I confirm that the work described in this thesis is my own except for some sources that support our research, which is appropriately cited and indicated. Parts of the work presented in this thesis have appeared in the following publications.

7th September 2020

Charith Silva

List of Publications Resulted from This Thesis

This research has been presented and published in the preceding of several IEEE and ACM conferences, which are listed below:

Full Published Research Papers

- C. Silva and M. Saraee, "**Understanding Causes of Low Voltage (LV) Faults in Electricity Distribution Network Using Association Rule Mining and Text Clustering**," 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), Genova, Italy, 2019, pp. 1-6, doi: 10.1109/EEEIC.2019.8783949. (Appendix 2)
- C. Silva and M. Saraee, "**Electricity Distribution Network: Seasonality and the Dynamics of Equipment Failures Related Network Faults**," 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2020, pp. 1-6, doi: 10.1109/ASET48392.2020.9118274. (Appendix 2)
- C. Silva and M. Saraee, "**Predicting Average Annual Electricity Outage using Electricity Distribution Network Operator's Performance Indicators**," 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2020, pp. 1-6, doi: 10.1109/ASET48392.2020.9118383. (Appendix 2)
- C. Silva and M. Saraee, "**Understanding the Relationship Between Population Density and Low Voltage Faults Causes in Electricity Distribution Network**," 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2020, pp. 1-6, doi: 10.1109/ASET48392.2020.9118391. (Appendix 2)
- Mo Saraee and Charith Silva. 2018. **A new data science framework for analysing and mining geospatial big data**. In Proceedings of the International Conference on Geoinformatics and Data Analysis (ICGDA '18). Association for Computing Machinery, New York, NY, USA, 98–102. DOI:https://doi-org.salford.idm.oclc.org/10.1145/3220228.3220236. (Appendix 2)

Full Accepted Research Papers

- C. Silva and M. Saraee, " A New Electricity Distribution Network Fault Segmentation Framework" 2020 IEEE International Energy Conference (ENERGYCON), Tunisia.
- C. Silva and M. Saraee " Understanding Temporal Patterns in Electricity Distribution Network Faults Using 2D and 3D Time-Series Calendar Heatmaps" 2020 IEEE International Energy Conference (ENERGYCON), Tunisia.

Other Published Research Papers

- C. Silva, M. Saraee. "***Predicting road traffic accident severity using decision trees and time-series calendar heatmaps"***. 2019 IEEE Conference on Sustainable Utilization And Development In Engineering and Technology (CSUDET), Malaysia.
- C. Silva, M. Saraee and M. Saraee, "**Data Science in Public Mental Health: A New Analytic Framework**," 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain, 2019, pp. 1123-1128, doi: 10.1109/ISCC47284.2019.8969723. (Appendix 2)
- C. Silva, M. Saraee and M. Saraee, "***Predictive modelling in mental health: a data science approach"***. 2019 IEEE Conference on Sustainable Utilization And Development In Engineering and Technology (CSUDET), Malaysia.

Published Abstracts Only

- Silva, C (2016). ***Modelling Electricity Distribution Networks with Geospatial Big Data***. Proceedings of the CSE 2016 Annual PGR Symposium (CSE-PGSym 16).
  Link: https://usir.salford.ac.uk/id/eprint/39314/1/2016%20SPARC%20abstracts.pdf
- Silva, C (2018). ***Geospatial Big Data Analytics Framework for Electricity Distribution Networks***. SPARC 2018 Internationalisation and collaboration: Salford postgraduate annual research conference book of abstracts.
  http://usir.salford.ac.uk/id/eprint/47964/1/SPARC%202018%20book%20of%20abstracts.pdf

List of Presentations

As a part of research dissemination activity, this research has been presented in several UK and overseas workshops, which are listed below:

- Silva, C. (2016, November). ***Modelling Electricity Distribution Networks with Geospatial Big Data.*** In annual Manchester Electrical Energy and Power Systems Workshop - (MEEPS 2016): “Tackling the Challenges of Evolving Electrical Energy Systems” held at the University of Manchester.

  Link: https://www.ieee-ukandireland.org/event/ieee-4th-meeps-tackling-the-challenges-of-evolving-electrical-energy-systems/

- Silva, C. (2018, August). ***Modelling Electricity Distribution Network Faults with Geospatial Big Data***. In Connecting to the smart cities workshop: Developing a road-map to explore big data analytical solutions to address Mekong Delta’s current challenges at Can Tho University, Ninh Kiều, Can Tho, Vietnam.

  Link: https://cet.ctu.edu.vn/nghien-cuu/hoi-thao-hoi-nghi/64-uk-vietnam-researcher-links-workshop-6th-to-9th-august-2018.html

- Silva, C. (2018, September). ***New Data Science Framework for Analysing and Mining Big Data.*** In Data Science Conference 4.0 in Belgrade, Serbia.

  Link: https://www.youtube.com/watch?v=Xy0oR06Dg2k

- Silva, C. (2019, January). ***Identify The Road Traffic Accident Hotspots Using Geospatial Data Science And Visualisation Techniques***. In UK-Philippines Researcher Links Workshop: Smart Transport Systems and Asset Management. Manila, Philippines.

  Link: https://www.britishcouncil.ph/about/press/uk-philippines-smart-transport-system

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Introduction

The electricity distribution network system is a vital part of the power distribution infrastructure, enabling electricity to be distributed to homes and customers from a large power generation plant. The electricity generates from the power station at high voltage and is delivered at medium to low voltage levels. Any Electric power distribution system can be simplified into three main stages, power generation, electricity transmission and distribution. In the UK electricity transmission, from the power generating stations to Distribution Network Operators (DNO) [1] and large industrial customers, is carried out by a single company, the National Grid. This monopoly is regulated by the Office of Gas and Electricity Markets [2]. Figure 1.1 below shows the UK energy landscape.

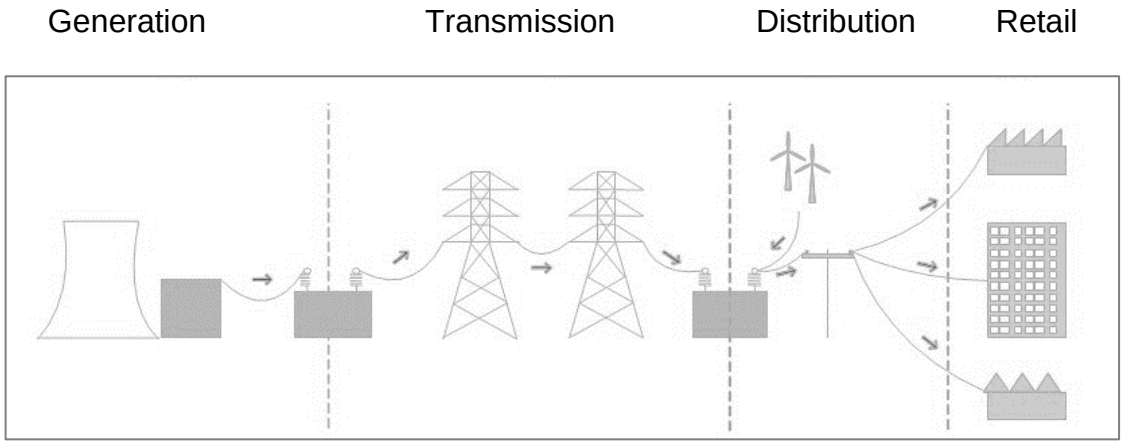


Figure 1.1: Pictorial representation of the UK energy landscape (Source: OFGEM [2])

Within regions, the distribution of energy is carried out by DNOs [3]. There are 14 DNOs in Great Britain, each of which has a license covering a defined geographical area [3]. They are accountable to the industry regulator, OFGEM. Seven holding companies own these 14 DNOs [3]. They work together where necessary through the Energy Networks Association [4]. The suppliers of electrical energy such as British

Gas, Eon, etc. are buying electricity from Electricity generators, and selling it to individual consumers. They pay the National Grid and the DNOs for the use of their assets to transport this energy from generators to consumers. Power systems are prone to frequent faults [5], which may occur in any power generating units, transformers, or power distribution media such as overhead and underground cables.

A fault in an electric power system can be defined as, any abnormal condition of the system that involves the electrical failure of the equipment, such as, transformers, generators, and busbars [6]. Electricity distribution network components are always vulnerable to frequent failures that may occur in any of the main components or sub-components. Faults that generally occur in transmission and distribution networks are short circuit transients caused predominantly by vegetation, animal and weather effects such as tree contact, large birds short-circuiting phases, creepage current through the path created by rain or moisture and the buildup of contaminants [4]. Weather is the single most influential factor that causes faults on the distribution network. Different weather parameters such as wind, temperature, snow and rainfall all have the potential to cause faults to different types of assets [4].

Any system malfunction causes significant supply disturbance, destabilising the entire system [4]. Detecting faults in electrical power grids is of paramount importance, both from the electricity operator and consumer point of view. For fault events, a customer satisfaction survey is a key component that determines the quality of customer service delivered by the DNO against a specific fault incident. Following an unplanned fault, the DNOs submit data to OFGEM's independent customer survey organization [7]. This company contacts relevant customers and ask a series of questions associated with the customer's experience during the interruption. The customer scores the DNO out of ten on ease of contact, politeness, the accuracy of information and usefulness of the information.

Usually, DNOs employee a large number of fault engineers, Cable jointers, and Overhead lineworkers to maintain and repair electricity distribution systems. The fault engineers specialise in working on substation equipment, underground assets or overhead assets, they also maintain equipment and carry out repairs as necessary. Cable jointers work on underground distribution cables, making connections to overhead lines or other parts of the distribution system. Overhead lineworkers build

and repair overhead electricity distribution lines, which are carried on wooden poles and steel pylons (Figure 1.2).

Figure 1.2: Overhead line workers fixing an overground cable fault (Source: SSE [93])

The National Fault and Interruption Reporting Scheme (NaFIRS) [8], was set up and is administered by the Energy Networks Association. Each DNO in Great Britain is required to report all faults which occur on their network, whether or not the fault results in loss of supply to customers. These reports are aggregated and analysed by cause and by voltage level. This Scheme was initially approved by the twenty-seventh Chief Engineers' Conference, held on 14th October 1964, and was subsequently revised several times [8].

The regulator agrees with each DNO an Interruption Incentive Scheme (IIS) annually over the price review period, which rewards or penalises DNOs depending on their interruption performance against their target [4]. The DNO is then penalised whenever a customer is off supply for more than 3 minutes, which is referred to as a customer interruption [1]. The performance of each DNO is reviewed annually and compared with previous years, with other DNOs, and with the individual targets set by OFGEM as part of the Interruption Incentive Scheme (IIS) framework and guaranteed standards of performance. This performance is publicly reported and is subject to financial rewards and penalties [3].

The financial penalties during fault outages can be significant, and understandably this has, where possible, driven DNOs to avoid interruptions and

restore customers quicker when a fault occurs. Initiatives to avoid interruptions can be targeted asset investment where the device has reached the end of its life; thereby, removing plant before failure occurs. Reducing interruption times via the introducing of new technology that improves fault prediction, fault segmentation, and fault forecasting has been instrumental in improving the DNO's interruption performance. Therefore, there is a business need to forecast faults accurately.

In recent years, few researchers have proposed methodologies for fault analysis purely focusing on electric current flow. However, this research is proposing a new analysis and prediction methodology for the network fault using data mining and machine learning techniques. In this research, the author also seeks to establish a relationship between environmental factors and fault causes. This research also builds a new data science model for fault forecasting and prediction in the electricity distribution network and validates it against the fault database. The research mainly uses data from the PCNaFIRS database. PCNaFIRS is a fault and interruption input and reporting system developed and supported by Devere Software Limited on behalf of the Energy Networks Association [4] for Distribution Network Operating Companies who subscribe to the ENA NaFIRS Scheme.

1.2 Research Motivations

There is a business need to reduce network fault to lower the operational expenditure in engineering departments and increase the performance of the electricity distribution network. Besides, it is also necessary to be able to predict future regional failures to obtain a clear understanding of any spatial distribution of network faults.

A comprehensive fault prediction and forecasting model provide opportunities for the electricity distribution industry to manage their engineering staff resources in a more controlled manner and reduce operational costs by reducing unnecessary call-outs and overtime charges. A sophisticated forecasting model will also help to improve the level of system availability by forecasting network faults and enabling precautionary measures to be taken. Making a valuable contribution to the electricity distribution industry is the primary motivation behind the conducted research.

There has been little discussion and research about Data Science adaptation in the utility distribution industry. Also, little attention has been paid to the use of Data

Science techniques for the fault analysis and forecasting in the electricity distribution network. Despite its high applicability, no research has examined the applicability of Data Science to forecast future network faults. No research has investigated the prediction of future faults in the electricity distribution network using National Fault and Interruption Reporting Scheme (NaFIRS).

Therefore, the scope of the study lies in utilising machine learning and Data Mining models for fault analysis and forecasting in the electricity distribution network. This study aims to develop a fault prediction model mainly from the industrial dataset in the fault management departments in DNOs, with higher prediction power than existing simulation methods.

1.3 Research Question

Understanding failures in the electricity distribution network are of the utmost importance for both the operators and the consumers. But it is challenging to be able to predict failures accurately for a given period due to the uncertain nature of the faults.

The main research question this study address is: **"Can data mining and machine learning approaches use to accurately predict and forecast faults in the Electricity Distribution network?"**.

Data mining covers the entire process of data analysis, including machine learning which aims at constructing programs that learn automatically from experiences. This research question is further divided into four sub-questions, as follows. The following key sub research questions are laid out here, which will be addressed in chapter 4, 5 and 6 of the thesis with a final discussion in the concluding Chapter 7.

1. Can industrial standard National Fault and Interruption Reporting Scheme (NaFIRS) data be used to predict and forecast potential faults in the network?

2. Can electricity distribution network operators accurately forecast the volume of future faults in the network and understand the seasonality? How can model performance be determined?

3. Can the predictions, forecast and new insights gained from data mining and machine learning analysis be used to enhance the functionality and the performance of the fault management process?

4. Do external factors such as population density influence the volume of future faults in the network?

1.4 Research Aims and Objective

This research aims to develop promising various multi-variant models in data mining and machine learning for prediction and forecasting electricity distribution network faults. Data mining and machine learning approaches were used in this research to execute characteristics similar to a decision support system of a human, and subsequently to apply it to imitate human anticipation. The objectives of the research are:

1. To gain a deeper understanding of Electricity Distribution Network faults and associated fault causes.
2. To gain an in-depth understanding of industry-standard National Fault and Interruption Reporting Scheme (NaFIRS) and the dataset.
3. Analyse the historical NaFIRS data to understand the association and correlation between the volume of the faults and the fault causes.
4. Develop models to understand the temporal patterns, seasonality and the dynamics of Network Faults. Also, Compare the performance of the alternative forecasting strategies.
5. Develop a multi-dimensional fault segmentation method using various fault characteristics.
6. Analyse the correlation and association between external factors such as local area population density, DNOs KPI and electricity distribution faults.
7. Use industrial case studies on real networks to demonstrate the concepts developed and their practical applicability and value.

1.5 Research Challenges

There are several challenges in this research relating to fault data:

1. Finding a required real industrial fault dataset such as NaFIRS dataset is challenging due to the commercial sensitivity of the data.
2. Dealing with the size and structure of the fault data. Also, attribute selection and the number of attributes to be considered in the analysis is also a significant challenge.
3. Data cleansing is difficult due to the nature of the data, which may be incomplete and noisy.
4. Dealing with the class imbalance problem in the fault data. In this study, the volume of some types of faults relatively very small in comparison with the frequent faults such as whether related fault data.
5. Some data mining and machine learning algorithms require noise-free data in a specific format. In most cases data sets contain invalid or incomplete data lead to complication in the analysis process, and some cases compromise the precision of the results.
6. The performance of data mining and machine learning techniques mainly depends on the efficiency of algorithms are using. In some algorithms and techniques may give less accurate results due to the nature of some characteristics of the input dataset.

All these challenges are discussed thoroughly in the individual technical chapters. The author's literature review suggests that no single data mining or machine learning technique has proved to be superior over the others in different experimental settings. Much depends on the underlying data population being tested, the set of explanatory variables available and the outcome.

1.6 Research Gap

Within the energy consumption research, several data mining and machine learning methods such as clustering, decision tree, and support vector machine were used to analyse the consumption data. One of the reasons for this is because these techniques allow researchers to measure the performance of the energy consumption forecasting model accurately, and the result is easily presented. However, based on the literature review, there is little to no research performed the same kind of analysis on fault data within the utility distribution industry. This research will investigate this area of research to understand and get more insight into this field.

Location finding of faults in the electricity distribution network has been widely researched in the last decade, following the rapid development of smart grid around the world. However, the literature review has been uncovered that the lack of scientific studies in the fault forecasting and prediction in the electricity distribution network.

As a data science researcher, the motivation for this work is the desire to bridge the gap created due to the shortage of data science adaptation in the fault management in utility distribution industry such as electricity, water and gas distribution. Also, no research has investigated the prediction of future faults in the electricity distribution network using National Fault and Interruption Reporting Scheme (NaFIRS).

This scientific research gap has been achieved by applying exploratory data mining and machine learning techniques to fault data to extract advanced analytic insights that would aid the understanding of possible relationships between internal and external factors and network faults. This would thus provide evidence-based literacy for network design engineers and industrial policymakers.

1.7 Research Focus Areas

Although analysing the fault data in the electricity distribution network is interesting; accurately predicting and forecasting future faults and identify any interesting patterns is still a difficult task. Predicting network faults is a complex process due to a number of faults, duration of the faults and number of customers affected is highly dependent on various internal and external factors. Few industrial studies were

conducted in this field, but only very few scientific studies have been conducted in this field. Hence, this thesis, with the objective to fill the knowledge gap, will focus on the predictive and time-series forecasting side of the domain with a preliminary analysis to develop a series of potential advanced decision-making models.

The four main research areas focused on in this thesis are identified as:

1. Temporal patterns and Seasonality of network faults to profoundly support the process of knowledge discovery within the fault data in any utility distribution network. Forecasting the trend of faults is a popular topic in the literature. Thus, forecasting with seasonally and detect hidden temporal patterns will be more attractive for implementation in the industry. This research focus area will be discussed in chapter 5.

2. Performing multidimensional segmentation using various fault characteristics. Fault segmentation is an essential tool for developing business intelligence in fault management department and maintaining competitive advantage among DNOs. This research focus area will be discussed in chapter 4.

3. Analyses of the historical fault data mainly NaFIRS database to understand the association and correlation between the volume of the faults and the fault causes. This research focus area will be discussed in chapter 4.

4. Analysing correlation and association between external factors such as local area population density and DNOs KPI. The ongoing growth of population followed by improvements in complex energy systems has raised challenges for distribution network providers and electrical engineers to assure system sustainability and efficiency. This research focus area will be discussed in chapter 6.

1.8 Thesis Structure

This thesis is structured into seven chapters giving an all-round view of the research problem, research methodology, various solutions, and discussion of contribution and future works.

Chapter 1: Introduction

This chapter gives a general overview of the background of the research work in this thesis, including the current energy landscape in the UK. Discussion of the network faults and the impact of the network faults on the distribution network operator's performance. Finally, the aim and objectives and research gaps of this project are discussed.

Chapter 2: Background and Literature Review

Reviews the existing literature on various topics, including Electricity Distribution Network configuration, NaFIRS schema, Network Faults, Data Science adaptation in the power and energy sector and all the algorithms used in the study.

Chapter 3: Research Methodology

In Chapter 3, the methodology of the research, which is the multimethod research method, is discussed. Also, in this chapter, a comprehensive discussion of the data analytics framework and its implementations are provided.

Chapter 4: Fault Cause Analysis and Prediction in Electricity Distribution Network

In this chapter discusses the methodology of in-depth understanding of causes of network Faults in Electricity Distribution Network using Association Rule Mining and explore the possibility of enhancing the knowledge gain from Association Rule Mining using Text Clustering. Also, this chapter discusses how segmentation can use to enhance the undesigning the network faults. Author of this thesis has proposed a new fault segmentation framework which DNOs can use to perform the fault segmentation.

This approach gives DNOs the option of performing multidimensional segmentation using various fault characteristics.

Chapter 5: Temporal Analysis of Faults in Electricity Distribution Network

In this chapter provides an in-depth discussion of equipment failure related network faults and compares the performance of a range of forecasting methods with a variety of accuracy measures. Also in this chapter provides an in-depth understanding of visual data mining concepts and discusses how 2D and 3D calendar heat map method can help provide a relatively new perspective in evaluating temporal patterns in electricity distribution network faults.

Chapter 6: Analysis of Impact of External Factors on Faults in Electricity Distribution Network

In this chapter, the author discusses how external factors like local population density affect the electricity distribution network faults. Various classification algorithms were used to build prediction models. Also, those models were validated and compare for the accuracy. Also, in this chapter, the author is trying to accurately understand the behaviour of CML performance indicator and trying to predict the annual Customer Minutes Lost (CML) figure using other annual financial and network performance indicators such as the number of customers affected, Totex, Network load.

Chapter 7: Conclusion

This chapter summarises the main accomplishment of this research. It describes the summary of the models developed in this thesis. The thesis concludes by reviewing the key contributions and directions for future work. It critically evaluates the presented research and discusses future work that could address its shortcomings or further extend or validate its contributions.

CHAPTER 2

Background and Literature Review

2.1 Introduction

As outlined in Chapter One, this research aims to develop various multi-variant models in data mining and machine learning to predict and forecast electricity distribution network faults. This chapter provides the background for this topic and highlights the gaps in the existing literature and the knowledge that the research aims to fill. This chapter begins by reviewing general Electricity Distribution Network configuration and discussing Low voltage and high voltage network. In section 2.2 Electricity Distribution Network architecture and in section 2.3, Faults in Electricity Distribution Network is discussed. In Section 2.3, the faults, fault causes, and National Fault and Interruption Reporting Scheme are discussed.

This chapter discusses literature from several publications from academic institutes, research organisations, governmental bodies, and utility companies which have focused on understanding the causes of power outages due to distribution network faults, and providing analysis of those events. Also, review the literature from the several academic journals which have focused on weather trends, and discussed the association between significant power outages and weather events. Also, in this section provide an overview of previously published literature about the research questions that have been stated in the introduction chapter. The main aim of this chapter is to conduct a comprehensive literature review to identify the gaps. The different data mining and machine learning methods were applied throughout this research; therefore, this chapter should present a literature review of machine learning and data mining, but the author has decided to discussed various data mining and machine learning literature within the individual chapters rather than in this chapter.

2.2 Electricity Distribution Network

In the United Kingdom and many other countries of the world, two primary infrastructure systems facilitate the transfer of electricity from where it is generated to where it is required such as industrial, commercial, or domestic consumers. The first system is the electricity transmission network, and the other system is the electricity distribution network.

In the UK, the first system which is the electricity transmission network, owned by the National Grid in England and Wales. This carries electricity from the generators to grid supply points situated at numerous locations around the country at high voltages between 400 and 275 kV [9]. The transmission system operates at typically 400,000 volts (400kV) or 275kV (and 132kV in Scotland), and the distribution system operates at voltages from 132kV to the average household voltage of 230V [9]. This is shown diagrammatically in below Figure 2.1 [4].

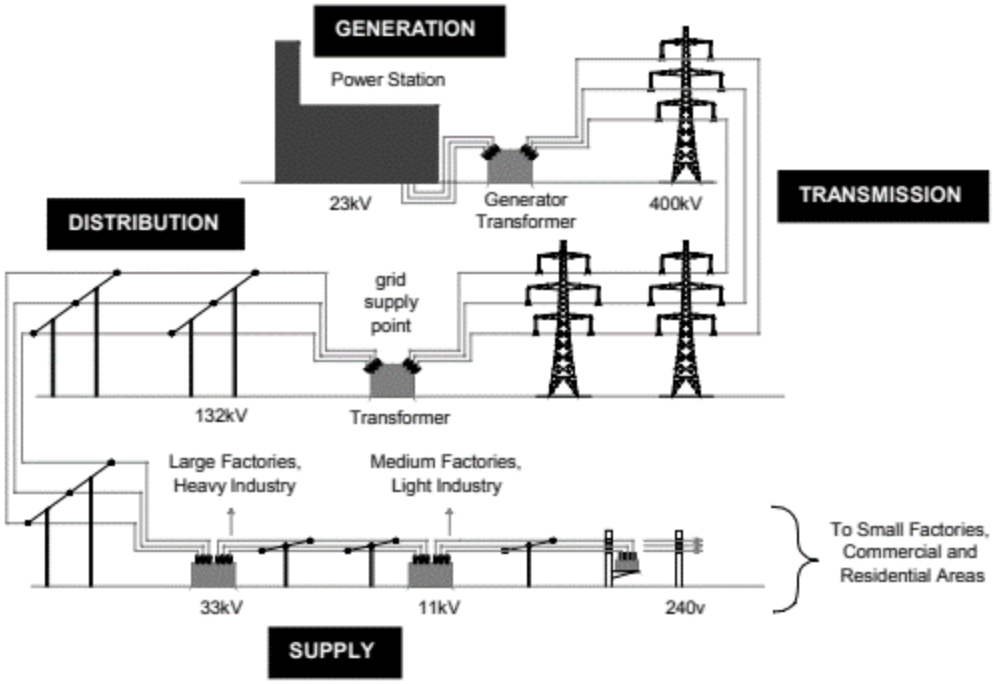


Figure 2.1: Different voltage levels in the UK energy landscape (Source: ENA [4])

The electricity distribution networks are part of the country's critical infrastructure, enabling electricity to be distributed to homes and non-domestic customers. The electricity distribution networks are regional grids that branch from the national grids to deliver power to industrial, commercial and domestic users. The UK

distribution network operators' regions are shown in Figure 2.2 below, together with those of independent distribution network operators (DNOs) who are ENA members [1].

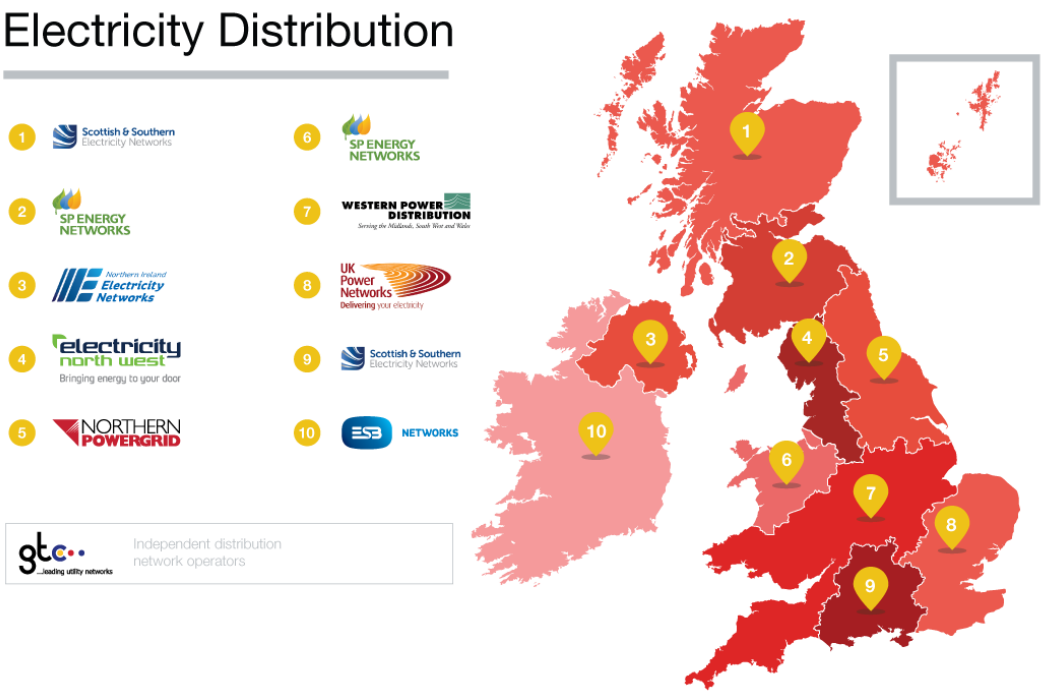


Figure 2.2: The UK Distribution Network Operators(DNOs) and their operational regions (Source: ENA [4])

DNOs are continually battling to meet rising demand from the consumers. Annually they have to invest a significant amount of money for replacing ageing or poorly performing assets, maintaining and improving network performance [10]. Also, as a regulatory business, they have to invest in maintaining regulatory requirements, such as reliability standards. As part of its regulatory contract, all the DNOs must comply with goals set out by the regulators. These goals also include meeting the financial expectations set by the regulators, constant supply uninterrupted power supply to the customers, provide secure network and maintain a high level of customer satisfaction.

Figure 2.1 illustrates the components of the UK electricity landscape. Together these parts work to safely and efficiently produce and supply electricity to meet the demand of commercial and domestic customers [11]. The Electricity Distribution Network comprises a combination of overhead lines and underground cables. There are also assets called substations, where voltage transformation occurs and where

switching, and control equipment are located [4]. These sites supply large numbers of customers, typically 5,000 to 30,000 customers at primary sites and 50,000 to 500,000 customers at grid sites [4]. In England and Wales, the National Grid owns and operate the Transmission System and, provides the interface between transmission and distribution systems which takes place within grid substations at 132kV.

In Scotland, the Transmission Networks are owned by Scottish Power and Scottish and Southern Energy but operated by National Grid. The interface between transmission and distribution systems takes place within grid substations at 33kV. The functions of various kinds of substations are described in Table 2.1 below.

Table 2.1: Types of Electricity Substations (Source: SSE [7])

Substation Type	Typical Voltage Transformation Levels	Approximate number nationally	Typical Size	Typical Number of Customers Supplied
Grid	400kV to 132kV	380	250m by 250m	200,000/500,000
	132kV to 33kV	1,000	75m by 75m	50,000/125,000
Primary	33kV to 11kV	4,800	25m by 25m	5,000/30,000
Distribution	11kV to 400/230V	230,000	4m by 5m	1/500

Then, the photographs in figures 2.3, 2.4 and 2.5 illustrate different types of substations and overhead line connections [7].

Figure 2.3: 132kV Grid Substation showing 132kV terminal tower and the start of 33kV wood pole overhead line (Source: SSE [7])

Figure 2.4: Primary substation showing equipment operating at 33kV and 11k (Source: SSE [7])

Figure 2.5: Distribution substation with equipment operating at 11kV and 400/230 volts (Source: SSE [7])

2.3 Faults in Electricity Distribution Network

Modern society is increasingly dependent on a reliable supply of electricity. Depending on where an electricity outage occurs and who is affected the consequences may range from a mere nuisance to significant economic losses or actual threats against the health and safety of citizens [12].

Due to the complexity of the electricity distribution systems, they are prone to frequent faults [3]. Especially the main equipment in the network are always vulnerable to numerous failures that may occur in any of the main components or subcomponents in leading equipment [3]. Any fault in the network component results in an outage of power not only to the area served by them but additionally throughout the neighbouring area. Any fault in the energy distribution system causes significant disturbance to the complete grid system. The financial penalties for the failures in the system can be significantly high. Understandably, where possible DNOs to avoid any supply interruptions and restore customer confidences quicker whenever a failure occurs to reduce financial penalties from the regulators. When a fault happens, locating the fault in the distribution systems is immensely important to the DNOs. Predicting distribution network faults specifically with their location, is very important.

All electricity network operators focus on offering a safe, secure, reliable and price-effective network to provide energy to customers. Whenever a customer loses electricity or has a power outage due to equipment failure, details of that interruption are recorded by either the transmission or distribution companies. Electricity distribution networks are much more affected by climate impacts than the transmission system [7].

Ofgem set overall standards of performance to provide additional incentives to DNOs to maintain availability and quality of supplies to individual customers. DNOs must pay specific penalties to customers related to restoring supplies after faults (within 18 hours of fault notification), giving notice of planned interruptions (5 days' notice required) and investigation of voltage complaints (within seven days) [13].

2.3.1 National Fault and Interruption Reporting Scheme (NaFIRS)

United Kingdom electricity companies, which are private organisations, collect data on faults in the NaFIRS database as part of the regulation criteria set out

by the government [3]. This database contains details of all the High Voltage (HV), and Low Voltage (LV) related faults which have occurred in the electrical distribution system, including date, time, number of consumers affected, and number of minutes lost[9].

The National Fault and Interruption Reporting Scheme (attached at Appendix A), set up and administered by the Energy Networks Association. Each DNO in Great Britain is required to report all faults which occur on their network, whether or not the fault results in loss of supply to customers. These reports are aggregated and analysed by cause and by voltage level.

This Scheme was initially approved by the twenty-seventh Chief Engineers' Conference, held on 14th October 1964, and was subsequently revised several times [8]. NaFIRS is designed to collect information relating to both network performance and equipment performance [2].

According to Ford, [8] the objectives of the NaFIRS Scheme are to :

A. obtain and disseminate information relating to the reliability in service of distribution system equipment;
B. provide information to permit the study of total distribution system performance, mainly at times when they fail;
C. provide information to permit the study of servicing organisations responsible for the operation, control, repair and maintenance of distribution systems and their components;
D. check the correctness of existing system design parameters, particularly those concerned with securing supplies;
E. indicate the need for further research and development;
F. provide a consistent and statistically sound indication of the standards of service being provided to customers.

All supply interruptions on distribution networks are recorded in the NaFIRS database [8]. An Interruption is when supply has been off or not adequate for a total of 3 minutes or longer. This information is shared nationally, and summaries are submitted to Ofgem. Data is available for over thirty years, but the quality of the data has improved significantly over the last fifteen years since the introduction of the Ofgem Interruptions Incentive Scheme (IIS)[3].

For each interruption, companies will capture a large amount of information, and up to 100 separate fields will be populated. Using data from the NaFIRS system, companies can monitor how their networks are performing, identify any trends in faults and respond accordingly [4]. These include:

- fault location
- number of customers affected
- duration
- type of equipment
- manufacturer of equipment
- cause of the fault.

Since 2010, DNOs periodically provide the full dataset to Ofgem. This allows Ofgem to perform their own analysis and publish national-level reports. Although the data is aggregated at this level, companies capture data to a more detailed level, attributing faults to one of 99 different direct causes specified in ENA Engineering Recommendation G43-3 (Instructions for Reporting to the National Fault and Interruption Reporting Scheme). Eleven of these causes are weather-related [4].

- lightning
- rain
- snow, sleet, blizzard
- ice
- freezing fog and frost
- wind and gale (excluding windborne material)
- solar heat
- airborne deposits (excluding windborne material)
- condensation
- flooding
- windborne materials

Each financial year OFGEM provides all DNO’s budget based on their CI (Customers Interrupted) & CML (Customer Minutes Lost). This money is paid upfront, and whatever is leftover, they may be retained and reinvested in their network. However, they are required to maintain a balance since reporting outstanding

performance will result in receiving a lower budget for the next period, and a poor performance will result in paying back the budget they received [3].

2.3.2 Elements of Fault Reporting

The management of faults is one evident demonstration of how a DNO is meeting some of its stakeholder expectations. Therefore, accurate and in-depth analysis of faults and fault reporting is key to the electricity distribution industry. NaFIRS data provide a vital platform to analyse how the DNOs has performed in the past, and that can be used in regionally and countrywide the forecasting of supply restoration. Also, countrywide Knowledge of the resources available to the DNOs is vital in being able to forecast restoration performance. Along with the individual company policies (DNO), and this data can be used to ensure that the forecast information given to customers is as accurate as possible. There may also be company-specific data available. This data will vary dependant on the various policies and procedures adopted by distribution companies. So, that OFGEM has introduced a set of key measures that can be used in fault reporting.

- CI (Customers Interrupted) is a one-off penalty that occurs every time a customer is interrupted
- CML (Customer Minutes Lost) is the duration of outage multiplied by the CI
- SDI (Short Duration Interruption) currently only used for HV figures and captured through CRMS (Control Room Management System)
- RI (Re-Interruptions) is customers being off again within 3 hours of a permanent restoration or 18hours of a temporary restoration (loop, bunch or generator). This does not carry an additional CI penalty.
- ONI (Occurrence Not Incentivised) fault work that does not require NaFIRS. This can range from anything such as changing a cut out to securing a substation and even spiking a cable.

Based on the historical annual electricity interruption reports, OFGEM annually sets targets for the number of customers interrupted (CIs) and duration (CMLs) of both planned and unplanned interruptions [3]. DNOs are rewarded if they meet or exceed these targets and are penalised if they fail to meet them. DNOs must continue to invest in network assets to reduce the number of Customer Interruptions, and Customer Minutes Lost.

Network operators may receive a considerable monetary reward or incur a significant monetary penalty depending on their performance against annual target for both the number and length of their network supply disruption. It is compulsory for high maintenance levels of customer services; also maintain and improving service level as defined by the regulators also key responsibilities of the DNO. Annually OFGEM conducts a customer satisfaction survey to measure DNOs customer satisfaction performance.

Since 2010, DNOs are periodically providing the full dataset to Ofgem. So, Ofgem can perform their own analysis and publish national-level reports. Although the data are aggregated at this level, companies capture data to a more detailed level, attributing faults to different direct causes specified in ENA Engineering Recommendation G43-3 (Instructions for Reporting to the National Fault and Interruption Reporting Scheme), Eleven of these are weather-related [14], e.g., lightning, rain, snow, sleet, blizzard, ice, etc. Each financial year OFGEM gives all DNO, CI (Customers Interrupted) & CML (Customer Minutes Lost) budget, this money is paid upfront, and whatever is leftover they get to keep and reinvest in their network.

2.4 Data Mining Methods

The advancement in computer technology has resulted in an increase in the collection, storage and manipulation of data. As a result of this, data collection has grown in size and complexity. Thus, the need for automatic data processing is increased which is supported by other technology developments, such as genetic algorithms in the 1950s, clustering, decision trees in 1960s and database management systems in 1970s that can store and query petabytes of data.

Data mining is the process of extracting data from large data sets using techniques that include machine learning, statistics, database management systems and algorithms. Traditional methods of data analysis often involve slow, expensive and highly subjective manual work and data interpretation. Rapid technological changes in many organisations have resulted in the collection and processing of a massive amount of data; the extraction and analysis of a large amount of data is a challenging and complicated process. The complexity can be reduced by using data mining techniques.

Extracting useful information from vast amounts of data can be challenging and requires innovative methods, algorithms, and statistical approaches. Data mining combines traditional data analysis methods with superior and sophisticated algorithms for processing a large amount of data. It is an interdisciplinary field merging concepts from database systems, statistics, machine learning, computing, information theory, and pattern recognition [16]. Data mining can be an interdisciplinary field that includes other subject matter, as shown in figure 2.6 below. Two of the areas shown are explored in this research. These are machine learning and visualisation.

Most organisations today use data mining in various ways, for example, in the media and entertainment sector, banking, retail and logistics, telecommunications, insurance, manufacturing and engineering, medicine, science and transport industries. Data mining is a process that uses intelligence tools such as predictive analysis, neural computing and advanced statistical methods to search for unknown relationships or information from large databases [17].

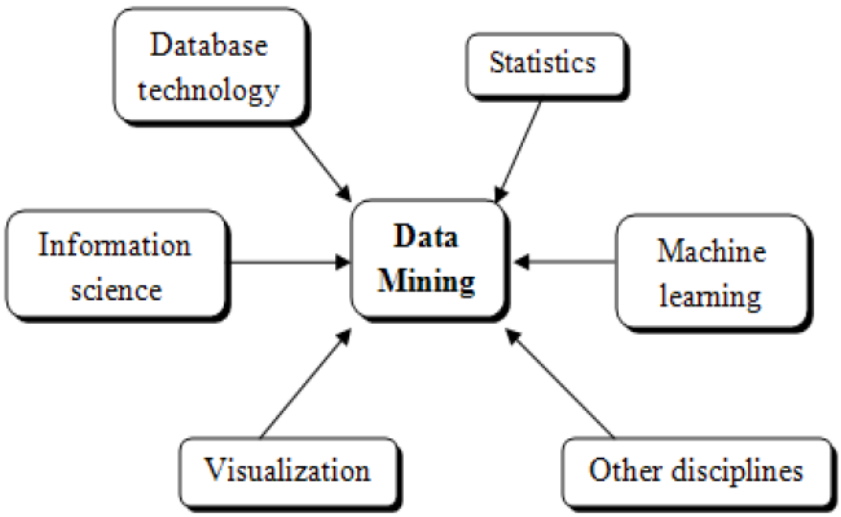


Figure 2.6: Data mining as an interdisciplinary field (Adapted from Han and Kamber, 2006) [16]

Data mining also helps to discover price relationships, purchasing behaviours, demographics and economic impact, and the influence of social media company, its income, profit and costs factors. Understanding these factors, in turn, helps to predict the future results of the company.

Data mining techniques can be broadly divided into three general categories:

- **Discovering** – inspecting a database to identify possible hidden patterns without a predetermined hypothesis as to what those may be.
- **Predictive Modelling** – using the discovered patterns to predict future results.
- **Analysis** – comparing the information derived against established patterns to detect unusual elements.

Each of the categories mentioned above involves specific data mining techniques.

Choosing the right data mining technology depends on the sector, company nature and business challenges being faced. Some of the most common techniques for data mining are association rule, classification and decision trees, regression, and neural networks.

The author of this study used a Knowledge Discovery (KDD) process conjunction with data mining to identify any hidden patterns in the data. Knowledge Discovery (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data [17]. It is a complete process of mining deeply into and discovering useful previously unknown knowledge from data. The knowledge discovery process consists of a set of clearly defined sub-steps to be followed by practitioners when executing a knowledge discovery project [17].

The knowledge obtained from the proposed model and its findings can be analysed by coinciding the context information. The impact and relationships of knowledge can then be correlated. Thus, this knowledge can be merged through its interlinkages and relationships, creating comprehensive knowledge. There are several challenges exist in the KDD process, including data acquisition, data extraction, data store, data cleansing and filtering and visualising.

Silva et al. [18] have introduced A New Data Science Framework for Analysing and Mining Geospatial Big Data. This framework can be used to plan, design and execute entire data science project. This framework is used as a KDD process in this study.

Data mining techniques come in two primary forms: supervised and unsupervised. Both categories encompass functions capable of finding different hidden patterns in large data sets.

2.4.1 Unsupervised Learning Techniques

Unsupervised machine learning does not require labelled historical data to detect patterns in data [17]. The data is put into the algorithm to recognise the groups with similar characteristics automatically. The goal is to explore the data beyond the human imagination and come up with novel discoveries that are interesting. In other words, Unsupervised learning is a type of algorithm that doesn't require any response variables at all. In this case, the model will learn patterns from the data by itself. You may ask what kind of pattern it can find if there is no target specified beforehand. This type of algorithm usually can detect similarities between variables or records, so it will try to group those that are very close to each other. In this research, two main unsupervised learning techniques will be used for the prediction, which is clustering and association rule mining.

2.4.1.1 Clustering

Cluster analysis is an unsupervised data mining technique employed in the process of separating data objects based on similarities to each other and dissimilarities with data objects grouped into other clusters [19]. It involves grouping a set of objects into different clusters such that objects in each cluster share similarities (have high intra-cluster affinity) among themselves but are dissimilar to objects in other clusters (have low inter-cluster affinity). It is a statistical method that helps in the measurement of homogeneity among members of a group. Clustering is a beneficial exploratory data mining technique as its application can lead to the unravelling of previously unknown groups within a dataset [17]. In this study, the clustering technique is used to group temporal data into clusters using a distance measure that calculates nearness to a cluster mean based on attribute values.

In this study, the K-Means clustering method, which is one of the main clustering algorithms, has been used to cluster the electricity distribution faults. K-Means groups unlabelled data into groups or categories which were not before apparent; it is an excellent method for extracting hidden patterns within a dataset. K-means is a classical clustering technique. This is a statistical approach that can help in the measurement of homogeneity among participants of a group. K-means algorithm aims at minimising

an objective function, in this case, a squared error function. The objective function can be defined as:

$$\mathrm{J} = \sum_{j=1}^{k} \sum_{i=1}^{x} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres [20].

In K-mean clustering parameter k specifies how many clusters are to be created. As cluster centres, the k points are selected in random. This k-parameter is specified in advance. As per Euclidean distance metric, all the instances are allocated to their nearest cluster. The mean of each instance with each cluster is calculated. This centroid or mean then become the new centre value for that particular cluster. The whole process is repeated until the same values are allocated to all the clusters in the repeated round.

In the final phase, the clusters will remain the same and will have their centres stabilised at this point each instance are allocated to its closest cluster centre. This minimises the overall squared distance from all points to their cluster centres [21].

The techniques involved methods to calculate a distance measure, ascertain the cluster-ability of the dataset, and the optimum number of clusters that can be obtained from the dataset using the Hopkins-Statistic and Elbow plot, respectively [17].

With a Hopkins-Stat value, the decision can be made that the data set contained meaningful clusters. The *Hopkins statistic* is used to assess the clustering tendency of a data set by measuring the probability that a uniform data distribution generates a given data set [17]. If the data is *d* dimensional, the Hopkins statistic is defined as [17],

$$H = \frac{\sum_{j=1}^{m} u_j^d}{\sum_{j=1}^{m} u_j^d + \sum_{j=1}^{m} w_j^d}$$

Two types of distances are defined:

u_j*:* as the minimum distance from y_j, to its nearest pattern in X

w_j *:* The minimum distance from a randomly selected pattern in X to its nearest neighbour *(m* out of the available *n* patterns are marked at random for this purpose).

Hopkins statistics check the uniform distribution of the data. The uniform distribution is a continuous probability distribution and is concerned with events that are equally likely to occur.

The null and alternative hypotheses can be defined as follows:

- *Null hypothesis*: the dataset is uniformly distributed. So, no meaningful clusters can be created
- *Alternative hypothesis*: the dataset is not uniformly distributed. So, meaningful clusters can be created

If the value of Hopkins statistic is close to 1, then we can reject the null hypothesis and conclude that the dataset is significantly a clusterable data.

One of the well-known methods of determining the optimal value of K in K-means is the Elbow Method. This method involves the drawing of a curve between the WSS (within the sum of squares) and the number of K. As K increases, the WSS will reduce. For k=n (with n being the total number of observations in the dataset) the WSS will become zero.

The value of *k* is added one by one, and the within Sum Square Error (WSS) value is recorded. WSS can be defined as [22]:

$$\text{WSS} = \sum_{i=1}^{k} \sum_{x \in C_i} dist\left(X, \bar{X}_{C_i}\right)^2$$

- C_i is the cluster
- x is a data point in cluster C_i
- $\bar{X}_{C_i}$ is the cluster centroid
- $\bar{X}$ is the sample mean

When Hopkins statistics (*H*) perform against to the research dataset, it shows *H* value is 0.15, which is very low.

- For well-defined clustered data., *H* value should be larger than 0.5.

- *H* is supposed to be much less than 0.5 for data that are neither clustered nor random.

The Hopkins statistic is known to be a fair estimator of randomness in a data set. However, in some cause outliers can pollute the dataset. So, it is better to remove outliers from the research dataset to improve the Hopkins statistics value for better clustering.

2.4.1.2 Association Rules Mining

Association rules is an unsupervised learning technique to discover the association of items. Association rule mining has been mainly applied to analysing customer's shopping baskets. This helps retailers identify items that are likely to be purchase at the same instance. Association rules identify a combination of items purchase that frequently occurs together [23].

A most popular technique for discovering association rules is the Apriori algorithm. The Apriori algorithm is used to discover association rules and finding frequent itemsets [24]. The Apriori algorithm is credited to Agrawal, Imieliński and Swami who applied it to market basket data to generate association rules [23].

Association Rules is also an important task used in data mining as is Classification and Clustering. An association rules is widely used in various fields and can help to produce general and qualitative knowledge that assists scientists to make better decisions [25]. Moreover, Association rules deals with transactions with both binary values and quantitative data [26]. Binary attributes databases have been used to create traditional algorithms in association rules, especially where many real transactions contain quantitative attributes. Hence, quantitative data use of association rules is a prevalent area of study between researchers [25].

Association rules are also evoking correlations from vast amounts of data. This can reveal any data dependences with respect to correlation, facilitating the receipt of object information from another data object. Association rules can discover any correlation between inconsistent item-sets from a database and can also discover any implicit information from data [27].

Association rules mining is governed by some defined rules that help to quantify the threshold level of association between items. As shown in Figure 2.7, these rules are classified as support, confidence and the lift.

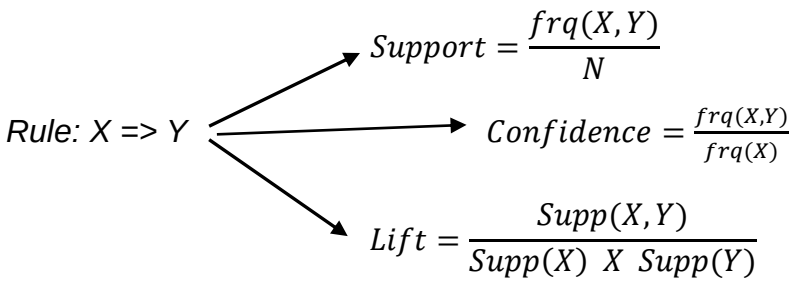


Figure 2.7: Three key parameters applied to association rules mining

The rules are created with an antecedent (X) and a consequent (Y), as shown in Figure 2.7. Rules can have multiple antecedents and consequents and do not have to be of length two.

Support: The support of item X, denoted as Supp (X) or the Supp (X ⇒ Y) is defined as the proportion of transactions in a dataset which contains the item X is also considered as the percentage of all transactions both in X and Y. It is a measure of how frequently the rules occurs in the dataset.

Confidence: The confidence of a rule is given as

Conf (X ⇒ Y) = Supp (X ∪ Y) / Supp(X)

This is the percentage of all transactions that include both X and Y divided by the number of transactions of X. This can be explained as the probability of Y to occur given that X has occurred.

Lift: The lift of a rule is given as

lift (X ⇒ Y) = Supp (X ∪ Y)/(Supp(X) * Supp (Y))

The value of the Lift measures the strength or significance of the association rule. Lift in association rules is considered as the ratio of the confidence of a rule to the expected confidence of the rule. The expected confidence is calculated with the assumption that the left-hand side rule is independent of the right-hand side rule. Therefore, it can be stated that the lift is a measure of association between the left-hand side and right-hand side rules. The greater the value of the lift, the stronger is the association of the left-hand side to the right- hand side.

In rules where the resulting value of the Lift is less than 1, the occurrence of LHS is said to be negatively correlated with RHS, which means that the occurrence of one of the events will likely result in the absence of the other event. However, If the Lift is greater than 1, then the two events are positively correlated, meaning there is a mutual relationship between the two events [17]. In a scenario where the Lift is equal to 1, then the two events are termed independent as no relationship exists between them.

Using this data mining procedure, it is essential to understand that the reliability of the results is directly correlated to the size of the transaction data, whereby the higher number of transactions the more reliable are the relationships produced. Often rules require the support of several hundred transactions before the rule is measured as statistically significant [28]. Because of the nature of the association rules algorithms, they only accept categorical data. Consequently, all continuous data need to be converted into discrete bins or categories.

The Apriori algorithm is the renowned algorithm to mine association rules. Given a set of transactions, the Apriori algorithm attempts to find subsets that are common to at least a minimum number of item sets. Apriori uses an iterative approach known as a level-wise search where k-itemsets are used to explore (k + 1)-item sets.

First, the set of frequent 1-itemsets is found by scanning the database to collect the count for each item and accumulating those items that satisfy minimum support. L1 denotes the resulting set. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found [29]. The Pseudo-code of the Apriori algorithm is shown below:

C_k: Candidate itemset of size k

L_k: frequent itemset of size k

L_1= {frequent items};

for (k= 1; L_k!=∅; k++) **do begin**

C_{k+1}= candidates generated from L_k;

for each transaction t in database do

increment the count of all candidates in C_{k+1}

that are contained in t

L_{k+1}= candidates in $C_{k+1}$with min_support

end

return $C_k L_k$;

2.4.2 Supervised Learning Techniques

Supervised machine learning involves learning by example from previously classified data that has been labelled [17]. Supervised learning refers to a type of task where an algorithm is trained to learn patterns based on prior knowledge. That means this kind of learning requires the labelling of the outcome (also called the response variable, dependent variable, or target variable) to be predicted beforehand. The supervision comes from the concept that learning is guided through the learning outcomes of the training data. The learning algorithm receives a set of input with predetermined correct outputs from which it learns the pattern and modifies accordingly. In this research, two main supervised learning techniques will be used for the prediction, which is classification and regression.

2.4.2.1 Classification Methods

Data mining algorithms which carry out the assigning of data objects into related classes are called classifiers [17]. There are different methods for data classification, such as Logistic Regression, k-Nearest Neighbour, Naïve Bayes, Decision Trees,

Random Forest, Support Vector Machine, Artificial Neural Networks, and so forth. Comparing the output of different classifiers and using the most accurate predictive classifier is essential in an exploratory study. Each of the classification methods shows different efficacy and accuracy based on the characteristics of the datasets.

Besides, there are various classification evaluation metrics for comparing the classification output as part of the methodology for creating a classification model. Different models are tested to find the optimal model to use. These models are checked for accuracy, sensitivity, specificity and versatility. Some models perform better with a certain amount of class types are specific data. Model testing is only initiated once the data has been thoroughly cleaned and prepared.

There are various classification methods available to use for this study. However, only four different classification methods have been taken into consideration Logistic Regression, Decision Trees, Random Forest and Support Vector Machine. Also, the classification performance was compared to identify the best classification method. All the methods were carefully validated using a different kind of accuracy measures.

2.4.2.1.1 Logistic Regression

Logistic Regression is the regression analysis which models the probability that a dependent variable, the response, belongs to a particular category. With this method, based on the information gained from the independent variables, the predictors, a quantitative prediction of the probability of the occurrence of each class of the response is calculated. Logistic regression is used to obtain the odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial.

Logistic regression is a powerful tool, especially in classification studies, allowing multiple explanatory variables being analysed simultaneously, meanwhile reducing the effect of confounding factors. Logistic regression will model the chance of an outcome based on individual characteristics. Because chance is a ratio, what will be modelled is the logarithm of the chance given by [31]:

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m$$

where

π indicates the probability of an event

β_i are the regression coefficients associated with the reference group

x_i are the explanatory variables

2.4.2.1.2 Decision Tree

The decision tree is a supervised classification algorithm that learns from historical data and applies the inference to classify new data into their appropriate categories. There are different classification algorithms available; however, Classification and Regression Algorithm(CART) has been chosen for this research because it has proven to be more accurate than other existing algorithms such as ID3 and C4.5. Classification and regression tree, or CART, is a classification method that builds a model from historical data. CART was firstly developed by Breiman, Freidman, Olshen and Stone in 1984 Pseudo-code of the CART algorithm is shown in figure 2.7 [32]. ==========

(1) Start at the root node.

(2) For each ordered variable X,
convert it to an unordered variable X' by grouping its values
in the node into a small number of intervals
if X is unordered, then set X' = X.

(3) Perform a chi-squared test of independence of each X' variable
versus Y on the data in the node and compute its significance
probability.

(4) Choose the variable $X*$ associated with the X' that has the smallest
significance probability.

(5) Find the split set $\{X* \in S*\}$ that minimizes the sum of Gini indexes
and use it to split the node into two child nodes.

(6) If a stopping criterion is reached, exit.
Otherwise, apply steps 2–5 to each child node.

(7) Prune the tree with the CART method.

==========

Figure 2.7: Pseudo-code of the CART algorithm (Source: Hongquan Guo et al. [32])

2.4.2.1.3 Predictive Modelling using Random Forest

Random forests can be defined as a collection of tree classifiers. It can be said that the random forest model is made up of many decision tree models. Each tree depends on the random sampling of the data where the distribution of each sample for each tree is the same. However, in the random forest algorithm, each node is split using the best variable amongst a subset of features rather than all features in the dataset. In practice, random forests are often found to be the most accurate learning algorithms to date. Pseudo-code of the Random Forest algorithm is shown in figure 2.8 [32].

==========

To generate c classifiers:
for i = 1 to c **do**
  Randomly sample the training data D with replacement to produce D_i
  Create a root node, N_i containing D_i
  Call BuildTree(N_i)
end for

BuildTree(N):
if N contains instances of only one class **then**
  return
else
  Randomly select x% of the possible splitting features in N
  Select the feature F with the highest information gain to split on
  Create f child nodes of N, N_1,..., N_f, where F has f possible values (F_1, ... , F_f)
  for i = 1 to f **do**
    Set the contents of N_i to D_i, where D_i is all instances in N that match
    F_i
    Call BuildTree(N_i)
  end for
end if

==========

Figure 2.8: Pseudo-code of the Random Forest algorithm (Source: Hongquan Guo et al. [32])

The random forest algorithm uses the bagging technique for building an ensemble of decision trees. Bagging is known to reduce the variance of the algorithm.

2.4.2.1.4 Predictive Modelling using Support Vector Machine

Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92 [94]. The Support Vector Machine algorithm can be implemented for classification as well as regression. It also tends to be used for binary classifications. SVMs were originally developed to solve the classification problem, but recently they have been extended to solve regression problem.

The SVM algorithm derives a separating hyperplane from categorising or classifying data points that are labelled. They belong to a family of generalized linear classifiers. In another term, Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximise predictive accuracy while automatically avoiding over-fit to the data.

SVM has been found to be successful when used for pattern classification problems. Training an SVM can be illustrated with the following pseudo-code (Figure 2.9) [33].

==========

Require: X and y loaded with training labeled data, $\alpha \Leftarrow 0$ or $\alpha \Leftarrow$ partially trained SVM

1: $C \Leftarrow$ some value (10 for example)
2: **repeat**
3: **for all** $\{x_i, y_i\}, \{x_j, y_j\}$ **do**
4: Optimize α_i and α_j
5: **end for**
6: **until** no changes in α or other resource constraint criteria met

Ensure: Retain only the support vectors ($\alpha_i > 0$)

==========

Figure 2.9: Pseudo-code of the SVM algorithm (Source: Pedersen et al. [33])

2.4.2.1.5 Classification Accuracy Measures

Once a classification model has been built, an estimate of accuracy on how a classifier predicts would be required to measure. In addition to the primary measure, known as accuracy, additional measures like specificity and sensitivity, will be used in this study. The accuracy evaluates the average efficiency of the algorithm, while the other two measurements estimate the classifier's behaviour on different groups. The outcomes predicted by classifier and the actual outcomes can only have four combinations. Those four combinations can be represented in the form of a matrix. That matrix called a confusion matrix. A confusion matrix is a summary of prediction results on a classification problem. Figure 2.10 shows the sample of a confusion matrix.

		Predicted Class	
Actual Class		**True**	**False**
	True	*True-positive (TP)*	*False-negative (FN)*
	False	*False-positive (FP)*	*True-negative (TN)*
	Total	*Total Positive*	*Total Negative*

Figure 2.10: Sample of a confusion matrix

- True-positive (TP) = Correct positive prediction
- False-positive (FP) = Incorrect positive prediction
- True-negative (TN) = Correct negative prediction
- False-negative (FN) = Incorrect negative prediction

Classification accuracy is defined as the percentage of correct predictions. It can also be defined as the ratio of the sum of true positives (TP) and true negatives (TN) to the total number of data points [sum of TP, false positives (FP), false negatives (FN), and TN]. Mathematically, this can be stated as:

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

Sensitivity is described as the percentage of true positives that the algorithm accurately observed. Mathematically, this can be stated as:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity is determined by the percentage of true negatives that are correctly recognised. Mathematically, this can be stated as:

$$Specificity = \frac{TN}{TN + FP}$$

Sensitivity and specificity are inversely proportional, meaning that as the sensitivity increases, the specificity decreases and vice versa.

2.4.2.2 Regression Analysis

Regression is a supervised machine learning method that predicts a numerical outcome (dependent variable) based on one or more inputs (independent variables). Regression analysis is referred to as statistical technique used for the estimation of the relationship between variables, mainly the relationship between the dependent variable and the independent variable. Regression analysis mainly is used for predicting, finding and forecasting the effect of one variable on the other.

It can be used to evaluate the strength and the modelling of the future relationship between the variables. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeller might want to relate the weights of individuals to their heights using a linear regression model.

The analyses of regression include several variations, for example linear, multiple linear and nonlinear. Simple linear and linear are the most common models. Nonlinear regression analysis is often used for complex data sets with nonlinear relationships between the dependent and independent variables.

There are several types of Regressions models available :

- Linear regression
- Logistic regression
- Polynomial regression
- Stepwise regression
- Ridge regression
- Lasso regression

Linear regression is the most common form of regression. Simple Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. This is based around one dependent variable on the y-axis that is influenced by a single or collection of independent variables (predictor) on the x-axis. However, the output should always be a continuous value.

The simple linear regression model is presented, as shown in the equation given below [34].

$$Y = \alpha + \beta X + \epsilon$$

In this equation :

α, β are the coefficients of the dependent variable

ϵ stands for the residuals

The model aims to minimise the sum of squared errors by fitting different values to different observations.

Multiple linear regression analysis is basically similar to the simple linear model, except that the model uses multiple independent variables. The mathematical representation of multiple linear regression is [34]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m + \varepsilon_1$$

y is the dependent variable.

β_0 is the intercept; it is the theoretical value of the dependent variable if the independent variables were zero.

β_1 is the effect size (parameter) of the first independent variable.

x_1 is the value of the first independent variable.

β_2 is the effect size of the second independent variable

x_2 is the value of the second independent variable.

ε is the residual error of the model; this is the difference between the predicted outcome and the actual value.

2.4.3 Time-series Forecasting

Time series forecasting is a technique for the prediction of events through a sequence of time. Time-series forecasting decomposes the historical data into the baseline, trend and seasonality if any. Generally, the movement of the data over time may be due to many independent factors. There is a number of time-series forecasting methods. In this research, the author is using three main forecasting models.

2.4.3.1 Holt-Winters' Additive Model

This method produces exponentially smoothed values for the level of the forecast, the trend of the forecast, and the seasonal adjustment to the forecast. This seasonal additive method adds the seasonality factor to the trended forecast, producing the Holt-Winters' additive forecast. Use of an additive model when the magnitude of the data is significant and does not affect its seasonal pattern. There are three basic formulas for the seasonal additive model [35].

$$Level \quad L_t = \alpha(y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1})$$

$$Trend \quad b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

$$Seasonal \quad S_t = \gamma(y_t - L_t) + (1-\gamma)S_{t-s}$$

$$Forecast \quad F_{t+k} = L_t + kb_t + S_{t+k-s})$$

where :

S is the seasonal change smoothing factor $0 < S < 1$

for $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$ and $0 \leq \gamma \leq 1$

L_t is the level at time t

b_t is the trend at time t

S_t is the seasonal component at time t

y_t defined the data value at time t

2.4.3.2 ARIMA Model

Auto-Regressive Integrated Moving Average is a typical time series model for solving forecasting problems which were first introduced by Box Jenkins [35]. ARIMA works well on complex relationships as it takes error terms and observations of lagged terms, making it suitable for the forecasting required in this research. The ARIMA model is one of the forecasting technique of time series, which is based only on the observed behaviour of variable data.

ARIMA models completely ignore the independent variable for this model using the present value and past values of the dependent variables to produce accurate short-term forecasting. The ARIMA model is a combination of the Autoregressive (AR) and Moving Average (MA) models ARIMA) [36].

ARIMA can be described in the general form:

$$(1 - \phi_1 B - \ldots - \phi_p B^p)(1-B)^d X_t = \theta_0 + (1 - \theta_1 B - \ldots - \theta_1 B^q)\, a_t$$

where

B denotes the backward shift operator

d is the order of backwards-difference operator ($1-B$)

X_t is a random variable at instant t.

$\phi 1 \ldots \phi_p$ are autoregressive parameters.

θ_0 is a constant term.

$\theta_1 \ldots \theta_q$ are moving average parameters and a_t is white noise process.

Constant term θ_0 is shown as $\theta_0 = (1 - \theta_1 - \ldots - \theta_p)\mu$

where μ is the mean of the actual values.

ARIMA models used with p,d,q values

where

p is AR parameter (ϕ_p)

d is the order of backwards-difference

q is MA parameter (θ_q) and shown as ARIMA$_{(p,d,q)}$

2.4.3.3 SARIMA Model

The Seasonal Autoregressive Integrated Moving Average Model (SARIMA), with additional parameters P, D and Q is an extension of an Auto-Regressive Integrated Moving Average ARIMA model [37]. The SARIMA model is used when a seasonal behaviour is present in the time series. A SARIMA model is represented as

$$\text{SARIMA}_{(p,\,d,\,q)(P,\,D,\,Q)m}$$

This model adds four new parameters to the original ARIMA model as follows:

P - the seasonal order of the autoregressive part

D - the seasonal order of the differencing part

Q - seasonal order of the moving average part

m - the number of time steps for a single seasonal period [37].

2.4.3.4 Measures of Forecasting Accuracy

The more accurate the knowledge of future network faults are, the better the results of the planning will be. The main challenge of this research work was to find an accurate forecasting technique among a large number of forecasting techniques to forecast future network faults.

This research study proposes to use a ranked matrix of Error Measuring Parameters to determine the best forecasting technique. Forecasting accuracy-test was performed on the selected Error Measuring Parameters [38].

A forecast error is a difference between the forecast and actual value for a given period.

$$E_t = A_t - F_t$$

where

E_t = forecast error for period t

A_t = actual value for period t

F_t = forecast for period t

The error measuring techniques used are as follows [39]:

- The Mean Error is referred to as **ME**
- The Root Mean Absolute Error is referred to as **RMSE**
- The Mean Absolute Error is referred to as **MAE**
- The Mean Percentage Error is referred to as **MPE**
- The Mean Absolute Percentage Error is referred to as **MAPE**
- The Mean Absolute Scaled Error is referred to as **MASE**

ME

The mean error (ME) is the simple average of errors.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)$$

RMSE

RMSE is the square root of the average of squared differences between prediction and actual observation. The RMSE depends on the scale of the dependent variable. It should be used as a relative measure to compare forecasts for the same series across different models. The smaller the error, the better the forecasting ability of that model according to the RMSE criterion.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}}$$

MAE

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. MAE is calculated by taking the absolute value of the difference between the estimated forecast and the actual value at the same time so that the negative values do not cancel the positive values. The average of these absolute values is taken to obtain the mean absolute error. The Mean absolute error MAE is also dependent on the scale of the dependent variable, but it is less sensitive to large deviations than the usual squared loss.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i|$$

The RMSE result will always be larger or equal to the MAE. If all the errors have the same magnitude, then RMSE will equal to MAE.

MPE

The MPE value provides the percentage deviation between the predicted and measured data. Its ideal value is also zero. The percentage error (MPE is the proportion of error at a particular point of time in the series. This measure adds up all the percentage errors at each time point and divides them by the number of time points.

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \hat{x}_i}{x_i}\right)100$$

MAPE

MAPE is the more objective statistical indicator because the measure is in relative percentage terms and will not be affected by the unit of the forecasting series. The closer MAPE approaches zero, the better the forecasting results. The mean absolute percentage error MAPE is the most useful measure to compare the accuracy of forecasts between different items or products since it measures relative performance [40].

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right|100$$

The following recommendations proposed by Hyndman [41] will be used to determine the best forecasting technique to forecast LV and HV faults [41].

- Always calculate forecast accuracy measures using test data that was not used when computing the forecasts.
- Use the MAE or RMSE if all your forecasts are on the same scale.
- Use the MAPE if you need to compare forecast accuracy on several series with different scales, unless the data contains zeros or small values, or are not measuring a quantity.
- Use the MASE if you need to compare forecast accuracy on several series with different scales, especially when the MAPE is inappropriate.

2.5 Data Science in Power and Energy Distribution Industry

Big data Analytics allow the massive amounts of data generated by electronic sensors, smart grid technologies, electricity supply, grid operations, and customer demands to be coordinated, analysed, understood, and effectively utilised. According to [42] Big data Analytics can be used to:

- Develop models and simulations of the electrical grid and infrastructure to improve their reliability, resilience, technology adoption, and energy demand management.
- Predict equipment failures and power outages, allowing utilities to optimise their maintenance budgets.
- Improve the operating efficiency of electrical generation, transmission, and distribution.
- Integrate intermittent power sources (i.e., renewables) more efficiently and effectively.
- Help managers, employees, and consumers to make better decisions, founded on data and empirical investigation, rather than on intuition or past-practice.
- Better target and tailor services to different customers.

In the past few years, energy analytic is emerging lots of research on electricity consumption analysis like consumer segmentation, characterisation, predictions and knowledge extraction from smart meter had been done. The data mining techniques mostly used classification, clustering of electricity demand patterns and cluster analysis of smart metering data [43]. However, not many data science researchers have investigated faults patterns in the distribution network.

2.6 Review of the Existing and Previous Works in the Power and Energy Network Faults Predictions

Fault management is one of the most demanding tasks for distribution network operators. Systems which are able to filter the incoming information burst and supply the operator with diagnosis information and recommendations for remedial actions relieve an enormous burden for the operator. In 2001, Apel, R [44] discussed fault management in electrical distribution networks. In this work, a fault management system is described, which automates the fault localisation and the determination of isolation and restoration measures to relief the operator from these demanding tasks during network failures. Also, Generation of outage reports discussed.

Very interesting research conducted by Al-Aomar et al. [45] presented a Six Sigma approach to reduce the frequency and duration of power interruption in a local utility company. The objective is to improve the overall service, reduce operating costs, and to increase customer satisfaction. A case study of electricity interruption in power distribution at a local utility company is used to illustrate the Six Sigma application. Such improvement in the critical service of power distribution is also expected to reduce the operating and maintenance costs of the utility company along with higher service level and customer satisfaction. The paper has only presented samples of the analyses, and full details were not shown for both confidentiality. This research conducted by Al-Aomar et al. [45] influences the research carried out in this thesis to understand the process of reducing power interruption in DNOs.

As stated in chapter 1, there are five main research areas focused on in this thesis.

- Temporal patterns, Seasonality and the Dynamics of Network Faults.
- Multidimensional fault segmentation.
- Analyses the historical fault data mainly NaFIRS database.
- Analysing correlation and association between external factors and electricity distribution faults.

Review of the previous work has been reviewed under these four research focus areas.

2.6.1 Temporal Patterns and Seasonality Prediction in Electricity Distribution Network Faults

In 2014, Haomin et al. [46] proposed a correlation rules-based distribution network failure prediction method to overcomes the difficulties of sophisticated modelling and prior parameter determination of the previous model-based method. The proposed method is based on extensive historical monitoring data to find out the correlation rules for confidence sets. In this study, Haomin et al. have used time series decomposition method to predict future failures. The results from this research allow the user to take anti-accident measures for the core equipment in advance based on the previous fault events. However, the depth of the analysis is not enough to understand the temporal patterns and seasonality in the fault occurrences. Overall, the proposed method is to predict failures from the historical fault data according to the characteristics of electrical Equipment; so that distribution operators can timely take preventive measures. This study greatly influences the thesis outcome, and it has helped to lay the foundation for this thesis.

Prediction of power outages caused by extreme weather events such as storms, which are highly localised in space and time, is of crucial importance to power grid operators. Tervo, Roope, et al. [47] propose a new machine learning approach to predict the damage caused by storms. The research studied the application of various classifiers to the problem of predicting power grid outages caused by hazardous storm cells. The classification method was based on the characteristics of the storm cell extracted from weather radar images, related ground weather observations, and lightning detection information.

Bai, Yuling, et al. [48] has proposed a short-term distribution network failure prediction method. The proposed method is based on weather and seasonal factors. The research has also analysed fault data to determine the most influential factors. The classification model such as Support vector machine (SVM) algorithm and several meteorological factors, has been used to conduct the regional wise predict the number of failures in the distribution network and establishes sub-region fault classification and forecasting. The researchers found out that the main influence factors are temperature, precipitation, wind and other meteorological factors. However, the number of failure factors are numerous, and many factors cannot be quantified, and in some cases, data cannot be made at present, so the research has faced low forecasting accuracy.

2.6.2 Multidimensional Fault Segmentation in Electricity Distribution Network Faults

Fault segmentation is the process of dividing a fault into groups that reflect similarity among faults in each logical group. The goal of segmenting faults is to decide how to relate to faults in each segment in order to understand the characteristics and similarities. The identified fault segments may assist in better fault modelling and predictive analytics and are also are used to understand the typical behaviour of other fault related factors.

The author has conducted numerous literature searches in electronic databases (e.g. IEEE, Science Direct, Elsevier,etc…) to find any literature related to fault segmentation in the utility industry, but it seems no scientific research has been carried out in this subject. However, the author has done a literature review on various other multidimensional segmentation techniques in other domains.

Traffic accident data are often heterogeneous, which can cause certain relationships to remain hidden. Benoît et al. [49] studied the effectiveness of a clustering technique for identifying homogenous traffic accident types. Author of this research selected latent class clustering as the applied cluster analysis and found that it succeeds in finding various clusters in a heterogeneous traffic accident data set.

The research shows that applying latent class clustering as a preliminary analysis can reveal hidden relationships and can help the domain expert or traffic safety researcher to segment traffic accidents. Furthermore, the research indicated that the traffic accident types, identified by the seven clusters, make sense and add value to subsequent injury analyses.

This Benoît et al. [49] research has dramatically helped to design the proposed Multidimensional fault segmentation framework in this thesis. Because of network fault, dataset and Traffic accident data are both often heterogeneous and share similar characteristics.

2.6.3 Analyses of the Historical Fault Data: Case of NaFIRS database

Electricity companies are under increasing pressure to deliver high-quality customer service. Ever tightening regulatory regimes require fault restoration performance to be improved and maintained at a high level. In 1972, Ford introduced

the National Fault And Interruption Reporting Scheme [50]. Most DNOs use the National Fault and Interruption Reporting Scheme (NaFIRS) which is administered by the Energy Networks Association (ENA). Others use an equivalent system. These systems collect information on the number of Customers interrupted and duration of interruptions to supply. Nevertheless, after numerous literature searches in electronic databases (e.g. IEEE, Science Direct, Elsevier,etc…) author of the research could find very few scientific studies on NaFIRS database. The author could not find any scientific study of electricity distribution faults analysis using NaFIRS database.

Newis, D, et al. [51] has presented an optimising Customer Information and the Fault Management Process and created a decision support tool for use in the Network Operations Centre environment. It provides a systematic way of estimating the time to restore customer supplies with significant improvements inaccuracy. Known as eaSTORM, the system estimates the restoration time of faults under emergency conditions and on regular fault days.

The historical fault data based upon the UK Nafirs (National Fault & Interruption Reporting Scheme) database was used for eaSTORM. This enables enhanced information to be delivered to customers, the media and other interested parties. The decision support tool eaSTORM will assist companies in achieving regulatory goals and assist in the achievement of enhanced customer information and fault management by Giving incident managers a more significant time to deal with overall restoration management.

Blake, Simon Richard, 2010 [52] has used data from NaFIRS database for his PhD thesis, which discussed Methodologies for the Evaluation and Mitigation of Distribution Network Risk. The objective of the research was to gain a deeper understanding of existing sub-transmission and extra-high voltage distribution networks, adopting a systems approach to classifying the inherent causes and consequences of circuit failure. Also, to develop technically accurate models which also reflect the present UK regulatory environment, while being sufficiently versatile to be adapted to different regulatory environments elsewhere, or in the future.

A composite methodology is also developed, to consider combinations of scenarios and combinations of mitigation strategies. The thesis concludes by considering issues likely to affect the extent and possible increase of network risk over the period 2010-2030. Author of this thesis has extensively reviewed the literature

provided in the Blake, Simon Richard, 2010 [52] work and worked on fill any research gap in the field.

According to the SINTEF, which is one of Europe's largest independent research organisations, the project called EarlyWarn use predictive models using big data is being developed based on historic power quality data. Preliminary results show that it is possible to predict with relatively high accuracy events in the power system purely based on the development of power quality data in the period before the actual fault event occurs. The preliminary results from EarlyWarn have been published in two scientific papers, in the AMPS and CIRED (to be published – June 2019) conferences, respectively [96]. Also, they have claimed that the preliminary results are promising and prove that further work should be done on testing different machine learning methods on power quality data, with the aim of increasing the performance and forecast horizon of the predictive models. But full details of the project or detailed finding were not yet released [96].

Joaquim et al. [97] has proposed an intelligent system that predict events using supervisory control and data acquisition (SCADA) and automated metering infrastructure (AMI) systems. The future occurrence of interruptions in distribution transformers is predicted based on the sequence of events priory generated andexogenous variables, such as weather and asset characteristics. In the presented use case, based on real data from a Portuguese utility, the system is able to achieve up to 75% accuracy. But dataset used is highly imbalanced, the proposed system shows promise and adequacy for rare event prediction.

2.6.4 Analysing Correlation and Association Between External Factors and Electricity Distribution Faults

Wang, Li 2017 [53] has presented research describing typical fault causes in the distribution system. The fault in the overhead line and underground are discussed separately. The data in the research are from the surveys by different agencies all around the world. Some typical fault causes (trees, animals, lightning, and vehicles) are subdivided and analysed in detail. Research has identified six types of fault causes in the urban distribution system, including external factors, natural factors, improper maintenance and operation, improper installation, equipment failure and customer

cause. This literature has helped to understand the fault causes in the distribution networks.

Zhanjun et al. [54] have presented a technique that can diagnose distribution network faults based on data mining methods. This technique synthetically analyses the spatial and temporal features of fault data produced in the distribution network. The study has used the APRIORI algorithm to generate association rules. The generated association rules have then been used to establish a strong association between spatial and temporal features. The distribution network fault is then diagnosed by using the association rules. However, this method has some limitation when it is applied to an extensive database such as National Fault and Interruption Reporting Scheme (NAFIRS).

2.7 Conclusions from the Literature Review

This section summarises the conclusions from the literature review to show the gaps identified and how these gaps are addressed in the thesis.

This literature review identified six main research gaps:

a) There has been little discussion about the Data Science adaptation in the utility industry. This literature review has also identified that there has been no comprehensive research which has examined the Data Science process flow.

b) Little attention has been paid to the use of data mining and machine learning techniques for fault analysis and forecasting in the electricity distribution network faults. Despite its high applicability, to the best of our knowledge, up to this date, no research has examined the applicability of data mining and machine learning techniques to forecast future LV and HV faults.

c) Up to this date, no research has investigated the prediction of future LV and HV faults in the electricity distribution network using the National Fault and Interruption Reporting Scheme (NAFIRS).

d) Up to this date, no research has investigated the fault segmentation methodology with the fault data.

e) There has been only little discussion about analysing correlation and association between external factors and electricity distribution faults.

f) There is a lack of study in understanding temporal patterns, seasonality and the Dynamics of Network Faults.

This chapter has reviewed previous studies to discover if any have considered data science adaptation in fault management but identified that, up to this date, no comprehensive scientific research had investigated accurate forecasting and prediction of the future LV and HV faults in an electricity distribution network using the National Fault and Interruption Reporting Scheme (NAFIRS).

In this chapter, literature related to this research was reviewed, and the research gaps in the literature were identified. It was demonstrated that the research questions were derived from the research gaps. Then, the contributions which will be achieved as results of answering the research questions were identified after reviewing the literature. The next step is to design the research. So, the next chapter provides a research methodology which will be applied in this research.

CHAPTER 3

Research Methodology

3.1 Introduction

This chapter outlines the research methodology. It starts by explaining the research design in Section 3.2. In Section 3.3. The data analytic framework used in this research is outlined. Finally, this chapter is summarised in Section 3.4.

3.2 Research Design

As outlined by Saunders et al. [16], the purposes of research could be categorised as exploratory, descriptive and explanatory. An exploratory study can be described as a valuable means of finding out what is happening to seek new insights [16]. It can be particularly useful in helping to understand a problem, clarify the nature of a problem or define the problems involved. It also enables us to develop propositions and hypotheses for further research, to discover new insights or to reach a greater understanding of an issue. The objective of descriptive research is 'to portray an accurate profile of persons, events or situations [16]. It is necessary to have a clear picture of the phenomena before the collection of the data. Descriptive research is planned, and structures. It requires precise specifications of who, what, when, where, why, and how. Explanatory studies that establish causal relationships between variables may be termed explanatory research [16]. The emphasis here is on studying a situation or a problem in order to explain the relationships between variables.

The main research questions this study address is: **"Can data mining and machine learning approaches use to accurately predict and forecast faults in the Electricity Distribution network?"**. This research question is further divided into four sub-questions. This requires establishing the nature of LV and HV faults and identifying appropriate data mining and machine learning techniques (e.g. Time series forecasting, clustering, association rule mining) to address the central research aim. This also requires establishing the relationships between analytic methods and the

performance of the models. Therefore, this research might be in line with an exploratory study as well as an explanatory study.

The research strategy is a general plan of how researchers go ahead and answer the research question that has been set by the researcher. The choice of research strategy will be guided by the research question(s), aim and objectives, the extent of existing knowledge, the amount of time and other resources available, as well as philosophical underpinnings [16]. Research strategies may include experiment, survey, case study, grounded theory, ethnography and practitioner-researcher. These strategies should not be thought of as being mutually exclusive. For example, it is quite possible to use a survey strategy as part of a case study [16].

Robson [55] defines a case study as ‘a strategy for doing research which involves an empirical investigation of a particular contemporary phenomenon within its real-life context using multiple sources of evidence’.

Yin [56] defined a case study as "an empirical inquiry that investigates a contemporary phenomenon within its real-life context". Saunders [16] noted that “a case study strategy is most often used in explanatory and exploratory research”. This research has been stated to be an exploratory study as well as an explanatory study. Therefore, a case study is suitable for achieving these research purposes.

There are several advantages to using case studies. Firstly, the examination of the data is most often conducted within the context of its use [56]. Second, variations in terms of intrinsic, instrumental and collective approaches to case studies allow for both quantitative and qualitative analyses of the data [57]. Third, the detailed qualitative accounts often produced in case of studies not only help to explore or describe the data in a real-life environment but also help to explain the complexities of real-life situations which may not be captured through experimental or survey research [57]. However, the case study method has always been criticised for its lack of rigour and the tendency for a researcher to possibly have a biased interpretation of the data.

Yin [56] writes that a case study can contain either a single study or multiple studies. The researcher, therefore, has to consider if it is wise to construct a single case study or if it is more sensible to construct a multiple case study to help understand the phenomenon. There are several different opinions as to whether a single case

study or a multiple case study is the better choice. According to Baxter & Jack [58], the evidence that is generated from a multiple case study is robust and reliable.

The research strategy of this study is the exploratory case study research that focuses on understanding the dynamics present within single settings. The exploratory case study investigates distinct phenomena characterised by a lack of detailed preliminary research, especially formulated hypotheses that can be tested, and by a specific research environment that limits the choice of methodology [59]. Criteria identification, data collection, and the validation stages of the research have been carried out qualitatively, whereas the data analysis and some parts of conceptualisation stages have been carried out using quantitative techniques.

If a study contains more than a single case, then a multiple-case study is required. This is often equated with multiple experiments. This study has two primary case studies which are one UK DNO and one Australian DNO. Therefore, two separate case studies will be developed.

After having decided on an approach (qualitative, quantitative, mixed-methods), preliminary literature review and a format for the research, the next step in the process is to design or plan the study [60]. Research design is the logical sequence that links the empirical data to a study's initial research questions. That is, the design discourages the situation in which the evidence is disconnected from the initial research questions [56]. Creswell defined Research designs as plans and the procedures for research that span the decisions from broad assumptions to specific methods of data collection and analysis. It involves the intersection of philosophical assumptions, strategies of inquiry, and specific methods [60].

Saunders et al. [16] describe the research process within the concept of a ‘Research Onion’ (illustrated in Figure 3.1) that comprises different layers where each layer of the onion refers to a research aspect beginning from the outer layer of Philosophy and narrowing the research down until the data collection and analysis stage which is the centre layer. Saunders et al. [16] research onion has been used to define the overall research methodology.

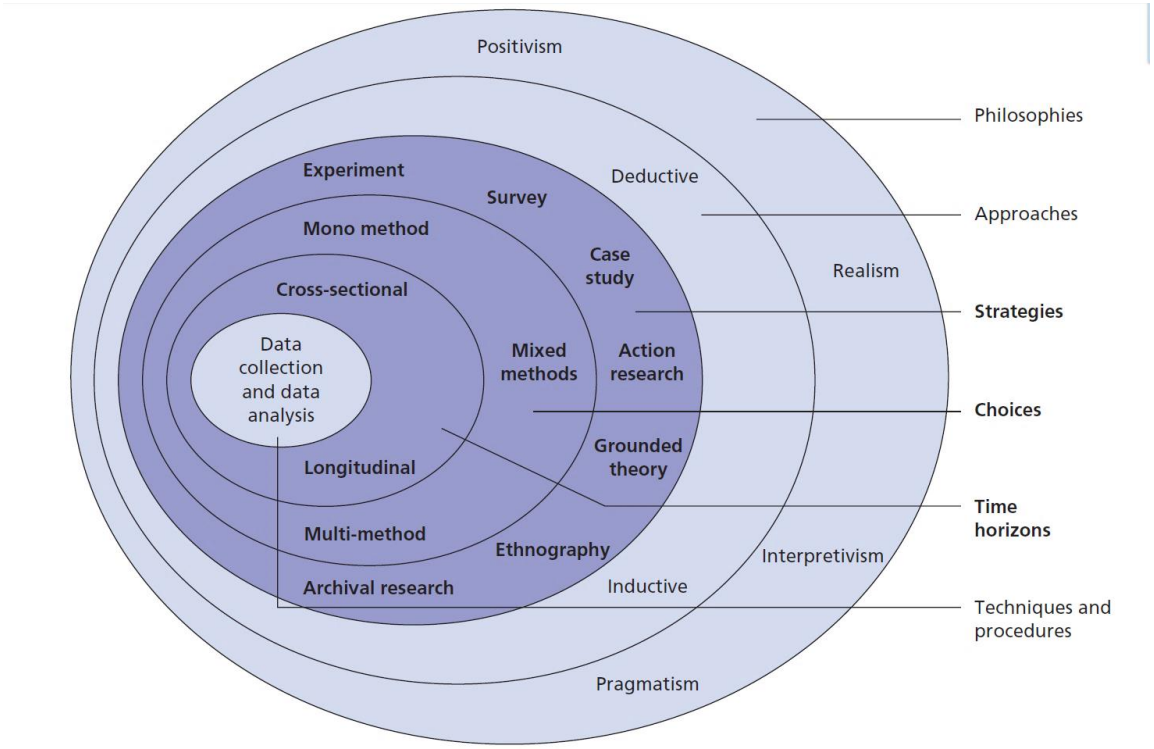


Figure 3.1: Research Onion - Explanation of the research process (Source: Saunders et al. [16])

This study will use the following approach to design the research:

- Philosophy**: Positivism**
- Approach**: Mix of Inductive and Deductive**
- Strategy**: Multiple exploratory case studies**
- Choice**: Multi-method quantitative study**
- Time horizons**: Cross-sectional**
- Techniques and procedures**:**
    - Data collection**: Real-world industrial dataset**
    - Data Analysis**: Statistics, Machine Learning and Data Mining**

In line with the above-introduced research methodology, a research design that the author of this research has developed is implemented throughout the research.

Furthermore, after conducting a comprehensive general literature review in chapter 2 and after conducting inclusive literature review about the research design methodology in chapter 3, the researcher has adopted research philosophy, approach and method are also implemented throughout the research.

The conducted research is mapped out in below Thesis Map (Figure 3.2), which comprises seven main phases. Below diagram shows each phase and relevant and incorporate chapter number. This may help the readers to understand the structure of the thesis and relevant components.

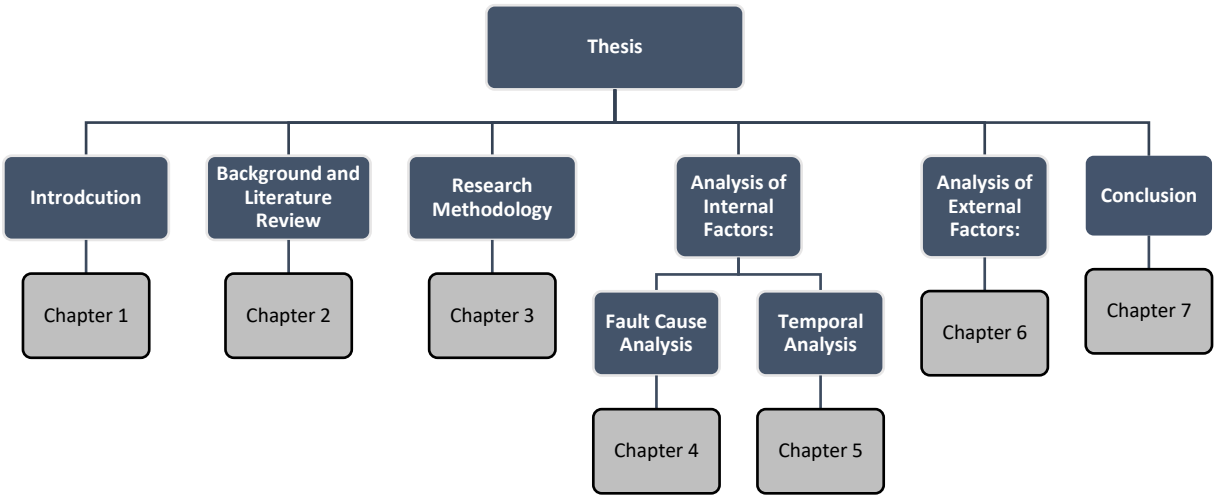


Figure 3.2: Thesis Map: The structure of the thesis and relevant components

3.3 Data Science Project Process Flow

[1]The literature review has identified that there has been no comprehensive research, examined the full Data Science project process flow. Therefore, a new tailor-made Data Science project process flow has been introduced in this section of the research for data acquisition, data fusion, data storing, processing, analysing, and modelling.

This process flow will be used in all the technical chapters in this thesis. One of the primary motivation for using this Data Science project process flow is to streamline the data mining and machine learning modelling process. Without a properly coordinated and structured framework, there is likely to be much overlap and duplication amongst project phases.

The data science field is a combined field of Big Data, Data Mining, Machine Learning and Statistics which seeks to provide meaningful information from large complex datasets. Data science combines different fields of work in statistics and computation in order to interpret data for the purpose of decision making. Data science can be applied to all sorts of organisations; businesses, government, not-for-profit and so on while the application of pattern recognition technology to large datasets has revolutionised the digital economy.

According to the Royal Statistical Society (RSS) in the UK, Data Science (DS) is still a cottage industry with small teams of artisan DS crafting bespoke prototypes to their own standards. Data science projects differ from most traditional data analysis projects because, in comparison to traditional data analysis project, they could be both complex and in need of advanced technologies.

[1] This section has been published in below research papers
Mo Saraee and Charith Silva. 2018. A new data science framework for analysing and mining geospatial big data. In Proceedings of the International Conference on Geoinformatics and Data Analysis (ICGDA '18). Association for Computing Machinery, New York, NY, USA, 98–102. DOI:https://doi-org.salford.idm.oclc.org/10.1145/3220228.3220236. (Appendix 2)

For this reason, it is essential to have a process to govern the project and ensure that the project participants are competent enough to carry on the process. Often DS projects are ill-defined throughout, and project participants might not be sure that application can even be built successfully until a late stage. This thesis presents a new tailor-made Data Science project process flow. Having a good process for data analysis and clear guidelines for comprehensive analysis is always a plus point for any data science project. It also helps to predict the required time and resources early in the process to get a clear idea of the business problem to be solved. Below section will explain each section of the proposed Data Science project process flow.

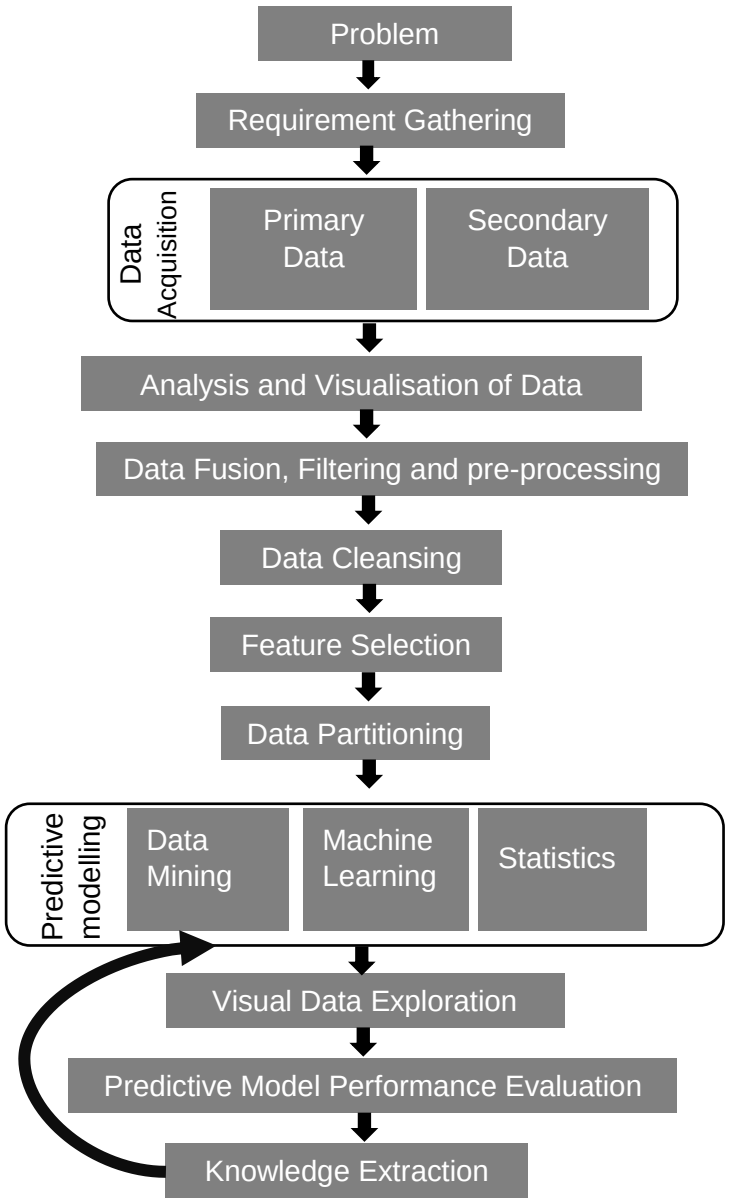


Figure 3.3: A new tailor-made data science project process flow

3.3.1 Problem Definition

A problem definition is a brief description of the subjects that need to be addressed by the data analytics project. In this step, it does not need to describe the approach of addressing the problem. The 5W1H technique can be used to describe the problem. The five W's and the H are acronyms for Who? What? Where? When? Why? And How? It is an excellent means of gathering information methodically in a challenging situation. When considering the problem, it is essential to stay focused and challenge the assumptions.

3.3.2 Requirement Gathering

Gathering requirements is a vital first step for any type of project. Requirement gathering is an integral part of any data analytics project. Poor requirements gathering techniques are the cause of many project failures. Gathering incomplete requirements are the cause of many design flaws. The development of a wide range of requirements early in the project will allow for accurate cost estimates; shorter project time periods; greater customer satisfaction and improved final solution accuracy. It is always best to avoid discussing technology or solutions until the project owners and participants fully understand the requirements. It is essential to create a clear, and complete requirements document to share with all key project participants.

3.3.3 Data Acquisition

The acquisition of complex, resource-rich data set is essential for any data projects and applications. Data collection and database maintenance are the most cost-effective and time-consuming aspects of data analytics projects. Research datasets can be available in different file formats. As a result, it may be challenging to obtain the correct type of data for the project.

Integrated analysis of different types of data from various repositories improves the ability to identify hidden patterns, trends and relationships. Therefore, it is always advantageous to acquire many different related datasets.

3.3.4 Analysis and Visualisation of Data Attributes

Data visualisation helps project participants to understand the data by engaging it in a visual context. Correlations, Patterns, trends, and links that may not be detected in the data are detected and identified more easily if users can use appropriate data visualisation techniques. Data Attributes can be described as data fields that represent the characteristics or features of a data object or dataset. Data attribute visualisation in the early stage of the project can be beneficial throughout the entire project. Early data attribute visualisation can identify any correlations and can easily identify any attributes which have strong correlations amongst themselves. In this stage, the data attributes acquired from the previous data acquisition stage which has been analysed and designed can be visualised it in a way that project participants can take advantage of the analysis without mining deeper into the data. Basic pair-wise scatter plots can be used in this stage.

3.3.5 Data Fusion, Filtering and Pre-processing

The aim of data fusion (data integration) process is to maximise the useful information content acquired by heterogeneous sources in order to infer relevant situations and events related to the observed environment [61]. Data integration from heterogeneous data sources helps to improve the capability to identify hidden patterns, trends, and relationships. Data pre-processing converts data to a format that will be processed more quickly and efficiently.

Data filtering is one of the steps used to explore, filter, and standardize the data before moving on to the modelling process. Data filtering and processing steps may occupy much of the allocated project time. A good understanding of data content is necessary to determine the best way to filter and pre-treat. For example, the untreated NULL value can destroy any modelling activities in the future.

3.3.6 Data Cleansing

Real-world data are often incomplete, inconsistent, and noisy [62]. The purpose of data cleansing is to improve the quality of data that will be processed by the data analyst. High accuracy of the prediction model will be achieved if the data is comprehensive, complete, consistent and accurate. Statistical methods can be used

to treat any missing data. Data cleansing is a tedious process that takes much time and cost, primarily when large amounts of data are used during the data cleansing process. Domain knowledge is vital when carrying out data cleansing activities.

3.3.7 Feature Selection

In the data analytics field, feature selection is the process of selecting a subset of relevant attributes to be used in model building. It provides the mechanism of determining useful patterns in the data, which then decreases the execution time, and decreases the overall size of the data with an improvement in execution performance [63]. Feature selection techniques are primarily intended to improve model prediction performances and runtimes, to reduce model overfitting, and to increase generalisation.

3.3.8 Data Partitioning

Dividing data into training and testing sets is an essential part of building data mining and machine learning models. The training set is used to train or build a model. Once created a predictive model using the training set, it is essential to validate the performance of the predictive model using the new dataset. This data set is known as the test dataset or validation dataset. Poor data partitioning may cause poor inference results. Thus, the data analyst should consider data partitioning methods before building the models data.

3.3.9 Predictive modelling

Predictive modelling is a process that uses data, mathematics and statistics to predict outcomes with data models [64]. Predictive modelling is used to build, test and validate a model to predict the probability of an outcome. This is an iterative process and often involve training the model. Testing the model on the same data set and finally finding the best fit model based on business requirements [65]. Building predictive models are useful in any industry because they provide hidden insight into most of the complex questions they face and allow users to create predictions with a high level of

probability. To maintain a competitive advantage in any industry, organisations must have insight into future events and opportunities that challenge key assumptions.

3.3.10 Visual Data Exploration

Visual analytics combines automated analysis techniques with interactive visualisations leading to sufficient understanding, reasoning and decision making based on enormous and complex datasets [62]. A key difference between data mining and visual data exploration is that visual data exploration is an entirely human-guided process. Visual exploration of the data and the results from the models have been considered as a new application and attracted attention from both the academic and industry communities. Visual analytics methods can be selected in a variety of ways, ranging from simple bar plot too complex geo visualisation plots. Domain knowledge is vital for visual data exploration and will add much knowledge to this step to help understand and interpret the results.

3.3.11 Predictive Model Performance Evaluation

To accomplish the real value of a predictive model, it is vital to know how good the model fits the data. Therefore, performance assessment plays a dominant role in predictive modelling technology. Predictive model performance is calculated and compared by selecting the correct metrics. Therefore, it is vital to choose the correct measurements for a given predictive model to achieve an accurate result. It is also essential to evaluate appropriate predictive models because several types of data sets will be used for the same predictive model. Confusion Matrix and ROC Curve can be used for the necessary Performance Evaluation for given prediction models.

3.3.12 Knowledge Extraction

Knowledge extraction is a complex process allowing the identification of previously unknown structures and potentially useful original information from a large amount of data [66]. Knowledge Discovery is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from extensive data

collections. Knowledge Discovery and Knowledge Extraction is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data.

3.4 Summary of Chapter and Conclusion

This chapter presented a tailor-made methodological research framework just for this research. The objective of these studies is likely to be in line with an exploratory study along with an explanatory study. Since a case study is appropriate for "how" and "what" questions along with exploratory research and explanatory research, which are the cases with this research, this research employed a case study as the research strategy. Theories relevant to these studies can be obtained from the existing theories, even though the necessary theories are not tested with the context of the research. However, theories from similar contexts are most likely, the theoretical foundation of this research.

This research employed the multiple case research design due to the fact this allows more opportunities for multiple experiments and cross observation, and the multiple case design is appropriate for complex cases, and typical cases precisely what are the case associated with these studies.

Finally, the overall research procedure and new tailor-made Data Science project process flow which will be used in technical chapters were described. Having a good process for data mining and machine learning and clear guidelines is always plus point for any data science project. It also helps to focus the required time and resources early in the process to get a clear idea of the problem to be solved. Hence, the process flow is proposed to support data science project lifecycle and bridge the gap with business needs and technical realities.

CHAPTER 4

Fault Cause Analysis and Prediction in Electricity Distribution Network

4.1 Introduction

An in-depth understanding of the cause of fault in electricity distribution networks has always been of paramount importance to Distributed Network Operators in order to achieve a reliable power supply. Faults in the network have a direct effect on its stability, availability, and maintenance. Consequently, the quick elimination, prevention, and avoidance of faults and the causes that generated them are of particular interest. An understanding of the causes and correlation of the factors where future faults may arise can significantly help electricity distribution operators who are accountable for the detection and repair of such faults. Every distribution network asset has a different level of performance and reliability, which may vary depending on environmental conditions. Fault identification in distribution network has a rich literature, but very few studies into using data mining and machine learning techniques have been carried out to understand the factors that contribute to faults.

Given the lack of studies on identifying distribution network faults using data mining, this study will formulate a starting point. This section of the research aims to use association rules mining and clustering techniques to understand the various hidden patterns within the faults database. The uncovered relationships can be represented in the form of Association rules and clusters.

Any component malfunction in the power system causes significant disturbance to the supply or destabilising the entire system [3]. Detecting faults in an electrical power distribution system is a primary importance to both DNOs and consumers. The customer satisfaction survey measures the customer service delivered by DNOs in relation to fault-related incidents [3]. Following an unplanned fault, the DNOs must submit data to OFGEM. OFGEM then pass that information to independent customer

survey organisations [10]. These organisations periodically contact relevant affected customers and ask a series of questions linked to the customers' experience during the power disruption. The survey results will be used to generate a series of performance evaluation reports. Any unplanned outages can harm a DNO's reputation, which might then receive penalties from the regulators.

Financial penalties for power outages can be substantial. It is understood that, where possible, DNOs are targeted to avoid power outages and restore customer confidence more quickly when a power failure occurs [2]. If the electricity distribution company fails to meet the level of service required by OFGEM, individual customers could be entitled to compensation [10]. Therefore, there is a business need to understand the faults and causes of the faults accurately.

In recent years, few researchers have proposed methodologies for fault analysis purely focusing on electric current flow. Nevertheless, this research seeks to introduce a new analysis approach by using data mining techniques. In this research, the author also seeks to establish a relationship between environmental features and fault causes.

From both the DNO and consumer point of view, an in-depth understanding of faults caused in the electricity distribution network is of paramount importance. This new proposed data mining model will help improve the level of system availability by identifying both the cause of any faults and reducing the number of network faults. There is also a business need to reduce operational expenditure in engineering departments. This model may, therefore, help DNOs to avoid faults before they happen. The industry should benefit from the study, through improved levels of system availability by reducing both the network faults and the number of complaints due to fewer network faults and hence avoid the costs and potential fines associated with future network faults from the regulators.

This section of the research used an exploratory cause study methodology to conduct the research. This approach is always helpful to understand a problem, to clarify the nature of a problem and to describe any problems involved. It also enables researchers to develop hypotheses for further research, to discover new insights and to understand the issue from a different dimension.

In this study, the author seeks to understand the most significant factors that contribute to distribution network faults using Association Rules Mining and to explore

the possibility of enhancing the knowledge gain from Association Rules Mining using Term Clustering. Association Rules Mining and Text Clustering techniques were performed to achieve the objectives. The study has addressed one of the main challenges in the Electricity Distribution Industry, which is managing and addressing faults in the network. The outcomes of this research should also help to support policy formulation in engineering departments.

4.2 Analysis of Associations Between Fault Causes using Association Rules Mining and Text Clustering

4.2.1 Case Study: NaFIRS Database from a UK DNO

[2]The dataset used in this research has been extracted from a real NaFIRS database of a DNO in the UK. Due to the nature of commercial sensitivity of the data, some of the data fields are not included in this research study, e.g., postcode, asset numbers, etc.

These data will allow data mining to be carried out more efficiently, and the author is confident that the real industrial dataset will be highly useful to produce real analysis results. The attributes of the data are a mix of numerical and categorical. Data transformation has been applied to some attributes to simplify the analysis. Similarly, other attributes also have been transformed into a suitable form for a better study of the data. Table 4.1 shows the attributes' explanation of the dataset.

In this analysis, the researcher aims to find the association between fault causes. Two of the main variables of the research dataset range from zero to many thousands. This range is quite large. A data discretisation process is critical in this instance because a broad range of continuous variables is being used. It is also vital to categorise attribute variables. Discretisation aims to reduce the number of values a continuous variable assumes by grouping them into several intervals.

[2] This section has been published in below research paper
C. Silva and M. Saraee, "Understanding Causes of Low Voltage (LV) Faults in Electricity Distribution Network Using Association Rule Mining and Text Clustering," 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), Genova, Italy, 2019, pp. 1-6, doi: 10.1109/EEEIC.2019.8783949. (Appendix 2)

Table 4.1: Attributes explanation of the NaFIRS dataset's contents

Attribute	Type	Description
Hour	*Factor*	*Hour of fault occurred*
Weekday	*Factor*	*Weekday of fault occurred*
Month	*Factor*	*The month of fault occurred*
Cause	*Factor*	*Direct Cause*
Equipment	*Factor*	*Equipment Involved*
Components	*Factor*	*Component Involved*
Customers	*Factor*	*No. of Customers*
Minutes_Lost	*Factor*	*Customer Minutes Lost*

Data discretisation is defined as a process of converting continuous data attribute values into a finite set of intervals with minimal loss of information [67]. In this research study, a number of minutes lost and a number of customers affected which have numeric attributes have been discretised by dividing the data into meaningful categories. There are many possible ways to discretise the data, and different choices may result in different discretised data sets. The goal is to optimise between a small number of possible states, in order to limit the size of the model search space, and a sufficiently good resolution of the data to preserve key dynamic features.

A data discretisation technique is able to transform quantitative data into qualitative information. For example, numerical attributes into nominal characteristics with a limited number of ranges, resulting in non-overlapping ongoing domain segregation. Each numerical quantity is then recognised at each interval using the interval limits.

Data are presented in different format such as, numerical, categorical, discrete and continuous. Numerical data, either continuous or discrete, assumes that there is an order among the values. However, in categorical data, no order can be assumed.

The primary variables of the research dataset and their contents are described in the below section.

Cause (Direct Cause) - Cause of the outage is mainly used to describe the responsibility and reason that caused an outage. There are 101 direct causes identified by the NAFIRS. Some samples are:

- *Lightning striking on assets*
- *Snow, Sleet and blizzard*
- *Vermin, Wild Animals and Insects*
- *Trees growing through the lines*
- *Metal Theft*

Equipment - (Main Equipment Involved -MEI). This is that part of the network which has been affected by the fault and is broken down into four main categories, Overhead, Underground Main, Underground Service & Switchgear/Fusegear/Link-box/Cut-out. There are more than 40 different types of equipment identified by the NAFIRS. Some samples are:

- *Overhead Main Insulated Conductors*
- *Overhead Service (Metered) Insulated Conductors*
- *Underground Main Districable*
- *Switchgear/Fusegear*
- *Un-Metered Service Underground*

Component - Component is directly linked to the Main Equipment Involved. It describes the equipment that has faulted. There are more than 50 different types of equipment identified by the NAFIRS. Some samples are :

- *Conductor – the cable*
- *Insulator – this part of the circuit*
- *Jointed Termination Compression*
- *Heat Shrink Termination Pole Mounted*
- *Main Contacts – LV board Jaws*

Customer - This variable indicates the number of customers affected. The discretisation is carried out to replace the raw value of the numeric attribute by the interval levels. Some samples are:

- *No_customer_involved*
- *Only_one_customer_involved*
- *Between_2_and_10*
- *Between_11_and_25*
- *Between_26_and_50*
- *Between_51_and_100*
- *Between_101_and_250*
- *Between_251_and_500*
- *Between_501_and_1000*

Minutes Lost - This variable indicates the number of minutes lost due to the fault. The discretisation is carried out to replace the raw value of the numeric attribute by the interval levels. Some examples are:

- *No_minutes_lost*
- *Between_1_and_15*
- *Between_16_and_30*
- *Between_31_and_60*
- *Between_61_and_120*
- *Between_121_and_360*
- *Between_361_and_720*
- *Between_721_and_1440*
- *More_than_1_day*

4.2.2 Data Cleansing and Quality Assurance on NaFIRS Dataset

The success of data mining and machine learning projects depend on the cleanliness of the data. Real-world data are often incomplete, inconsistent, and noisy [62]. Therefore, data cleansing and quality assurance are paramount to data mining and machine learning projects.

Data Cleansing also referred to as Data Scrubbing, is the action of identifying and then removing or amending any data within a database that is: incorrect, incomplete, duplicated and irrelevant. The purpose of data cleansing and quality assurance is to improve the quality of data that will be used by the data analyst. High accuracy of any prediction model will be achieved if the data are comprehensive, complete, consistent and accurate. Statistical methods can be used to treat missing data. Data cleansing is a tedious process that takes much time and cost, primarily

when large amounts of data are used during the data cleansing process. Domain knowledge is vital for data cleansing activities.

The cleaning process starts with analysing the data for any missing values and any duplicated records. The dataset provided by NAFIRS dataset contained the duplicated records for some network faults. Thus the repeated records were removed using pre-processing data methods. Any missing values were replaced using statistical methods.

4.2.3 Predictive Modelling using Association Rules Apriori Algorithm

The Apriori algorithm is the renowned algorithm to mine association rules. Given a set of transactions, the Apriori algorithm attempts to find subsets that are common to at least a minimum number of the item set. Apriori uses an iterative approach known as a level-wise search where k-itemsets are used to explore (k + 1)-item sets. In summary, the process of extracting association rules from faults datasets using data mining methods involves the following sequential steps:

Step 1: Process the data to a format suitable for association rules mining (association rule algorithms require that input data be in a character format).

Step 2: Extract the most frequently occurring itemsets. Itemsets are a combination of items which occur together in transactions in the dataset. This step is dependent on the minimum support value inputted.

Step 3: The third step involves generating strong association rules from the frequent itemsets identified in step 2. The mining algorithm filters out the rules based on the interestingness measure set by the analyst.

4.2.4 Predictive Modelling using NaFIRS Case Study

The faults dataset consists of more than 50,000 faults, each described by eight attributes. The first experiment would be to identify all the associated factors that contribute to the ***no minutes lost faults***.

The first set of rules were obtained for ***minutes_lost=no_minutes_lost*** in Right Hand Side (RHS) with the support set to 0.01, and confidence set to 0.8. In total, these settings generated 50 rules. Table 4.2 shows the breakdown.

Table 4.2: Breakdown of the association rules generated by the apriori algorithm (support set to 0.01, and confidence set to 0.8)

	2 Items	3 Items	4 Items	5 Items
No. of rules	*1*	*30*	*17*	*2*

Even though the dataset contains over 50,000 faults, and some faults may include rare events, the minimum support threshold must be reduced to identify any less-common faults. Consequently, the second set of rules were obtained with support set to 0.001 and confidence set to 0.8. In total, these settings generated 972 rules. Table 4.3 shows the number of rules. The complexity of the rules increases significantly.

Table 4.3: Breakdown of the association rules generated by the apriori algorithm (support set to 0.001, and confidence set to 0.8)

	2 items	3 items	4 items	5 items	6 items
No. of Rules	*1*	*84*	*484*	*348*	*55*

A scatter plot can be used to visualise the generated association rules. Support and Confidence can be used as X and Y axes. The third measure, Lift, is shown using colour (the red levels) of the points.

Figure 4.2 illustrates the relationship between Confidence, Support and Lift for 972 rules. The optimal rules are those which have high Support, Confidence and Lift. Fig. 4.2 shows that rules with high Lift have low Support, but the 972 rules on the scatterplot distort the graph and the interpretation. It is, therefore, worthwhile to increase the Confidence to reduce the number of rules.

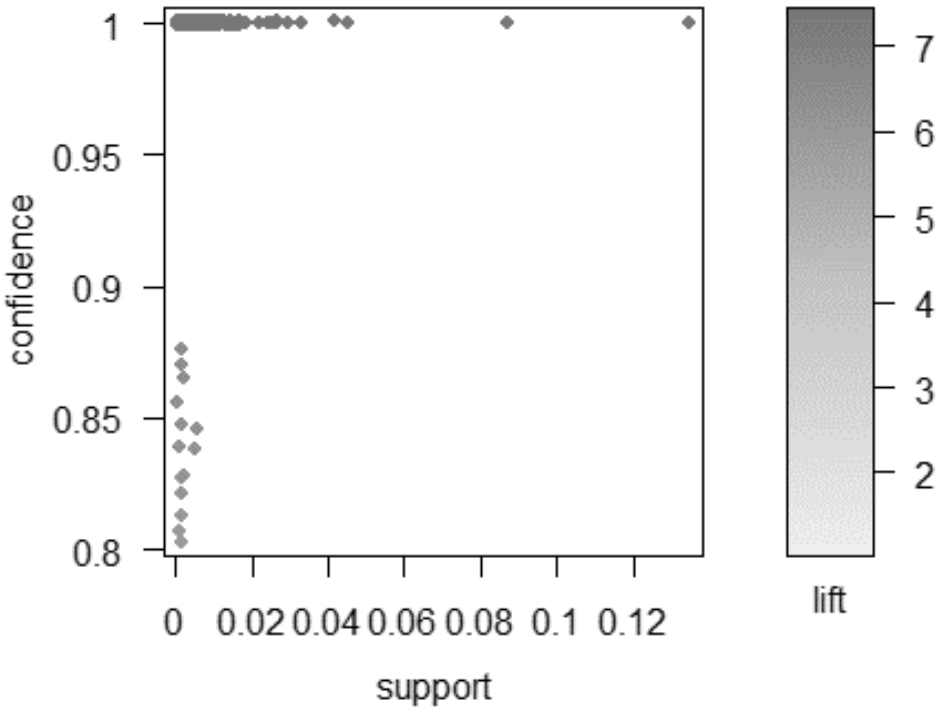


Figure 4.2: Visualising association rules on scatter plot (support set to 0.001, and confidence set to 0.8)

With minimum confidence of 90%, the algorithms produce 958 rules. So, it shows that a number of rules cannot be reduced by increasing the Confidence level. Table 4.4 shows the number of rules generated with different confidence thresholds.

Table 4.4: Breakdown of the association rules with different confidence levels but support set to 0.001

Confidence	2 Items	3 Items	4 Items	5 Items
95%	*1*	*80*	*477*	*345*
90%	*1*	*80*	*477*	*345*
85%	*1*	*81*	*479*	*346*
80%	*1*	*84*	*484*	*348*
75%	*1*	*88*	*487*	*348*
70%	*2*	*97*	*494*	*348*

However, each rule should be examined and finally confirmed by an industry expert. Clearly, there are too many rules to be considered and the effort to reduce the number of rules has failed. Therefore, the researcher is, proposing a new method by combining association rules mining and text clustering to address this issue. Before

explaining this new approach, the author will present ways to group the variables in the next section. Some interesting rules mined using the Apriori algorithm are shown in table 4.5.

Table 4.5: Association rules generated by Apriori algorithm using NAFIRS dataset (support set to 0.001, and confidence set to 0.8)

#	Rules	support	confidence	lift	count
1	*{Weekday=Friday,Cause=by_Highway_Authorities_or_their_Contractors,Components=UnderGround_-_Cable_other_than_joints_&_terminations} => {Minutes_Lost=No_Minutes_Lost}*	*0.002*	*0.8*	*6.41*	*69*
2	*{Weekday=Thursday,Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types,Components=UnderGround_-_Cable_other_than_joints_&_terminations} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.81*	*6.44*	*50*
3	*{Month=October,Cause=by_Highway_Authorities_or_their_Contractors} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.81*	*6.44*	*46*
4	*{Weekday=Friday,Cause=by_Highway_Authorities_or_their_Contractors} => {Minutes_Lost=No_Minutes_Lost}*	*0.002*	*0.81*	*6.46*	*72*
5	*{Weekday=Thursday,Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.82*	*6.51*	*53*
6	*{Weekday=Tuesday,Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types,Components=UnderGround_-_Cable_other_than_joints_&_terminations} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.83*	*6.62*	*58*
7	*{Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types,Components=UnderGround_-_Cable_other_than_joints_&_terminations} => {Minutes_Lost=No_Minutes_Lost}*	*0.006*	*0.83*	*6.63*	*249*
8	*{Weekday=Wednesday,Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types,Components=UnderGround_-_Cable_other_than_joints_&_terminations} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.83*	*6.65*	*45*

9	*{Weekday=Tuesday,Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.84*	*6.69*	*62*
10	*{Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types} => {Minutes_Lost=No_Minutes_Lost}*	*0.006*	*0.84*	*6.7*	*265*
11	*{Weekday=Wednesday,Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.84*	*6.72*	*48*
12	*{Equipment=Unmetered_service_-_Underground,Components=UnderGround_-_Cable_other_than_joints_&_terminations} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.85*	*6.82*	*47*
13	*{Weekday=Friday,Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types,Components=UnderGround_-_Cable_other_than_joints_&_terminations} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.92*	*7.32*	*44*
14	*{Weekday=Friday,Cause=by_Highway_Authorities_or_their_Contractors,Equipment=Underground_Service_(metered)_-_Plastics_insulated_concentric_types} => {Minutes_Lost=No_Minutes_Lost}*	*0.001*	*0.92*	*7.33*	*45*
15	*{Customers=no_customer_involved} => {Minutes_Lost=No_Minutes_Lost}*	*0.125*	*1*	*7.99*	*5296*

4.2.5 Enhance the Results Generated by Apriori Algorithm using Text Clustering

Text clustering is an unsupervised process forming its basis solely on finding the similarity relationship between documents with the output as a set of clusters [68]. Text clustering can be used to automatically group textual documents (for example, documents in plain text, web pages, and emails) into clusters based on their content similarity.

Frequent Term Based Text Clustering proposed by Beil et al. [69] sought to solve the problems of applying conventional clustering methods on text datasets. Such problems included not suitable for high dimensionality large size of the database, and unable to give cluster descriptions. The concept of the frequent term set is based on the frequent itemset of the transaction data set [70]. Figure 4.3 shows steps to enhance the results using Text Clustering.

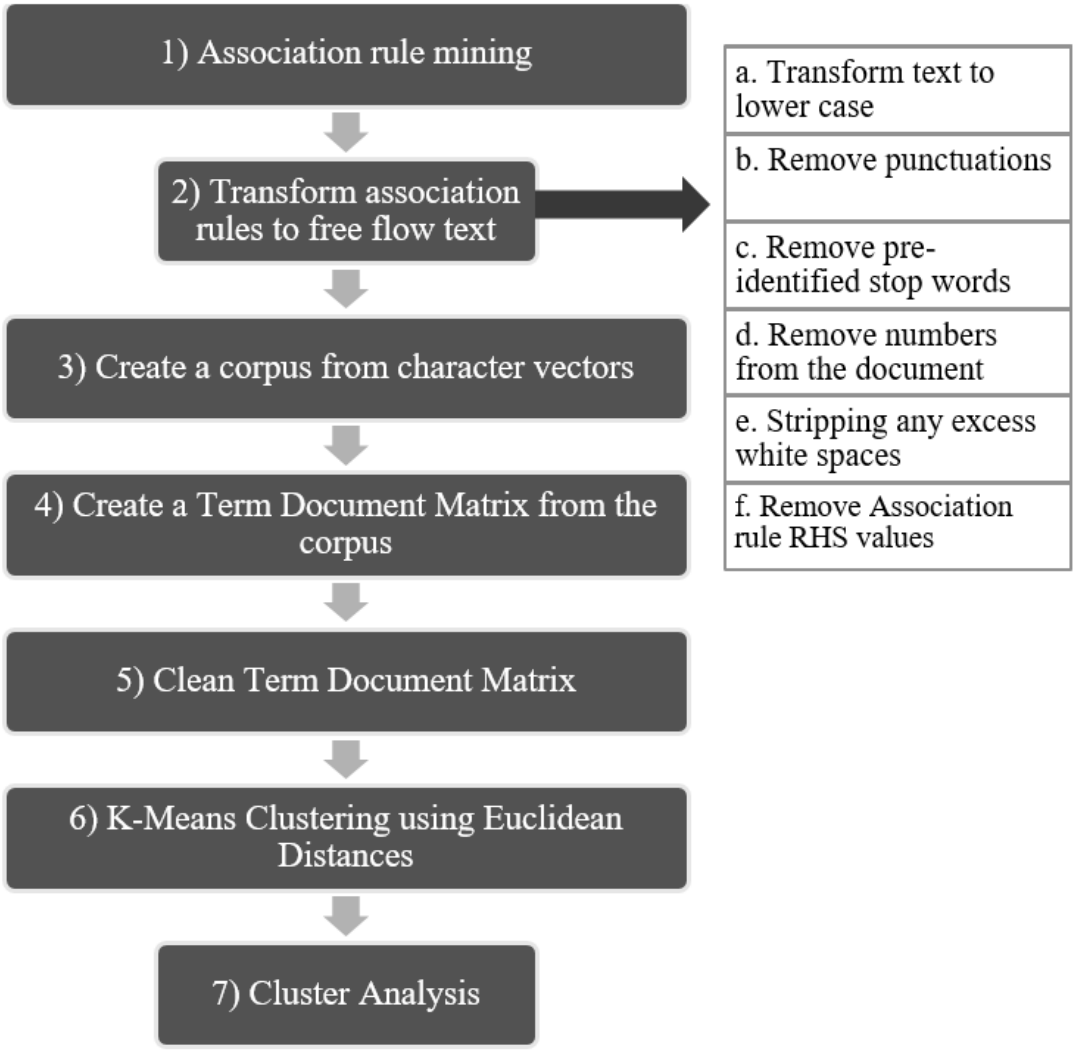


Figure 4.3: Steps to enhance the results generated by the apriori algorithm using text clustering

1) ASSOCIATION RULE MINING

In this section, the same previously generated 972 association rules were used as input. This set of rules were obtained with the support set to 0.001, and confidence set to 0.8.

2) TRANSFORM ASSOCIATION RULES TO FREE FLOW TEXT

The cleaning process of the dataset was carried out through various steps. Generated Text document has a collection of sentences. In this case, a group of

association rules. This step divides the whole statement into words by removing spaces, commas, numbers etc. Stemming is usually also part of this section. Stemming is the process of converting a word to its stem. As in this method, the author seeks to isolate association rules. Stemming has been ignored from the text transformation process. Below are the sub-steps which were required to transform association rules to free flow text.

a. Transform text to lower case

b. Remove punctuations

c. Remove pre-identified stop words

d. Remove numbers from the document

e. Stripping any excess white spaces

f. Remove Association rule RHS values from the text

e.g., *minutes_lost=no_minutes_lost*

3) CREATE A CORPUS FROM CHARACTER VECTORS

A corpus is a collection of texts used for linguistic analyses, usually stored in an electronic database so that the data can be accessed easily [71]. A corpus can be created from various available sources. In this study, the corpus has been created using character vector consisting of one document (One association rule) per element. There are 972 documents in the corpus.

4) CREATE A TERM DOCUMENT MATRIX FROM THE CORPUS

The Term Document Matrix is a matrix that defines the frequency of terms that occur in a collection of documents, and it is a numerical representation of the documents in the corpus. The process of creating a Term Document Matrix from a corpus is an integral part of the text mining process.

Table 4.6 shows part of the Term Document Matrix created using association rules. In the Term Document Matrix, if the text response contains the word or phrase, then the corresponding cell of the table will contain a value of 1. Otherwise, it will contain a value of 0.

Table 4.6: Sample of the term-document matrix created using association rules

	1	2	3	4	5	6	7	8	9	10
cause=by_private_developers_or_their_contractors	0	0	0	0	0	0	0	0	0	0
cause=corrosion	0	0	0	0	0	0	0	0	0	0
cause=deterioration_due_to_ageing_or_wear_(excluding...	0	0	0	0	0	0	0	0	0	0
components=underground_cable__other_than_joints_&_t..	0	0	1	0	0	0	0	0	0	0
customers=no_customer_involved	1	1	0	1	1	1	1	1	1	1
equipment=underground_main_plcs_(armoured_or_unar...	0	0	0	0	0	0	0	0	0	0
equipment=underground_service_(metered)_plastics_ins...	0	0	0	0	0	0	0	0	0	0
equipment=underground_service_(metered)_plcs	0	0	0	0	0	0	0	0	0	0
hour=10	0	0	0	0	0	0	0	0	0	0
hour=11	0	0	0	0	0	0	0	0	0	0

5) CLEAN TERM DOCUMENT MATRIX

The Term Document Matrix is output as a sparse matrix since many values are likely to be zero. This step examines each row of the matrix and determines if all the values are zero. If they are, remove the row from the matrix.

6) K-MEANS CLUSTERING USING EUCLIDEAN DISTANCES.

Several algorithms have been proposed in the literature for clustering. The k-means clustering algorithm is the most commonly used because of its simplicity and relatively simple to implement, and it can scale to large data sets [95]. Also, has been used in many studies and produced reliable results.

But centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Consider removing or clipping outliers before use k-means clustering algorithm. K-means clustering, as a generic algorithm for finding groups or clusters in multivariate data, has found wide application in biology, psychology and economics. Therefore, the author has used the K-means clustering algorithm in this study.

K-Means clustering algorithm is implemented in six steps:

a) Choose a value of k (number of clusters to be formed).
b) Partition objects into k nonempty subsets.
c) Randomly select k data points from the data set as the initial cluster centroids.
d) For each data point, compute the distance between the data point and the cluster centroid.
e) Assign the data point to the closest centroid
f) Repeat d & e steps until the mean of the clusters stop changing.

In this study, k=3 has been used.

Figure 4.4 shows the clusters created by the K-Means clustering algorithm using the previously generated term-document matrix. Cluster 2 and 3 have only one item, and cluster 1 has seven items. The items included in the clusters are shown below.

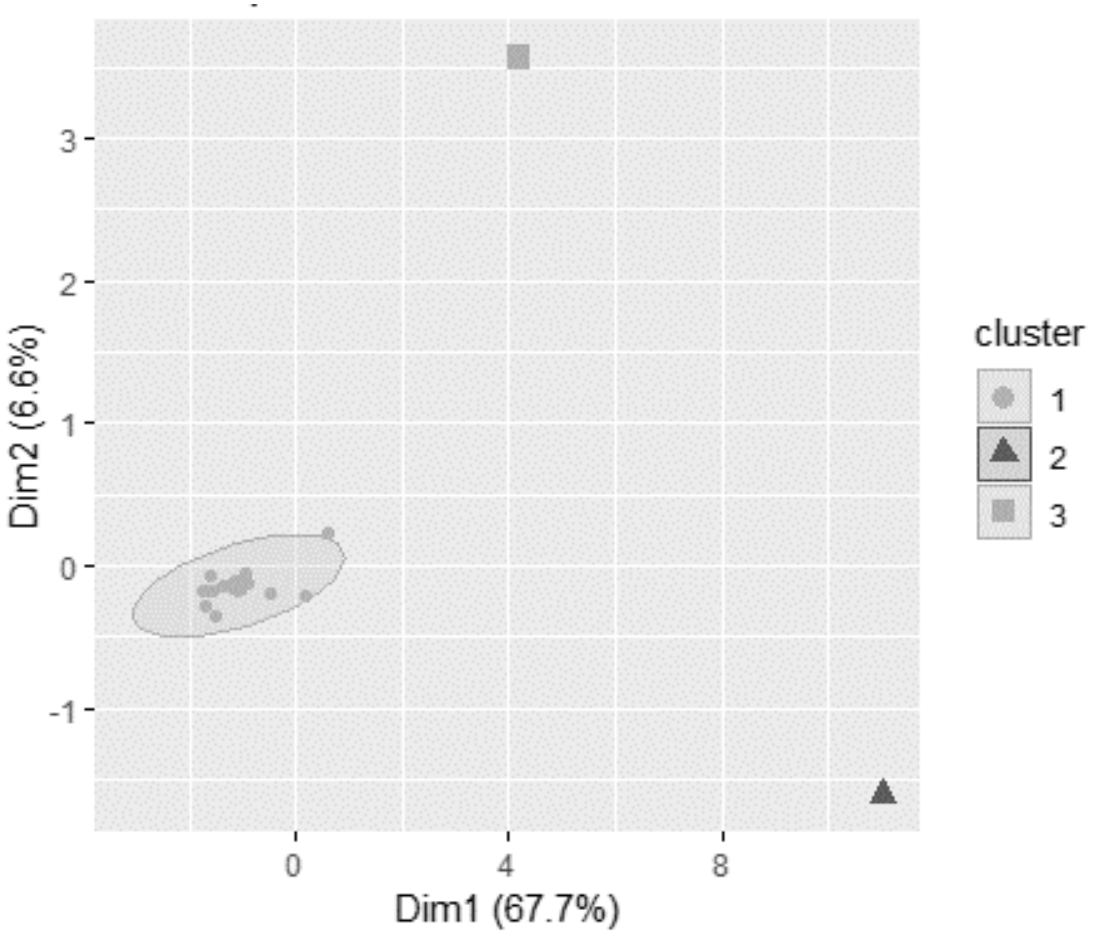


Figure 4.4: Cluster plot created by the K-Means clustering algorithm using the previously generated term-document matrix

Items in the cluster 1

- *cause=by_private_developers*
- *cause=corrosion*
- *cause=deterioration_due_to_ageing*
- *equipment=underground_main_plcs*
- *equipment=underground_service_(metered)*
- *hour=10*
- *hour=11*
- *hour=14*
- *weekday=monday*
- *weekday=tuesday*
- *weekday=wednesday*
- *weekday=thursday*
- *weekday=friday*

Items in the cluster 2

- *customers=no_customer_involved*

Items in the cluster 3

- *components= underground_cable_other_than_joints &_terminations*

7) CLUSTER ANALYSIS

Cluster analysis is an exploratory analysis method that attempts to identify a similar group of a data point within a dataset. More specifically, cluster analysis is a unique purpose technique that is used to classify objects into related groups called clusters. More importantly, the researcher must be able to interpret the group classification based on their understanding of the data points to determine whether the results of the analysis are essential.

Text clustering has been used to group similar documents and make meaningful groups. In this study, text clustering has been used to group factors that are contained in the association rules. According to Association Rules mining and text cluster analysis, most of the *no minutes lost* LV faults happened due to *underground cable other than joints &_terminations* and those faults had not impacted any customers.

4.2.6 Results Validation

A set of rules were obtained for ***minutes_lost= more_than_1_day*** in RHS with the Support set to 0.001, and Confidence set to 0.8. In total, these settings generated 1384 rules. As Table 4.7 shows, the number of rules and the complexity of rules increases significantly.

Table 4.7: Breakdown of the association rules for minutes_lost= more_than_1_day in RHS

	2 Items	3 Items	4 Items	5 Items	6 Items
No of rules	*3*	*161*	*715*	*424*	*81*

Some interesting rules mined using the Apriori algorithm are shown in table 4.8.

Table 4.8: Association rules generated by Apriori algorithm using NAFIRS dataset for minutes_lost= more_than_1_day in RHS (support = 0.001 and confidence = 0.8)

#	Rules	support	confidence	lift	count
1	*{Hour=12,Cause=Deterioration_due_to_Ageing_or_Wear_(excluding_corrosion),Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.002*	*0.8*	*2.14*	*136*
2	*{Month=August,Cause=Corrosion,Equipment=Underground_Main_Consac} => {Minutes_Lost=More_than_1_Days}*	*0.001*	*0.8*	*2.14*	*88*
3	*{Hour=7,Cause=Cause_Unknown,Equipment=Underground_Main_PLCS_(armoured_or_unarmoured),Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.002*	*0.8*	*2.14*	*128*
4	*{Hour=8,Cause=Deterioration_due_to_Ageing_or_Wear_(excluding_corrosion),Equipment=Underground_Main_PLCS_(armoured_or_unarmoured),Customers=BETWEEN_11_AND_25} => {Minutes_Lost=More_than_1_Days}*	*0.001*	*0.8*	*2.14*	*96*
5	*{Hour=12,Cause=Deterioration_due_to_Ageing_or_Wear_(excluding_corrosion),Equipment=Underground_Main_PLCS_(armoured_or_unarmoured),Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.001*	*0.8*	*2.14*	*92*
6	*{Weekday=Tuesday,Cause=Cause_Unknown,Equipment=Underground_Main_Consac,Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.001*	*0.8*	*2.14*	*88*

7	*{Weekday=Monday,Equipment=Underground_Main_Mixed_or_unclassified,Components=UnderGround_No_component_identified,Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.003*	*0.8*	*2.14*	*200*
8	*{Weekday=Monday,Cause=Cause_Unknown,Equipment=Underground_Main_Mixed_or_unclassified,Components=UnderGround_No_component_identified,Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.003*	*0.8*	*2.14*	*200*
9	*{Weekday=Tuesday,Month=June,Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.001*	*0.8*	*2.14*	*109*
10	*{Month=June,Equipment=Underground_Main_Mixed_or_unclassified,Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.001*	*0.8*	*2.14*	*105*
11	*{Hour=4,Cause=Deterioration_due_to_Ageing_or_Wear_(excluding_corrosion)} => {Minutes_Lost=More_than_1_Days}*	*0.003*	*0.8*	*2.14*	*206*
12	*{Weekday=Saturday,Month=February,Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.001*	*0.8*	*2.14*	*93*
13	*{Month=January,Equipment=Underground_Main_Consac,Components=UnderGround_Main_joints_other_than_1_above} => {Minutes_Lost=More_than_1_Days}*	*0.001*	*0.8*	*2.14*	*81*
14	*{Weekday=Thursday,Month=December,Cause=Cause_Unknown,Components=UnderGround_No_component_identified,Customers=BETWEEN_26_AND_50} => {Minutes_Lost=More_than_1_Days}*	*0.001*	*0.8*	*2.14*	*81*
15	*{Weekday=Friday,Equipment=Underground_Main_Consac,Components=UnderGround_Main_joints_other_than_1_above} => {Minutes_Lost=More_than_1_Days}*	*0.002*	*0.8*	*2.14*	*150*

According to the Association Rules mining and text cluster analysis, most of the LV faults that caused an outage more than one-day affected more than 50 customers but less than 100 customers. A cluster plot is shown in Figure 4.5.

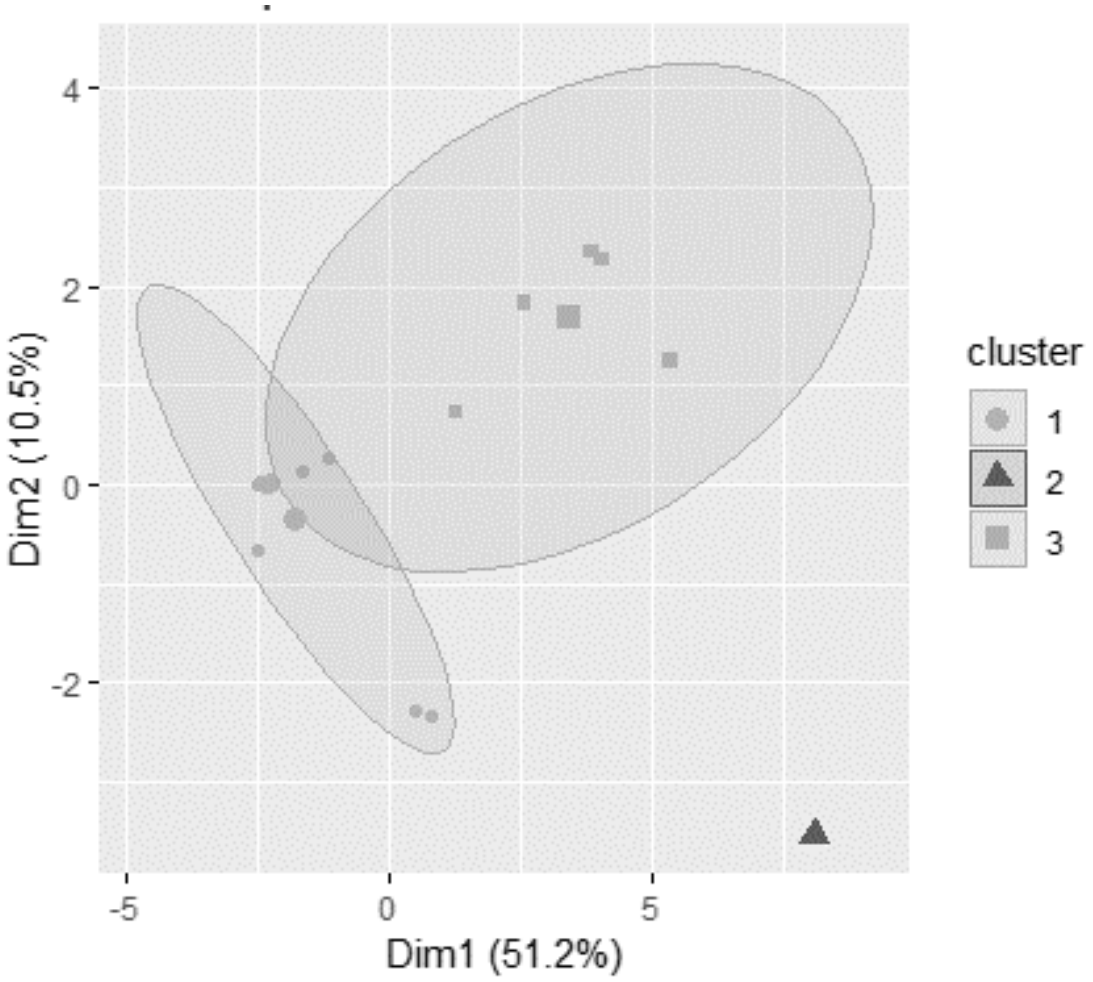


Figure 4.5: Cluster plot created by the K-Means clustering algorithm using the previously generated term-document matrix for minutes_lost= more_than_1_day in RHS

4.2.7 Results Analysis and Discussion

This study seeks to build a model which can be used to analyse, understand and predict the potential causes for low voltage network faults. This research, therefore, proposed a new method that analyses historical fault data and seeks to understand the impact of the fault with other factors such as Main Equipment Involved, Component, and Direct Cause. This proposed data mining model may also be used to safeguard the electrical power distribution system key equipment which may be damaged or destroyed by some future faults.

In this research, Association rules mining has been used; Association Rules mining typically use when it requires to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository. In this study association rules mining provided an understanding into the patterns within the dataset which are correlated with the number of customer minutes lost. The Apriori algorithm generates if-then rules which show the probability of relationships within the dataset. As described previously in chapter 2.4.1.2, the association rules are comprised of two parts, and these include and antecedent and a consequent. The antecedent is the item which is found within the dataset, and the consequent is the combination found with the antecedent. In some scenarios antecedent called as a left-hand side or LHS and consequent called as a right-hand side or RHS of the rule.

The first set of rules were obtained for the condition of **minutes lost=no minutes lost** in Right Hand Side (RHS) with the Support set to 0.001, and Confidence set to 0.8. In total, these conditions generated 972 rules. The second set of rules were obtained for the condition of **minutes lost= more than one day** in RHS with the Support set to 0.001, and Confidence set to 0.8. In total, these settings generated 1384 rules.

Clearly, there are too many rules to be considered and the effort to reduce the number of rules by increasing Confidence has been failed. Therefore, the researcher is, proposing a new method by combining association rules mining and text clustering to address this issue. Text clustering is an unsupervised process forming its basis solely on finding the similarity relationship between documents with the output as a set of clusters. Text clustering can be used to automatically group textual records into clusters based on their content similarity. For the low voltage network fault, which resulted in no customer minutes lost, text clustering process has identified three main groups (Clusters).

The first group identified is, low voltage network fault which resulted in no customer minutes lost not involved any customers, but it is obvious that it is highly likely that fault that has no customer minutes lost will not affect any customer. The second association identified, is low voltage network fault, which resulted in no customer minutes lost, highly likely caused by faults in the underground cable other

than joints and terminations. The third group have multiple associated factors, shown below:

- Fault cause is by private developers, corrosion and deterioration due to ageing
- Types of equipment involved are underground main and underground service (metered)
- The fault occurred hours are 10, 11,14
- The day of the week is either Monday, Tuesday, Wednesday, Thursday or Friday

Association rules mining provided a deeper understanding of the underlying relationships which occurred within the dataset, thus leading to a deeper understanding of the low voltage network fault causes. However, due to the majority of network fault within the dataset having **minutes_lost=no_minutes_lost** and **minutes_lost= more_than_1_day** in RHS, rules could not be obtained for other conditions. Therefore further research should include a well-balanced dataset in order to gain a much deeper understanding of all levels of network faults.

Experimental results showed that the proposed method could identify significant underlying relationships among low voltage network fault causes. The model can be further improved by applying advanced association rules mining and clustering algorithms.

4.3 Electricity Distribution Network Fault Segmentation using Clustering-Based Segmentation Techniques

[3]Infrastructure systems are the backbone of modern economies, and critically infrastructure resilience is essential to sustainable development. Disruption of electricity supplies could have serious adverse effects on the performance and security of the economy and citizens' everyday life. DNOs are committed to providing a safe, reliable and affordable network to deliver energy to customers. It is of paramount importance to understand network faults patterns and trends in order to be able to predict and, wherever possible, prevent them. This section of the research presents a new framework for electricity distribution network fault segmentation. Electricity distribution network operators (DNOs) can use this proposed framework to analyses and understand the network faults better.

According to the statistics from Ofgem, the UK annual actual average minutes lost per customer per year is 37. In comparison internationally, the UK has an above-average continuity of electricity supply. However, since the beginning of 2015, customer interruptions have fallen by 11%, and the duration of interruptions has fallen by around 9% [10].

Fault segmentation is the practice of dividing a fault into groups that reflect similarity among faults in each logical group. The goal of segmenting faults is to decide how to relate to faults in each segment in order to understand the characteristics and similarities. The identified fault segments may assist in better fault modelling and predictive analytics and are also are used to understand the typical behaviour of other fault related factors.

The main aim of fault segmentation is to understand failures better and to use that understanding to improve performance, reliability and availability of the distribution network. Although segmentation is widely used in the customer space, there are significant opportunities for DNOs to enhance their approach to fault segmentation. In particular, by adding different fault measures and by expanding the use of

[3] This section has been published in below research paper
C. Silva and M. Saraee, " A New Electricity Distribution Network Fault Segmentation Framework" 2020 IEEE International Energy Conference (ENERGYCON), Tunisia.

segmentation beyond traditional methods. This study develops a framework which can be used to apply exploratory data mining techniques to fault data to extract advanced analytic insights. That new knowledge would aid the understanding of network faults and provide evidence-based information that can then be used by DNOs in policymaking and network design.

Comprehensive analysis of fault types and fault causes in the electricity distribution network is absolutely essential for DNOs. Furthermore, there is a business need to reduce operational expenditure due to network faults. Faults segmentation has the potential to allow engineers to address each fault segment most effectively and understand the interrelationship between failures. Having a large amount of data available on a faults database, a fault segmentation analysis allows engineers to identify discrete groups of faults with a high degree of accuracy based on the equipment and components involved and other indicators. In the fault segmentation framework, cluster examination uses a mathematical model to find sets of similar faults predicated on locating the smallest modifications among failures within each group.

Accurately segment faults using clustering in order to identify discrete groups of faults with a high degree of accuracy based on equipment and components involved. Enhance the accuracy of the fault segmentation by treating outliers. Develop a framework for electricity distribution network fault segmentation. Fault segmentation is an essential tool for developing business intelligence in fault management department and maintaining competitive advantage among DNOs. The use of knowledge gain from fault segmentation will develop clarity among fault analysis. Appropriate remedial action can be developed for each identified fault segment. It may eventually reduce the number of customer complaints due to fewer network.

In the electricity distribution industry, fault data are gathered from engineers on the ground, IoT sensors and automatically triggered alarms. With so much data being collected by a fault management database system so quickly, it can be challenging to determine the right fault analysis dimensions, which metrics to create, and which relationships matters. Fault segmentation can be a useful tool to simplify fault analysis. Segmenting faults starts with grouping together network faults with similar qualities.

However, if the segmentation process uses outdated, duplicate or erroneous data, inaccurate segmentation will result. Hence, data preprocessing is an essential segmentation task. In this study, the author has used different clustering methods to

achieve the research objective. The author has also studied the well-established customer segmentation process, and its characteristics to understand the segmentation process and have applied relevant customer segmentation principles when establishing fault segmentation.

4.3.1 Customer Segmentation

Customer segmentation is one of the essential tasks for every business. With increasing competition in the business and too many options for customers in the market, a significant activity for every industry (including energy companies) is to categorise their customers and target them accordingly.

In the past, there was little customer segmentation and collection of data because manual methods were both times consuming and expensive [72]. However, with the development of big data technologies, there is now a great deal of data available which can be used to model customer demographics and behaviours and segment them accordingly. Customer segmentation is the process by which customers are divided into different groups based on their characteristics in order to derive potential value to the business. By focusing on specific market segment instead of all customers, companies can ideally, quickly modify their products and services to meet the needs of those customers and even design future products and services with them in mind.

Segmenting the customers based on their profiles or buying patterns is a good starting point, but no segment should be used for commercial activities before validating using their most recent data because customer buying patterns and profile may change rapidly. Many companies (including energy companies) are now using segmentation as one of the critical tools for predicting customer behaviour because this allows the business to communicate, engage and interact with their customers in more effective ways.

4.3.2 Customer Segmentation using Clustering

Customers may vary in terms of buying patterns, profiles, needs, and characteristics. Clustering techniques expose internally homogeneous and externally heterogeneous groups from the dataset. The main goal of using clustering techniques

for customer segmentation is to identify different types of customers and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently. Both hierarchical and partitioning clustering algorithms are widely used in customer segmentation, the most prominent among them being K-Means and Agglomerative Hierarchical Clustering [72].

Cluster analysis is an unsupervised data mining strategy employed in the procedure of separating data objects predicated on similarities to other data points and dissimilarities with data points grouped into various other clusters. It includes grouping a couple of objects into different clusters in a way that things in each cluster show similarities among themselves but are dissimilar to things in other clusters. Clustering is a beneficial exploratory data mining strategy as its application can bring about the unravelling of previously unidentified groups within a dataset.

In this study, the clustering technique is used to group LV faults into clusters using a distance measure that calculates nearness to a cluster mean based on attribute values. Hopkins statistic is used to assess the clustering tendency of a dataset by measuring the probability that a given dataset is generated by a uniform data distribution [17]. If the value of Hopkins statistic is close to zero, then the null hypothesis can be rejected, and it can be concluded that the dataset is significantly a clusterable data [17].

Clustering is a very effective method of market segmentation, which helps to identify item classes so that items in a cluster are grouped with familiar characters. The goal of cluster analysis is to sort events, individuals or patterns into groups so that the level of association among members of the similar group or cluster is strong. Clustering is one of the statistical methods used to identify items that are homogeneous or common in nature; each item within a group will have a structure that is inherent but different from another. Clustering uses the “distance measure” to compute the distance between pair of items including the Euclidian distance or Manhattan distance. An excellent method of clustering has a high similarity level within the cluster and a low inter-cluster level. There are different types of clustering methods which includes -

- Hierarchical clustering
- Density-based clustering

Partitioning methods Fuzzy clustering

- Model-based clustering

The two significant methods of clustering are Partitioning methods and Hierarchical methods.

With the introduction of the smart meter in the electricity industry, clustering has been extensively used to perform customer segmentation based on customers half-hourly power consumption profiles. Different clustering algorithms have been used, such as dynamic, hierarchical, and k-means clustering. The characteristic profile for each cluster is obtained by averaging the profiles of the cluster's members. Once classes are created, cluster assignment can be predicted from the covariate vector by applying any classification algorithm.

4.3.3 Case Study: NaFIRS Database from a UK DNO

The dataset used in this study has been extracted from a real NaFIRS database of a DNO in the UK. Due to the nature of commercial sensitivity of the data, some of the data fields are not included in this research study, e.g., postcode, asset numbers, etc.

Data discretisation has been applied to some attributes to simplify data analysis (Fault weekday, Fault time, etc.). Similarly, other attributes have been transformed into an appropriate form for a better analysis of the data. Table 4.9: shows the attributes' explanation of the dataset.

Table 4.9: Attributes explanation of the NaFIRS Database contents used for Electricity Distribution Network Fault Segmentation

Attribute	Type	Description
Direct Cause	*Factor*	*Direct Cause*
Equipment	*Factor*	*Equipment Involved*
Components	*Factor*	*Component Involved*
Hour	*Factor*	*Fault occurred hour*
Weekday	*Factor*	*Fault occurred weekday*
Month	*Factor*	*Fault occurred month*
Customers	*Factor*	*No. of Customers Interrupted*
Minutes_Lost	*Factor*	*Customer Minutes Lost (CML)*

- *Direct Cause* - is the reason that causes the outage/interruption. Approximately a hundred direct causes identified by the NAFIRS schema [10].
- *Equipment* - is the central hardware part of the network which has been affected by the fault. There are about 40 different types of equipment that have been recognised by the NAFIRS schema [8][10].
- *Component* - is the main component in the central hardware part of the network which has been affected by the fault. There are about fifty different types of equipment identified by the NAFIRS schema [8][10].

4.3.4 Data Transformation on NaFIRS Database Contents used for Electricity Distribution Network Fault Segmentation

In the data transformation phase, the first steps taken were to clean the data in the research dataset. Irrelevant columns that were not needed for the modelling stage were removed before filtering through the data to find missing values. Once irrelevant columns were excluded, the data was scanned for any missing values.

Before proceeding to build the model, it is necessary to clean the data and make any necessary changes so that it will be ready to give accurate analytical results. To increase the performance of the model, it is also crucial to transforming some of the dataset attributes. The full research dataset includes data relating to around 50,000 faults. Fewer than 500 data entries with missing values were found, which accounts for 1% of the given data set.

Due to the percentage being small, these data entries were eliminated from the data set. As part of the data transformation, the dataset was grouped by the Direct Cause of the faults, and the aggregated dataset for cluster analysis shown in Table 4.10 was created.

Table 4.10: Sample of aggregated NaFIRS dataset for cluster analysis

Direct Cause	Count	Average Customers Involved	Average Minutes Lost
by Private Developers	*x1*	*y1*	*z1*
Corrosion	*x2*	*y2*	*z2*
Deterioration due to Ageing or Wear	*x3*	*y3*	*z3*

Following data transformation and aggregation, the extracted dataset is comprised of 51 rows of data objects and four attributes. It contained data for aggregated direct fault cause for the period of five years. Table 4.11 shows the fault types and assigned label used for this analysis.

Table 4.11: Fault type and assigned label for the NaFIRS dataset

Label	Fault Cause
F1	*Accidental Contact, Damage or Interference by DNOC or their Contractors (including Live Line Work)*
F2	*Birds (including Swans and Geese)*
F3	*By Cable TV Companies or their contractors*
F4	*by Gas Company or their Contractors*
F5	*by Highway Authorities or their Contractors*
F6	*by Local Building Authorities or their Contractors*
F7	*by Other Third Parties*
F8	*by Private Developers or their Contractors*
F9	*by Private Individuals (excluding 49 and 56)*
F10	*by Public Telecommunications Operator or their Contractors (e.g. BT or Mercury)*
F11	*by Unknown Third Parties*
F12	*by Water/Sewage Company or their Contractors*
F13	*Cause Unknown*
F14	*Causes Unclassified in this Table*
F15	*Condensation*
F16	*Corrosion*
F17	*Deterioration due to Ageing or Wear (excluding corrosion)*
F18	*DNO Equipment affected by Private Generator or Authorised Electricity Operator (other than National Grid)*
F19	*Falling live trees (not felled)*
F20	*Farm and Domestic Animals*
F21	*Fault on Equipment Faulting Adjacent Equipment*
F22	*Faulty Installation or Construction by DNOC Staff*
F23	*Faulty Manufacturing, Design, Assembly or Materials*
F24	*Fire not due to Faults*
F25	*Flooding*
F26	*Ground Subsidence*
F27	*Growing or Falling Trees (not felled)*
F28	*Growing Trees*
F29	*Inadequate or Faulty Maintenance*
F30	*Inadequate Rupturing or Short Circuit Capacity*
F31	*Incorrect Application of DNOC Equipment*
F32	*Incorrect or Inadequate System Records, Circuit Labelling or Identification*
F33	*Incorrect or Unsuitable protection Settings or Fuse Rating*
F34	*Interruption to remove local generator or restore temporary connection (wherein use > 18 hours)*
F35	*involving Farm Workers or Farm Implements*
F36	*Lightning*
F37	*Load Current above Previous Assessment*
F38	*Local Generation Failure (Isolated System)*

F39	*Mechanical Shock or Vibration*
F40	*Metal Theft*
F41	*Operational or Safety Restriction*
F42	*Rain*
F43	*Snow, Sleet and Blizzard*
F44	*Switching Error by DNOC Personnel*
F45	*Transient Fault - No Repair*
F46	*Unsuitable Paralleling Conditions*
F47	*Unsuitable Protection Characteristics*
F48	*Vermin, Wild Animals and Insects*
F49	*Wilful Damage, Interference (not including Metal Theft)*
F50	*Wind and Gale (excluding Windborne Material)*
F51	*Windborne Materials*

Due to the commercial sensitivity of the data, actual fault count, the number of customers affected, and the number of customer minutes lost can not be displayed. However, Table 4.12 below is a tabular presentation of a sample of the top ten of the ranked dataset and its percentage from the total data points.

Table 4.12: Top ten rows of the ranked NaFIRS dataset and their percentage

Rank	Direct Cause	Fault Count %
1	*Deterioration due to Ageing or Wear (excluding corrosion)*	*35%*
2	*Cause Unknown*	*28%*
3	*Corrosion*	*13%*
4	*by Private Developers or their Contractors*	*5%*
5	*Transient Fault - No Repair*	*3%*
6	*by Private Individuals*	*2%*
7	*by Local Building Authorities or their Contractors*	*1.50%*
8	*by Highway Authorities or their Contractors*	*1.50%*
9	*by Unknown Third Parties*	*1%*
10	*Wind and Gale*	*1%*

4.3.5 Electricity Distribution Network Fault Segmentation

Distribution network operators have financial incentives to minimise the number and duration of interruptions, including those caused by climate impacts. Ofgem introduced the Interruptions Incentive Scheme (IIS) in April 2002 [10]. Under this scheme, distribution companies are set a target for the number of interruptions each year. If they achieve these targets, they are rewarded. Equally, they are penalised if

they do not achieve their targets [10]. In order to achieve these Ofgem targets, network faults need to be analysed to identify possible risk factors and their effects on faults severity levels.

A fault segmentation method can be used to identify homogenous network fault types. Most often, the segmentation of network fault data is based on an engineer's expert domain knowledge. Although expert knowledge can lead to a workable segmentation of faults data, it does not guarantee that each segment comprises a homogenous group of network faults. Therefore, faults analysis could benefit from using a data mining technique, which would assist the process of distribution faults segmentation [73].

4.3.5.1 Electricity Distribution Network Fault Segmentation using K-means Clustering

In this study, the approach is to use unsupervised machine learning algorithm for segmentation. In the electricity industry, typically clustering has been extensively used to perform customer segmentation based on their customer's electricity consumption. Different clustering algorithms have been used, such as hierarchical and k-means clustering.

In this study, the K-Means clustering method, which is an unsupervised machine learning algorithm, has been used to segment the electricity distribution faults. Segment boundaries were defined by the levels of similarity between faults concerning the data attributes. The techniques involved methods to calculate a distance measure, ascertain the cluster-ability of the dataset, and the optimum number of clusters that can be obtained from the dataset using the Hopkins-Statistic and Elbow plot methods, respectively.

The Hopkins Statistic is known to be a fair estimator of randomness in a data set. However, in some cases, outliers can pollute the dataset. It is better, therefore, to remove outliers from the research dataset to improve the Hopkins Statistics value for better clustering.

4.3.5.2 Outliers Detection

Hawkins has defined an outlier as an observation that deviates so much from other observations as to arouse suspicion that a different mechanism generated it. Databases are likely to contain inaccurate and anomalous data due to different reasons such as observational and human error, poor data quality, and malfunctioning of equipment. Analysing a dataset for the presence of anomalies is an essential task in data mining.

If outliers are not detected and eliminated than the entire dataset is biased and building a classifier based on influenced records becomes difficult. Outlying observations can have a considerable influence on any analytical results. Anomaly and outlier have a similar meaning. However, analysts prefer to use outliers as anomalies may be present in different domains. For example, outliers are associated with intrusion detection and fraud detection, whereas anomalies are associated with time-series data.

Anomalies and Outliers Detection Techniques

Data labels associated with data observations shows whether an observation belongs to standard data or anomalous data. Based on the availability of labels for data instance, anomalies and Outliers detection techniques can be divided into three types

- Supervised Detection

  There are two phases in the supervised anomaly detection technique. These are the training phase and the testing phase. In the training phase labels are present for both standard and abnormal observations. The model applied during training learns from the patterns and later use that knowledge to label the records in the testing phase.

- Semi-supervised Detection

  The semi-supervised Anomaly detection technique is similar to the Supervised technique as it also has a training phase. However, the training data set only comprises of standard records. A record is labelled as an outlier only if

it deviates from the standard observations that were learned by the model during the training phase.

- Unsupervised Detection

  This technique does not require training data as it does not expects labels to be present in the training set. It assumes that anomalies are much less common than standard data in the dataset.

After reviewing the literature, a Density-Based Spatial Clustering of Applications with a Noise algorithm-based outlier detection method, which is an unsupervised Anomaly Detection approach has been chosen and implemented.

4.3.5.3 Outlier Detection using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Algorithm

Density-Based Spatial Clustering of Application with Noise (DBSCAN) is a data clustering algorithm proposed by Ester et al. [74]. Unlike K-Means, DBSCAN does not require the number of clusters as a parameter. Instead, it infers the number of clusters based on the data, and it can discover clusters of arbitrary shape. DBSCAN is very useful when separating high-density clusters from low-density clusters.

This algorithm finds high-density core samples and expands clusters from them. The literature shows that the DBSCAN algorithm can discover anomalies even if they are not extreme values. The author has, therefore, used this algorithm to remove outliers from the dataset. Two parameters required ε (eps), the maximum distance between two samples for them to be considered as in the same neighbourhood and the minimum number of points required to form a dense region (minPts).

The DBSCAN algorithm can be abstracted into the following steps: [75]

1. Find the ε (eps) neighbours of every point, and identify the core points with more than minPts neighbours.

2. Find the connected components of the core points on the neighbour graph, ignoring all non-core points.

3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbour, otherwise, assign it to noise.

The author has applied the DBSCAN algorithm to the research dataset with the value of ε (eps) =1000, and minPts value is 3.

As shown in figure 4.6, the results have created one large cluster and six noise points which have been regarded as outliers and removed from the original dataset.

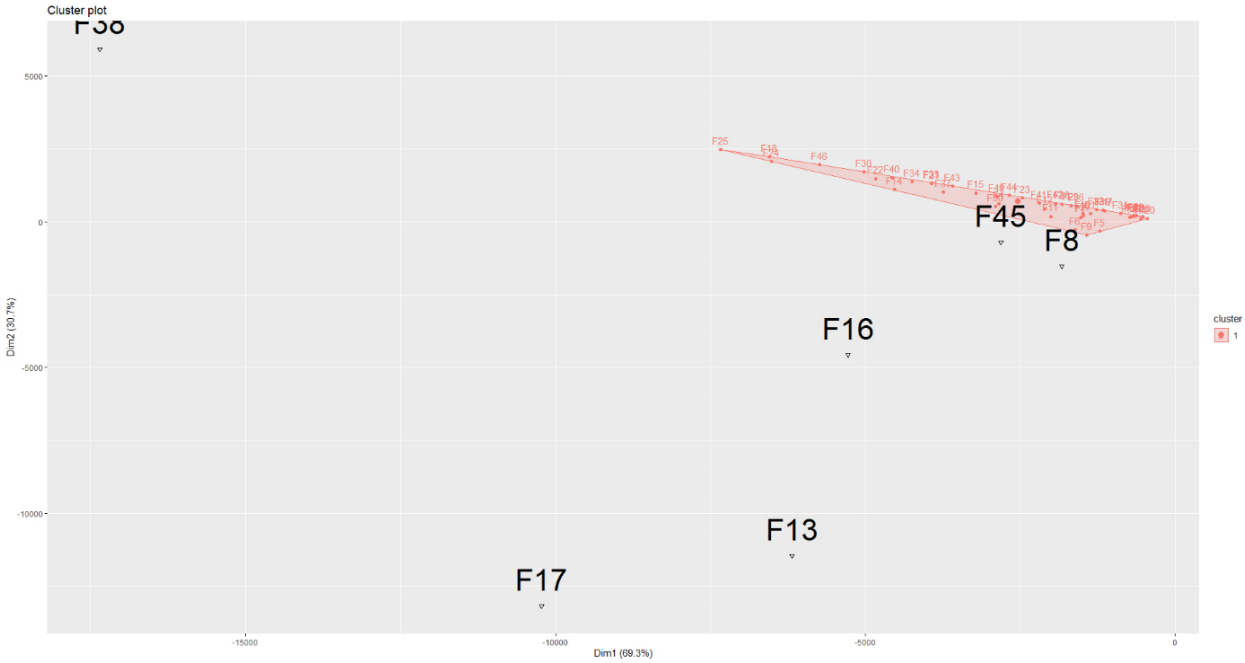


Figure 4.6: DBSCAN cluster analysis using aggregated NaFIRS dataset

Anomalous points are not only extreme points but are also the data that does not occur frequently. However, it is advisable to create a new dataset using that outlier dataset for further analysis. Table 4.13 shows the outlier fault types.

Table 4.13: Identified outlier fault types using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Algorithm

Label	Direct Cause
F8	*by Private Developers or their Contractors*
F13	*Cause Unknown*
F16	*Corrosion*
F17	*Deterioration due to Ageing or Wear (excluding corrosion)*
F38	*Local Generation Failure (Isolated System)*
F45	*Transient Fault - No Repair*

After removing outliers, the dataset analysed by Hopkins Statistics (*H*) for the research dataset is 0.62, which is relatively high. *H* should be more significant than 0.5 and almost equal to 1.0 for very well defined clustered data. However, the H value of 0.62 has been used as significant enough to pass the dataset to clustering.

The new dataset will now be analysed using the elbow method. The elbow point, as shown in Figure 4.7, is where the graph tends to flatten. That is, increasing the number of clusters beyond three is not significantly reducing the WSS. This inflation point is usually chosen to be the value of K for K-means.

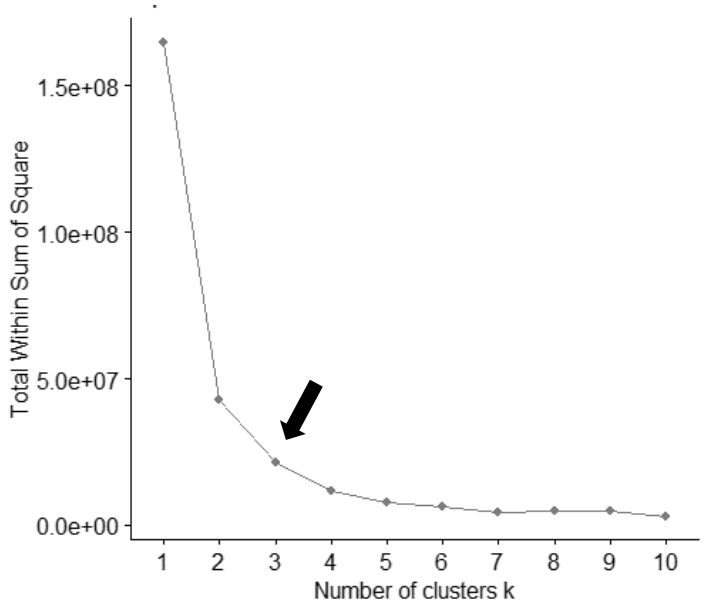


Figure 4.7: Optimal number of clusters for aggregated NaFIRS dataset after remove identified outlier fault types

The K-Means clustering of the LV faults data using the optimum cluster value reveals the distribution of the faults data by the direct cause into 3 clusters. Results

show that the first, second and third clusters contain 5, 26 and 14 direct fault causes respectively. Figure 4.8 gives a pictorial representation of the clusters formed using the K-means cluster analysis.

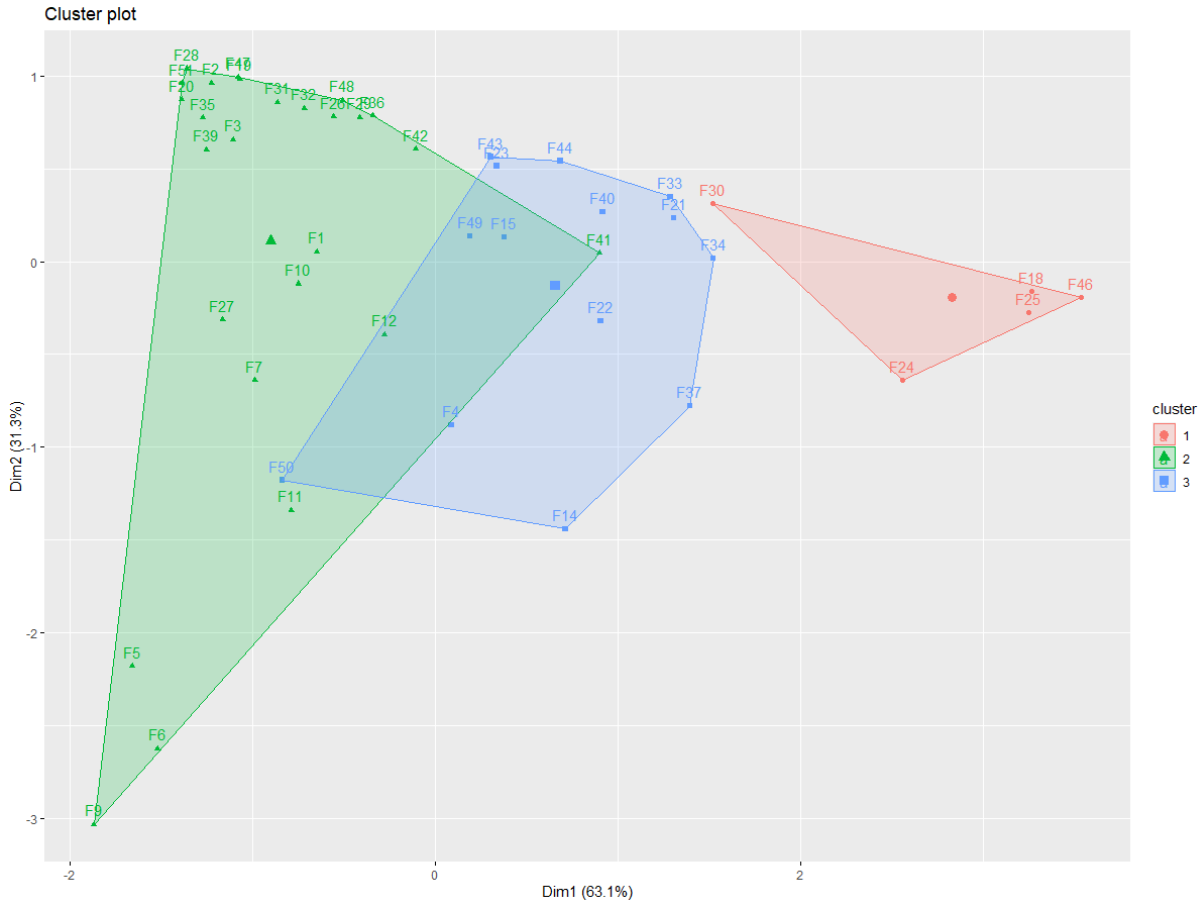


Figure 4.8: K-means cluster analysis for aggregated NaFIRS dataset after remove identified outlier fault types

Table 4.14, 4,15 and 4.16 shows the faults types in each cluster formed by the K-means cluster analysis.

Table 4.14: Faults types contain in cluster 1 which is generated by the K-means cluster analysis for the aggregated NaFIRS dataset

Label	Fault Cause
F18	*DNO Equipment affected by Private Generator or Authorised Electricity Operator (other than National Grid)*
F24	*Fire not due to Faults*
F25	*Flooding*
F30	*Inadequate Rupturing or Short Circuit Capacity*
F46	*Unsuitable Paralleling Conditions*

Table 4.15: Faults types contain in cluster 2 which is generated by the K-means cluster analysis for the aggregated NaFIRS dataset

Label	Fault Cause
F1	*Accidental Contact, Damage or Interference by DNOC or their Contractors (including Live Line Work)*
F2	*Birds (including Swans and Geese)*
F3	*By Cable TV Companies or their contractors*
F5	*by Highway Authorities or their Contractors*
F6	*by Local Building Authorities or their Contractors*
F7	*by Other Third Parties*
F9	*by Private Individuals (excluding 49 and 56)*
F10	*by Public Telecommunications Operator or their Contractors (e.g. BT or Mercury)*
F11	*by Unknown Third Parties*
F12	*by Water/Sewage Company or their Contractors*
F19	*Falling live trees (not felled)*
F20	*Farm and Domestic Animals*
F26	*Ground Subsidence*
F27	*Growing or Falling Trees (not felled)*
F28	*Growing Trees*
F29	*Inadequate or Faulty Maintenance*
F31	*Incorrect Application of DNOC Equipment*
F32	*Incorrect or Inadequate System Records, Circuit Labelling or Identification*
F35	*involving Farm Workers or Farm Implements*
F36	*Lightning*
F39	*Mechanical Shock or Vibration*
F41	*Operational or Safety Restriction*
F42	*Rain*
F47	*Unsuitable Protection Characteristics*
F48	*Vermin, Wild Animals and Insects*
F51	*Windborne Materials*

Table 4.16: Faults types contain in cluster 3 which is generated by the K-means cluster analysis for the aggregated NaFIRS dataset

Label	Fault Cause
F4	*by Gas Company or their Contractors*
F14	*Causes Unclassified in this Table*
F15	*Condensation*
F21	*Fault on Equipment Faulting Adjacent Equipment*
F22	*Faulty Installation or Construction by DNOC Staff*
F23	*Faulty Manufacturing, Design, Assembly or Materials*
F33	*Incorrect or Unsuitable protection Settings or Fuse Rating*
F34	*Interruption to remove local generator or restore temporary connection (wherein use > 18 hours)*
F37	*Load Current above Previous Assessment*
F40	*Metal Theft*
F43	*Snow, Sleet and Blizzard*
F44	*Switching Error by DNOC Personnel*
F49	*Wilful Damage, Interference (not including Metal Theft)*
F50	*Wind and Gale (excluding Windborne Material)*

The author created a new dataset with previously removed outliers and used K-Means clustering to identify the clusters with the outliers dataset. Initially, a new outlier dataset was analysed by Hopkins Statistics (H); for the outlier dataset. The H value showed as 0.498, which is relatively high compared to previous H values.

The outlier dataset was analysed again using the elbow method. The Elbow method suggests that three can be the best value for K. The results showed that the first, second and third clusters contain 3, 1 and 2 direct fault causes respectively (Figure 4.9).

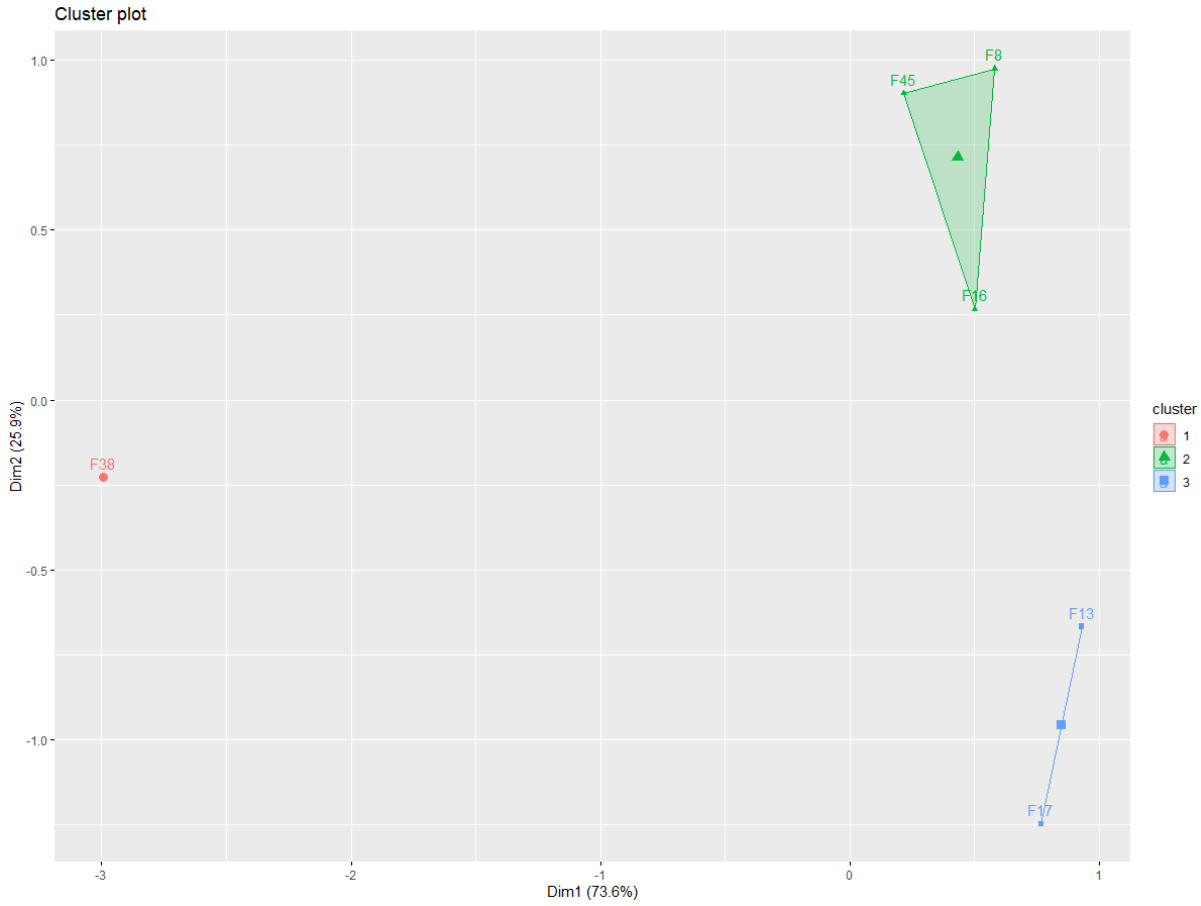


Figure 4.9: K-means cluster analysis for outliers dataset

Table 4.17, 4,18 and 4.19 shows the faults types in each cluster formed using the K-means cluster analysis for outliers dataset.

Table 4.17: Faults types contain in cluster 1 which is generated by the K-means cluster analysis for the outliers dataset

Label	Fault Cause
F38	*Local Generation Failure (Isolated System)*

Table 4.18: Faults types contain in cluster 2 which is generated by the K-means cluster analysis for the outliers dataset

Label	Fault Cause
F8	*by Private Developers or their Contractors*
F16	*Corrosion*
F45	*Transient Fault - No Repair*

Table 4.19: Faults types contain in cluster 3 which is generated by the K-means cluster analysis for the outliers dataset

Label	Fault Cause
F13	*Cause Unknown*
F17	*Deterioration due to Ageing or Wear (excluding corrosion)*

In total, six clusters were identified based on the two-cluster analysis carried out in the previous step. So, for this experiment, six distinct fault segments were identified. The clustering-based segmentation technique described successfully revealed internally homogeneous and externally heterogeneous fault groups. Network faults varied in terms of the number of customers affected, the number of minutes lost and other environmental characteristics. The primary goal of the clustering techniques was to identify different fault types and segment the network faults into clusters of similar profiles so that the process of taking measures to prevent those faults could be executed more efficiently. Based on the findings from the experiment evaluation, as shown in Figure 4.10, the author has introduced a new Fault Segmentation Framework to analysis network faults.

4.3.6 Proposed Fault Segmentation Framework

Once a viable and robust segmentation framework is in place, DNOs can gain new insights from fault data and should be able to define fault prevention strategies. For example, where there is a fault group that does not affect the performance, but there is also a group that contributes highly network reliability and performance, it is critical to identify this group and then decide on the best strategy for handling that group of faults.

Fault segmentation is not only concerned with clearly identifying different type of faults, but also tracking their changes. Multidimensional segmentation becomes critical in the electricity distribution network sector because it provides DNOs with a

deeper understanding of fault behaviours and fault causes. This enables the development of new strategies and new network designs that reduce the fault counts and the number of minutes lost due to network faults. Using a multidimensional segmentation framework also allows DNOs to decide on the optimum allocation of engineering resources, for each identified fault segment.

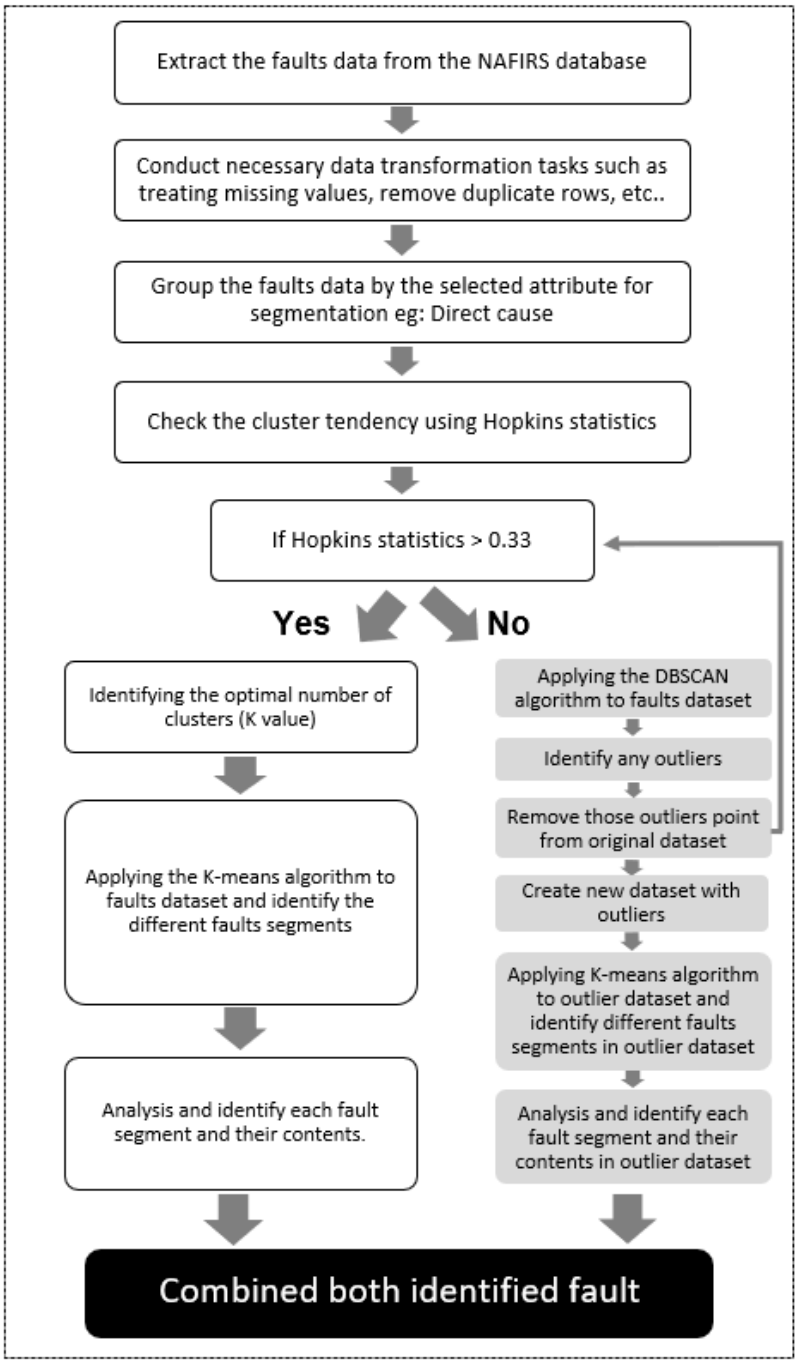


Figure 4.10: Proposed Electricity Distribution Network Fault Segmentation Framework

Fault segmentation in an Electricity Distribution Network is an essential pre-processing step for the early diagnosis and elimination of faults in the network. Fault segmentation also has a most significant role to play within a DNO engineering department, for effectively detecting or preventing future failures to increase the stability and availability of the network.

In this study, two different clustering methods have been used as a segmentation technique. The K-means clustering method was used to perform the segmentation, and the DBSCAN algorithm used to identify outliers. In this study, the author has introduced a new fault segmentation framework which DNOs can use to perform fault segmentation. This approach gives DNOs the option of performing multidimensional segmentation using various fault characteristics such as the number of faults, the number of minutes lost, and the number of customers affected. Multidimensional segmentation is, therefore, a very powerful conceptual model for the analysis of large and complex datasets. It will allow fault engineers in DNOs to understand the similar groups of faults in the network and their similarities.

4.4 Summary of Chapter and Conclusion

Power systems are prone to frequent faults, which may occur in any of power generating units, transformers, and power distribution media such as overhead and underground cables. Electricity distribution network components are always vulnerable to frequent failures that may occur in any of the main components or sub-components. Faults that generally occur in transmission and distribution networks are short circuit transients caused predominantly by vegetation, animals and weather effects such as tree damage. Other causes include large birds short-circuiting, current creepage through the path created by rain or moisture and the build-up of contaminants. According to the Office of Gas and Electricity Markets (OFGEM) report, severe weather conditions are the most influential factors that cause faults in the DNO network. Different weather conditions such as rainfall, snow, wind, humidity, and temperature all have the possibility of causing faults in assets in the distribution network.

Because of the high demand in the electricity market, the management of electricity distribution networks needs to manage a fast-changing profit originated business environment. The primary financial goal of any DNO is to reduce their operational costs to be competitive in their market. Fault management is one of the most demanding responsibilities for distribution network operators. The distributed network operator needs an excellent understanding of fault patterns in their network in order to reduce the incidence of faults which in turn will help to lessen their operational costs.

However, any fault in the energy distribution system causes significant disturbance to the complete grid system. The financial penalties for the failures in the system can be significantly high. Understandably, where possible DNOs seek to avoid any supply interruptions and restore customer confidence quickly whenever a failure occurs to reduce any financial penalties imposed by the regulators. Therefore, there is a business need to understand the details of faults and the factors behind the causes of faults.

In this section of the study, the author has attempted to understand most significant factors that contribute to distribution network faults using Association Rule Mining and explore the possibility of enhancing the knowledge gain from Association Rule Mining using Term Clustering. This study address one of the main challenges in

the Electricity Distribution Industry, which is faults cause identification in the network. The outcomes of this research will support the policy formulation in engineering departments to reduce the network faults. This section of the research used exploratory case study methodology to conduct the research. Exploratory case study methodology predominantly beneficial in helping to understand a problem, clarify the nature of a problem or describe the problems involved. It also enables the author to develop hypotheses for further research, to discover new insights or to understanding the issue from a different dimension.

DNOs are committed to providing a safe, reliable and affordable network to deliver energy to customers. It is of paramount importance to understand the network faults patterns and trend to prevent them. This section of the research presented a new framework for electricity distribution network fault segmentation. Electricity distribution network operators can use this proposed framework to analyses and understand the network faults better. Fault segmentation method can be used to identifying homogenous network faults types. Most often, the segmentation of network fault data is based on an engineer's expert domain knowledge.

The main aim of fault segmentation is to understand the failures better and to use that understanding to improve performance, reliability and availability of the distribution network. Fault segmentation is an essential tool for developing business intelligence in a fault management department and for maintaining competitive advantage among DNOs. The use of knowledge gained from fault segmentation will develop fault analysis clarity. Appropriate remedial action can be developed for each identified fault segment. It may eventually reduce the number of customer complaints due to fewer network outages.

CHAPTER 5

Temporal Analysis of Faults in Electricity Distribution Network

5.1 Introduction

Modern industry is almost dependent on electric power for its operation. Homes and office buildings are lighted, heated, cooled and ventilated by electric power. Electrical distribution systems are an essential part of the electrical power system. In general, the distribution system is the electrical system between the substation fed by the transmission system and the consumer end. Electrical distribution systems are inclined to have frequent failures due to various reasons, such as extreme weather events, equipment failures, human activities, etc. Any failures in a distribution network directly affect the network stability, availability and reliability. Understanding the electricity distribution network failure patterns is a challenging task.

Power systems are inclined to frequent failures due to equipment malfunctions in the network. Equipment malfunctions can occur in any network equipment, including transformers, switchgear, over ground cables or underground cables. Therefore, quick elimination and prevention of network faults are important for the DNOs. Any malfunction of the equipment causes significant interruptions in supply and destabilises the entire system. Understanding failures in the network are of the utmost importance for the operator even for the consumer.

It is challenging to be able to predict equipment failure accurately for a given period due to the uncertain nature of the fault forecasting process. Equipment in electricity distribution network depreciates with age and usage. Less aged equipment is less like to get faults than older equipment, and equipment with lower usage are less like to get faults than those with higher usage. Economic and weather factors such as consumer demand, temperature, rainfall, etc. also affect the reliability of the equipment. For example, faults in underground cables can be higher during summer

than during winter because of the high temperature. Also, faults in the overground cables can be higher during winter than during summer because of the high wind and snow. To further complicate matters, different equipment types, make and model, depreciate at different rates. This section of the study aims to predict the monthly distribution network faults caused by equipment failures with the highest possible accuracy using the three-different time-series algorithms. Those three models were implemented in each category of data sets to find the most efficient algorithm based on Mean Absolute Percentage Error as the selected accuracy metrics.

To make better business decisions, the DNO community also needs to understand the role of the seasonality in the network faults. This study will investigate seasonality using the time-series seasonal decomposition method. Accurate fault prediction at the distribution level and correct understanding of seasonality will help distributed network operators manage and plan network maintenance work. This research also may influence the DNOs in engineering staff management, setting up asset investment priorities and future design strategy.

DNOs always strives to provide customers with the highest levels of power supply and reliable customer service. There are times when a supply is interrupted due to planned or unintended interruptions. As part of maintaining and improving the network, DNOs may need to turn off the power supply in some distribution areas to provide safe access to power pylons and cables. Nevertheless, unplanned outages that result in loss of power in any distribution network cannot be prevented. They can be caused by storms, lightning strikes, falling trees, birds or equipment malfunctions. When there is any unplanned interruption, a DNO always tries to restore supplies as quickly as possible, as a high priority [10].

All the DNOs are committed to providing a safe, secure, reliable and cost-effective network to supply energy to customers [10]. Each time a customer loses supply, details of that interruption are recorded by transmission and distribution companies. The performance of each distribution network operator is reviewed annually and compared with previous years, with other DNOs, and with the individual targets set by OFGEM as part of the Interruption Incentive Scheme (IIS) framework and guaranteed standards of performance [10]. The financial penalties during any unplanned interruption can be significantly high. DNOs always try their best to avoid unplanned interruptions and restore electricity supply quickly when a fault occurs.

Initiatives to avoid interruptions can be targeted by asset investment when an asset has reached the end of its life. Therefore, removing an asset before a failure occurs may reduce the number of faults in the network [10]. Significant reduction of unplanned interruption may not be entirely possible, but reducing unplanned interruption times may reduce the effect of unforeseen network faults. The introduction of new technology that improves fault diagnostics, classification, segmentation and fault prevention can be instrumental in enhancing the DNO's interruption performance.

An accurate fault forecasting plays a crucial role in maintaining an electricity power supply system. It helps to plan asset maintenance scheduling, design new facilities, budget allocation and purchasing of new assets. Any extensive overestimation of the number of anticipated faults could result in significant additional investment for asset replacement and maintenance, while underestimation could result in poor customer satisfaction. Low network performance can severely affect the financial status and reputation of the DNO.

Therefore, there is a business need to accurately forecast faults and identify the location of the vulnerable assets. The time horizon for short term fault forecasting ranges between one day to one week. The time horizon for midterm fault forecasting ranges between one month to one year. It is challenging to forecast network faults exactly over a certain period due to the uncertain nature of fault occurrence. There are a large number of dominant features that characterise and directly or indirectly affect the underlying fault forecasting process. Most are uncertain and unmanageable. Therefore, any short-term or mid-term fault forecast, by nature, can be a challenging task. Furthermore, the highest care should be taken when forecasting future faults. The accuracy and reliability of the forecasting model may have a significant effect on the DNO’s engineering department.

There is a business need to reduce the operating costs of engineering departments [23]. It is essential to predict future faults of the electricity distribution network accurately. It is also necessary to predict future failures with seasonality in order to have a more informed set of data on which to base decisions. A reliable and accurate predictive modelling mechanism allows energy distribution companies to more efficiently manage engineering staff resources and lower their operating costs by reducing unnecessary callouts and overtime pay [23].

The author believes that having carefully studying faults data, hidden patterns can be identified as well as the temporal factors that correlate with distribution network faults. This study will help improve system availability by more accurately predicting network faults.

5.2 Predicting in Electricity Distribution Network Faults Caused by the Equipment Failures using Time-series Analysis

[4]In this section of the research, the author is investigating the daily faults caused by equipment failures. Daily fault datasets are considered complex because of the higher frequency of data recording and the occurrence of multiple seasonality. The essential steps in this kind of complex time series are to decompose the time series into different parts to obtain a deep understanding and get insight from the dataset. The decomposition helps to detect the overall structure of the data. This also helps to identify the appropriate technique for the time series analysis.

Three algorithms are considered for implementation. These are the Holt-Winters' Additive method, ARIMA and SARIMA. These three algorithms have a proven history of supporting complex datasets. The performance for the different models was compared using the Mean Absolute Percentage Error (MAPE) of the prediction against the actual data. Therefore, the main purpose of this section of the research is to compare the different forecasting techniques to achieve a better prediction which will help to predict distribution network faults caused by equipment failures.

After building the forecasting and prediction models, the accuracy and efficiency of the built models must be evaluated to find out how useful they are in solving the problems they are intended for. The results from all the models implemented are provided with detailed visualisations. In forecasting, the accuracy of a model is

[4] This section has been published in below research paper
C. Silva and M. Saraee, "Electricity Distribution Network: Seasonality and the Dynamics of Equipment Failures Related Network Faults," 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2020, pp. 1-6, doi: 10.1109/ASET48392.2020.9118274. (Appendix 2)

determined by how minimal the error is when the forecasted value is compared with the actual value known as the forecast error.

There are different error metrics, but the Mean Absolute Percentage Error (MAPE) has been chosen because they are scale-independent and are suitable for comparing accuracy across different models. MAPE is the more objective statistical indicator because the measure is in relative percentage terms and will not be affected by the unit of the forecasting series. The closer MAPE approaches zero, the better the forecasting results. MAPE is the most useful measure to compare the accuracy of forecasts between different items or products since it measures relative performance.

Despite its useful application, it has some drawbacks such as singularities occurring. For example, when zeros occur in the actual values, not having an upper limit when the forecast is too high and being unbalanced due to a more significant penalty being placed on negative errors.

5.2.1 Case Study: Ausgrid

Ausgrid is the largest electricity distributor on Australia's east coast, providing power to over 1.7 million customers. The Ausgrid distribution network spans some 22,275km^2 throughout Sydney, the Central Coast and the Hunter Valley [77]. They frequently published their data for research purpose. This research has used Ausgrid's past power, interruption quarterly data. Whilst data is available for over seven years, the quality of the power outage data has improved over the last five years. The research dataset contains power outages affecting fifty customers or more which last for longer than 5 minutes [77]. It does not include planned interruptions such as maintenance work, emergency work; or interruptions on significant event days when the network is affected by extreme weather events [78].

Data has been extracted for five years across seven variables. The first four years data was used for training and the remaining one year for testing. This was to ensure that the experiments always respected the time ordering of the time series data, which increases the reliability of the estimates. The variables description for all the variables is shown in table 5.1 as follows.

Table 5.1: Attributes explanation of the Ausgrid dataset

Attribute	Type	Description
Event ID	*Numeric*	*Fault ID*
LGA	*Character*	*Local Government Area*
Start Date	*Date*	*Date of fault occurred*
Start Time	*Time*	*Time of fault occurred*
Customers	*Numeric*	*No. of Customers*
Ave Dur. (min)	*Numeric*	*Customer Minutes Lost*
Reason	*Character*	*Fault Cause*

Similarly, other attributes have been transformed into a suitable form for a better study of the data. The fault dataset has been explored using different statistical methods to gain an overall picture of the variables and the underlying trends in the data. The original dataset contained seven variables and 8,112 observations. After selecting the essential variables for this research, a new dataset was formed with four variables and 8,112 observations from 1st Jan 2014 to 31st Dec 2018. The variables description for the selected variables is shown below.

Attribute	Type	Description
Start Date	*Date*	*Date of fault occurred*
Customers	*Numeric*	*No. of Customers*
Ave Dur. (min)	*Numeric*	*Customer Minutes Lost*
Reason	*Character*	*Fault Cause*

5.2.2 Data Exploration of the Ausgrid Case Study

The initial analysis of the dataset (Table 5.2), shows that more than 50% of the power outages were caused by equipment failures and 38% of faults occurred due to environmental conditions.

Table 5.2: Primary fault cause analysis of the Ausgrid case study

Fault Cause	Count %	No. of Customers %	Customer Minutes Lost %
3rd Party	*0.01%*	*0.00%*	*0.02%*
Cable dig	*1.17%*	*1.79%*	*1.09%*
Customer installation	*0.35%*	*0.07%*	*0.42%*
Directed to interrupt	*0.12%*	*0.15%*	*0.09%*
Environmental	*38.30%*	*38.33%*	*35.02%*
Equipment fault	*50.12%*	*47.61%*	*54.16%*
Lightning	*1.47%*	*4.19%*	*1.34%*
Operating fault	*1.26%*	*2.14%*	*0.74%*
Third party	*6.93%*	*4.38%*	*6.94%*
Third party- upstream	*0.05%*	*1.20%*	*0.01%*
Vandalism	*0.22%*	*0.14%*	*0.17%*

In the following analysis, the author considers only equipment failure related power outages to understand the seasonal patterns and dynamics.

After selecting the essential variables and filtering only for equipment failure related power outage for this research, the new dataset was formed with four variables and 4,066 observations from 1st Jan 2014 to 31st Dec 2018.

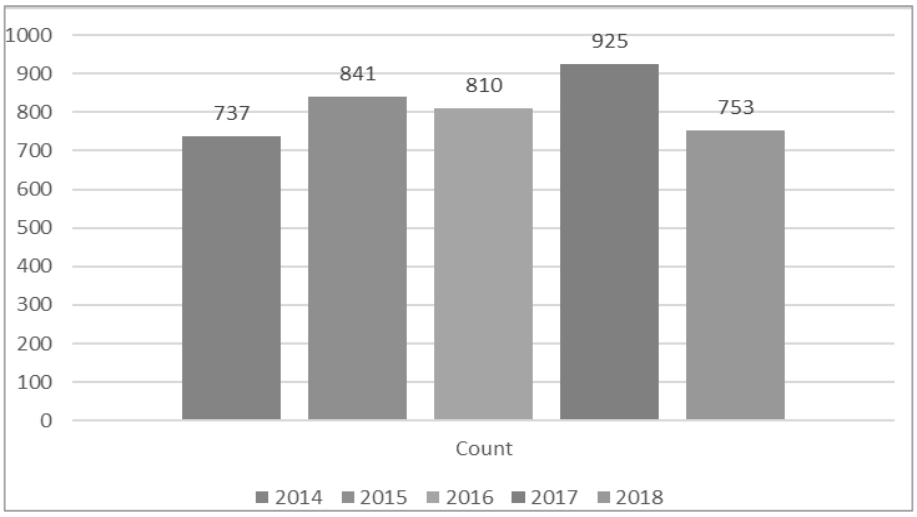


Figure 5.1: Annual equipment failure related power outages of the Ausgrid case study

The bar plot of the equipment failure related faults shown in Figure 5.1 above, shows that there is overall a slightly increasing trend pattern during 2014 to 17. Nevertheless, the gradient of the trend line is low. However, a significant decrease in 2018 is also shown.

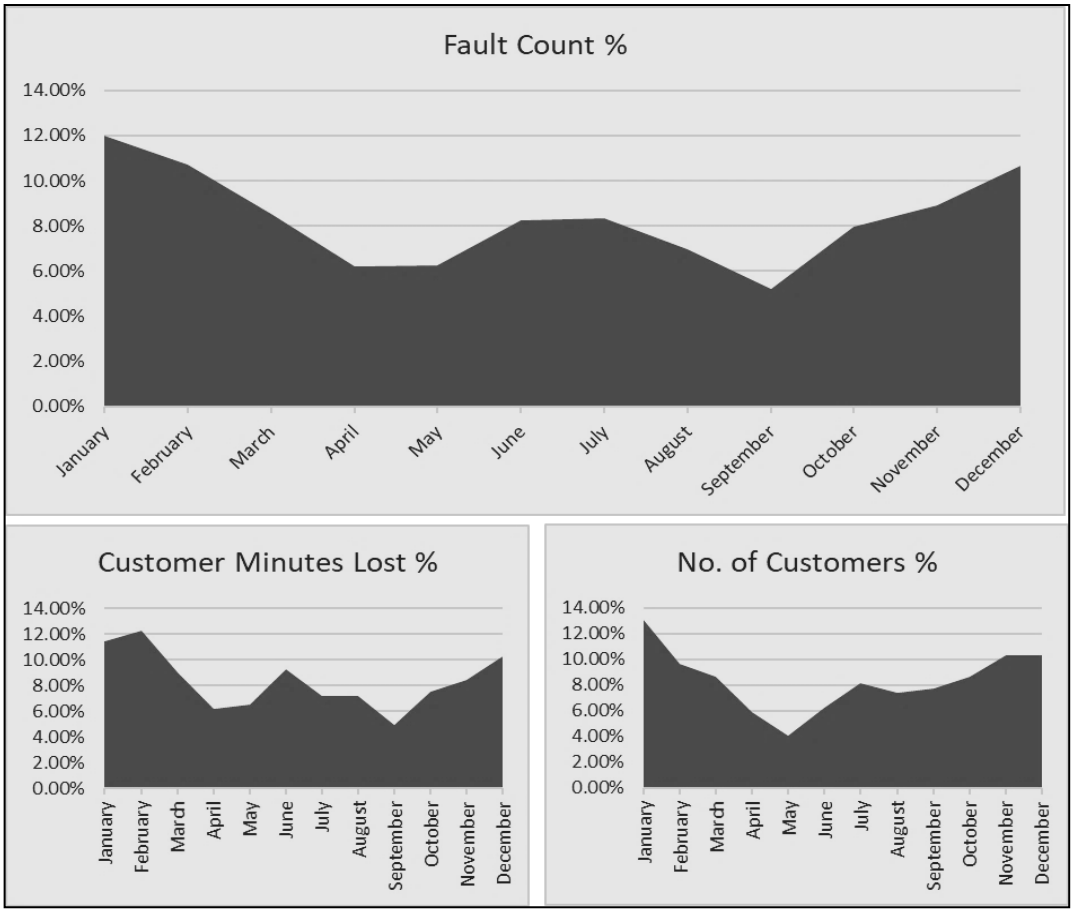


Figure 5.2: Average monthly changes in faults due to equipment failures in the Ausgrid case study

Initial data exploration shows the seasonality present in the dataset. It is important to check the stability of the time series as any seasonality makes the time series unstable. Data can be captured in time series at weekly, monthly and yearly intervals. Therefore, the data will be drilled down to identify “time” patterns in faults across the monthly, hourly and days of the week periods.

Furthermore, to understand the underlying seasonal pattern, an analysis of the monthly faults was conducted. The results in figure 5.2 show the average monthly changes in faults due to equipment failures with two peaks per year. First peak period occurred between January and February. The second peak is between November and

December. Both peaks are reasonable due to their occurrence during the hottest months in the area.

Drilling further to understand the underlying pattern, an analysis of the hourly figures of the faults was conducted. The results in figure 5.3 show the peak period to be between 5 pm and 8 pm. This is again reasonable considering that most consumers are using electricity within these times, leading to higher electricity demands and causing stress to the network equipment.

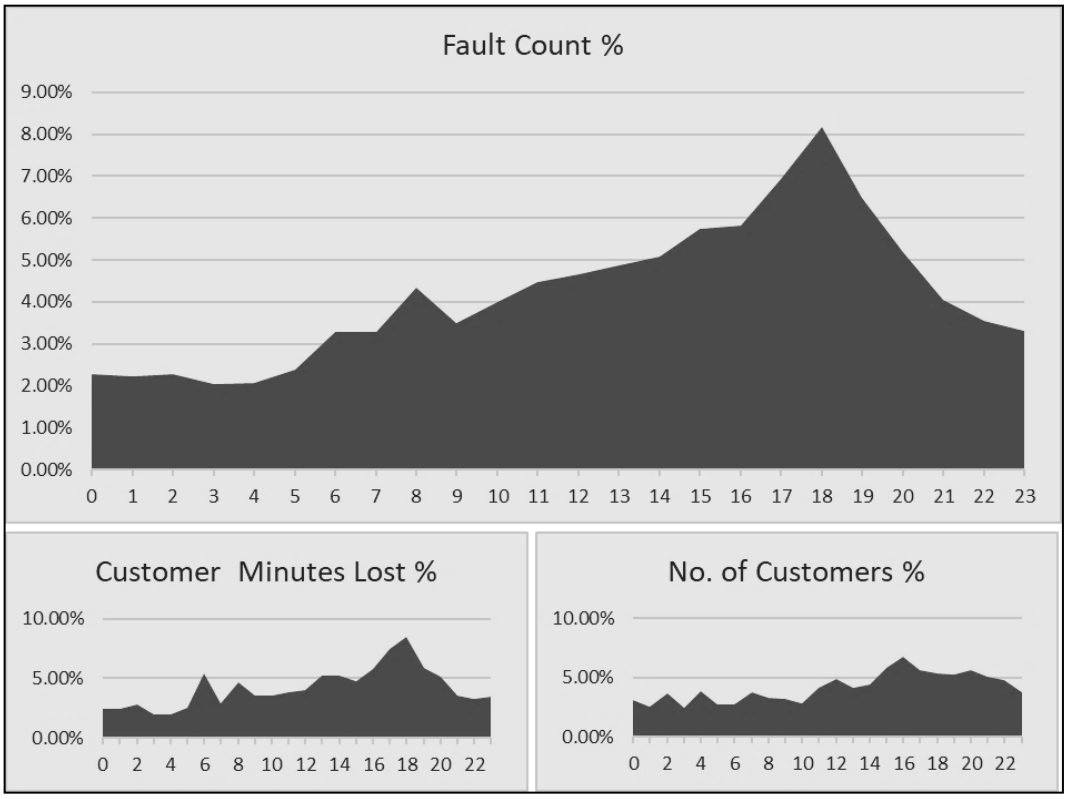


Figure 5.3: Average hourly changes in faults due to equipment failures in the Ausgrid case study

The above basic data exploration shows that the seasonality exists in the dataset. Therefore, time-series decomposition was conducted a clearer picture.

5.2.3 Time-series Decomposition on Ausgrid Fault data

Time-series decomposition is a procedure which transforms a time-series dataset into a level, trend, seasonality, and noise components. Each component represents one of the underlying pattern categories.

- ***Trend:*** This is the long-term behaviour perpetuated during several periods, and an upward or downward sloping curve represents it.

- ***Seasonal:*** Seasonality is the repetitive behaviour occurring with the same frequency on the time-series, i.e. repeated patterns appearing over and over again. It could occur hourly, daily, weekly, monthly, quarterly or yearly depending on the data. It could be additive or multiplicative.

- ***Reminder:*** This is the residuals of the original time series after the seasonal and trend series are removed.

The time series decomposition plot shown in Figure 5.4 has been created using a full fault dataset. The decomposition plot shown has four graphical components. These are the original data, the trend and seasonality as well as the one non-systematic component called remainder (residuals). The trend and seasonality component represent the movement of the model and how it is affected by seasonality. It also shows that seasonality exists in equipment related network faults.

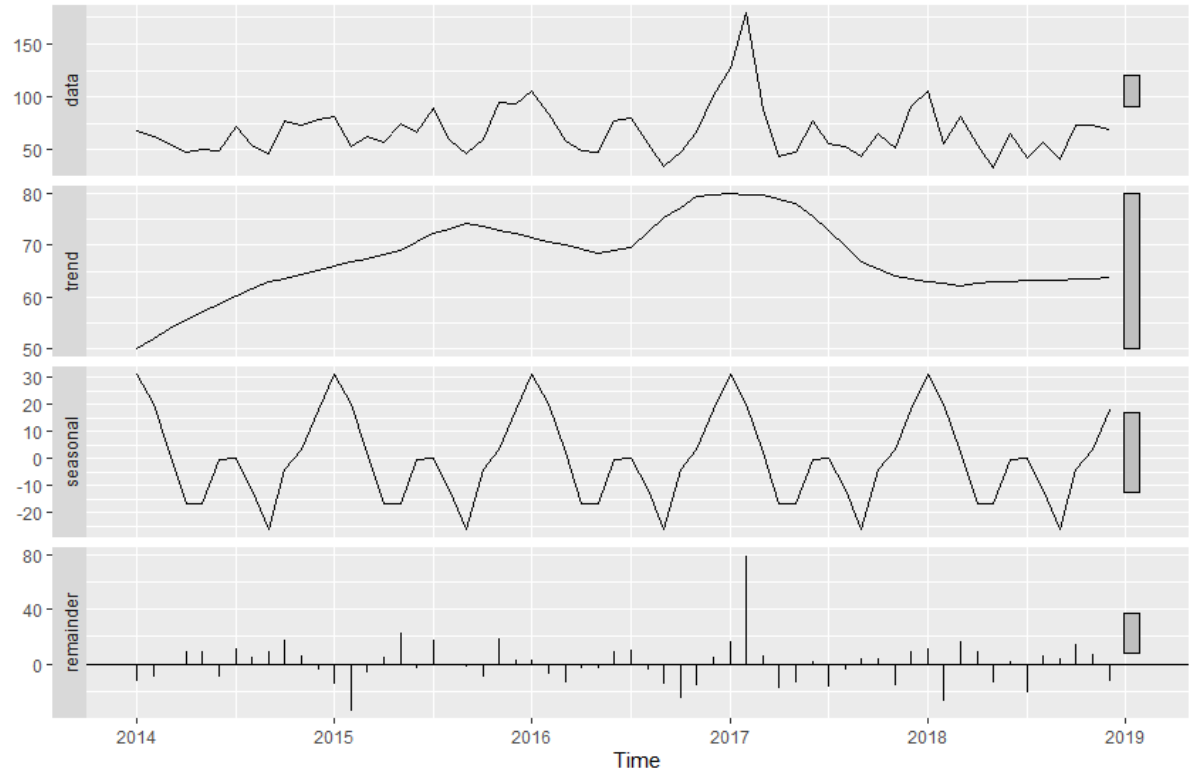


Figure 5.4: Timeseries decomposition plot of the Ausgrid case study

5.2.4 Analysing Seasonality using Seasonal Box Plot: The Ausgrid Case Study

A seasonal box plot can be used to display the monthly variations of equipment related distribution network faults studied. The interquartile spread is represented by the range between the 25th and 75th quantiles. The dots show the outlier values. The seasonal box plot shown in figure 5.5 shows that September has the least equipment related faults, while January has the highest number of faults. As with the seasonal box plot, a distinct seasonal pattern is apparent.

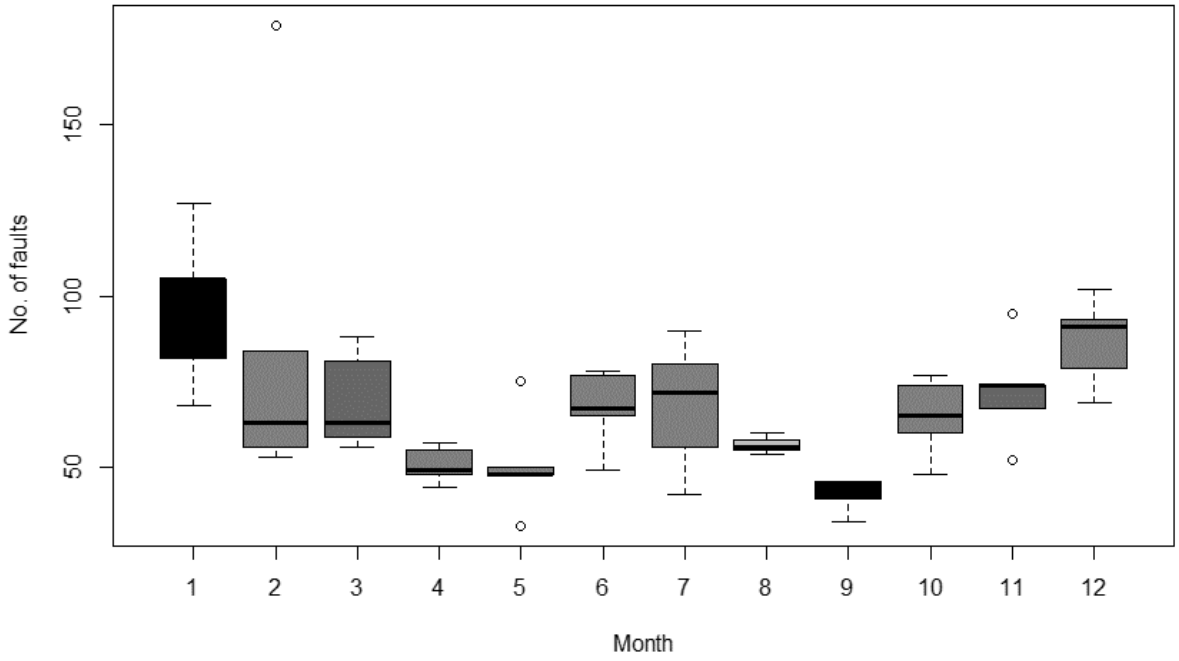


Figure 5.5: Seasonal box plot – Number of equipment related distribution network faults in the Ausgrid case study

5.2.5 Timeseries Forecasting on Fault Count: The Ausgrid Case Study

The implementation of time-series modelling starts by splitting the dataset into training and testing parts to compare the prediction of the model with the actual dataset. The train and the test split were carried out using five years of data from 2014 to 2017 for the training set, and the 2018 data for the test set. Holt-Winters' Additive method, ARIMA and SARIMA have been used as time series forecasting models.

The time series forecast carried out using three different models are presented in the section below. The Holt-Winters Additive model trained on the training dataset and predicted the test set, as shown in Figure 5.6.

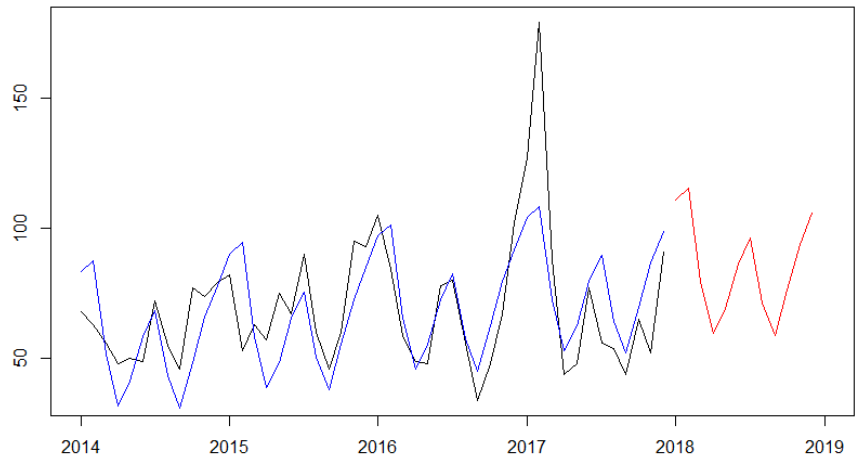


Figure 5.6: Fault count forecasting with Holt-Winters Additive method using Ausgrid Dataset

Overall, the model accuracy is measured using the Mean Absolute Percentage Error (MAPE). MAPE expresses accuracy as a percentage of the error. Because the MAPE is a percentage, it can be easier to understand than other accuracy measure statistics. Smaller values indicate a better fit. The MAPE value for the training dataset is 20.32%, and for the test dataset is 45.34%. Its mean, forecast on the test dataset is off by 45.34%. The ARIMA model was trained on the training dataset and predicted the test set, as shown in Figure 5.7.

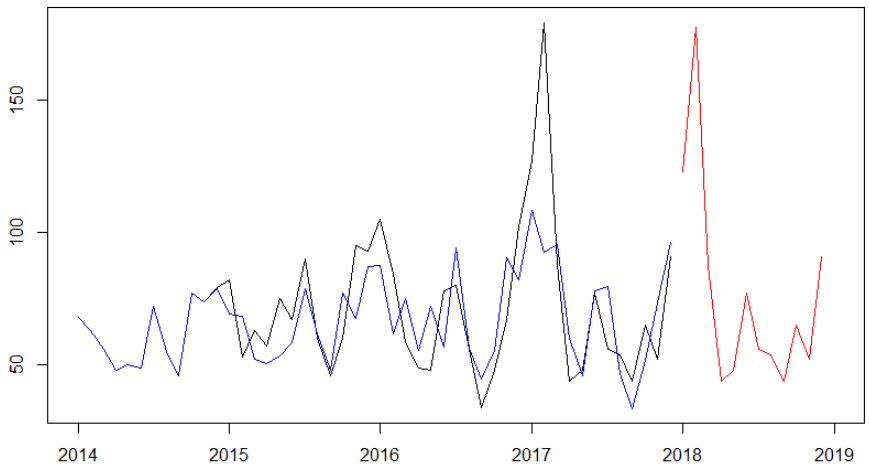


Figure 5.7: Fault count forecasting with ARIMA model using Ausgrid Dataset

The ARIMA model performance was measured using the Mean Absolute Percentage Error. The MAPE value for the training dataset is 15.38%, and for the test dataset is 37.24%. It is mean, time-series forecast on the test dataset is off by 37.24%.

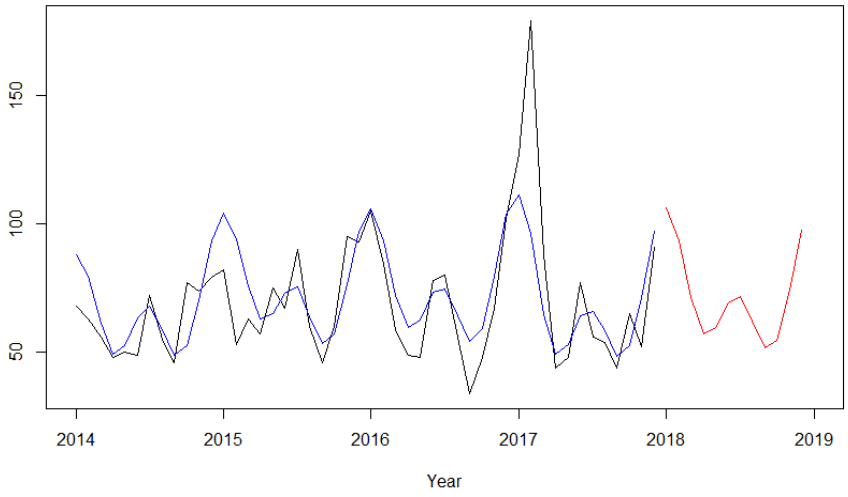


Figure 5.8: Fault count forecasting with the SARIMA model using Ausgrid Dataset

The SARIMA model trained on the training dataset and predicted the test set, as shown in Figure 5.8. The MAPE value for training dataset is 17.51%, and for the test dataset is 31.41%.

The results obtained from the time-series models for the number of distributed network faults caused by equipment failures have been compared in Figure 5.9.

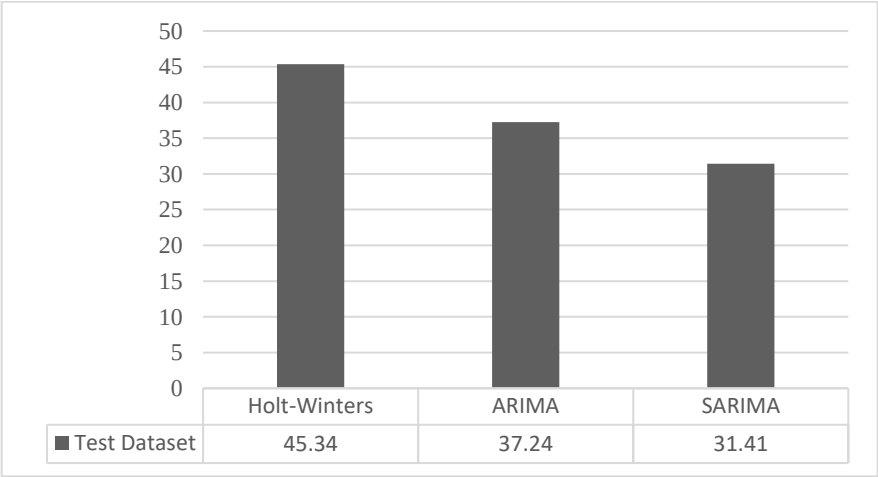


Figure 5.9: Comparison of fault count forecasting model accuracy using MAPE

It is apparent from the chart that the SARIMA model performs better on the dataset than the Holt-Winters Additive model and ARIMA model. The margin between the metrics for the three models is significant, showing that the Holt-Winters Additive model is not well suited for this type of fault prediction.

To further validate the findings of the research, the number of customers involved and the number of customer minutes lost was also analysed in the same way as the number of faults.

5.2.6 Timeseries Forecasting on the Number of Customers Affected Due to the Equipment Failures: The Ausgrid Case Study

Decomposing a time series means separating it into its components. These components are often a trend component, seasonal component, and a random component. The observed component from the plot below represents the normal time-series distribution of the dataset and plot is generally affected by the other components of the time series. The seasonal composition is only present in the decomposition of the data is a seasonal data object. It is apparent from figure 5.10 that the number of customers affected by the equipment related network failures also has seasonality factors. The decomposition chart below shows the time-series broken down into its components. It is also useful to study the underlying patterns that exist in the data related to the number of customers involved. After decomposing the time series data, more understanding of the underlying patterns is gained from which future predictions can be made.

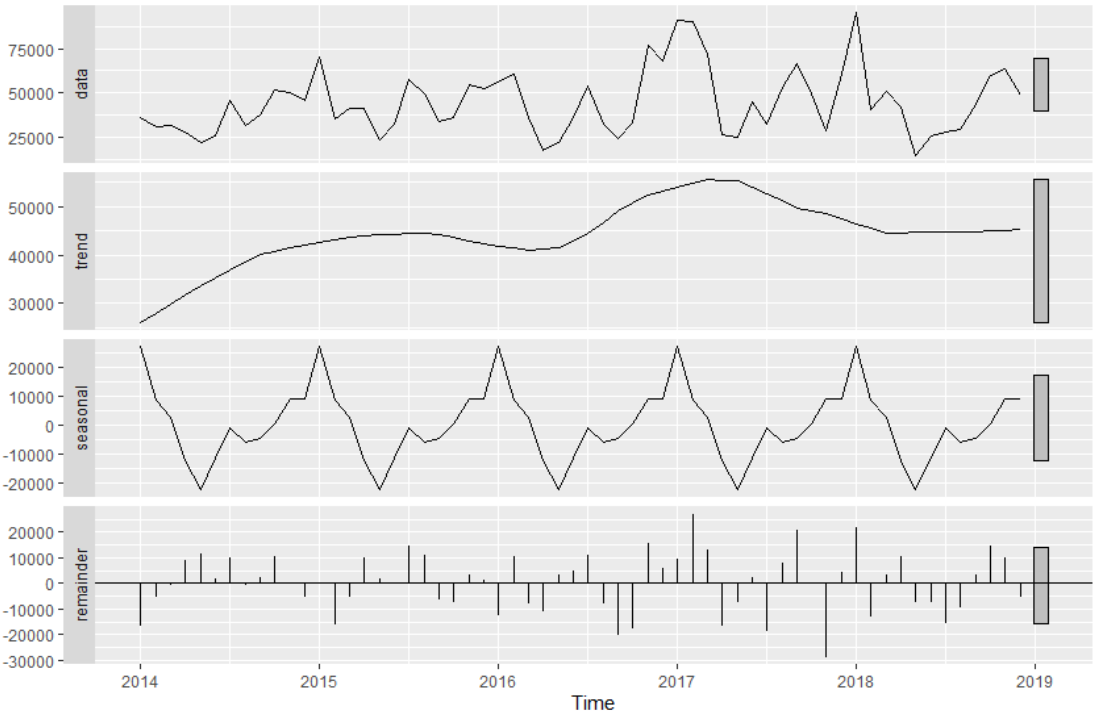


Figure 5.10: Decomposition plot - number of customers involved in the Ausgrid case study

Figure 5.11, indicates that the number of customers involved by the equipment related network failures also has seasonality factors.

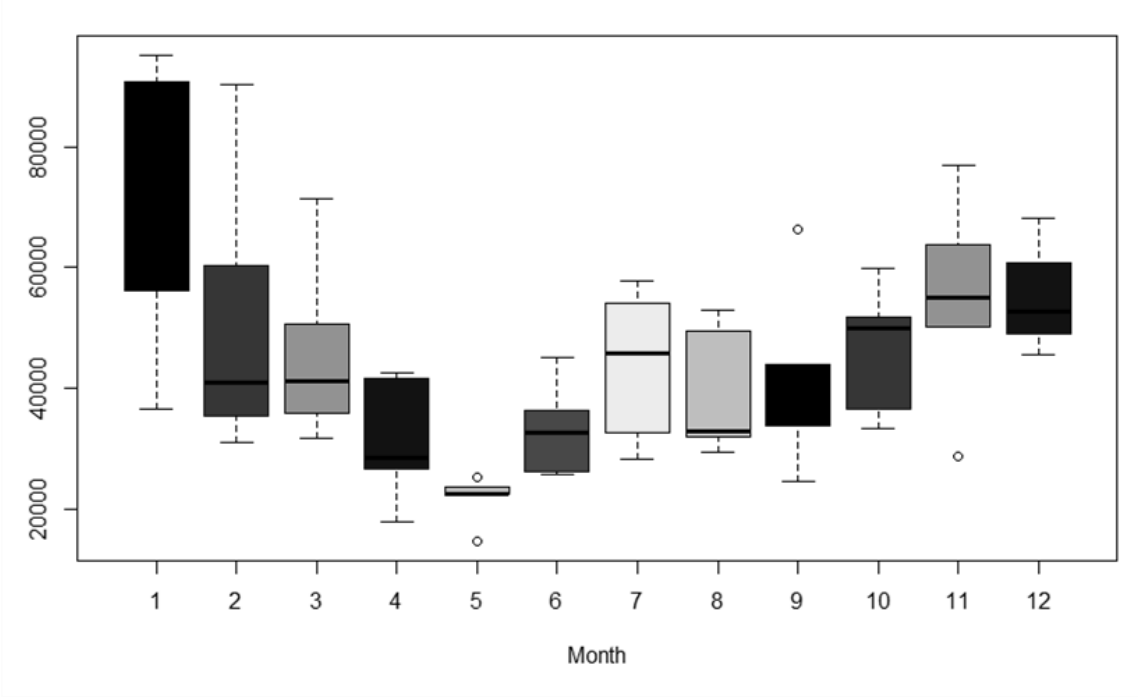


Figure 5.11: Seasonal box plot - number of customers involved in equipment related distribution network faults in the Ausgrid case study

The number of customers involved by the equipment related network failure related time series data shows seasonality (Fig 5.10). By seasonality, it is mean periodic fluctuations. The run sequence plot is a recommended first step for analyzing any time series. Although seasonality can sometimes be indicated with this plot, seasonality is shown more clearly by the box plot. The box plot shows (Fig 5.11) the seasonal difference (between-group patterns) quite well, but it does not show within-group patterns. However, for large data sets, the box plot is usually easier to read than the seasonal subseries plot.

The seasonal box plot shows the seasonal pattern more clearly. In this case, the number of customers involved by the equipment related network failure is at a minimum in May. From there, steadily the concentrations increase until January and then begin declining until April. The interquartile spread is represented by the range between the 25^{th} and 75^{th} quantiles. The dots show the outlier values.

The time series forecast carried out using three different models (Holt-Winters' Additive method, ARIMA and SARIMA) are presented in the below section using the number of customers involved by the equipment related network failure.

Same as the number of fault analysis done in 5.2.5 section, the analysis of the number of customers involved by the equipment related network also analyse by Holt-Winters' Additive method, ARIMA and SARIMA. The train and the test split were carried out with five years of data from 2014 to 2017 for the training set; and 2018 data for the test set.

The time series forecast carried out using those three different models are presented in the below section. The Holt-Winters Additive model trained on the training dataset and predicted the test set, as shown in Figure 5.12.

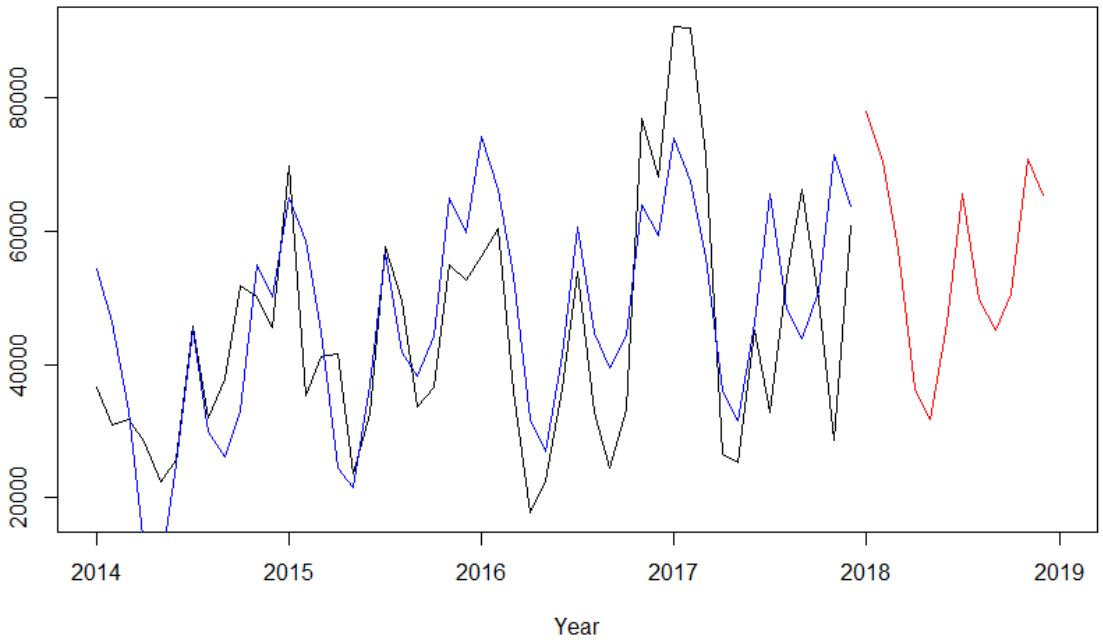


Figure 5.12: Forecasting of the number of customers involved with the Holt-Winters Additive method using Ausgrid Dataset

Overall model accuracy will be measured using the Mean Absolute Percentage Error (MAPE). The MAPE value for the Holt-Winters Additive method training dataset is 28.00%, and for the test dataset is 47.97%.

The ARIMA model trained on the training dataset and predicted the test set, as shown in Figure 5.13.

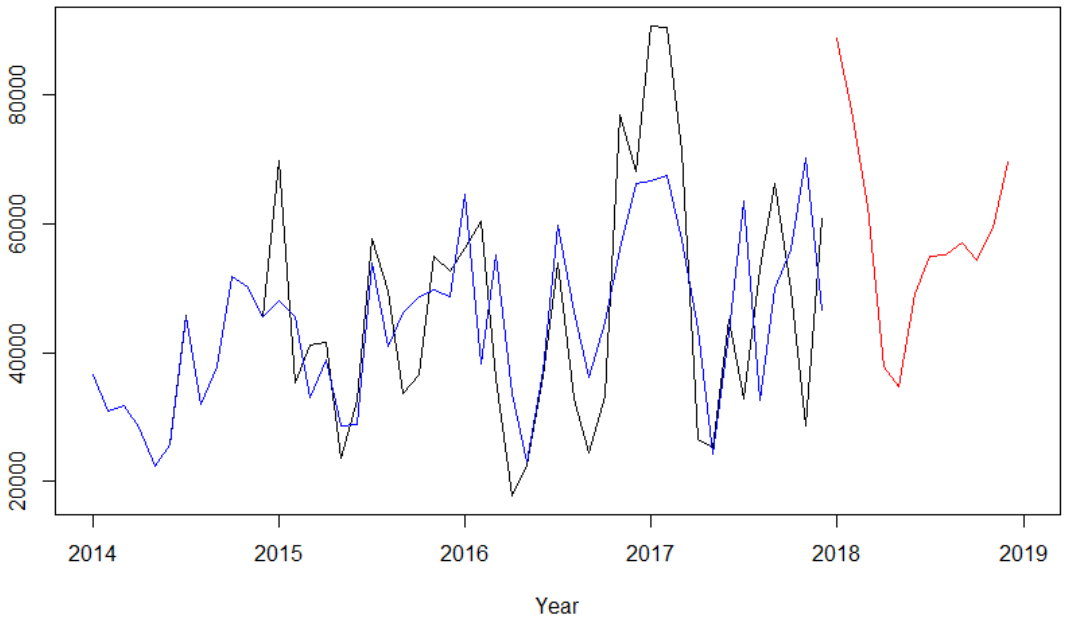


Figure 5.13: Forecasting of the number of customers involved with the ARIMA model using Ausgrid Dataset

The MAPE value for the ARIMA training dataset is 22.39%, and for the test dataset is 51.96%.

The SARIMA model trained on the training dataset and predicted dataset, as shown in Figure 5.13.

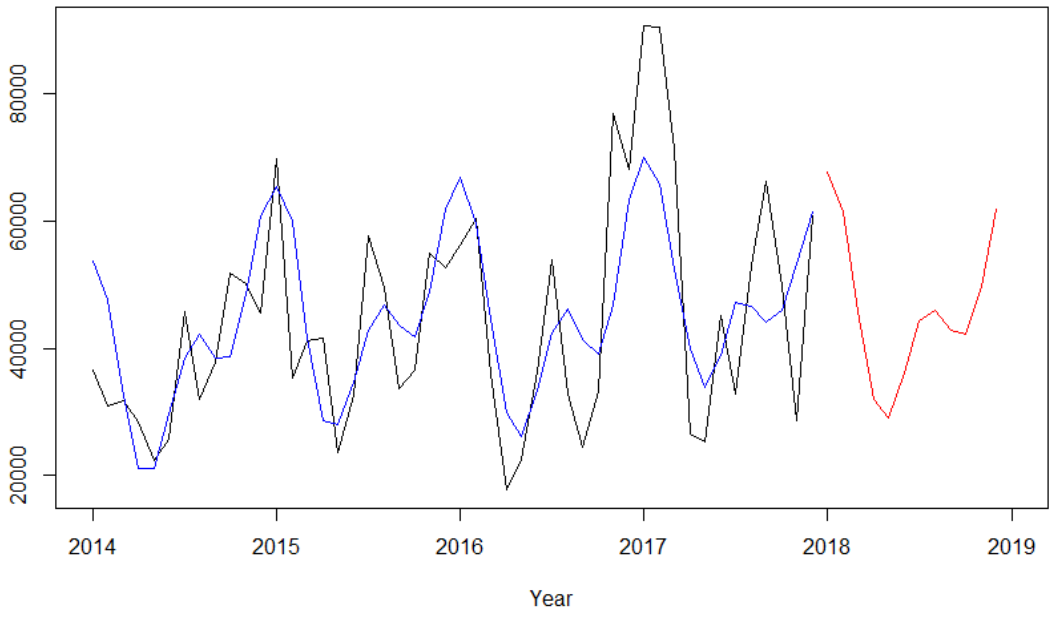


Figure 5.14: Forecasting of the number of customers involved with the SARIMA model using Ausgrid Dataset

The MAPE value for the SARIMA training dataset is 24.74%, and for the test dataset is 36.89%.

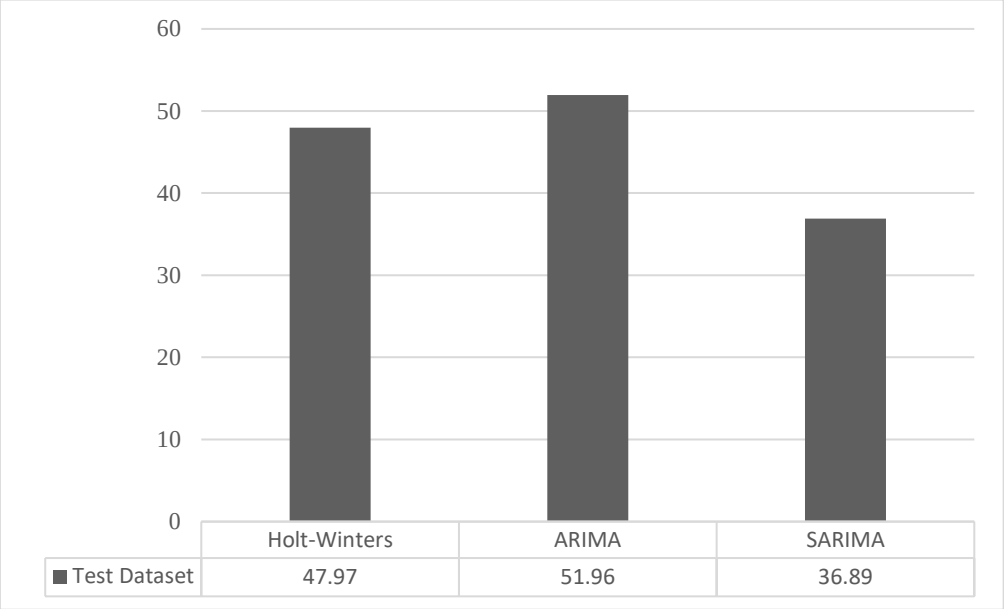


Figure 5.15: Comparison of model accuracy using MAPE- number of customers involved in the Ausgrid case study

The results obtained from the time-series models for the number of customers involved by the equipment failure's dataset has been compared in figure 5.15.

It is apparent from the chart that the SARIMA model performs better on the dataset than the Holt-Winters Additive model and ARIMA model. The margin between the metrics for the three models is significant, showing that the ARIMA is not well suited for this type of fault prediction.

5.2.7 Timeseries Forecasting on the Number of Minutes Lost Due to the Equipment Failures: The Ausgrid Case Study

In this section number of minutes lost due to the equipment failure's will be discussed. The analysis will start from the time-series decomposition. Decomposition is the breaking down of time series into its components. This is useful to study the underlying patterns that exist in the data. After decomposing a time series data, more insight is gained into understanding the underlying patterns with which future predictions can be made. Decomposition can be additive or multiplicative based on the nature of the components of the time-series data. In additive decomposition, the

components are decomposed in such a way that when they are added together, the original time series can be obtained while in multiplicative decomposition.

It is apparent from figure 5.16; the number of minutes lost by the equipment related network failures also has seasonality factors. Below decomposition chart is the breaking down of time series into its components.

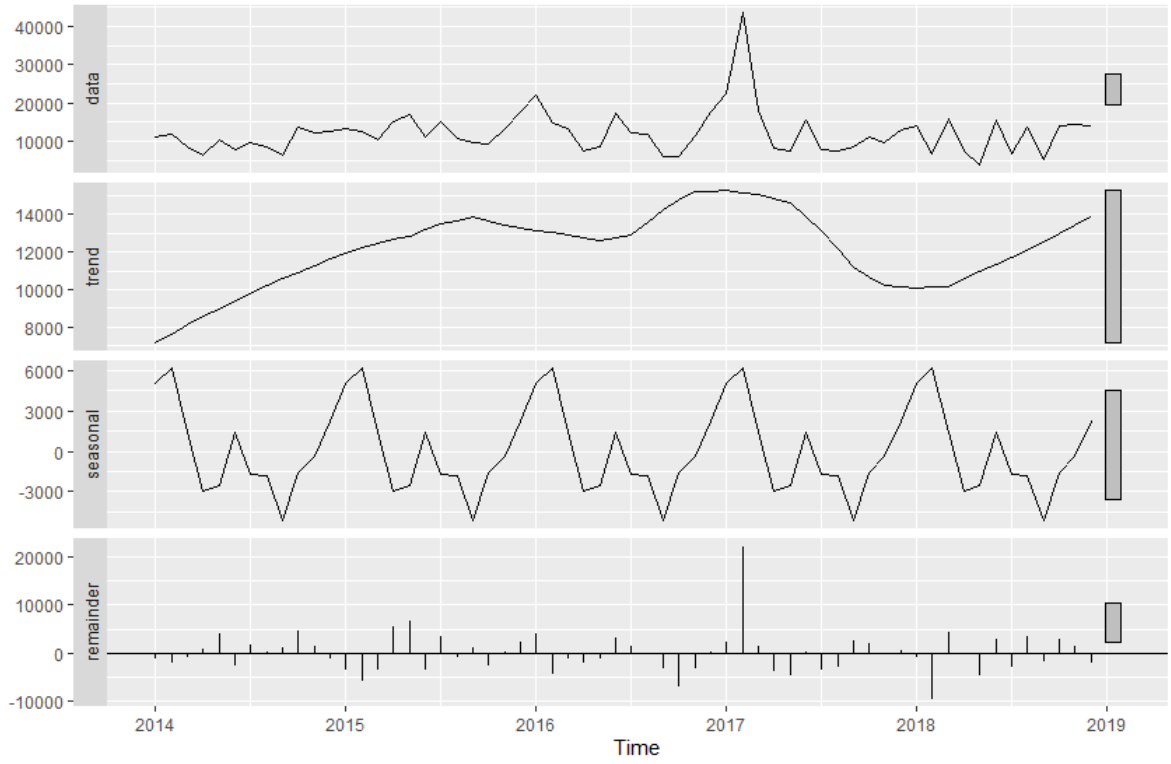


Figure 5.16: Decomposition plot - number of minutes lost in the Ausgrid case study

A seasonal plot enables for a clearer view of the underlying seasonal pattern and is particularly helpful in identifying a month when the pattern changes. Boxplot is particularly used to show the spread of within the months. It helps in determining whether there is an increase in the number of minutes lost in particular months. Thus, it provides information to reduce the number of minutes lost of certain months that are exceptionally high minutes lost.

The seasonal box plot shown in figure 5.17 shows the seasonal pattern more clearly. In this case, the number of minutes lost by the equipment related network failure is at a minimum in April and September. The number of minutes lost is high in the month of December and January. The interquartile spread is represented by the range between the 25th and 75th quantiles. The dots show the outlier values.

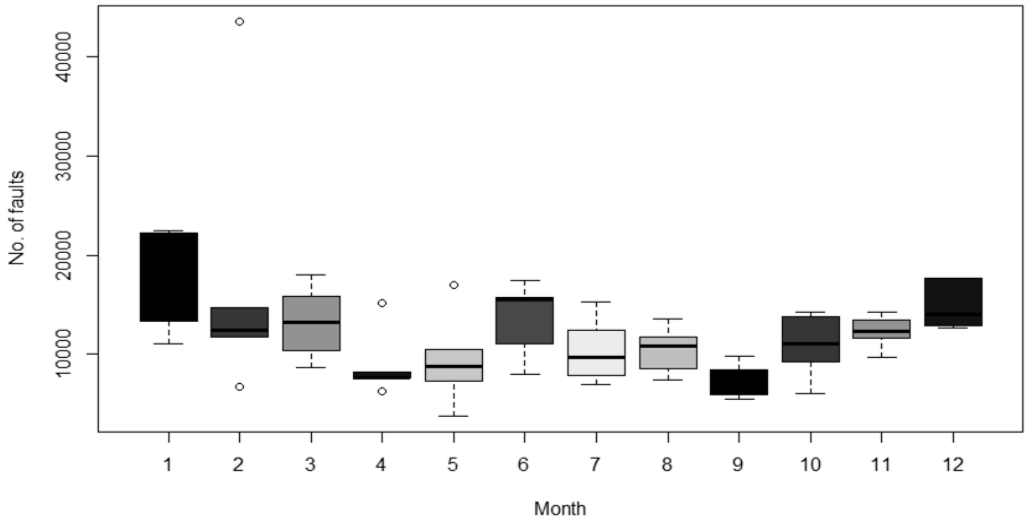


Figure 5.17: Seasonal box plot- number of minutes lost in equipment related distribution network faults in the Ausgrid case study

Same as the number of fault analysis done in 5.2.5 section, the analysis of the number of minutes lost by the equipment related network also analyse by Holt-Winters' Additive method, ARIMA and SARIMA. The train and the test split were carried out with five years of data from 2014 to 2017 for the training set; and 2018 data for the test set. The time series forecast carried out using those three different models are presented in the below section. The Holt-Winters Additive model trained on the training dataset and predicted the test set, as shown in Figure 5.12.

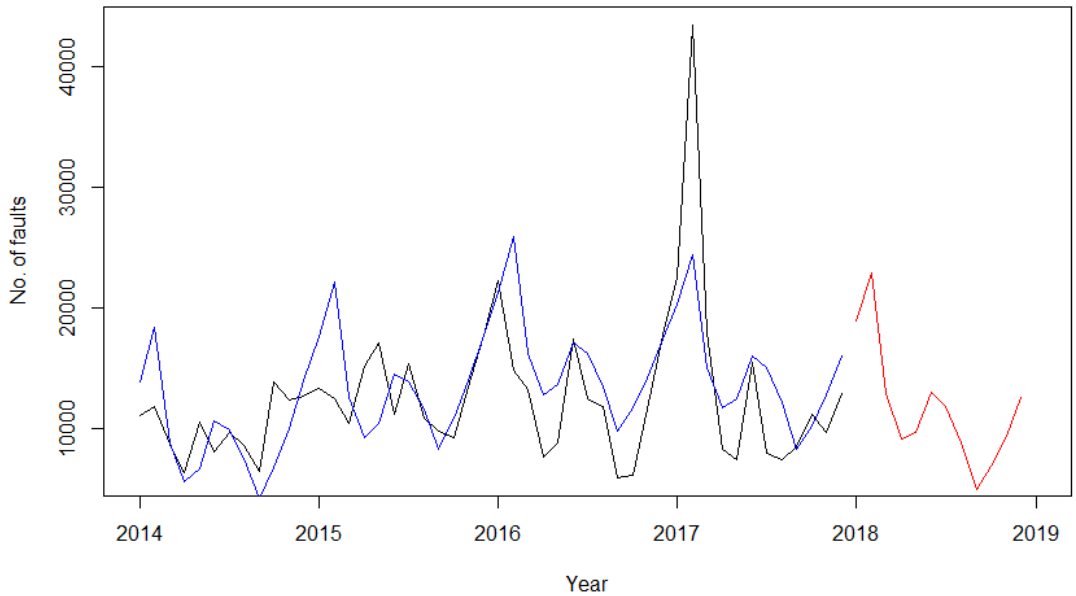


Figure 5.18: Forecasting number of minutes lost with the Holt-Winters Additive method using Ausgrid Dataset

Overall model accuracy will be measured using the Mean Absolute Percentage Error (MAPE). The MAPE value for the training dataset is 30.11%, and for the test dataset is 57.84%. The ARIMA model trained on the training dataset and predicted the test set, as shown in Figure 5.19.

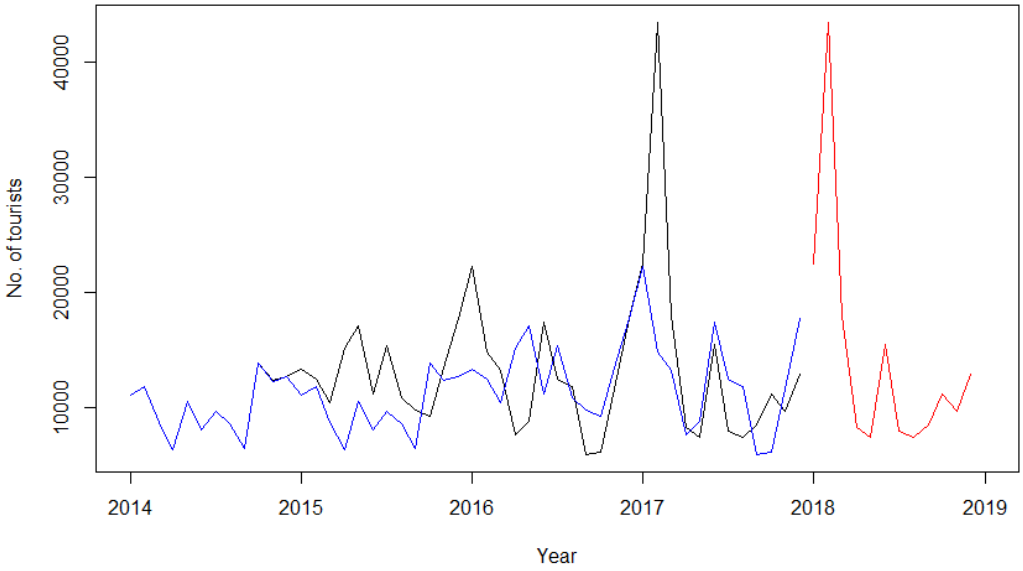


Figure 5.19: Forecasting number of minutes lost with the ARIMA model using Ausgrid Dataset

The MAPE value for the ARIMA models for the number of minutes lost by the equipment related network failures training dataset is 24.90%, and for the test dataset is 74.64%. The SARIMA model trained on the training dataset and predictions on the test set, as shown in Figure 5.20.

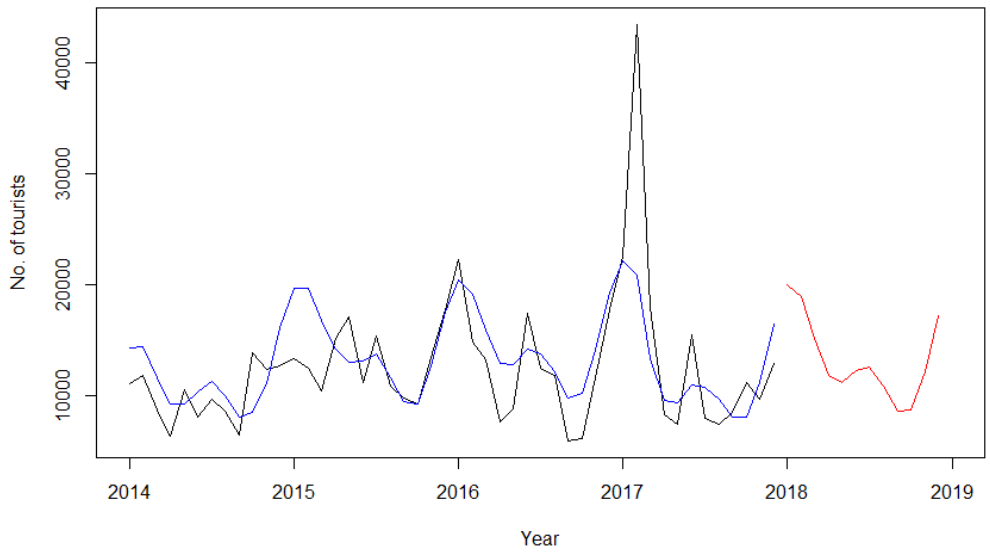


Figure 5.20: Forecasting number of minutes lost with the SARIMA model

The MAPE value for SARIMA training dataset is 25.20%, and for the test dataset is 60.61%.

As status above, three different time series models were used to analyze and forecast the number of minutes lost by the equipment related network failures. The models have fitted accordingly to the data used for this study, and their accuracy of fit was compared. The strategy used for testing the accuracy of the number of minutes lost by the equipment related network failures predictions were obtained by assessing the developed models by their forecast errors by using Mean Absolute Percentage Error (MAPE). The table below (figure 5.21) portrays the comparison of the forecast errors of the developed models for the test dataset.

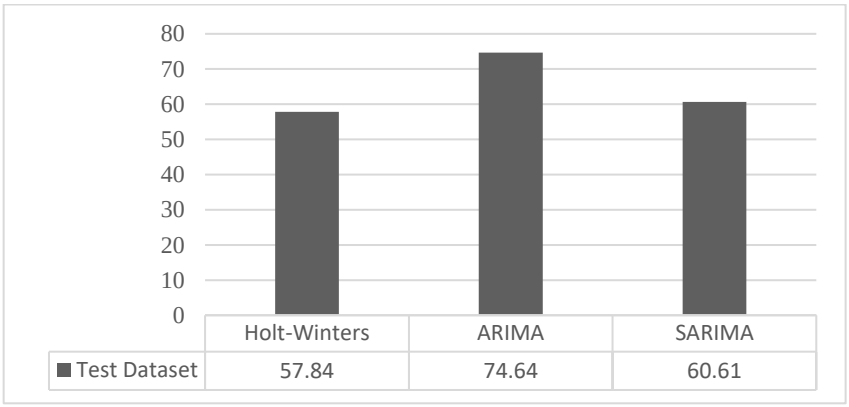


Figure 5.21: Comparison of model accuracy using MAPE - number of minutes lost in the Ausgrid case study

It is apparent from the chart that all the models perform poorly. But the Holt-Winters Additive model performed better than the other two models.

5.2.8 Results Analysis and Discussion

This section of the research is focussed on identifying the seasonality of equipment failures related to electricity distribution faults. It is also investigating the best forecasting models to predict future faults in the network. The forecasting was carried out after exploring data with different attributes of the time series. The seasonal decomposition method was used to distinguish the different components of the time series. Accurate fault analysis and forecasting are essential for the electricity distribution industry. This is because the results of fault forecasting influence the DNOs in various ways. These include engineering staff management, setting investment priorities and setting strategies.

In this study, the performance of three different types of time series forecasting methods (Holt-Winters' Additive method, ARIMA and SARIMA) was measured by the forecasting accuracy measures with an industrial dataset from Ausgrid which is the largest electricity distributor on Australia's east coast. Table 5.3 shows the comparison of the results on the training dataset, and table 5.4 shows the comparison of the results on the test dataset. In both occasions, SARIMA model shows the best results.

Table 5.3: Forecasting results comparison of the training dataset in the Ausgrid case study

	Holt-Winters Additive Method	ARIMA	SARIMA
No. Of Faults	*20.32%*	*15.38%*	*17.51%*
No. Of Customers Affected	*28.00%*	*22.39%*	*24.74%*
No. Of Minutes Lost	*30.11%*	*24.90%*	*25.20%*

Table 5.4: Forecasting results comparison on the test dataset in the Ausgrid case study

	Holt-Winters Additive Method	ARIMA	SARIMA
No. Of Faults	*45.34%*	*37.24%*	*31.41%*
No. Of Customers Affected	*47.97%*	*51.96%*	*36.89%*
No. Of Minutes Lost	*57.84%*	*74.64%*	*60.61%*

This section of the research concluded that equipment failures related to network faults have seasonality. Even faults which have a very uncertainty nature can be predicted using a time series forecasting model which supports seasonality. This work, therefore, makes a major contribution to knowledge by exploring the performance of time series forecasting models on faults data from DNOs.

Although it has been demonstrated that the proposed forecasting methods work on this particular network configuration, it was observed that unexpected severe weather conditions could jeopardise the whole process of accurate fault forecasting. Therefore, additional tests need to be performed to assess the robustness of the developed method. When outliers have been detected due to unexpected events such as wildfire and, storms, it is better to apply appropriate outlier detection technique to clean the data before it is passed to time series models.

5.3 Visual Data Mining Approach for Electricity Distribution Network Faults Triggered by the Equipment Failures using 2D and 3D Calendar Heatmaps

[5] In this section of the study, the author investigates temporal patterns of unplanned interruptions occurred in lowe voltage distribution network. Visualisation techniques such as 2D and 3D Time-Series Calendar Heatmap and unsupervised data mining techniques such as K-mean clustering were performed in order to achieve the objectives. An understanding temporal pattern in unplanned interruptions can be an essential tool for developing business intelligence in the fault management department. If there any interesting temporal pattern exists, appropriate remedial action can be taken to avoid unplanned interruptions. Remedial action may reduce the volume of unplanned interruptions, and that may improve the level of system availability, and it may reduce potential fines associated with network faults from the regulators.

Every Electricity Distribution Network Operator is responsible for recording the fault details, including start date and time and duration of the fault. The fault report received time is the earliest time that a DNO became aware of a loss of supply, an abnormality or a suspected abnormality in the distribution network [79]. It must be the earliest of the date and time at which the Customer informed the DNO about the fault, or an alert was received by automatic alarm systems or Control engineers identifying the issue. In the case where DNO advises a customer to isolate their supply, then the time of the advice being given will be noted as the beginning time of the incident. A pre-arranged incident which requires an interruption to supply to Customers must be treated as a planned incident, and the outage start time should within the period stated on the notification provided to the Customer(s). Network operators are legally bound to inform customers about the date and time of the planned interruption before the interruption, as early as possible and by appropriate means.

[5] This section has been published in below research paper
C. Silva and M. Saraee " Understanding Temporal Patterns in Electricity Distribution Network Faults Using 2D and 3D Time-Series Calendar Heatmaps" 2020 IEEE International Energy Conference (ENERGYCON), Tunisia.

However, to report the incident, the start time must be based on the time the first report was received. The particular date and time of interruption and the particular date and time of repair must be recorded for every single repair stage. The numbers of Customers involved and the total time spent in each restoration stage will be used to calculate the number of Customers interrupted, and duration of interruptions to supply [79]. Where fault started time and resolved time span across a period of two reporting years, the fault must be assigned to the year in which it started.

5.3.1 Experimental Evaluation using Ausgrid Case Study

Previously used Ausgrid case study and dataset in section 5.2 has been used in this study as well. Data has been extracted for five years across seven variables. Some attributes also have been transformed into a suitable form for a better study of the data. The fault dataset will be explored using different visualisation and data mining methods to get an overall idea of the variables and the underlying patterns in the data.

The original dataset contained seven variables and 8,112 observations. After selecting the essential variables for this research, the new dataset was formed with five variables and 8,112 observations from 1st Jan 2014 to 31st Dec 2018. The variables description for the selected variables are shown below (Table 5.5).

Table 5.5: Selected variable description of the Ausgrid dataset for the temporal pattern analysis

Attribute	Type	Description
Incident Month	*Numeric*	*The month of fault occurred*
Incident Hour	*Numeric*	*Hour of fault occurred*
Customers	*Numeric*	*No. of Customers*
Ave Dur. (min)	*Numeric*	*Customer Minutes Lost*
Reason	*Character*	*Fault Cause*

Data preparation includes dealing with data and simplifying the attributes. Data preparation not only simplifies the dataset but also helps in data mining and analysis of the critical findings from the dataset. Different data mining techniques need data to be formatted in a particular way to get a useful outcome from the data. Handling

missing values, reducing the noise, scaling the data are some of the steps that are carried out in this research for the Ausgrid Case Study datasets.

Study about the annual seasonal variations in network faults has already been in literature. However, annual seasonality is not the only important to DNOs; its always better to understand any short-term seasonality or short-term temporal patterns in the fault data. Within the context of electricity distribution network faults, this study aims to make an in-depth analysis of intra-hour faults patterns using time-series calendar heatmap which is a visual data mining approach as an alternative approach to the traditional time series analysis in order to provide new tools to manage the faults in the network.

5.3.2 Visualising Temporal Data: Calendar Based Visualisation

In this study, calendar layout has been used to understand any monthly, daily and hourly temporal patterns. The purpose of the calendar-based visualisation is to provide insights into network faults related to temporal intervals such as hourly, monthly, etc.

Data visualisation is designed to help users understand the significance of their data by placing it in a visual context. Examples are patterns, trends and correlations that might go undetected in text-based data. The data can be shown and understood more easily using data visualisation.

A primary goal of data visualisation is to communicate information clearly and professionally using statistical graphics, plots and information graphics. Numerical data may be encoded using dots, lines, or bars, to communicate a quantitative message visually. It makes complex data more accessible, understandable and usable. Tables are generally used where users will examine a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables. The vital part of data visualisation is choosing the colour for data representation. Three main types matter when choosing colour schemes for data. These are Sequential, Diverging, and Qualitative colour schemes.

Sequential colour schemes are used to organise quantitative data from high to low using a gradient effect. Quantitative data typically needs to show a progression

rather than a contrast. Using a gradient-based colour scheme allows this progression to be shown without causing any misperception.

Diverging colour schemes help to highlight the middle range/extremes of quantitative data by using two contrasting hues on the extremes and a lighter tinted mixture to highlight the middle range. Qualitative colour schemes are used to highlight qualitative categories.

Although data visualisation is about presenting data in a more meaningful and sensible way, it is often imperative to carefully select adequate representational forms to showcase data in the most straightforward but understanding way visibly. Successful data visualisation is only possible if it adheres to some basic principles and guidelines using a critical assessment of a dataset while considering critical questions from the perspective of the targeted audience. Data visualisation can be viewed as an approach which should provide rich and previously unidentified knowledge from the dataset that may provide multiple answers to its audience without making unreasonable demand in order to understand the information being presented.

As technology and society advances, information also advances at an exponential rate, inundating users with a larger and larger volume of data, some of which will be embedded with valuable information that could be helpful to society and human knowledge. Data visualisation, as earlier mentioned, is the use of computer-based interactive visual representation of elaborate and non-interactive based data to augment human cognition. A further significant challenging aspect which needs to be addressed is dataset dimensionality. The dimensionality of data visualisation refers to the variables or attributes that are present in the dataset to be visualised. This leads to the classification of visual presentation of data into different categories depending on their dimensionalities.

- One-dimensional data are dataset which consists of only one variable or attribute. They are also known as Univariate data. They can be visually presented effectively by simple data tools such as tables, Bar and Pie charts, and histogram.
- Two-dimensional data, also known as Bivariate, are dataset with two variables, they often can be displayed using the x-y axis of a 2D space. They equally can be presented as one-dimensional data and in a scattered plot.

- Multidimensional or Multivariate data are defined as a dataset of higher dimensions of three and above. Three-dimensional data can be displayed in the 3D space of the x-y-z axis.

Visual data mining (VDM) is the process of interaction and analytical reasoning with one or more visual representations of abstract data [80]. The process may lead to the visual discovery of robust patterns in these data or provide some guidance for the application of other data mining and analytics techniques. It facilitates analysts in obtaining a deeper understanding of the underlying structures in a data set. The process relies on the tight interconnectedness of tasks, selection of visual representations, the corresponding set of interactive manipulations, and several analytical techniques. Discovered patterns then form the information and knowledge utilised in decision making.

Understanding the temporal aspects of electricity distribution interruptions are important when considering possible remedial actions. Temporal data can be expressed in days, weeks, months, or years. The most common way to visualise temporal data is to use a simple line chart, where the horizontal axis plots the increments of time and the vertical axis plots the variable that is being measured. Nevertheless, in this study, the Time-Series Calendar Heatmap has been used to visualise and identify and any interesting electricity distribution interruption temporal patterns. Heatmaps are a popular visualisation technique that encodes 2D density distributions using colour or brightness. The essence of the calendar heatmap is viewing data overtime at a glance. It is a variation of a traditional heatmap where data is organised using calendar format, and colour encoding are represented using a metric which varies by specified temporal indicators such as day, weekday, month and year.

The calendar heatmap is an alternate visualisation approach used to analyse time-series data. It is usually used to show activity throughout an extended period, such as several months, or multiple years. They are best used when there is a need to illustrate how some quantity varies depending on the day of the week, or how it trends over time. Aggregating and visualising hourly fault data using a calendar heatmap provides a much more granular overview of faults patterns across 24-hour windows over 12 months period. Table 5.6: shows a sample of the Ausgrid AER dataset aggregated by the fault occurred Month and Hour.

Table 5.6: Sample of the Ausgrid AER dataset aggregated by the fault occurred Month and Hour

Fault Occurred Month	Fault Occurred Hour	Fault count	Number of Customers Affected	Customer Minutes Lost
1	*0*	*22*	*14031*	*2801*
1	*1*	*31*	*24676*	*6450*
1	*2*	*23*	*10042*	*5850*
1	*3*	*16*	*10836*	*2298*
1	*4*	*24*	*32296*	*3097*
1	*5*	*26*	*12859*	*3576*
1	*6*	*24*	*14804*	*7570*
1	*7*	*26*	*11756*	*3563*
1	*8*	*44*	*68842*	*6135*
1	*9*	*38*	*16058*	*4407*
1	*10*	*35*	*41183*	*4744*
1	*11*	*38*	*13932*	*5296*
1	*12*	*44*	*24573*	*7516*
1	*13*	*46*	*22186*	*9833*
1	*14*	*49*	*27389*	*8590*
1	*15*	*65*	*39032*	*10969*
1	*16*	*59*	*44580*	*10670*
1	*17*	*67*	*44389*	*10493*
1	*18*	*73*	*36191*	*12634*
1	*19*	*43*	*26679*	*6998*
1	*20*	*55*	*19419*	*7145*
1	*21*	*42*	*52594*	*6131*
1	*22*	*36*	*37504*	*3997*
1	*23*	*22*	*11800*	*4890*
2	*0*	*22*	*27735*	*3692*
2	*1*	*11*	*9922*	*1465*
2	*2*	*18*	*22484*	*2095*
2	*3*	*22*	*12184*	*3863*
2	*4*	*16*	*14054*	*1665*
2	*5*	*16*	*24938*	*2920*
2	*6*	*16*	*7089*	*6937*
2	*7*	*25*	*14325*	*4948*
2	*8*	*28*	*11218*	*7982*
2	*9*	*26*	*35524*	*4909*
2	*10*	*24*	*8590*	*3327*
2	*11*	*38*	*26803*	*6112*
2	*12*	*35*	*22729*	*4702*
2	*13*	*30*	*13732*	*4035*
2	*14*	*38*	*22301*	*9137*
2	*15*	*42*	*15237*	*5858*
2	*16*	*54*	*42007*	*9792*
2	*17*	*72*	*35754*	*11591*
2	*18*	*68*	*27689*	*13130*
2	*19*	*51*	*20256*	*7707*
2	*20*	*54*	*36594*	*12977*
2	*21*	*31*	*17355*	*5297*

As an initial step, the author has used calendar-based visualisation on aggregated monthly data over five years (shown in figure 5.21). This type of monthly temporal pattern comparison may identify monthly patterns of network faults. It may also assist engineers in understanding the monthly temporal patterns and in taking action to prevent future network failures.

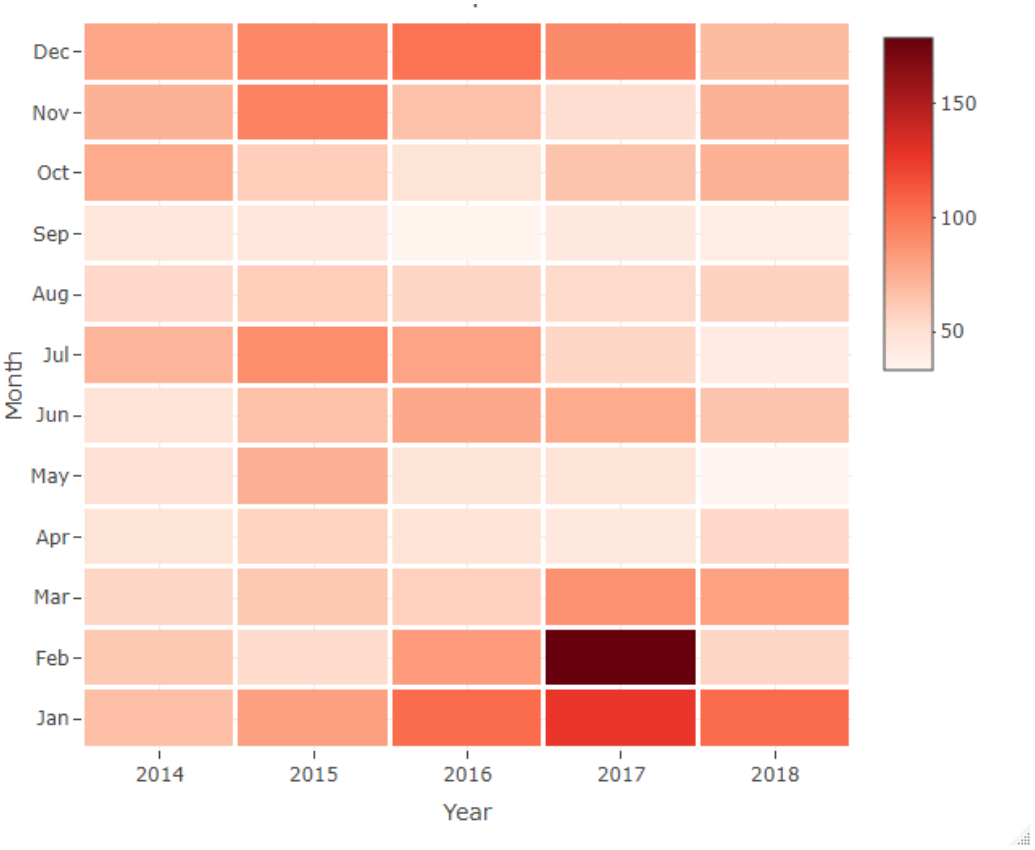


Figure 5.22: Calendar-based visualisation on the Ausgrid aggregated monthly data

5.3.3 Data Visualisation using Hourly-Monthly Calendar Heatmaps: The Ausgrid Case Study

Heatmaps can be defined as a graphical representation of data that employ colour-coded systems. The primary purpose of heatmaps is to visualise the volume of measures within a dataset and have the potential to enable improved user understanding of the data presented. Data visualising using calendar heatmaps are straightforward to understand and most importantly are useful. Even though they represent much data in a compact visual pallet the use of colour gradients and a recognisable arrangement of the data providing a unique identity for Calendar a heatmaps. Aggregating and visualising hourly faults data on a calendar heatmap gives

much more granular overview of the faults patterns across the 24hour windows. The colour scale represents the number of faults. In figure 5.23 shown below, black represents the low numbers, and red represents the high number of faults. However, a human-guided process is required to evaluate the distribution and magnitude of colours. The same colour scale across a given hour shows changes in a number of faults over that given hour. In figure 5.23, substantial faults increase between 3 pm and 9 pm. Between midnight and 6 am consistent, a low fault count is shown throughout five years. The number of network faults is higher between 3 pm and 9 pm because of higher energy consumption. The lower faults numbers occur during the night, especially during the period after midnight till 6 am.

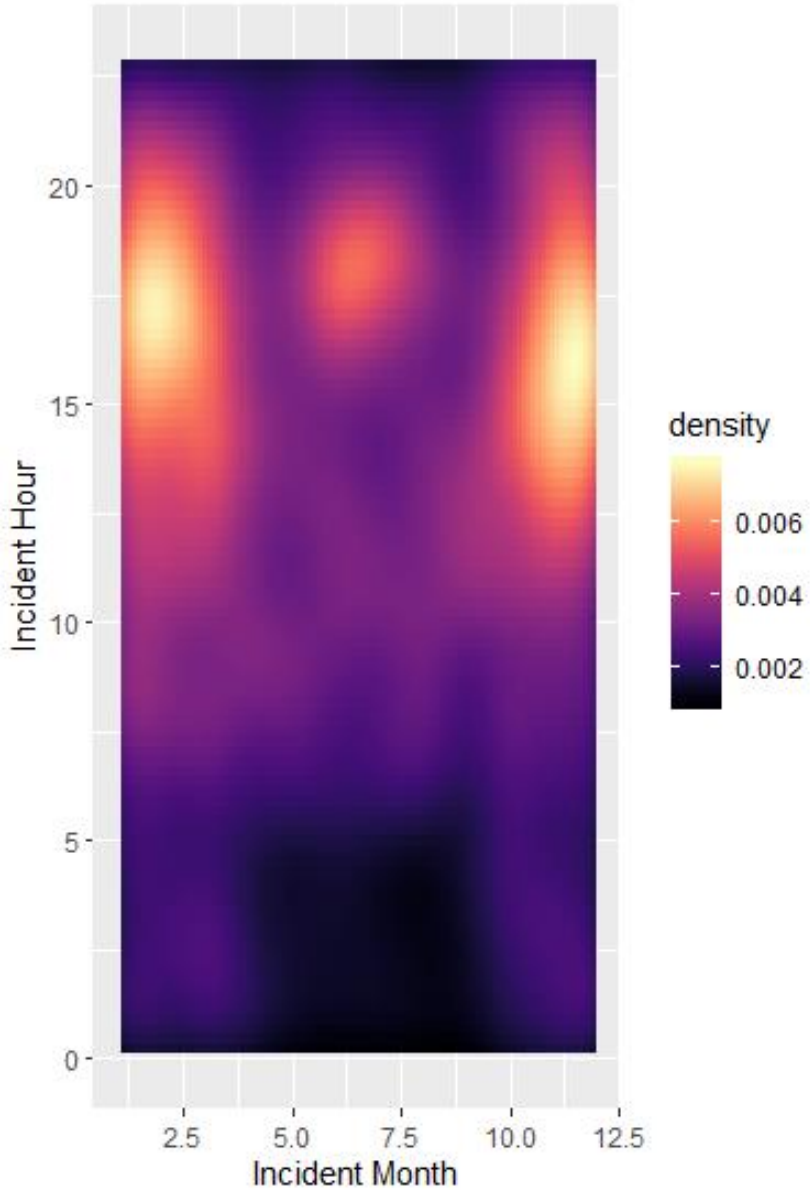


Figure 5.23: Visualising Number of Network Faults on 2D Hourly Calendar Heatmap using Equipment Related Distribution Network Faults in the Ausgrid Dataset

The figure showed unsurprising peaks of network faults during the evening hours, and a far lower rate during the night than the day, but the pattern varies throughout the year. It showed that January, February, June, July, November and December follow the same pattern. However, the month of March, April, May, August, September and October follow a different pattern.

Even though above 2D heatmap is very useful to understand the temporal patterns, the author has used 3D heatmaps to enhance heatmap usability. In 2D and 3D, the “D” specifies the dimensions involved in the shape. So, the primary difference between 2D and 3D shapes is that a 2D shape comprised of two dimensions that are length and width. As against, a 3D shape incorporates three dimensions that are length, width and height.

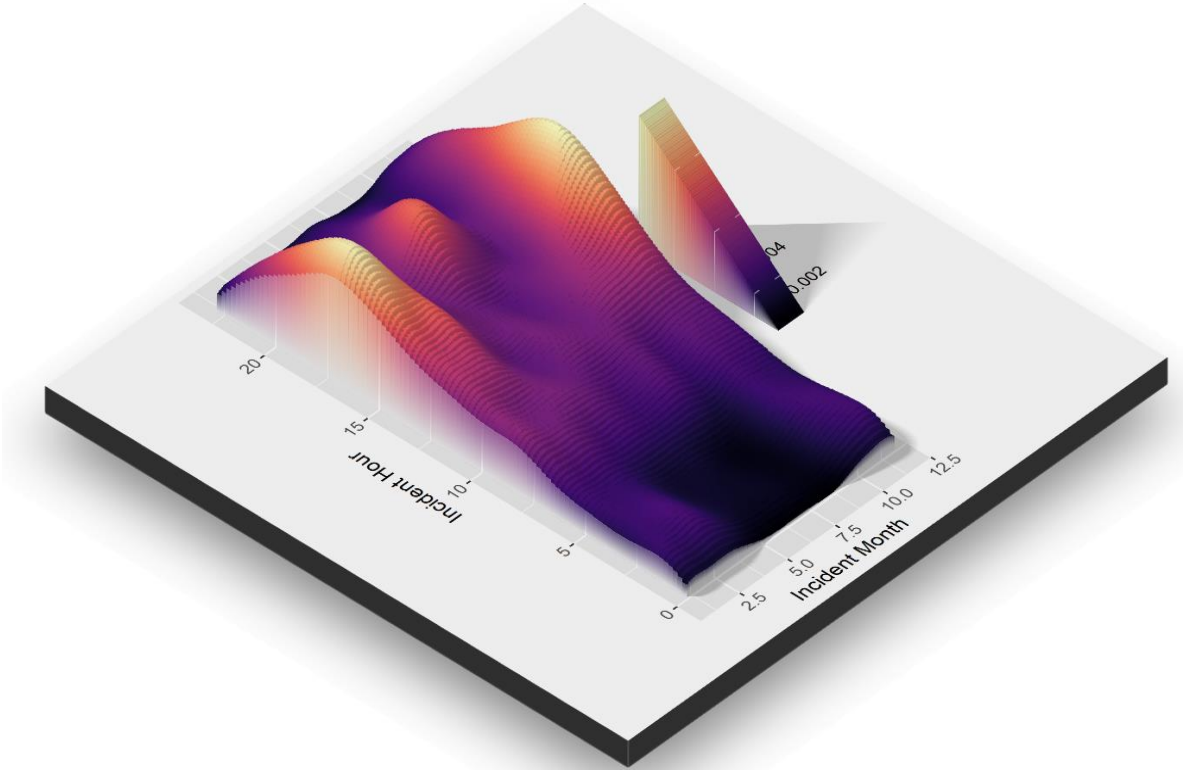


Figure 5.24: Visualising Number of Network Faults on 3D Hourly Calendar Heatmap using Equipment Related Distribution Network Faults in the Ausgrid Dataset

As shown in figure 5.24, data is plotted on a Heatmap chart in terms of month, hour, and fault count. The data is plotted as a three-dimension surface. The colour and height are representing the fault count number and its magnitude. In a 3D heat map graph, the number of faults is a numeric scale, while month and hour are represented as a label, so month and hour values will be evenly spaced according to the number

of rows or columns in the data source. While other kinds of analytics have their strengths, heatmaps draw instant attention to data distribution and the outliers on the data sources.

By comparing 2D and 3D heatmaps, 3D heatmaps are certainly better than the 2D heatmaps. 3D heatmaps can help fault engineers to understand the fault distribution better and ultimately give them a better visualisation experience. However, that does not mean 3D heatmaps provide all the answers to the data analytics problem. They should be seen as a vital tool that should be in data analytics toolkit. They work best when combined with other data analytics and visualisation techniques.

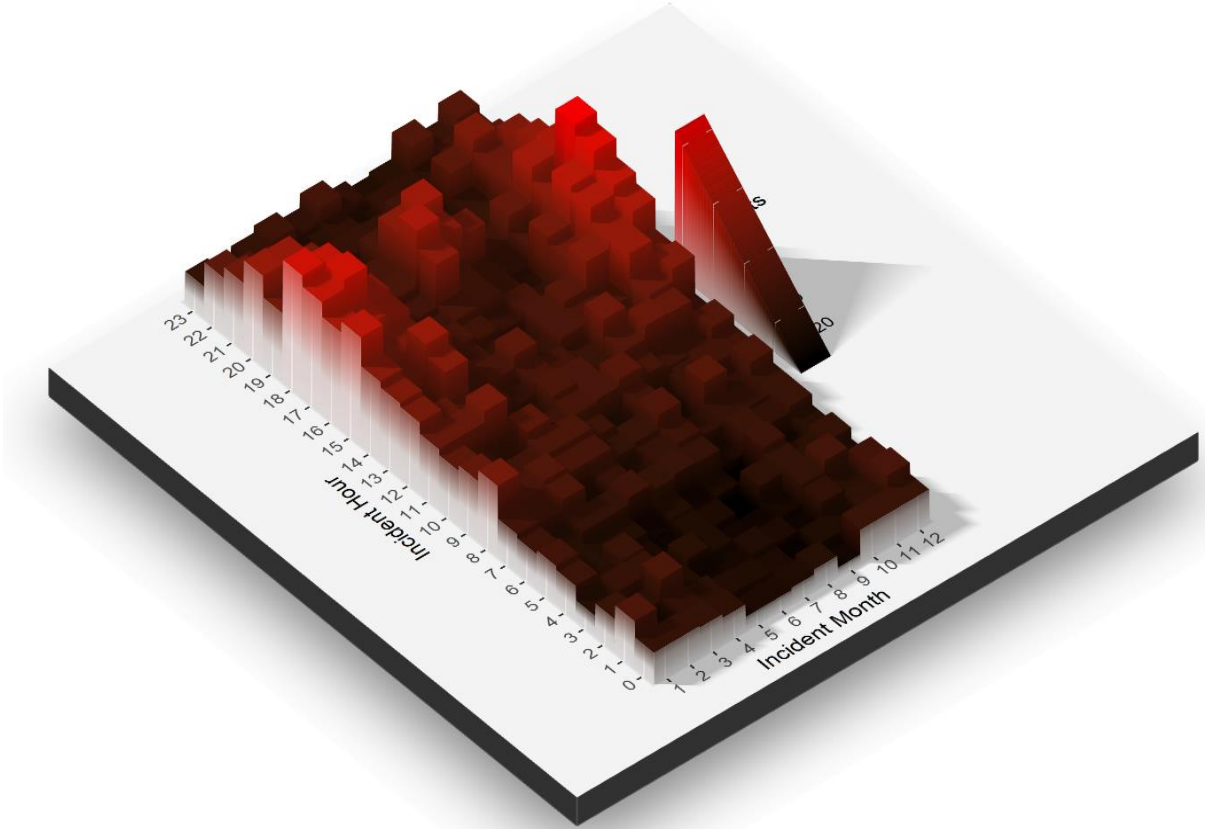


Figure 5.25: Visualising Number of Network Faults on 3D Hourly Calendar Heatmap - Bivariate Histograms

Figure 5.25 shows the same 3D data graph that was shown in figure 5.24. It has used histogram style pillars to represent the data objects. This type of graph is called a Bivariate Histograms. 3D bivariate histograms are used to visualise crosstabulations of values in two variables. They can be considered to be a conjunction of two simple histograms, combined such that the frequencies of co-occurrences of

values on the two analysed variables can be examined. Although specific frequency data are easier to read in a table, the overall shape and global descriptive characteristics of bivariate distributions may be more straightforward to explore in a graph. This is particularly useful for quickly modifying the properties of the bins or changing the display.

In some cases, the Figure 5.25 style 3D calendar heatmap is a better representation of the fault data, enabling the user to understand the actual magnitude of the data points. Histogram style 3D heatmaps are incredibly adaptable and resourceful in drawing attention to trends and patterns. Therefore, they have become increasingly popular within the analytics community.

To validate further the findings of this research, the number of customers involved, and the number of customer minutes lost has also been analysed in the same way as the number of faults. The results are shown in Figures 5.26 – 5.29.

5.3.4 Visualising Number of Customers Involved using Hourly-Monthly Calendar Heatmaps: The Ausgrid Case Study

In this section of the study, a number of customers affected by the equipment related network failure time series data has been used to visualise on Hourly-Monthly Calendar Heatmaps. Aggregating and visualising hourly number of customers affected data on calendar heatmap gives much more granular overview of the temporal patterns across the 24hour windows over 12 months period. The colour scale represents the number of customers affected. In figure 5.26 shown below, black represents the low numbers, and red represents the high number of customers affected.

The colour scale is useful to study the underlying patterns that exist in the data related number of customers affected—however, the human-guided process required to evaluate the distribution and magnitude of colours. The same colour scale across a given hour shows changes in a number of faults over give an hour.

In figure 5.26 shown below, a substantial number of customers affected inflation between 3 pm and 10 pm. Also, in most months, between midnight and 7 am show a consistently low number of customers affected throughout five years. Even

figure showed unsurprising peaks of a number of customers affected during the evening hours and a far lower rate during the night than the day but did not follow the same pattern throughout the year.

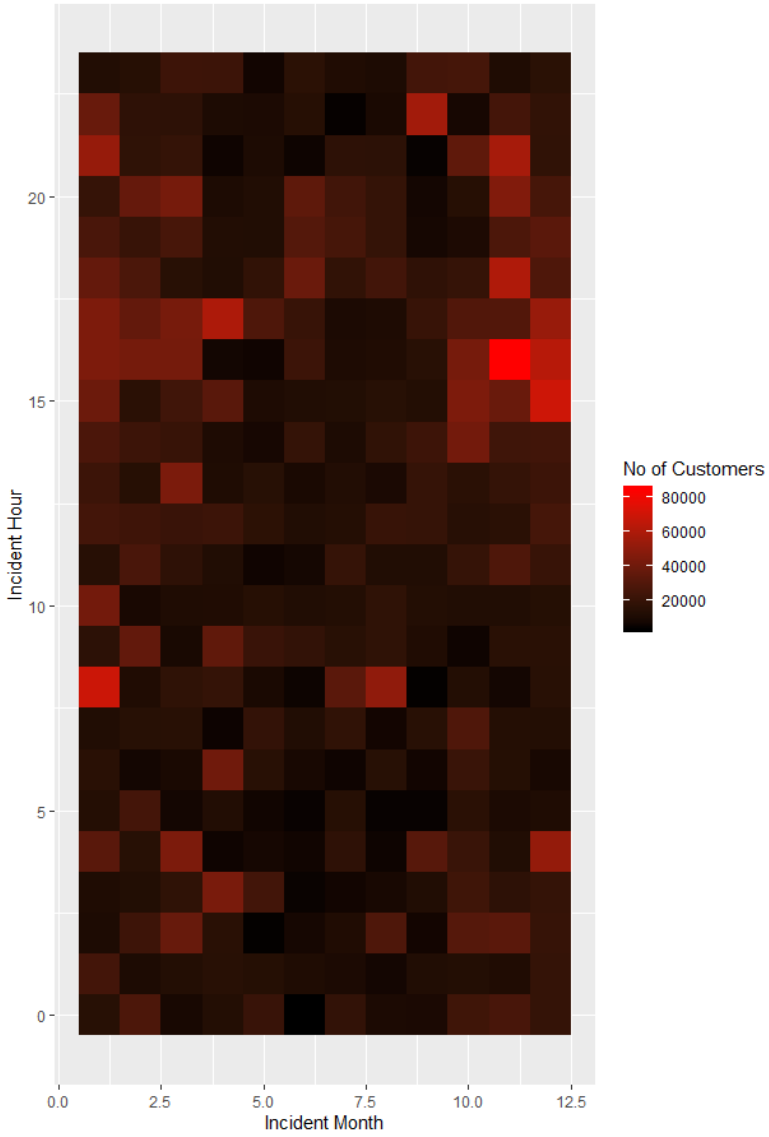


Figure 5.26: Visualising number of customers affected on 2D Hourly-Monthly Calendar Heatmap

Even though above 2D heatmap (Figure 5.26) is useful to understand the temporal patterns of a number of customers affected, the author has proposed to used 3D heatmaps to enhance the usability of the heatmap.

As shown in figure 5.27, data is plotted on a Heatmap in terms of month, hour, and a number of customers affected. The data is plotted as a surface, with a three-dimension—the colour and height representing a number of fault count and its magnitude. By comparing 2D and 3D calendar heatmap, the main difference between

the 2D and 3D shapes is that in 2D shapes only two axes are incorporated x and y-axis. On the other hand, in 3D shapes, the three-axis x, y and z-axis are covered. 3D calendar heatmap is providing better visualisation to understand the actual magnitude of the data points. But in some cases, 3D heatmaps have disadvantages such as in 2D shapes it shows all the edges of that shape, but in 3D shapes, these edges could be hidden. For example, in figure 5.26, all the data points are visible. However, take an example of figure 5.27; then, it is not possible to display all of its data points from one angle.

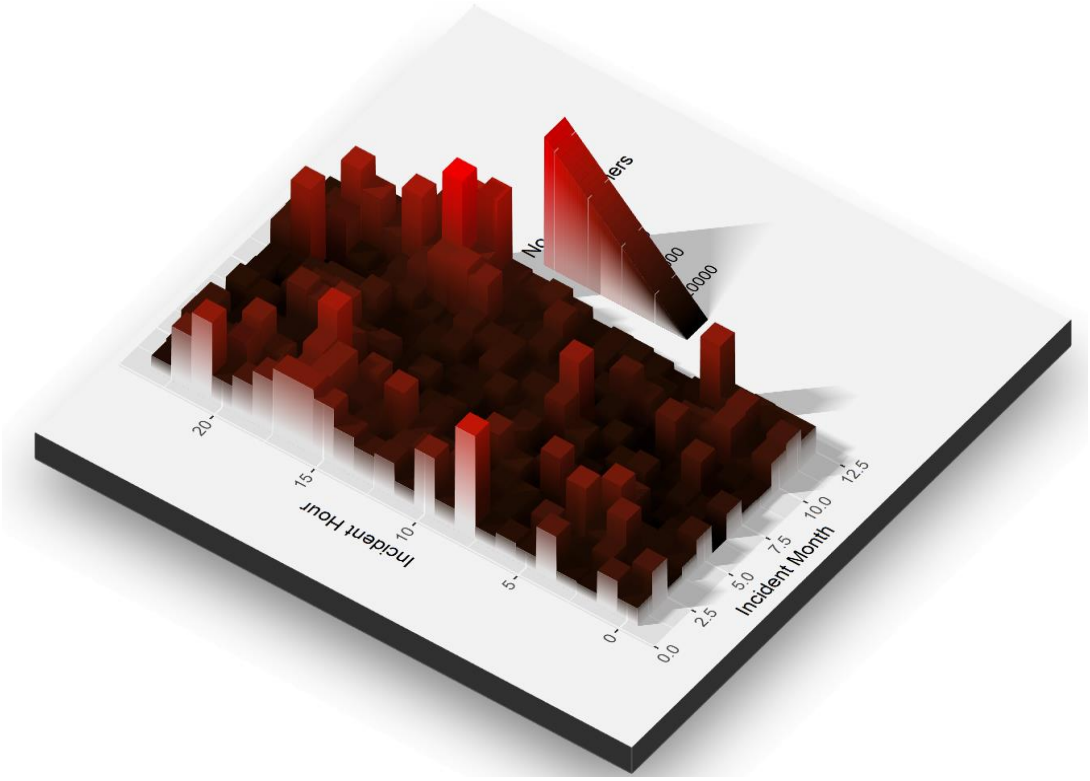


Figure 5.27: Visualising number of customers affected on 3D Hourly-Monthly Calendar Heatmap

5.3.5 Visualising Number of Customer Minutes Lost using Hourly-Monthly Calendar Heatmaps: The Ausgrid Case Study

In this section of the study, a number of customer minutes lost due to the equipment related network failure data has been used to visualise on Hourly-Monthly Calendar Heatmaps. Heatmaps show data point by representing data density with a colour overlay. Heatmaps are ideal for helping end-users to make sense of big data sets, aiding rapid comprehension insights that often remain buried in a large dataset.

Visualising a number of customer minutes lost on calendar heatmap gives a quick overview of the temporal patterns across the 24hour windows over 12 months period.

In figure 5.28 shown below, a substantial number of customer minutes lost increase can be identified between 2 pm and 10 pm. Also, in most months, between 10 pm and 6 am show a steady low number of customer minutes lost throughout five years. Even figure showed unsurprising peaks of a number of customer minutes lost during the evening hours and a far lower rate during the night than the day but did not follow the same pattern throughout the year.

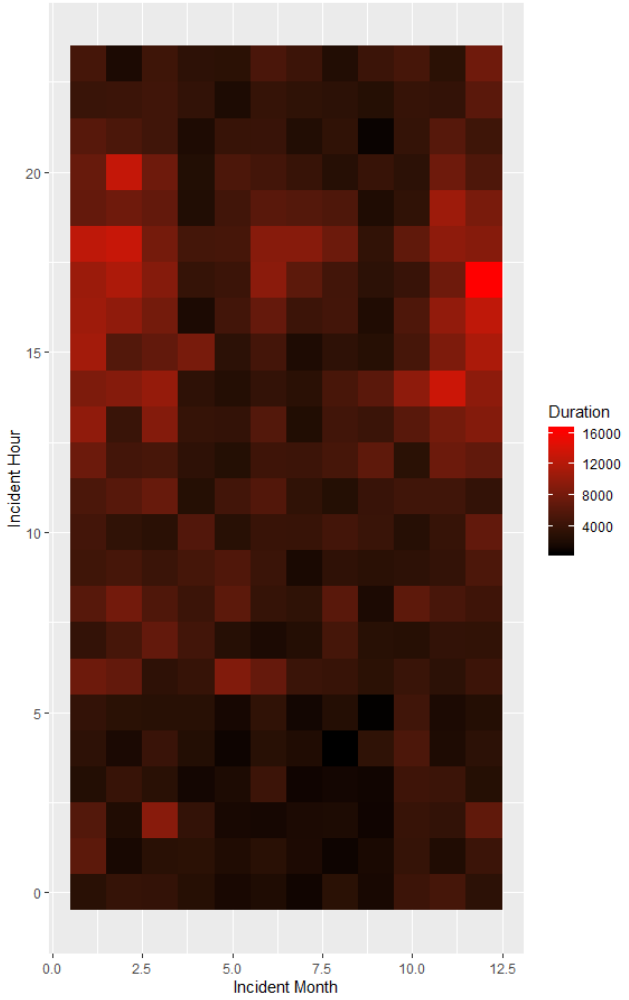


Figure 5.28: Visualising number of customers minutes lost on 2D Hourly-Monthly Calendar Heatmap

As shown in figure 5.29, the same dataset used to create figure 5.28 has been plotted on a 3D calendar Heatmap. Calendar heatmap is great for circumstances where users need to spot patterns and trends in a large temporal dataset. Using calendar heatmap to visualize this data on 3D view reduces cognitive load for users to

help them comprehend complex data more efficiently. 3D visualization takes the data beyond a flat, top-down overlay, allowing users to view their data from different perspectives.

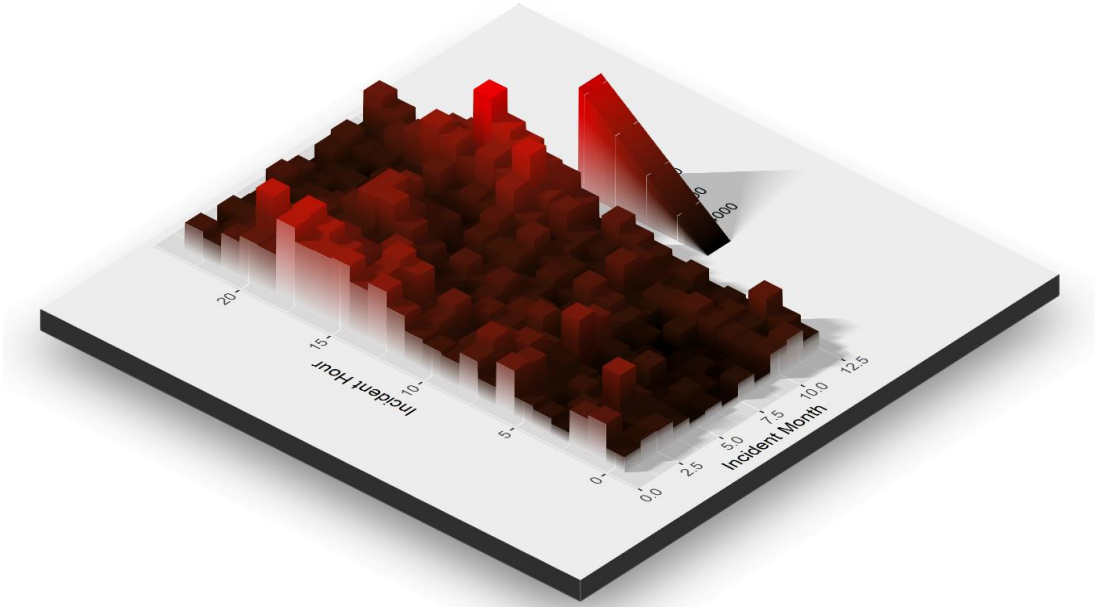


Figure 5.29: Visualising number of customers minutes lost on 3D Hourly-Monthly Calendar Heatmap

5.3.6 K-Means Cluster Analysis using Aggregated Ausgrid Temporal Data

In this study, the K-Means clustering method, which is an unsupervised machine learning algorithm, has been used to cluster the aggregated temporal faults data. Cluster boundaries were defined by the levels of similarity between faults using the data attributes. The techniques involved methods to calculate a distance measure, ascertain the cluster-ability of the dataset, and the optimum number of clusters that can be obtained from the dataset using Hopkins-Statistics and Elbow plot, respectively.

Using Hopkins Statistics on the aggregated temporal faults, data shows a value of 0.21, which is low. For well-defined clustered data, the value should be larger than 0.5. The value is expected to be much less than 0.5 for data that are neither clustered nor random. Hopkins Statistics is known to be a fair estimator of randomness in a data set. Nevertheless, in some cases, outliers can pollute the dataset. It is better, therefore, to remove outliers from the research dataset to improve the Hopkins Statistics value to achieve better clustering.

A well-known method of determining the optimal value of K in K-means is the Elbow Method. This method involves drawing a curve between the WSS (within the sum of squares) and the number of K. As K increases, the WSS will reduce. For k=n (with n being the total number of observations) the WSS will become zero. Using the elbow method to analyse the aggregated temporal faults dataset, produces. The results are shown in Figure 5.30, where the graph tends to flatten. Increasing the number of clusters beyond three does not significantly reduce the WSS. This inflation point is usually chosen to be the value of K for K-means.

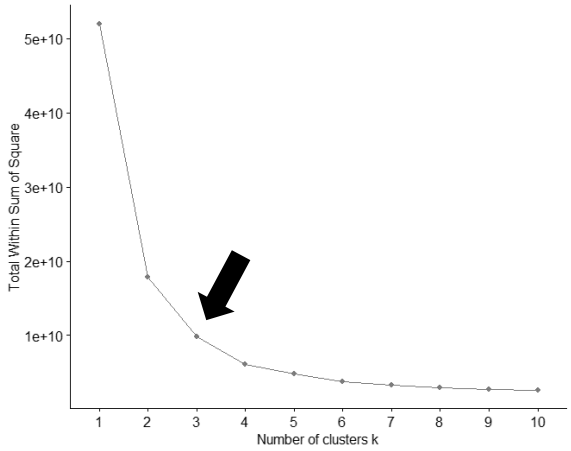


Figure 5.30: Optimal number of clusters for aggregated temporal faults data

The K-Means clustering of the faults data using the optimum cluster value reveals the distribution of the temporal faults data into 3 clusters. Figure 5.31 gives a pictorial representation of the clusters formed using the K-means cluster analysis.

Figure 5.31: K-Means cluster analysis for aggregated temporal faults data

For a better visualisation perspective, temporal clusters can be visualised on the Hourly-Monthly Calendar Heatmap. As is shown in Figure 5.32, aggregating and visualising hourly fault data on a calendar heatmap gives a much-improved granular overview of the fault patterns across the 24-hour windows over a 12 months window. The three-colour scale (Bright Red, Brown and Black) represents the different clusters. Temporal clusters can be clearly identified by examining the calendar heatmap shown in Figure 5.32.

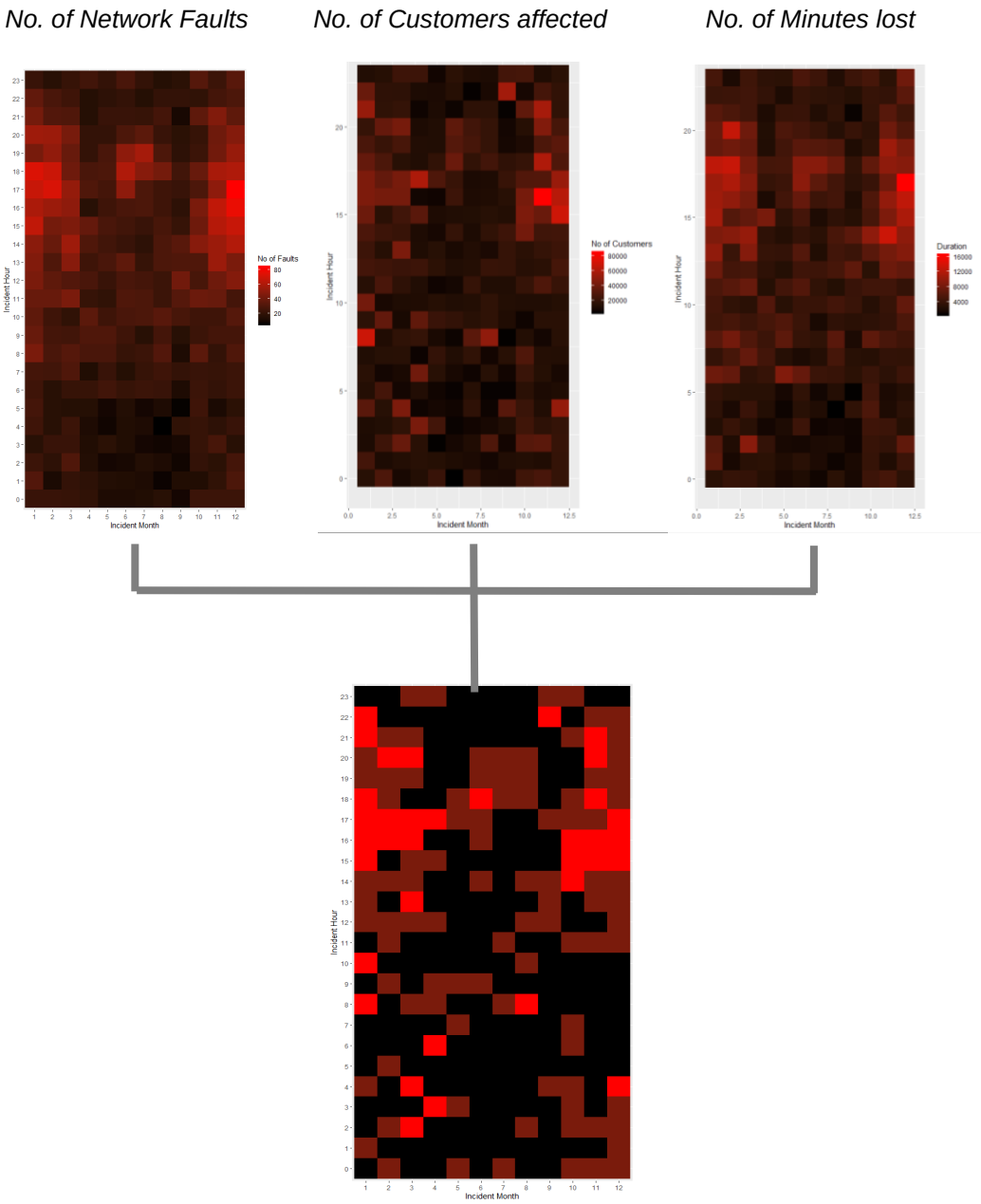


Figure 5.32: Clustered Hourly-Monthly Calendar Heatmap for aggregated temporal faults data

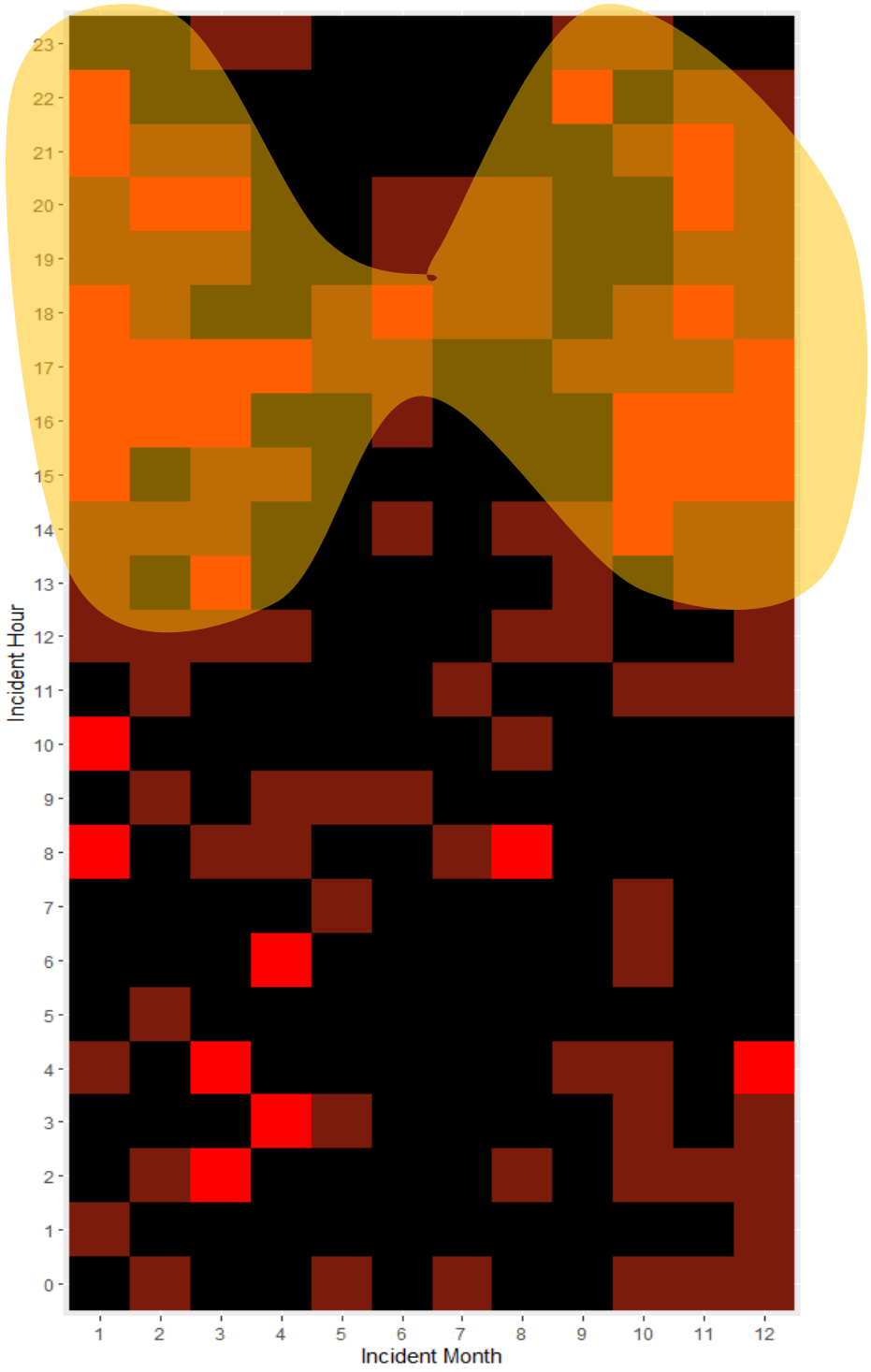


Figure 5.33: Cluster analysis results on Hourly-Monthly Calendar Heatmap

Figure 5.33 shows the Cluster analysis results on Hourly-Monthly Calendar Heatmap. This heatmap shows clustered data pointed with three different colours (red, black and brown), visualise by hour cross the months.

5.3.7 Results analysis and discussion

In this section of the research, the author investigated the temporal and seasonal pattern of electricity distribution network faults by examining the start time of the faults. To achieve this, 2D/3D time-series calendar heatmap and k-means clustering have been used as a new approach to analysing the temporal patterns.

The results obtained from the new approaches identified seasonality, and that temporal patterns exist in the fault data. Cluster represented in red colour in Figure 5.33 shows that more network faults are centred between 3 pm and 8 pm and cluster represented in back colour is more centred between 12 am and 12 pm. Clear pattern cannot be identified in cluster represented in brown colour. Table 5.7 shows the cluster visualised in red in Figure 5.33 and its represented values.

Table 5.7: One of the identified cluster and its represented values for the AER Case study.

Fault Occurred Month	Fault Occurred Hour	Cluster (Marked in Red)
1	8	3
1	10	3
1	15	3
1	16	3
1	17	3
1	18	3
1	21	3
1	22	3
2	16	3
2	17	3
2	20	3
3	2	3
3	4	3
3	13	3
3	16	3
3	17	3
3	20	3
4	3	3
4	6	3
4	17	3
6	18	3
8	8	3
9	22	3
10	14	3
10	15	3
10	16	3
11	15	3
11	16	3
11	18	3
11	20	3
11	21	3
12	4	3
12	15	3
12	16	3
12	17	3

The purpose of clustered heatmap visualisation is to recognise patterns and trends unseen to the human eye, which will allow fault engineers in DNOs to interpret the specific faults in the network occur monthly or on certain hours.

The findings are generally consistent with daily consumer power consumption patterns and seasonal weather patterns. The results also show that faults can be clustered using faults measures such as the number of customers affected, number of minutes lost and the fault count. These results demonstrate how the 2D and 3D calendar heat map method can help provide a relatively new perspective when evaluating temporal patterns in electricity distribution network faults.

In present days, smart metering and IoT devices produce a huge number of real-time temporal data in the distribution network infrastructure. With this large number of temporal data, visualisation and data mining integration in the dynamic Decision Support Systems in electricity distribution network reduces subjectivity and provides a new interesting visual knowledge for decision-makers. In this context, dynamic decision-making, visualisation and data mining are important trends and acquire a more and more significant place in the research field.

Although it has been demonstrated that the proposed visualisation methods work on this particular network configuration, it was observed that unexpected severe weather conditions could result in the high peak of fault counts. Therefore, additional analysis needs to be performed to identify any anomalies in the data. If outliers are detected, it is recommended that an appropriate outlier detection technique is applied to clean the data before it is passed to time-series calendar heatmaps.

.

5.4 Summary of Chapter and Conclusion

The electrical power distribution system is growing in size and complexity in all around the world due to the high demand for electricity. Unplanned interruptions in the power system network result in severe economic losses and reduce the reliability of the electrical system. Unplanned interruptions can be caused by equipment failures such as transformers or substations, human errors, and adverse environmental conditions, etc.. In this section of the research mainly focused on the equipment related unplanned interruptions.

Equipment malfunctions can occur in any of the equipment in the network such as transformers, switchgear, overground cables, underground cables or substation. Therefore, quick elimination and prevention of network faults are essential for the DNOs. Understanding failures in the network asserts are of the highest importance for the operator. It is challenging to predict equipment failure accurately for a given period due to the uncertain nature of the equipment failures. This issued the need to research and investigates how to detect faults as accurate and as early as possible. An accurate fault forecasting plays a crucial role in maintaining an electric power system. The fault forecasting model is based on historically observed fault patterns and, therefore, able to accurately translate data into expected fault volumes.

In this section of the research, the author is investigating the daily faults caused by equipment failures. Research is targeting to finding the seasonality of the equipment failures related to the electricity distribution network. Also, it is investigating to find the best forecasting models to predict future faults in the network.

In this study, the performance of three different types of time series forecasting methods (Holt-Winters' Additive method, ARIMA and SARIMA) was measured by the forecasting accuracy measures with an industrial dataset from Ausgrid which the largest electricity distributor on Australia's east coast. Results of the experiment show the SARIMA model shows the best results. This research can conclude that equipment failures related to network faults have seasonality—also, even faults caused by equipment failures. Have very uncertainty nature; they can be predicted using time series forecasting model which support the seasonality.

In this section of the study, the author also has attempted to understand the temporal patterns of unplanned interruptions in the distribution network. Visualisation

techniques such as 2D and 3D Time-Series Calendar Heatmap and unsupervised data mining techniques such as K-mean clustering were performed in order to achieve the objectives. Also, clustered heatmap has been used to visualise the clustered data.

The most common approach of visualising time series data is a basic line graph, but this research has adopted the 2D and 3D calendar heatmap to detect temporal patterns and trends in the distribution network. Heatmaps are employed to display extremely large data. They display a large number of data points and use colour and order to dictate and annotate similarities and/or differences between data points. Calendar heat maps are an alternative method of analysing time-series data.

Clustering analysis has been performed with aggregated temporal fault data. It is a process of partitioning a dataset into groups of meaningful subclasses where grouped objects share common traits. It is implemented to understand the natural hidden structure in the data that would otherwise remain unobserved. Usually, scatterplot will be used to displaying clustering results. But the author has proposed to use 2D and 3D calendar heatmap to visualise cluster results. The purpose of clustered heatmap visualisation is to recognise patterns and trends unseen to the human eye, which will allow fault engineers in DNOs to interpret the certain faults in the network occur monthly or on certain hours.

A combination of data mining and unique kind of visualisation may offer a solution, however in all likelihood modelling 2D and 3D approaches or even interactive interpretations of these plots would signify significant improvements upon the data mining and machine learning models. Ultimately future studies should focus on these alternative strategies of visualisation, as understanding and rapidly interpreting these data sets will help hasten our understanding in numerous scientific fields.

CHAPTER 6

Analysis of Impact of External Factors on Faults in Electricity Distribution Network

6.1 Introduction

The purpose of this section of the study is to discover the relationship between external factors like population density and faults causes in an electricity distribution network using machine learning classification models. Also, the study aims to identify the relationship between DNO's Key Performance Indicators and network faults. The study will use different classification models and comparing the results. The results of this study should identify the ideal classification model to use in understanding the relevant relationships.

In this study, the correlation method has been used for feature selection to select the most suitable variables to build the classification models. It should also give more insight into how the data should be prepared before being input into a machine learning classification models. Correlation analysis has revealed the multifaceted relationships that exist among the variables in multivariate fault data

The ongoing growth of population followed by improvements in complex energy systems has raised challenges for distribution network providers and electrical engineers to assure system sustainability and efficiency. One of the essential duties of any DNO is that customers must be guaranteed a reliable continuous supply of electricity. For this reason, the average period in which UK households were not supplied with power is calculated and analysed annually by the Ofgem regulator [3].

Based on various demographic factors, they set annual targets for customer minutes lost and customer interruptions for each DNO [81]. Each year DNOs must report on their network performance under their licencing conditions. This performance report allows regulators to assess if targets have been met and to reward or penalise DNOs appropriately.

According to the latest Ofgem's annual report, Network operators continue to perform well and exceed their overall targets for 2017-18; however, several network operators missed individual elements of their goals [82]. Ofgem incentivise distribution network operators to improve the reliability on their network, penalising underperformance and rewarding those who exceed their targets [82]. The data shown in the annual Ofgem report provides an overview of the financial results of DNOs and their performance. Two main areas where costs may exceed budgets are network error and operations support costs. It is estimated that in 2016-17, the impact of external factors such as weather phenomena (e.g. flood) have driven up these extra costs.

According to the statistics from Ofgem, the UK annual actual average minutes lost per customer per year is 37. By comparison internationally, the UK has an above-average continuity of electricity supply. In order to maintain this high level of system availability, it is necessary to ensure appropriate investment incentives for the expansion and maintenance of the grid infrastructure.

The deregulation and high competition in the electricity supply market put significant pressure on DNOs to improve the consistency and reliability of the services they provide. However, since the beginning of 2015, customer interruptions have fallen by 11%, and the duration of outages has fallen by around 9% [83]. Table 6.1 shows the customer interruptions and minutes lost from the UK main DNOs.

Table 6.1: Customer interruptions and minutes lost figures from the UK main DNOs

	Target for customers interrupted	Actual number of customers interrupted	Target average minutes lost per customer per year	Actual average minutes lost per customer per year
Electricity North West	47.45	33.23	44.23	34.63
Northern Powergrid	64.32	49.97	61.03	40.51
Western Power Distribution	62.17	53.27	42.17	31.57
UK Power Networks	53.12	36.47	45.8	32.03
SP Energy Networks	44.62	35.9	43.58	32.09
Scottish and Southern Electricity Networks	65.27	56.24	55.14	51.4

NAFIRS adopts six groups of Direct Cause classifications of fault causes. The direct cause is the prime reason for the occurrence of the incident. Many faults are coded as cause "unknown" due to genuine difficulty in identifying a "direct cause", particularly where there is no damage, and the circuit is restored without the fault being found. From this study results and analysis, the shows that there is a new relationship between local population density and fault causes which suggest fault causes has a strong relationship with population density. These findings may help DNOs in policy-making and network design. Also, this research may assist in Smart City planning projects.

6.2 Analysis of Identifying the Relationship Between Population Density and Network Faults

[6]The main objective of this section of the study is to explore and understand the relationship between population density and LV faults and to discover significant patterns that can be used to identify any specific population density area issues. The author has used an exploratory case study research approach as a research methodology. Case study research, usually allows researchers to explore and understand complex issues. It can be considered a robust research method, particularly when a holistic, in-depth investigation is required. The case study method also enables a researcher to carefully examine data within a specific context. Exploratory research can be described as research used to investigate a problem which is not clearly defined [9]. The exploratory case study research also enables the researcher to develop a hypothesis for further research and, to discover new insights or understanding the issue from different dimensions. This study will address the challenges identified by using data mining and machine learning approaches.

[6] This section has been published in below research paper
C. Silva and M. Saraee, "Understanding the Relationship Between Population Density and Low Voltage Faults Causes in Electricity Distribution Network," 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2020, pp. 1-6, doi: 10.1109/ASET48392.2020.9118391. (Appendix 2)

6.2.1 Population Growth and Energy Consumption

Energy consumption is influenced by the characteristics of households which include building size, household income, total energy cost, and building characteristics. Fast-growing urban populations consume more energy than rural areas. The energy consumptions pattern and its sustainability influence population growth in urban communities and cities. Many research studies address the forms of sustainability and energy efficiency in cities and urban environments. The widely available UK regional population density data has allowed the researchers to analyse how components of the UK energy system interact with and relate to centres of population.

Population density can be defined as the number of inhabitants per hectare (1 Hectare is equal to 0.01 square kilometre). While the average UK population density is about nine people per hectare, most people live in cities, which have a much higher density. Even among cities, density values can vary considerably from one city to another.

Among the DNOs, two main areas have high levels of overspend. These are network faults and operational costs. Therefore, there is a business need to reduce operational expenditure in engineering departments. An in-depth understanding of failures and the fault causes can, therefore, be vital for DNOs in order to reduce network faults. There are many studies, and the literature reviewed shows that there is a relationship between energy consumption and population density.

Nevertheless, the researcher has been unable to identify any study which has been carried to understand the correlation between population density and distribution network faults. The different types of network faults may have different effects on population density. The area-wise UK population density data allow researchers to analyse the correlation between the electricity distribution network faults and local area population density.

6.2.2 Case Study: NaFIRS Database from a UK DNO

The dataset used in this research has been extracted from a real NaFIRS database of a DNO in the UK. Due to the nature of commercial sensitivity of the data, some of the data fields are not included in this research study, e.g., postcode, asset numbers, etc.

The UK regional population density data has been extracted from the Local Government Association [84]. Table 6.2 shows the attributes' explanation of the dataset with area density data. Data exploration was carried out to gain a thorough understanding of the data.

Table 6.2: Attributes explanation of the NaFIRS dataset's contents

ATTRIBUTE NAME	TYPE
Area Code	*Factor*
Area Density	*Numeric*
Avg. Min Lost - Third Party	*Numeric*
Fault Count - Third Party	*Numeric*
Avg. Min Lost - Unclassified or Unknown	*Numeric*
Fault Count - Unclassified or Unknown	*Numeric*
Avg. Min Lost - Weather and Environment	*Numeric*
Fault Count - Weather and Environment	*Numeric*
Avg. Min Lost - Company	*Numeric*
Fault Count - Company	*Numeric*

Before the data can be loaded into a database and modelled, it must first be collected, explored, and summarised. By examining the data, its quality can be verified, and any missing values or inconsistent formats can be identified. In the data science community, a common phrase "garbage in, garbage out" describes the need for data exploration. If the data is not cleaned ahead of time, models created from that data may be potentially misleading or useless, depending on the structure of the original data. Data exploration also allows for the identification of attributes that may be useful in the data mining process, together with the selection of irrelevant attributes which can be disregarded. Table 6.3 shows the sample research dataset. The original dataset contained ten variables and 182 rows. For confidentiality purpose, all the real area identification attributes are anonymised and removed from the study and to preserve the anonymity of the area details, and all the records are pseudonymised.

Table 6.3: Sample of the NaFIRS dataset aggregated by the area code (top 30 observations)

Area Code	Area Density	Avg Minutes Lost - Third Party	Fault Count - Third Party	Avg Minutes Lost - Unknown	Fault Count - Unknown	Avg Minutes Lost - Weather	Fault Count - Weather		Avg Minutes Lost - Company	Fault Count - Company
Area 1	*13.3*	*84*	*68*	*76*	*386*	*210*	*69*		*201*	*450*
Area 2	*5.3*	*132*	*28*	*67*	*143*	*175*	*40*		*189*	*159*
Area 3	*7.8*	*159*	*35*	*73*	*169*	*215*	*46*		*176*	*186*
Area 4	*6.1*	*61*	*42*	*70*	*134*	*172*	*31*		*200*	*173*
Area 5	*12.1*	*187*	*57*	*77*	*368*	*227*	*78*		*206*	*439*
Area 6	*8.2*	*71*	*35*	*65*	*161*	*197*	*61*		*224*	*207*
Area 7	*5.5*	*125*	*64*	*80*	*192*	*350*	*74*		*213*	*264*
Area 8	*13.9*	*189*	*50*	*78*	*223*	*164*	*61*		*218*	*308*
Area 9	*6.3*	*107*	*25*	*59*	*73*	*189*	*20*		*191*	*112*
Area 10	*0.8*	*235*	*54*	*98*	*110*	*147*	*98*		*233*	*155*
Area 11	*3.8*	*210*	*30*	*78*	*138*	*278*	*29*		*196*	*141*
Area 12	*14.6*	*76*	*43*	*74*	*212*	*165*	*56*		*199*	*245*
Area 13	*0.4*	*209*	*8*	*187*	*8*	*141*	*24*		*124*	*16*
Area 14	*0.4*	*140*	*7*	*175*	*23*	*353*	*27*		*246*	*32*
Area 15	*6.6*	*152*	*15*	*167*	*36*	*241*	*61*		*165*	*48*
Area 16	*22.6*	*158*	*66*	*74*	*158*	*226*	*100*		*169*	*123*
Area 17	*21.1*	*110*	*46*	*96*	*128*	*165*	*104*		*208*	*91*
Area 18	*37.8*	*94*	*60*	*63*	*138*	*206*	*125*		*226*	*104*
Area 19	*32.4*	*53*	*39*	*66*	*76*	*240*	*69*		*196*	*86*
Area 20	*10.2*	*109*	*28*	*91*	*60*	*342*	*58*		*225*	*66*
Area 21	*7.1*	*91*	*35*	*84*	*111*	*238*	*84*		*193*	*102*
Area 22	*2.5*	*92*	*21*	*80*	*51*	*268*	*62*		*221*	*54*
Area 23	*13.6*	*151*	*35*	*79*	*125*	*195*	*103*		*213*	*96*
Area 24	*10.4*	*87*	*44*	*98*	*236*	*298*	*109*		*184*	*145*
Area 25	*26.2*	*107*	*24*	*70*	*84*	*88*	*13*		*161*	*86*
Area 26	*0.2*	*80*	*25*	*152*	*62*	*128*	*121*		*101*	*76*
Area 27	*0.5*	*41*	*32*	*76*	*155*	*226*	*113*		*112*	*124*
Area 28	*0.3*	*117*	*26*	*98*	*50*	*107*	*91*		*162*	*47*
Area 29	*0.6*	*71*	*26*	*67*	*78*	*182*	*137*		*89*	*50*
Area 30	*3.2*	*50*	*47*	*85*	*109*	*155*	*183*		*140*	*81*

6.2.3 Data Cleansing on NaFIRS Dataset

Data cleansing is mandatory to ensure the quality of the result and accuracy of the outcomes. In the classification process, data cleansing improves the accuracy of the results. Domain knowledge is vital for data cleansing activities. In this research, data cleansing has been carried out to populate any incomplete data. The following validations and modifications to the data were carried out before removing any information from the dataset. From the initial dataset, which has approximately 50,000 records, around 300 were removed for having null values in the fault cause group variable. Additionally, nearly 100 records were deleted for presenting inconsistencies between other main variables. Regarding other variables, around 30 records were deleted for having negative values in the number of minutes lost variable.

6.2.4 Data Transformation and Data Discretisation on NaFIRS Dataset

Data pre-processing, dealt with many issues that may affect the overall accuracy of classification models such as data transformation, data integration, data reduction, feature selection, data inconsistencies, and data discretisation. In this research study, area population density which is a numeric attribute has been discretised by dividing data into two categories as High Density and Low Density.

The median of area density has been used for discretisation the area density variable. The median is found by ordering the set from lowest to highest and finding the exact middle. If a data set has an odd number of observations, then the median is the middle value. If it has an even number of observations, the median is the average of the two middle values. The estimated median can be calculated with the following formula :

$$Median = \frac{\mathrm{n} + 1}{2}$$

n - number of observations

As previously stated, due to the sensitivity of the NAFIRS data and area details, their anonymity is protected by using pseudonymous names for area identity.

Table 6.4: Sample of the aggregated and discretised NaFIRS dataset

Area Code	Area Density Category	Avg Minutes Lost - Third Party	Fault Count - Third Party	Avg Minutes Lost - Unknown	Fault Count - Unknown	Avg Minutes Lost - Weather	Fault Count - Weather	Avg Minutes Lost - Company	Fault Count - Company
Area 100	*High*	*46*	*28*	*93*	*35*	*153*	*10*	*218*	*73*
Area 101	*Low*	*181*	*30*	*77*	*83*	*165*	*85*	*181*	*63*
Area 102	*High*	*105*	*20*	*66*	*80*	*251*	*65*	*175*	*84*
Area 103	*Low*	*95*	*35*	*70*	*133*	*162*	*89*	*202*	*110*
Area 104	*Low*	*93*	*25*	*82*	*43*	*114*	*47*	*188*	*42*
Area 105	*High*	*170*	*31*	*80*	*59*	*944*	*3*	*267*	*29*
Area 106	*High*	*50*	*37*	*65*	*44*	*145*	*16*	*170*	*67*
Area 107	*Low*	*82*	*11*	*85*	*23*	*120*	*2*	*112*	*17*
Area 108	*High*	*74*	*14*	*70*	*31*	*118*	*5*	*148*	*49*
Area 109	*High*	*126*	*32*	*86*	*79*	*101*	*20*	*165*	*124*

Table 6.4 shows the sample of discretised research dataset. Figure 6.1 and figure 6.2 shows annual average local area fault cause measurements in high and low-density areas on bar and box plots to provide the overall distribution of the data.

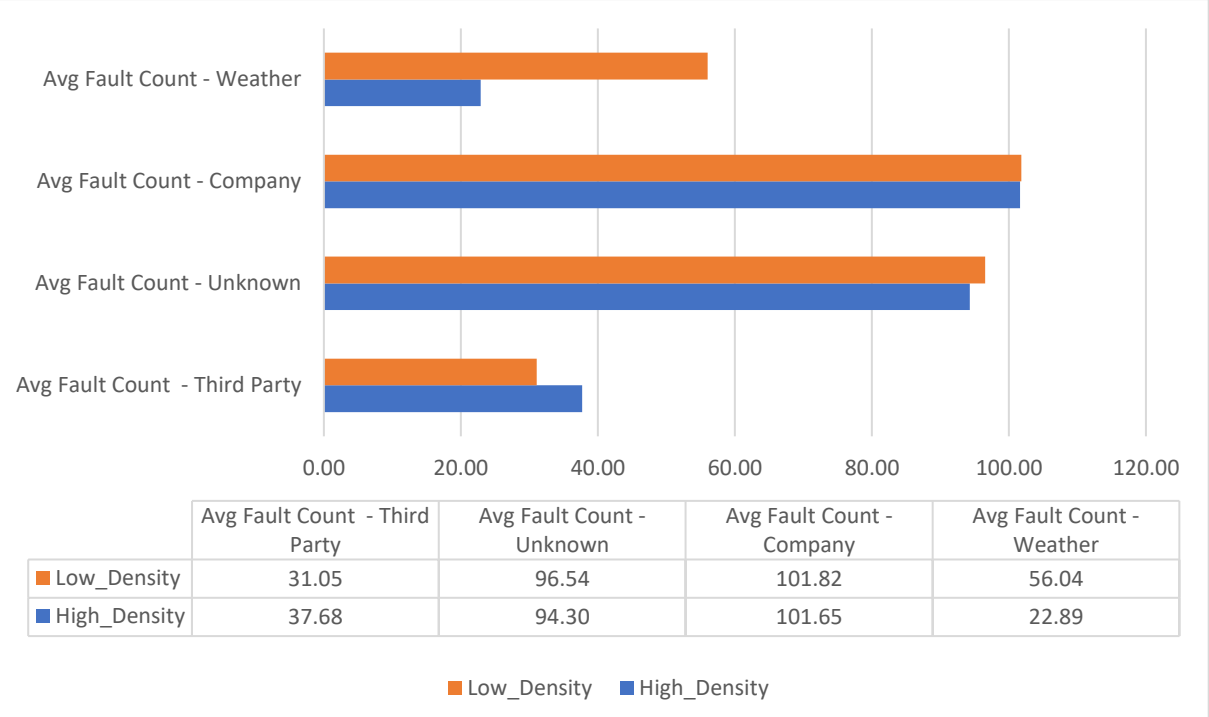


Figure 6.1: Annual average local area wise fault count figures in high and low-density areas on the bar plot

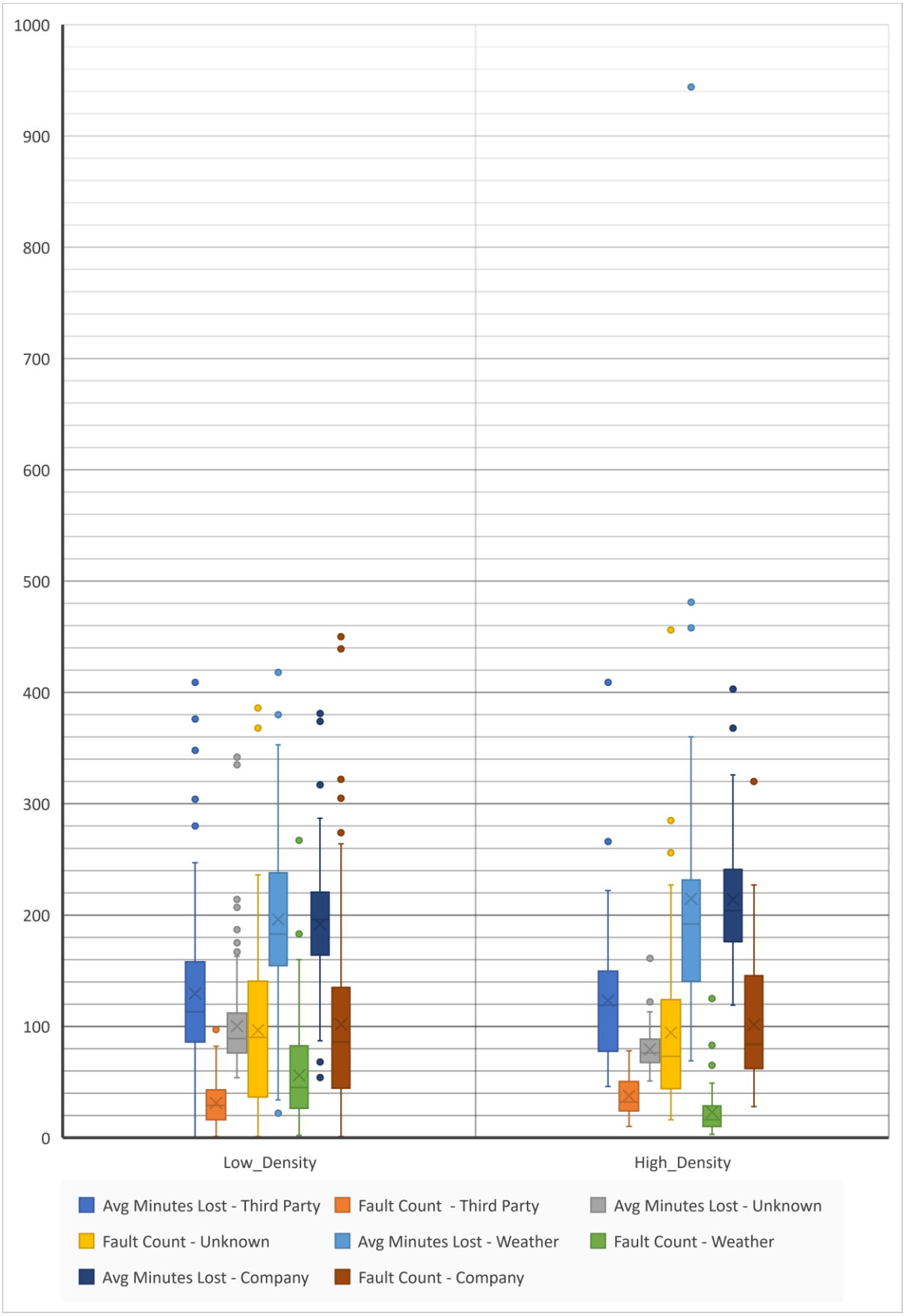


Figure 6.2: Comparison of annual average fault cause measurements in high and low-density areas on a box plot

6.2.5 Predictive Modelling using Classification Methods

There are various classification methods available to use for this study. However, only four different classification methods have been taken into consideration - Logistic Regression, Decision Trees, Random Forest and Support Vector Machine. The classification performance was compared to identify the best classification method. All the methods were carefully validated using a different kind of accuracy measures. Below shows the confusion matrix related to this section of the research study.

		Predicted Class	
Actual Class		**High Density**	**Low Density**
	High Density	*True-positive (TP)*	*False-negative (FN)*
	Low Density	*False-positive (FP)*	*True-negative (TN)*
	Total	*Total Positive*	*Total Negative*

- **True-positive (TP)** = the number of areas correctly identified as High Density
- **False-positive (FP)** = the number of areas incorrectly identified as High Density
- **True-negative (TN)** = the number of areas correctly identified as Low Density
- **False-negative (FN)** = the number of areas incorrectly identified as Low Density

6.2.5.1 Predictive Modelling using Logistic Regression: NaFIRS Case Study

The results of the confusion matrix shown below show that the predicted results from the Logistic Regression.

		Reference	
		0	*1*
Prediction	*0*	*25*	*3*
	1	*8*	*18*

Overall the classification accuracy for this model 79%, which is considered very high. The numerical value of accuracy represents the proportion of true positive results

(both true positives and true negative) in the selected population. Table 6.5 shows the Logistic Regression Modeling Accuracy, Sensitivity and Specificity measures.

Table 6.5: Logistic Regression Modeling Accuracy on NaFIRS Dataset

Prediction Method	Accuracy	Sensitivity	Specificity
Logistic Regression	*0.796*	*0.757*	*0.857*

The numerical values of sensitivity represent the probability of an area identified as a Low-Density area as being a Low-Density area. The higher the numerical value of sensitivity, the less likely the classification will return false-positive results. In this case, sensitivity was reported as 75%. This means that when DNOs use fault cause measures to check on an area of high population density, there is a 75% chance, that the area will be identified as positive. A test with low sensitivity tends to capture all possible positive conditions without missing any. The numerical value of specificity represents the probability of a classifying in a particular area without giving false-positive results. In this case, specificity was reported as 85%. This means that when conducting a classification test on a high-density area, there is an 85% chance; that area will be identified as high density. A classification test can be very specific without being sensitive, or it can be very sensitive without being specific. Both factors are equally important. A good classification test is a one has both high sensitivity and specificity.

6.2.5.2 Predictive Modelling using Decision Tree: NaFIRS Case Study

The classification accuracy from the decision tree model record was 72%, which is considered high, but it is less than the logistic regression model. Table 6.6 shows the Accuracy, Sensitivity and Specificity measures from the decision tree method.

Table 6.6: Decision Tree Modeling Accuracy on NaFIRS Dataset

Prediction Method	Accuracy	Sensitivity	Specificity
Decision Tree	*0.722*	*0.727*	*0.714*

6.2.5.3 Predictive Modelling using Random Forest: NaFIRS Case Study

The classification accuracy from the random forest model record was 74%, which is considered high but is less than the logistic regression model. Table 6.7 shows the Accuracy, Sensitivity and Specificity measures from the random forest method.

Table 6.7: Random Forest Modeling Accuracy on NaFIRS Dataset

Prediction Method	Accuracy	Sensitivity	Specificity
Random Forest	*0.740*	*0.757*	*0.714*

6.2.5.4 Predictive Modelling using Support Vector Machine: NaFIRS Case Study

The classification accuracy from the SVM model record was 70%, which is considered high, which is much less than the logistic regression model. Table 6.8 shows the Accuracy, Sensitivity and Specificity measures from the SVM method.

Table 6.8: SVM Modeling Accuracy on NaFIRS Dataset

Prediction Method	Accuracy	Sensitivity	Specificity
SVM	*0.703*	*0.636*	*0.809*

6.2.6 Enhance the Results using Feature Selection

Feature selection methods aim to reduce model dimensionality by reducing the number of irrelevant features in a model, thereby increasing the model's generalizability and predictive accuracy [85]. Irrelevant features are defined as features which do not affect a target outcome in a statistically significant way. Feature selection is seen as one of the most important aspects of a classification problem and can be the main reason for a successful or unsuccessful model. Too many features can cause over-training as the features are too specific for the training set. Over-training or overfitting is when the learning model performs too well and learns the specific outliers and noise for the dataset. The predictions will then be too specific for that training set. The issue with overtraining is that initially, it may be hard to detect.

In the real-world, feature selection usually requires substantial domain expertise. The benefits of feature selection are numerous. They include distinct data analysis elements, such as better visualisation and comprehension of information, reduced computer time and length, and better precision of forecast. There are several methods for feature selection. Some of those most used are statistical methods and tests like principal component analysis, chi-squared tests or Spearman's correlation test. However, these methods require the data to have certain assumptions such as normality, which usually requires making transformations to the data. Being able to use them in this research study, the author used the point-biserial correlation coefficient as a feature selection method.

Correlation is a statistical method used to assess a possible linear relationship between two continuous variables. Biserial correlation measures the relationship between quantitative variables and binary variables. There are two types of biserial correlations depending on the type of the dichotomous variable. In the point-biserial correlation, the dichotomous variable is discrete. In biserial correlation, some form of the continuum is inferred by the dichotomous variable. The point-biserial correlation can be defined as a particular case of correlation in which one variable is continuous, and the other variable is binary (dichotomous). The formula for the point-biserial correlation coefficient is

$$r_{pb} = \left(\frac{\bar{Y}_1 - \bar{Y}_0}{s_Y}\right)\sqrt{\frac{np_0(1-p_0)}{n-1}}$$

Where

$$s_Y = \sqrt{\frac{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}{n-1}}$$

$$\bar{Y} = \frac{\sum_{k=1}^{n} Y_k}{n}$$

$$p_1 = \frac{\sum_{k=1}^{n} X_k}{n}$$

$$p_0 = 1 - p_1$$

The Point-Biserial Correlation Coefficient measures the strength of association of two variables in a single measure ranging from -1 to +1, where -1 indicates a perfect negative association, +1 indicates a perfect positive association and 0 indicates no association at all. The statistical relationship between the dependent variable and the

independent variables are given in Table 6.9. The figure reflects the relations between the electricity distribution network faults causes measurements and area density.

Table 6.9: Point-Biserial Correlation Coefficient on NaFIRS Dataset

Attribute	Correlation Coefficient
Avg. Min Lost - Third Party	*-0.007*
Fault Count - Third Party	*-0.289*
Avg. Min Lost - Unclassified or Unknown	*0.368*
Fault Count - Unclassified or Unknown	*-0.204*
Avg. Min Lost - Weather and Environment	*-0.101*
Fault Count - Weather and Environment	*0.308*
Avg. Min Lost - Company	*-0.187*
Fault Count - Company	*-0.181*

Correlation can be positive (increase in one value of feature increases the value of the target variable) or negative (increase in one value of feature decreases the value of the target variable). The below bar chart shown in (Figure 6.3) makes it straightforward to identify which features are most related to the target variable.

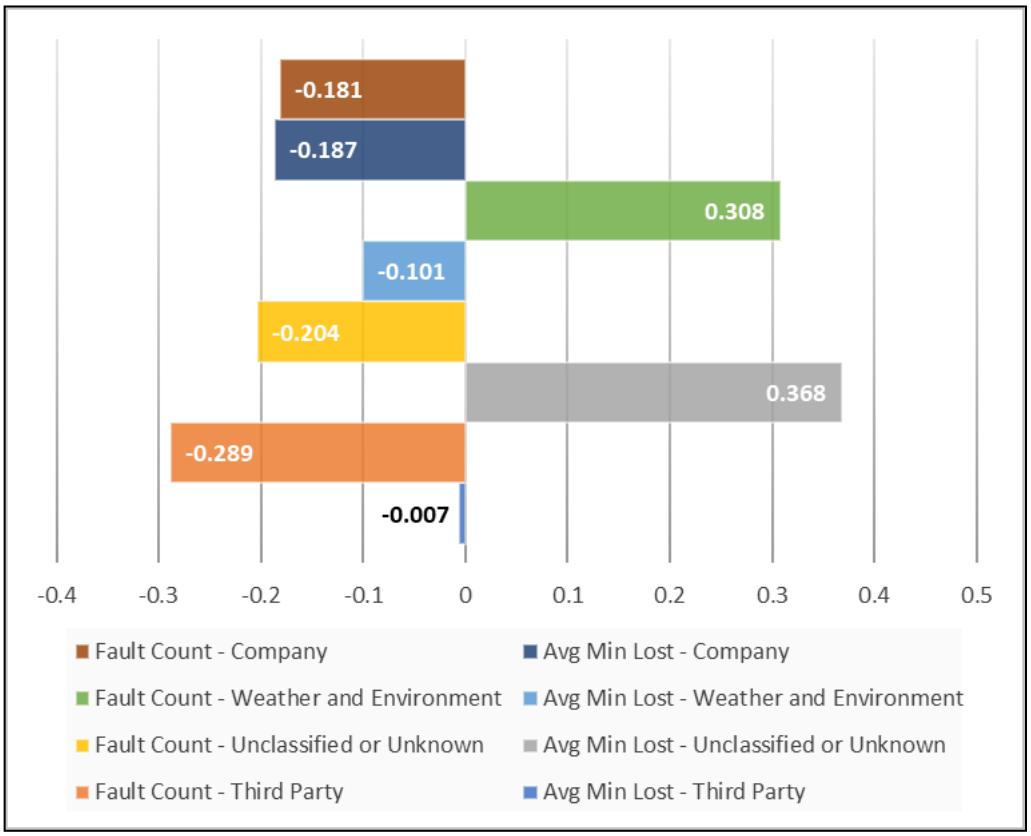


Figure 6.3: Point-Biserial correlation coefficient comparison graph

Identifying an appropriate set of predictors is essential for making efficient and accurate prediction models. In this study, the Point-Biserial Correlation Coefficient has been used as a feature selection method. This aims to reduce model dimensionality by reducing the number of irrelevant features in a model, thereby increasing the model's classification predictive accuracy. According to the Point-biserial correlation coefficient shown in (Table 6.10), four features have been identified as the highest correlated features to Area Density.

Table 6.10: Highest correlated features identified in the NaFIRS Dataset

ATTRIBUTE	CORRELATION COEFFICIENT
Fault Count -Third Party	*-0.289*
Avg. Min Lost -Unclassified Or Unknown	*0.368*
Fault Count -Unclassified Or Unknown	*-0.204*
Fault Count -Weather And Environment	*0.308*

The previously used four different classification methods (Logistic Regression, Decision Trees, Random Forest and Support Vector Machine) have been used again for reclassification. The classification performance was compared again in the same way to identify the best classification method. All the methods were again validated using different accuracy measures. Table 6.11 shows the difference between the performance evaluation measures of the four classification models. Based on the values in the table, the impact of the feature selection has contributed to increasing all the accuracy measures of the prediction models. The classification accuracy from the Logistic Regression model record is 83%, which is considered higher than the other three models.

Table 6.11: Reclassification performance evaluation measures on highest correlated features identified in the NaFIRS Dataset

Prediction Method	Accuracy	Sensitivity	Specificity
Logistic Regression	*0.833*	*0.787*	*0.904*
Decision Tree	*0.722*	*0.727*	*0.714*
Random Forest	*0.759*	*0.787*	*0.714*
SVM	*0.759*	*0.727*	*0.809*

The experimental results of all four classifiers showed adequate results, both (before and after feature selection) that can be used for accurate prediction. So, this section of the study can, therefore, conclude that the low voltage network faults have a strong readership between population density. The correlation study has also proved the association and correlation between faults causes in the distribution network and local area population density.

6.2.7 Result Analysis and Discussion

There is a business need to reduce operational expenditure in engineering departments in DNOs. This study sought to understand the relationship and correlation between different type of faults causes in the electricity distribution network and population density using data mining and statistical approaches. The objectives set to achieve this wear to extract relevant NAFIRS data and apply classification algorithms to build, validate and compare the accuracy of the models. The data was successfully analysed, and the main fault causes identified where some interesting relationships had been identified.

The experimental results of all four classifiers showed adequate results that can be used for accurate prediction modelling. The study also used the point-biserial correlation coefficient as a feature selection method, which aims to reduce model dimensionality by reducing the number of irrelevant features in a model, thereby increasing the model's classification predictive accuracy.

Four important LV fault causes were considered, using the customer minutes lost, and the number of customers affected to build classification models. The correlation between the fault causes is essential to determine which faults causes are particularly important for users in order to understand the relevance of population density.

In conclusion, the study has primarily designed to analyse the impact of external factors on faults in electricity distribution networks. This study has successfully achieved the objectives set because findings of the research have presented in Advances in Science and Engineering Technology International Conferences (ASET) 2020, Dubai, United Arab Emirates and has received much interest and attention from industry and academic experts. Also, research findings have already been published

in IEEE explore. While there are improvements that could be made to the study, such as testing more classification models, the study provides an excellent starting point for further research studies into understanding electricity network faults and impact of external factors such as social demographic factors. The study has explored the optimal machine learning models by comparing four well-known classification models. However, the research could have explored more models and perhaps explored more different types of machine learning models. Future work could extend and use of other available faults causes and measurements to study the impact of rare events.

6.3 Analysis of Relationship Between DNO's Key Performance Indicators and Network Faults

[7]The DNOs need to make a substantial investment in network infrastructure. DNOs connects customers to electricity network via a vast network of overhead and underground cables. The Electricity distribution network made of a large number of different components such as primary and secondary substations, underground cables, overground cables, IoT sensors, pylons and various monitoring equipment [86]. The total length of distribution networks in the UK is around 800,500 kilometres, crossing both urban and regional areas [83].

DNOs are continually battling to meet rising demand from the consumers. Annually they have to invest a significant amount of money for replacing ageing or poorly performing assets, maintaining and improving network performance [83]. Also, as a regulatory business, they have to invest in maintaining regulatory requirements, such as reliability standards. As part of its regulatory contract, all the DNOs must comply with goals set out by the regulators. These goals also include meeting the financial expectations set by the regulators, constant supply uninterrupted power supply to the customers, provide secure network and maintain a high level of customer satisfaction.

Modern power systems have become involved and challenging to operate, maintain and protect, as they consist of hundreds of thousands of components scattered across a wide geographic area. This makes them susceptible to multiple sources of failure, which are difficult to predict when designing a system accurately. The complexity of the network sometimes leads to a cascading series of events, starting with a small initial problem leading to severe regional failures. UK customers continue to benefit from reduced interruptions. Since the beginning of 2015, Customer interruptions have fallen by 11%, and the duration of interruptions has fallen by 9% [83].

[7] This section has been published in below research paper
C. Silva and M. Saraee, "Predicting Average Annual Electricity Outage using Electricity Distribution Network Operator's Performance Indicators," 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2020, pp. 1-6, doi: 10.1109/ASET48392.2020.9118383. (Appendix 2)

However, since 2018 the average duration of interruptions has slightly increased to 36 minutes. Most DNOs have delivered continued performance improvements, while others' performance has worsened since the beginning of 2015 [86]. It is significantly essential to comprehend any relationship, correlation, and the association between

The exploratory case study research methodology has been used for this study. The exploratory case study investigates unique phenomena that are characterised by a lack of detailed preliminary research, in particular hypotheses that can be tested, and by a specific research environment that limits the choice of method. Exploratory case study research methodology can be considered as a reliable research method, especially if a comprehensive and detailed investigation is needed. In this study, the author investigates two distinct case studies. The first case study is from the UK, and the other one is from Australia. Both cases studies have been analysed using Pearson's correlation, multiple linear regression, Pont Biserial Correlation and Logistic Regression.

6.3.1 Reliability of Power Supply

Reliability of the supply of electricity is a crucial service performance measure for a DNO [83]. Distribution outages account for over 95 per cent of electricity interruptions in the UK [87]. Interrupted supply of electricity to a DNO's customers can be planned or unplanned.

Planned interruptions occur when a DNO needs to disconnect supply to undertake maintenance or construction works. Such interruptions can be scheduled for minimal impact, and the DNO notifies the customer of its intention to interrupt supply. DNOs should give customers a minimum of four business days written notice of a planned outage. The mainly unplanned interruption occurred when equipment failures happened in the network. Equipment may fail due to equipment malfunction, system overload, ageing, corrosion or extreme weather events.

Underground cables usually require less maintenance, so maintenance costs are significantly less than for overhead cables. Usually, underground cables are protected from external factors (weather or human activities), so they are more sustainable and reliable [86]. In case of damage in underground cable in the network, it usually takes lengthier to find and repair the fault than the overground cable lines.

Underground lines are more frequent in the populated urban areas, where they are often located along with other infrastructures such as water, telecommunications and gas. Nowadays, as a standard, in most new residential homes and commercial property's supply lines are being built underground as usual.

Comparatively overhead cables are much cheaper to build than underground cable because the material is cheaper and requires less complicated design. However, it has higher maintenance costs than underground lines due to the higher structural costs and safety management procedures [86]. Overhead cables are more common in extensive rural areas networks because they are comparatively cheaper to build than underground cables.

Unplanned interruptions typically have a more significant effect on customers than planned interruptions because they do not provide customers with sufficient warning to act to manage the impact of the interruption. The key measures for supply reliability are [87]:

- CI - the number of customer interruptions per 100 customers on the network.
- CML- the average length of time customers without power per interruption.

The regulator sets targets for the number of customers interrupted (CIs) and duration (CMLs) of both planned and unplanned interruptions. DNOs are rewarded if they meet or exceed these targets and are penalised if they fail to meet them. DNOs must continue to invest in network assets to reduce the number of Customer Interruptions, as well as improve operational practices (such as locating and repairing faults) to reduce the Customer Minutes Lost.

Network operators may receive a considerable monetary reward or incur a significant monetary penalty depending on their performance against a target for both the number and length of their network supply disruptions [88]. It is compulsory for upholding high levels of customer services; also maintain and improving service level as defined by the regulators also key responsibilities of the DNO. Annually OFGEM conducts a customer satisfaction survey to measure DNOs customer satisfaction performance. The customer satisfaction survey is intended to capture customers' experiences of the interruption, minor connection and general enquiry services delivered by the DNOs. In this study, the author investigates to understand the behaviour of the CML. CML_t is the duration of interruptions to supply in the relevant year *t* and is derived from the following formula [89]:

$$CML_t = CMA_t + CMLB_t + CMLC_t + CMLD_t + CMLE_t$$

- CMLA is the duration of interruptions from unplanned incidents.
- CMLB is the duration of interruptions from pre-arranged incidents.
- CMLC is the duration of interruptions arising from incidents on the systems of transmission companies.
- CMLD is the duration of interruptions arising from incidents on the systems of distributed generators.
- CMLE is the duration of interruptions arising from incidents on any other connected systems.

Each of the terms $CMLA_t$, $CMLB_t$, $CMLC_t$, $CMLD_t$ and $CMLE_t$ should be separately calculated.

6.3.2 Case Study 1: Office of Gas and Electricity Markets (Ofgem)

This case study contains performance data for the UK electricity distribution businesses regulated by the Office of Gas and Electricity Markets (Ofgem). Ofgem is the government regulator for gas and electricity markets in Great Britain. Data are presented for a variety of monetary and performance measures. The data covers the regulatory years from 2011 to 2018 [83]. All the distributed network operators should report to regulators based on the fiscal year. All financial values have been recorded on Pound Sterling (GBP).

6.3.2.1 Ofgem Dataset

Table 6.12 shown below demonstrates the datasets attributes to give a clearer understanding of the data structure.

Table 6.12: Attributes explanation of the Ofgem dataset

Attribute	Type
Totex	*Numeric*
Customer_Numbers	*Numeric*
Network_length	*Numeric*
Units_distributed	*Numeric*
Peak_load	*Numeric*
CML	*Numeric*

- Totex – is the Total Expenditure. Its includes Capital expenditure (CAPEX) and operating expenditure (OPEX).
- Customer Numbers - is the number of customers connected to the distribution networks.
- Network Length – The length of the distributed network in kilometres. It includes both overhead lines and underground cables.
- Units Distributed - is a measure of Total units distributed (GWh) through the distribution networks in each year.
- Peak Load - Network-wide Peak load (MVA) is simply the highest electrical power demand that has occurred over a specified period
- CML - CML is the average length of time customers are without power per interruption.

Table 6.13 shows the sample of the research dataset.

Table 6.13: Sample of the Ofgem dataset

Totex	Customer Numbers	Network length	Units Distributed	Peak Load	CML
241	*2359*	*56952*	*23835*	*4389*	*47*
135	*1576*	*40160*	*15796*	*2997*	*71*
186	*2258*	*52783*	*23063*	*4285*	*68*
250	*2447*	*63459*	*24932*	*4662*	*90*
231	*2614*	*71700*	*27883*	*5042*	*55*
127	*1099*	*35162*	*12111*	*2153*	*32*
181	*1541*	*50183*	*14625*	*2904*	*43*
215	*2252*	*36628*	*29355*	*5203*	*42*
255	*2233*	*52200*	*21694*	*3976*	*73*
385	*3517*	*96266*	*34525*	*6966*	*72*
188	*1993*	*63848*	*16195*	*3320*	*49*
203	*1485*	*49669*	*16195*	*3320*	*47*
112	*741*	*47024*	*8521*	*1633*	*78*
244	*3026*	*76220*	*33646*	*6693*	*64*
258	*2364*	*57116*	*23367*	*4280*	*48*
138	*1581*	*40352*	*15272*	*2852*	*69*
197	*2266*	*52946*	*22571*	*4195*	*65*
263	*2462*	*64111*	*24248*	*4792*	*49*
239	*2623*	*71920*	*27074*	*5292*	*37*
127	*1103*	*35299*	*11776*	*2118*	*37*

183	*1551*	*50441*	*14078*	*2855*	*40*
193	*2267*	*36704*	*28492*	*5125*	*31*
233	*2248*	*52316*	*20813*	*4108*	*43*
339	*3537*	*96409*	*34093*	*6407*	*47*
211	*1994*	*63943*	*19341*	*3611*	*49*

6.3.2.2 Correlation Study for the Ofgem Case Study

Although, the correlation between different variables can be measured with different coefficients such as Pearson's (r), Spearman's rho (rs), and Kendall's tau coefficient. Nonetheless, Pearson's correlation was used in this research to evaluate the linear relationship between the continuous variable present in the dataset.

Therefore, the Pearson correlation coefficient is defined as a measure of the strength of the linear relationship between two variables represented by r [90]. Pearson correlation intends to draw a line of best fit via the data of two variables, and the coefficient, r, shows the distance of the data points away from the line of best fit. Thus, the value of r range from +1 to -1. When r = 0 it shows that there is no association between the two variables. R > 0 indicates that there is a positive association between the variable; therefore, an increase in one variable will increase the other. R < 0 shows a negative association between two variables; therefore, as one of the variable increases the other decreases.

For a feature with values x and classes y, the Pearson correlation coefficient is given by:

$$\hat{p}_{x,y} = \frac{\sum_{i=1}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt[2]{\sum_{i=1}(x_i - \bar{x}_n)}\ \sqrt[2]{\sum_{i=1}(y_i - \bar{y}_n)}}$$

Where :

$\bar{x}_n$ is the mean value of x

$\bar{y}_n$ is the mean value of y

The Pearson correlation is applicable to binary, continuous, or single value target variables. Irrelevant features should have a near-zero Pearson correlation; however; the Pearson correlation is limited by the fact that it does not capture non-

linear relationships between variables. Therefore, a non-linear relationship may exist, and features which have a near-zero Pearson correlation could still be relevant.

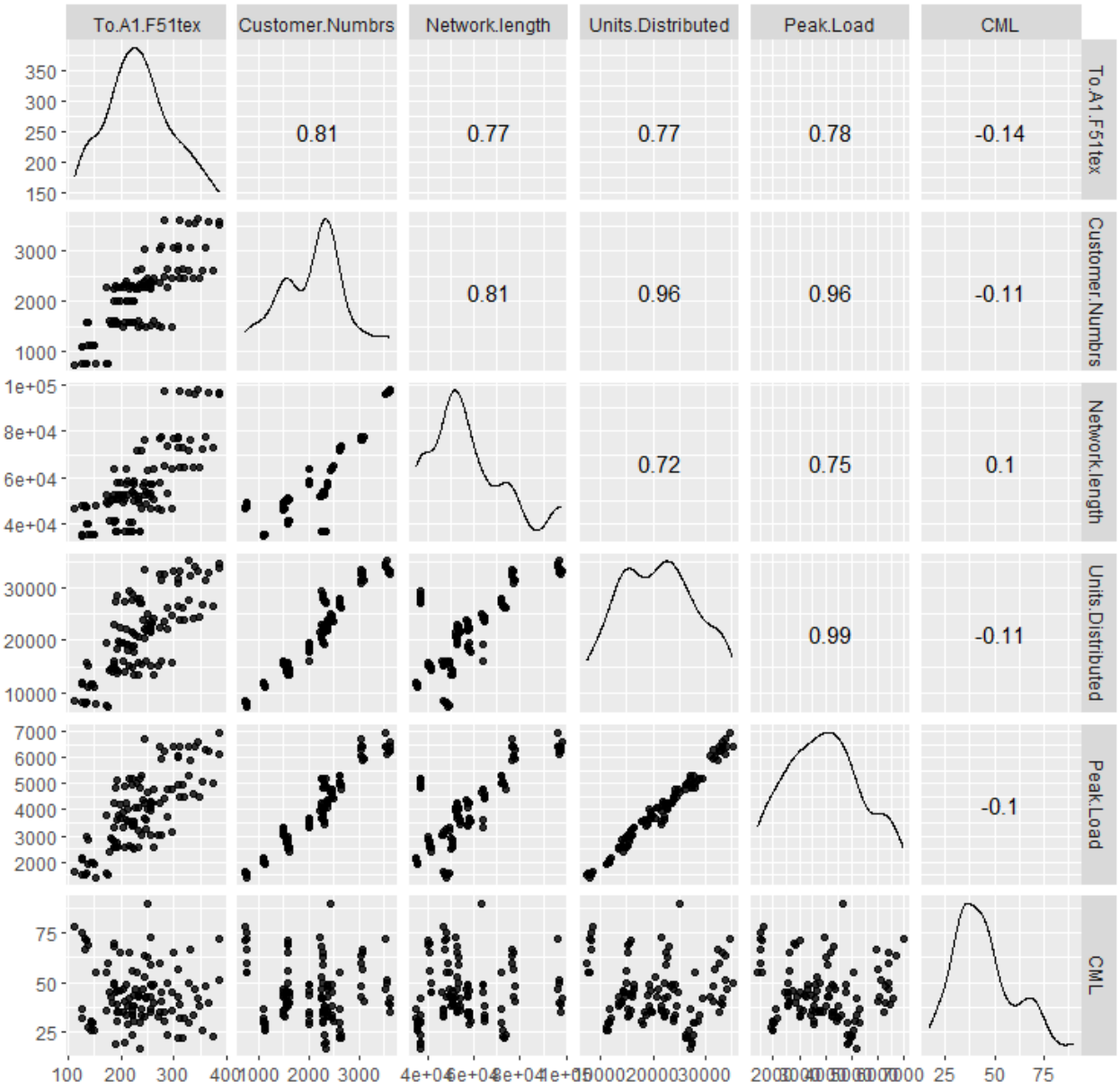


Figure 6.4: Pearson's Correlation Graph for the Ofgem dataset

Figure 6.4 shows the correlation between all the variables. There is a strong positive correlation between Totex, Customer Numbers, Network length, Units distributed and Peak-load.

The most considerable correlation is between the total number of customers and Total units distributed (GWh), with a correlation of 0.93. The smallest correlation is

between Network length (km) and CML, with a correlation of 0.06. There is little to no correlation between CML and all the other variables.

6.3.2.3 Predict CML in the Ofgem Dataset using Multiple Linear Regression

Multiple linear regression models were built and used to find the correlation and linear relationship between the CML and other indicators. Multiple linear regression model is a regression model with one dependent variable and more than one independent variables, i.e. the dependent variable y can be influenced by the one or more of the independent variable.

However, since multiple linear analysis contains several independent variables, the model has another mandatory condition, which is the Non-collinearity; there should be a minimum of correlation between independent variables. If the independent variables are highly correlated, the true relationships between the dependent and independent variables will be challenging to assess.

The output of multiple leaner regression shows that p value=0.000000053; from this result, the p-value significantly less than 0.05, P-value is also known as probability value indicate how likely a result can occur by chance alone. The results indicate that the test is statistically significant and recommend to reject the null hypothesis. Thus, all variables except the measure of Total Expenditure, together have a significant effect on customer minutes lost, and it concludes that there is a linear relationship between customer minutes lost and other measurements. However, R2 = 0.325, showing that the other indicators in the model only described approximately 32.5 % of the CMLs.

6.3.2.4 Data Discretisation on the Ofgem Case Study Dataset

The discretisation is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals that span the range of variable values [90]. The discretisation is done to replace the raw value of the numeric attribute by the interval levels. There are several approaches to transform continuous variables into discrete ones. Discretisation methods fall into two categories which are supervised and unsupervised. Unsupervised methods use the variable distribution, to create the contiguous bins in which the values will be placed. However, supervised methods typically use target information in order to create bins or intervals. In this

study, the author has used K means clustering, which is an unsupervised method as data discretisation method. In this study, CML has been discretised into two categories (Low and High) using K means clustering.

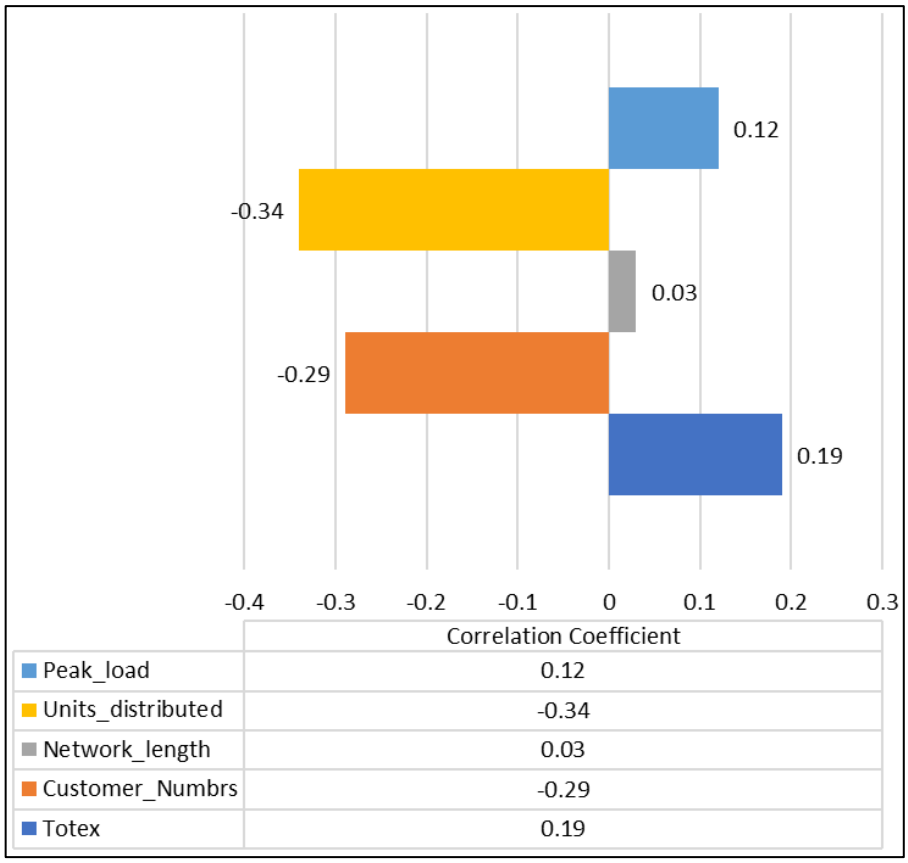
6.3.2.5 Calculating the Point Biserial Correlation using Ofgem Dataset

The point-biserial correlation can be defined as a particular case of correlation in which one variable is continuous, and the other variable is binary (dichotomous). The Point-Biserial Correlation Coefficient measures the strength of association of two variables in a single measure ranging from -1 to +1, where -1 indicates a perfect negative association, +1 indicates a perfect positive association and 0 indicates no association at all. The statistical relationship between the dependent variable and the independent variables are given in Table 6.14. The correlation figures reflect the relations between the CML and other indicators.

Table 6.14: Point-Biserial Correlation Coefficient for the Ofgem dataset

Attribute	Correlation
Totex	*0.19*
Customer_Numbrs	*-0.29*
Network_length	*0.03*
Units_distributed	*-0.34*
Peak_load	*0.12*

Table 6.14 and Figure 6.5 shows the Pont Biserial Correlation between CML and all the other variables. The most significant correlation is between CML and Total units distributed (GWh), with a correlation of -0.34. The smallest correlation is between CML and Network length (km), with a correlation of 0.03.



	Correlation Coefficient
Peak_load	0.12
Units_distributed	-0.34
Network_length	0.03
Customer_Numbrs	-0.29
Totex	0.19

Figure 6.5: Pont Biserial Correlation Comparison Graph for the Ofgem dataset

6.3.2.6 Predict CML Category in the Ofgem Dataset using Logistic Regression

Logistic regression is a regression technique to predict a dependent dichotomous variable [91]. Logistic regression is helpful in cases where predictions are to be made from a set of given variables. It is comparable to a model of linear regression but is suitable for a dataset with binary variables. Logistic regression is a popular classification technique which is well documented in the literature. Logistic regression is noted to perform poorly with small-sample biases, and this bias inversely correlates heavily with the number of rare cases in the sample.

The author has applied multi-level logistic regression analyses to investigate the association between CML and other variables. Although their research seeks association between the variables, it is context quite different from this study which hopes to identify patterns suggestive of a correlation and dependence among variables.

Predicting the accuracy of the logistic regression is looking at the diagonal of the confusion matrix generated by the model. The percentage of accuracy in the trained data is compared to the accuracy of the validated data. If the accuracy of both the trained and validated data is relatively the same, then It is concluded that the logistic regression accurately classifies that data. The confusion matrix, as shown in Table 6.15, find the percentage of classification, the following formula is used:

$$\frac{\sum(diag(matrix))}{\sum(matrix)}$$

Table 6.15: Trained Confusion Matrix

Predicted	*Actual*		
		Low	*High*
	Low	*55*	*11*
	High	*4*	*6*

Table 6.16 shows the trained confusion matrix, which shows the classification accuracy of 80% with an error of 20%.

Table 6.16: Validated Confusion Matrix

Predicted	*Actual*		
		Low	*High*
	Low	*24*	*3*
	High	*2*	*7*

Table 6.16 shows the validated confusion matrix, which shows the classification accuracy of 86% with an error of 14%. The two results show high accuracy rate with the logistic regression modelling.

To conclude, the logistic regression predicts the CML category from the other performance indicators such as Totex, Customer Numbers, Network length, Units distributed and Peak-load.

6.3.3 Case Study 2: Australian Energy Regulator (AER)

This case study contains performance data for the Australian electricity distribution businesses regulated by the Australian Energy Regulator (AER) [77]. AER is responsible for the economic regulation of electricity networks in the Australian National Electricity Market. Data are presented for a range of financial and network performance measures. The research dataset contains data between 2006 to 2017. Some of the distribution companies are reporting based on a calendar year and some based on a financial year. All financial values have been converted to June 2017 Australian dollars.

6.3.3.1 AER Dataset

Table 6.17 shown below demonstrate the datasets attributes to give a clearer understanding of the data structure.

Table 6.17: Attributes explanation of the AER dataset

Attribute	Type
Capex	*Numeric*
Opex	*Numeric*
Energy_delivered	*Numeric*
Customer_numbers	*Numeric*
Circuit_length_Overhead	*Numeric*
Circuit_length_Underground	*Numeric*
Network_utilisation	*Numeric*
Outage_duration	*Numeric*

- Capex - Capital expenditure is a measure of investment in the distribution networks.
- Opex – Operating expenditure includes network operation, maintenance and other non-capital costs incurred by the distribution businesses.
- Energy Delivered - Energy delivered is a measure of total energy transported through the distribution networks in each year.
- Customer Numbers - Customer numbers represent the number of customers connected to the distribution networks.
- Circuit length Overhead – Length of Overhead cables in the network.
- Circuit length Underground – Length of Underground cables in the network.

- Network utilisation - is a measure of the use of the network. Utilisation rates are derived by comparing maximum demand to the total capacity of the distribution network, at the zone substation level
- Outage duration - This shows the average length of time each customer was without supply when averaged over all customers in the distribution network.

Correlation study

6.3.3.2 Correlation Study for the AER Case Study

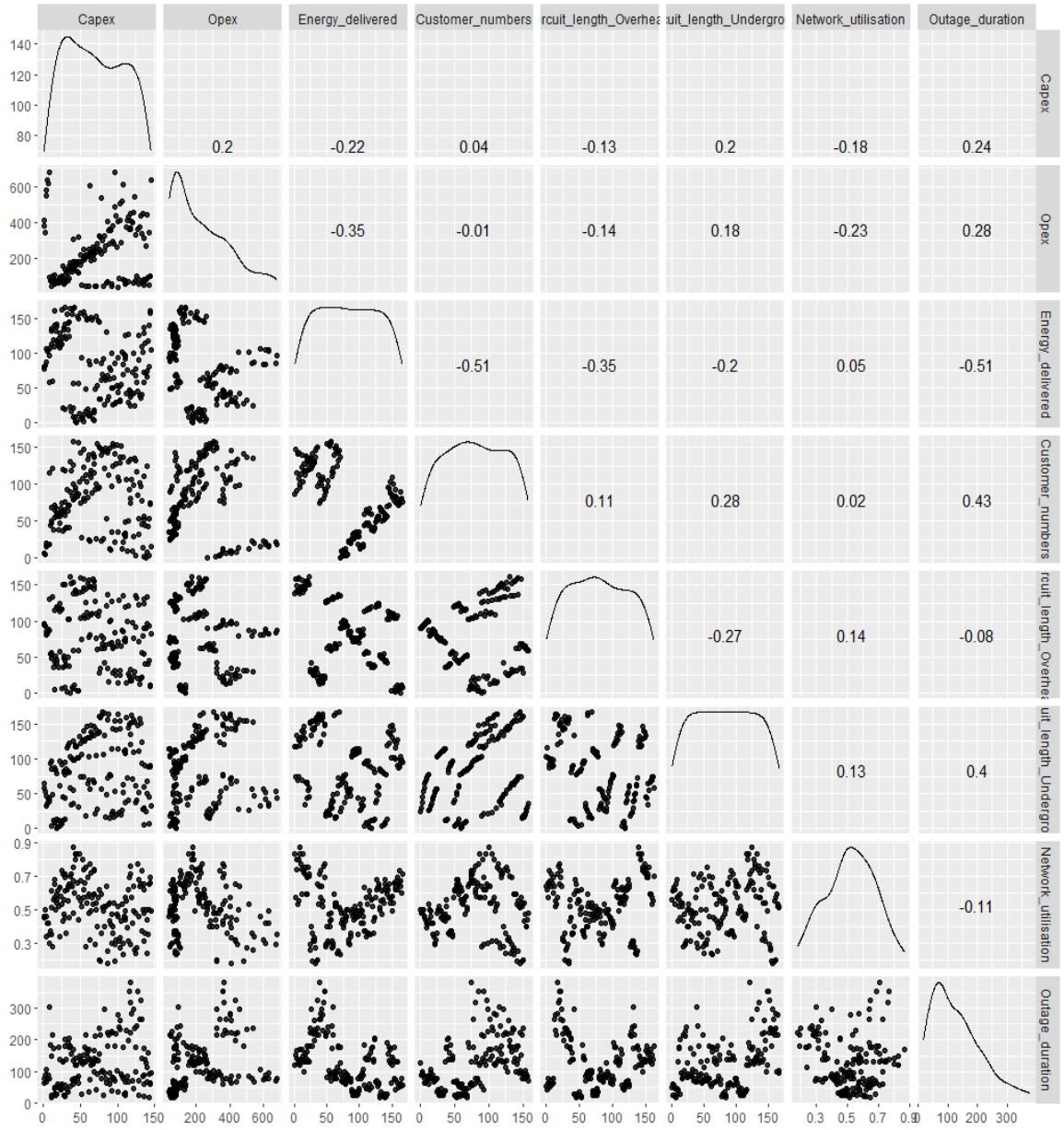


Figure 6.6: Pearson's Correlation Graph for the AER dataset

Figure 6.6 shows the correlation between all the variables. There is a strong negative correlation between Outage duration, Energy delivered and Customer numbers. The most significant correlation shows between Energy delivered and Outage duration, with a correlation of -0.51. The smallest correlation is between Outage duration and Network utilisation, with a correlation of 0.02. In general, there are not any significant correlation between variables like in the case study 1.

6.3.3.3 Predict CML in the AER Dataset using Multiple Linear Regression

The results from the multiple leaner regression model show that the p value=3.39e-16, which is significantly less than 0.05. This indicates that the test is statistically significant, and the null hypothesis should not be rejected.

Thus, four variables (Energy_delivered, Customer_numbers, Circuit length both Overhead and Underground) collectively have a significant effect on the Outage duration and conclude that there is a linear correlation between Outage duration and other variables in the dataset. However, R^2 = 0.392, showing that the other indicators in the model only described approximately 39.2 % of the Outage duration.

6.3.3.4 Data Discretisation on the AER Dataset

Data discretisation has been done in this research using K means clustering, which is a classical unsupervised data mining technique. Outage_duration variable has been discretised into two categories (Low and High) using K means clustering.

Table 6.18: Point-Biserial Correlation Coefficient for the AER dataset

Attribute	Correlation Coefficient
Capex	*-0.16*
Opex	*-0.11*
Energy_delivered	*0.36*
Customer_numbers	*-0.37*
Circuit_length_Overhead	*0.1*
Circuit_length_Underground	*-0.28*
Network_utilisation	*0.21*

Table 6.18 shows the Pont Biserial Correlation between Outage duration and all the other variables. The most considerable correlation is between Outage duration and Customer numbers, with a correlation of -0.37. The smallest correlation is between Outage duration and Circuit Length Overhead, with a correlation of 0.03.

6.3.3.5 Predict Outage Duration in the AER Dataset using Logistic Regression

Table 6.19: Confusion Matrix for the outage duration prediction using the AER training dataset using Logistic Regression

Predicted	*Actual*		
		Low	*High*
	Low	*69*	*15*
	High	*10*	*24*

Table 6.19 shows the trained confusion matrix, which shows the classification accuracy of 78% with an error of 22%.

Table 6.20: Confusion Matrix for the outage duration prediction using the AER validation dataset using Logistic Regression

Predicted	*Actual*		
		Low	*High*
	Low	*32*	*5*
	High	*4*	*9*

Table 6.20 shows the validated confusion matrix, which shows the classification accuracy of 82% with an error of 18%. The two results show that the logistic regression accurately.

To conclude, the logistic regression can accurately predict the Outage duration category from the other key performance indicators such as Capex, Opex, Energy Delivered, Customer Numbers, etc..

6.3.3.6 Outage Duration Prediction Results Comparison between Ofgem Dataset and AER Dataset

Table 6.21shows the outage duration prediction results in a comparison between Ofgem Dataset and AER Dataset using relevant confusion matrix values,

which shows the classification accuracy of training and validation dataset. The two datasets show a high accuracy rate with the logistic regression modelling.

Table 6.21: Outage Duration Prediction Results Comparison between Ofgem Dataset and AER Dataset

	Classification Accuracy using Logistic Regression	
	Training dataset	Validation dataset
Ofgem Dataset	80%	86%
AER Dataset	78%	82%

By looking at both results from the two case studies, we are able to determine that classification techniques like logistic regression able to predicts the customer minutes lost from the other performance indicators such as Totex, Customer Numbers, Network length, Units distributed and Peak-load.

6.4 Summary of Chapter and Conclusion

Each of the DNO in the UK legally obligated to report all the network faults whether the fault results in loss of power supply to customers such as domestic, commercial users or no impact [89]. Accurately understanding and ability to predict the annual Customer Minutes Lost (CML) figure is essential in fault management department in DNOs. In the DNOs perspective, it is vital to bring the annual CML figures down to sustain and perform in a highly completive energy distribution industry. This study aimed to improve DNOs annual CML figures. Also, the new knowledge gain from the study will develop clarity among fault analysis in fault management departments.

Various statistical approaches and two case studies have been used to achieve the objectives. Pearson's correlation, multiple linear regression, Point Biserial

Correlation and Logistic Regression were used to perform the analysis. In case study 1, statistical analysis using Pearson's correlation shows that there is no or weak correlation between CML and other performance variables. Nevertheless, the output of multiple leaner regression shows that p value=0.000000053 with R^2 of 0.325. So, the author has used K means clustering to discretise the CML into two categories (Low and High) and apply Pont Biserial Correlation between CML and all the other variables. The most considerable correlation is between discretised CML and Total units distributed (GWh), with a correlation of -0.34. Next step was to use multi-level logistic regression analyses to investigate the association between CML and other variables. Results show that the validated confusion matrix shows the classification accuracy of 86%. The results show that the logistic regression accurately predicts the CML category from the other performance indicators. After considering all the statistical results, conclude that there is a linear correlation between annual CML and other annual financial and network performance indicators. So, that DNOs can investigate the behaviour of the other indicators to understand the annual CML values.

This work can be further developed by using machine learning algorithms such as Decision tree, Neural Network. Based on the other similar research in the domain, hybrid models consisting of statistical and machine learning can be explored to benefit from the combined benefits of both models in a bid to improve accuracy. They were making the model more robust by providing more data attributes in the modelling, such as customer satisfaction score.

CHAPTER 7

Conclusion

7.1 Overview

This chapter summarises the main accomplishment of the research. It describes the summary of models which developed in this thesis. The thesis concludes by reviewing the key contributions and directions for future work. It critically evaluates the presented research and discusses future work that could address its shortcomings or further extend or validate its contributions. The summary of chapters and main findings of the research are given in the following sections.

7.2 Introduction

The electricity supply system includes a large-scale power generation installation and a large convoluted network of electrical circuits that work together to supply electricity to consumers efficiently and reliably. The Distributed Network Operators (DNOs) are responsible for carrying electricity from the high voltage transmission network to lower voltage industrial, commercial and domestic users. In some instances, the DNOs distribute energy from local generation sources which connect to distribution networks directly.

Due to the complexity of electricity distribution systems, they are prone to frequent faults. Especially, the leading equipment in the network is always vulnerable to numerous failures that may occur in any of the main components or subcomponents in leading equipment. The weather is the single most influential factor that causes failures in the distribution network. High wind (The speed of gale force), exceptionally low or high temperature and rainfall have potential in causing defects to different types of assets. The National Fault and Interruption Reporting Scheme [3], is set up and administered by the Energy Networks Association. Each DNO in Great Britain is required to report its every fault which occur on their network; regardless of the failure results in loss of electricity supply to customers.

The main research questions which address this study are: **“Can data mining and machine learning approaches use to predict and forecast faults in the Electricity Distribution network accurately?”**. This research question is further divided into four sub-questions, as follows.

1) Can industrial standard National Fault and Interruption Reporting Scheme (NaFIRS) data be used to predict and forecast potential faults in the network?
2) Can electricity distribution network industry forecast the volume of future faults in the network accurately and understand the seasonality? How can model performance be determined?
3) Can predictions, forecast and new insights gained from data mining and machine learning analysis be used to enhance the functionality and the performance of the fault management process?
4) Do external factors such as population density influence the volume of future faults in the network?

This research aims to develop multi-variant data mining and machine learning models to predict and forecast electricity distribution network faults. Multiple objectives are set to achieve this aim such as gaining a deeper understanding of Electricity Distribution Network faults and National Fault and Interruption Reporting Scheme (NaFIRS) and the dataset, Develop models to understand temporal patterns, Develop a multi-dimensional fault segmentation method and Analyse the correlation and association between external factors.

There are several challenges in this research, in particular in fault data: finding a required real industrial fault dataset such as NaFIRS dataset is challenging due to commercial sensitivity of the data and reliability of the data.

Based on the literature review, there is little to no research performed in data mining and machine learning methods such as clustering, decision tree, and support vector machine which used to analyse on fault data within the utility distribution industry. This research will investigate this area of research to understand and gain more insight into this field. Location finding of faults in the electricity distribution network has been widely researched in the last decade, following the rapid development of

smart grid around the world. However, the literature review has been uncovered that lack of scientific studies in fault forecasting and prediction in the electricity distribution network. Also, literature related to this research was reviewed, and the research gaps in the literature were identified and found out that no research has investigated the prediction of future faults in the electricity distribution network using National Fault and Interruption Reporting Scheme (NaFIRS). Also, there is a lack of studies in understanding temporal patterns, seasonality and Dynamics of Network Faults identified.

This scientific research gap has been achieved by applying exploratory data mining and machine learning techniques to fault data to extract advanced analytic insights that would aid the understanding of possible relationships between internal and external factors and network faults. This would thus provide evidence-based literacy for network design engineers and industrial policymakers.

A few industrial studies were conducted in this field, but only a very few systematic scientific studies have been conducted in this field. Hence, this thesis, with the objective to fill the knowledge gap, has been focused on the predictive and time-series forecasting side of the domain with a preliminary analysis to develop a series of potential advanced decision-making models. The five main research areas focused on in this thesis are identified as:

1. Temporal pattern mining and visualisation.
2. Seasonality and the Dynamics of Network faults.
3. Multidimensional segmentation.
4. Understand the association and correlation between the volume of faults and fault causes.
5. Analysing correlation and association between external factors such as local area population density, DNOs KPI and electricity distribution faults.

This research employed a multiple case study research design due to the fact this allows more opportunities for multiple experiments and cross observation, and the multiple case design is appropriate for solving complex issues.

7.3 Research Contributions to the Knowledge

The work presented within this thesis is expected to lead to the following original contributions to the scientific community who is working with data science and electricity distribution network.

1. This research presents a data mining based new fault segmentation framework which DNOs can use to perform fault segmentation. This approach provides an option of performing multidimensional segmentation using various fault characteristics to any utility distribution network such as electricity, water or gas.

2. This research presents a novel time-series calendar heatmap based on temporal pattern mining method to support the process of visual knowledge discovery with fault data in any utility distribution network.

3. This research presents a novel unsupervised data mining methodology that analyses historical fault data and trying to understand the impact of fault with other associated factors. This proposed unsupervised data mining methodology may use to safeguard any key equipment in the utility distribution network, which can be destroyed by upcoming faults.

4. This research presents a data mining and machine learning method of analysing correlation and association between external factors such as local area population density, DNOs KPI and electricity distribution faults.

5. Author of this thesis has proposed a tailor-made Data Science project process flow for data acquisition, data fusion, data storing, processing, analysing, and modelling. The primary motivation for introducing this new process flow is to streamline the data mining and machine learning modelling process.

In this study, the author has introduced a new fault segmentation framework using K-means clustering and DBSCAN algorithm which DNOs can use to perform the fault segmentation. Fault segmentation in Electricity Distribution Network is an essential pre-processing step for the early diagnosis and elimination of faults in the

network. The main aim of fault segmentation is to understand the failures better and to use that understanding to improve performance, reliability and availability of the distribution network. This approach gives companies an option of performing multidimensional segmentation using various fault characteristics such as a number of faults, a number of minutes lost, and a number of customers affected. Multidimensional segmentation is a powerful conceptual model for the analysis of large and complex datasets.

In chapter 4, it has attempted to analysis the most significant factors that contribute to distribution network faults using Association Rule Mining with industrial recognized NAFIRS dataset. The study also explores the possibility of enhancing the knowledge gain from Association Rule Mining using Document Term Clustering. The outcomes of this section of the research will support in policy formulation in engineering departments to reduce network faults.

In chapter 5, the study has investigated whether the equipment failures which related to network faults have seasonality. Even those faults have very uncertainty nature; the study has proved that those faults can be predicted using time series forecasting model which supports the seasonality. The seasonal decomposition method was used to distinguish the different components of the time series. This work also contributes to knowledge by exploring the performance of time series forecasting models on faults data from the DNOs.

Also, the study has demonstrated how the 2D and 3D calendar heat map method can help provide a relatively new perspective in evaluating temporal patterns in electricity distribution network faults. The clustering analysis has been performed with aggregated temporal fault data. It is a process of partitioning a dataset into groups of meaningful subclasses where grouped objects share common traits. It is implemented to understand the natural hidden structure in the data that would otherwise remain unobserved. Usually, scatterplot will be used to display clustering results. However, the author has proposed to use 2D and 3D calendar heatmap to visualise cluster results.

Ability to predict the annual Customer Minutes Lost (CML) figure is essential in fault management department in DNOs. In the DNOs perspective, it is vital to bring annual CML figures down to sustain and perform in a highly competitive energy distribution industry. This study aimed to improve DNOs annual CML figures by

introducing a new way of predicting CML figures. Various statistical approaches and two case studies have been used to achieve the objectives. Pearson's correlation, multiple linear regression, Point Biserial Correlation and Logistic Regression were used to perform the analysis. Also, the new knowledge gain from the study will develop clarity among fault analysis in fault management departments.

7.4 Future Work

1. Although it has been demonstrated that the proposed time-series forecasting methods perform well with research dataset, it was observed that unexpected severe weather conditions could jeopardise the whole process of accurate fault forecasting. Therefore, additional tests need to be performed to assess the robustness of the developed method.

2. When outliers have been detected due to unexpected event such as flood, wildfire, storms, it is better to apply appropriate outlier detection techniques to clean the data before they pass through to data mining and machine learning models.

3. The classification models developed in the study can be improved further by advanced machine learning algorithms such as Neural Network. Based on other similar research in the domain, hybrid models consisting of statistical and machine learning can be explored to benefit from the combined benefits of both models in a bid to improve accuracy.

4. Data Mining and machine learning models can make it more robust by providing more data points, so it would be better to re-run those proposed models with larger datasets (Subject to availability).

5. The 2D/3D calendar heatmap approach is described in this study which offers significant advantages over understanding the correlation between the hour and month of the network faults. Although this study is based on one DNO data in Australia, the approach can be applicable to DNO from the UK and globally with comparable data.

7.5 Potential Benefits of this Research for the Electricity Distribution Industry

All the DNOs are committed to providing a safe, secure, reliable and cost-effective network in order to supply energy to customers. The work presented within this report is expected to lead to the following original benefits to the field of fault management in the electricity distribution industry:

7.5.1 Quantitative Benefits

- Electricity distribution fault engineers usually work 40-hours per week. They are required to work at weekends and must be available for standby duties which could cost a substantial amount of money to the DNOs. Therefore, there is a business need to reduce operational expenditure in engineering departments. This new fault, forecasting and prediction model will provide opportunities to utilise the engineering staff resource in a more controlled manner and reduce call outs and overtime charges.

- Prioritise investments in new or replacement infrastructure, which will ensure that financial and human resources are used more efficiently.

- Reduction in the number of complaints by able to predict the faults in the network and, wherever possible, prevent them.

7.5.2 Qualitative Benefits

- Improved level of system availability by able to predict the faults in the network and, wherever possible, prevent them.
- Increased non-engineering staff productivity due to pre-planned engineering work.
- Improved fault reporting to senior leadership teams with more accurate forecasting figures.
- Avoidance of costs and potential fines associated with future network faults imposed by the regulators.

- Fault segmentation is an essential tool for developing business intelligence in a fault management department and for maintaining competitive advantage among DNOs. The use of knowledge gained from fault segmentation will develop fault analysis clarity.

REFERENCES

[1]. Association, E. (n.d.). Overview of DNOs. Retrieved June 03, 2016, from https://www.energynetworks.org/electricity/regulation/overview-of-dnos.html

[2]. Judson, R., Tebbs, J., & Clarke, G. (2000). Report on medium term network performance monitoring (Tech.). OFGEM.

[3]. OFGEM. (2017). RIIO electricity distribution annual report 2016-17 (Rep.).

[4]. ENA. (2015). Climate Change Adaptation Reporting Power Second Round. Version 1 ed.

[5]. Filomena, A.D., Resener, M., Salim, R. H. & Bretas, A.S. (2011). Distribution systems fault analysis considering fault resistance estimation. International Journal of Electrical Power & Energy Systems, 33, 1326-1335.

[6]. Types of Faults in Electrical Power Systems. (2017, December 24). Retrieved from https://www.electronicshub.org/types-of-faults-in-electrical-power-systems/

[7]. SSE. (2015). Climate Change Adaptation Report (Rep.).

doi:https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/478927/clim-adrep-sse-power-distribution-2015.pdf

[8]. FORD, D. V. (1972). The British Electricity Boards National Fault and Interruption Reporting Scheme-Objectives, Development and Operating Experience. IEEE Transactions on Power Apparatus and Systems, PAS-91, 2179-2188

[9]. Dunn, S., Wilkinson, S., Alderson, D., Fowler, H. & Galasso, C. (2018). Fragility Curves for Assessing the Resilience of Electricity Networks Constructed from an Extensive Fault Database. Natural Hazards Review, 19.

[10]. RIIO electricity distribution annual report 2016-17. (2018, March 20). Retrieved from https://www.ofgem.gov.uk/publications-and-updates/riio-electricity-distribution-annual-report-2016-17.

[11]. Distribution, S. E. P. (2013). About electricity distribution networks: now and in the future.

[12]. Landegren, Finn, et al. “Quality of Supply Regulations versus Societal Priorities Regarding Electricity Outage Consequences: Case Study in a Swedish Context.”

International Journal of Critical Infrastructure Protection, vol. 26, 2019, p. 100307., doi:10.1016/j.ijcip.2019.100307.

[13]. BUTLER, S. (2001). The nature of UK electricity transmission and distribution networks in an intermittent renewable and embedded electricity generation future, Centre for Environmental Technology.

[14]. Climate Change Adaptation Reporting Power Second Round (Rep.). (2015). Retrieved December 14, 2019, from ENERGY NETWORKS ASSOCIATION

[15]. RIIO-ED1 Annual Report 2016-17 (Rep.). (2017, December 19). Retrieved March 10, 2019, from OFGEM website:

[16]. Saunders, M. N., Lewis, P., & Thornhill, A. (2019). Research methods for business students. New York: Pearson.

[17]. Han, J., Kamber, M. and Pei, J. (2012). Data mining. 3rd ed. Haryana, India: Elsevier.

[18]. Turban, E., Delen, D., & Sharda, R. (2018). Business intelligence, analytics, and data science: A managerial perspective. Harlow ; Munich: Pearson Prentice Hall.

[19]. Bramer, M. (2013). Principles of Data Mining (2 ed.). London: Springer.

[20]. Saeed, K. (2018). Computer information systems and industrial management: 17th international conference, Cisim 2018, Olomouc, Czech Republic, September 27-29, 2018: proceedings. Cham: Springer.

[21]. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data Mining: Practical Machine Learning Tools and Techniques (Fourth ed.). Amsterdam: Morgan Kaufmann.

[22]. Assessing clustering tendency: A vital issue - Unsupervised Machine Learning. (n.d.). Retrieved October 26, 2019, from http://www.sthda.com/english/wiki/print.php?id=238.

[23]. Olson, d. L. (2019). Descriptive data mining. S.l.: springer verlag, singapor.

[24]. Turban, E., Delen, D., & Sharda, R. (2018). Business intelligence, analytics, and data science: A managerial perspective. Harlow ; Munich: Pearson Prentice Hall.

[25]. A. Gosain and M. Bhugra, 2013. "A comprehensive survey of association rules on quantitative data in data mining," 2013 IEEE Conference on Information & Communication Technologies, JeJu Island, pp. 1003-1008.

[26]. Zhong.Y, Liao.Y, 2012. Research of Mining effective and Weighted Association Rules Based on Dual Confidence, CPS Fourth International Conference on Computational and Information Sciences

[27]. G. Lei, M. Dai, Z. Tan and Y. Wang, 2011. "The Research of CMMB Wireless Network Analysis Based on Data Mining Association Rule," 2011 7th International Conference on Wireless Communications, Networking and Mobile Computing, Wuhan, pp. 1-4.

[28]. Gorunescu, F. (2011). Data Mining: Concepts, models and techniques (Vol. 12), 250-251. Springer Science & Business Media.

[29]. Han, Jiawei, and Micheline Kamber. Data Mining: Concepts and Techniques. Elsevier, 2012. [30].

[31]. Sperandei S. (2014). Understanding logistic regression analysis. Biochemia medica, 24(1), 12–18. doi:10.11613/BM.2014.003

[32]. Guo, Hongquan & Nguyen, Hoang & Vu, Diep-Anh & Bui, Xuan-Nam. (2019). Forecasting mining capital cost for open-pit mining projects based on an artificial neural network approach. Resources Policy. 10.1016/j.resourpol.2019.101474.

[33]. Pedersen, R. (2006). An Embedded Support Vector Machine. 2006 International Workshop on Intelligent Solutions in Embedded Systems. doi: 10.1109/wises.2006.237155

[34]. The Multiple Linear Regression Equation. (n.d.). Retrieved from http://sphweb.bumc.bu.edu/ otlt/MPH-Modules/BS/BS704-EP713_MultivariableMethods/ BS704-EP713_MultivariableMethods2.html

[35]. Yang, Y.-M., Yu, H., & Sun, Z. (2017). Aircraft failure rate forecasting method based on Holt-Winters seasonal model. 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). doi: 10.1109/icccbda.2017.7951969

[36]. Qonita, A., Pertiwi, A. G., & Widiyaningtyas, T. (2017). Prediction of rupiah against US dollar by using ARIMA. 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). doi: 10.1109/eecsi.2017.8239205

[37]. Y. Yang, H. Zheng and R. Zhang, "Prediction and analysis of aircraft failure rate based on SARIMA model," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, 2017, pp. 567-571.

[38]. CHATFIELD, C. 2000. Time-Series Forecasting, Taylor & Francis Ltd.

[39]. ADHIKARI, R. & K. AGRAWAL, R. 2013. An Introductory Study on Time series Modeling and Forecasting.

[40]. MAKRIDAKIS, S., C. WHEELWRIGHT, S. & HYNDMAN, R. 1984. Forecasting: Methods and Applications.

[41]. HYNDMAN, R. J. Measuring forecast accuracy. 2014.

[42]. SCHUELKE-LEECH, B.-A., BARRY, B., MURATORI, M. & YURKOVICH, B. J. 2015. Big Data issues and opportunities for electric utilities. Renewable and Sustainable Energy Reviews, 52, 937-947.

[43]. ALI, U., BUCCELLA, C. & CECATI, C. Households electricity consumption analysis with data mining techniques. IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, 23-26 Oct. 2016 2016. 3966-3971.

[44]. Apel, R. “Fault Management in Electrical Distribution Networks.” 16th International Conference and Exhibition on Electricity Distribution (CIRED 2001), 2001, doi:10.1049/cp:20010792.

[45]. Al-Aomar, Raid, et al. “Reducing the Interruption of Power Distribution: A Six Sigma Application.” 2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA), 2017, doi:10.1109/ickea.2017.8169934.

[46]. Haomin, Chen, et al. “Fault Prediction for Power System Based on Multidimensional Time Series Correlation Analysis.” 2014 China International

Conference on Electricity Distribution (CICED), 2014, doi:10.1109/ciced.2014.6991916.

[47]. Tervo, Roope, et al. “Short-Term Prediction of Electricity Outages Caused by Convective Storms.” IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 11, 2019, pp. 8618–8626., doi:10.1109/tgrs.2019.2921809.

[48]. Bai, Yuling, et al. “Short-Term Prediction of Distribution Network Faults Based on Support Vector Machine.” 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2017, doi:10.1109/iciea.2017.8283062.

[49]. Depaire, Benoît, et al. “Traffic Accident Segmentation by Means of Latent Class Clustering.” Accident Analysis & Prevention, vol. 40, no. 4, 2008, pp. 1257–1266., doi:10.1016/j.aap.2008.01.007.

[50]. Ford, D. “The British Electricity Boards National Fault and Interruption Reporting Scheme-Objectives, Development and Operating Experience.” IEEE

[51]. Newis, D, et al. “Optimising Customer Information and the Fault Management Process.” 16th International Conference and Exhibition on Electricity Distribution (CIRED 2001), 2001, doi:10.1049/cp:20010787.

[52]. Blake, Simon Richard. 2010. “Methodologies for the Evaluation and Mitigation of Distribution Network Risk”. Durham University, 2010

[53]. Wang, Li. “The Fault Causes of Overhead Lines in Distribution Network.” MATEC Web of Conferences, vol. 61, 2016, p. 02017., doi:10.1051/matecconf/20166102017.

[54]. Zhanjun, Gao, et al. “A Distribution Network Fault Data Analysis Method Based on Association Rule Mining.” 2014 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), 2014, doi:10.1109/appeec.2014.7066121.

[55]. ROBSON, C. 2002. Real World Research : A Resource for Social Scientists and Practitioner-Researchers / C. Robson.

[56]. YIN, R. K. & SAGE. 2003. Case Study Research: Design and Methods, SAGE Publications.

[57]. ZAINAL, Z. 2007. Case study as a research method.

[58]. BAXTER, P. & JACK, S. 2008. Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. The Qualitative Report. BEYER, M. A. & LANEY, D. 2012. The Importance of

[59]. CHRISTOPH, S. K. 2010. Encyclopedia of Case Study Research. Thousand OaksThousand Oaks, California: SAGE Publications, Inc

[60]. CRESWELL, J. 2009. Research Design: Qualitative, Quantitative, and Mixed-Method Approaches.

[61]. F. Mastrogiovanni, A. Sgorbissa and R. Zaccaria. A Distributed Architecture for Symbolic Data Fusion. In IJCAI-07, pp 2153-2158. 2007.

[62]. Saraee, M., & Silva, C. (2018). A new data science framework for analysing and mining geospatial big data. Proceedings of the International Conference on Geoinformatics and Data Analysis - ICGDA 18. doi:10.1145/3220228.3220236

[63]. Abdullah, A. S., Ramya, C., Priyadharsini, V., Reshma, C., & Selvakumar, S. (2017). A survey on evolutionary techniques for feature selection. 2017 Conference on Emerging Devices and Smart Systems

[64]. I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182 . http://dl.acm.org/citation.cfm?id= 944919.944 96 8 .

[65]. Predictive Modeling: The Only Guide You Need. (n.d.). Retrieved from https://www.microstrategy.com/us/resources/introductory-guides/predictive-modeling-the-only-guide-you-need

[66]. D.A. Keim, J. Kohlhammer, G. Ellis, F. Mannsmann (Eds.), Mastering the Information Age. Solving Problems with Visual Analytics, Eurographics Association, Goslar, 2010.

[67]. Entezarimaleki, R., Rezaei, A., & Minaeibidgoli, B. (2009). Comparison of Classification Methods Based on the Type of Attributes and Sample Size.

Journalof Convergence Information Technology, 4(3), 94–102. doi: 10.4156/jcit.vol4.issue3.14

[68]. Reddy, G., Rajinikanth, T., & Rao, A. (2014). A frequent term based text clustering approach using novel similarity measure. 2014 IEEE International Advance Computing Conference (IACC). doi:10.1109/iadcc.2014.6779374

[69]. Beil, F., Ester, M., & Xu, X. (2002). Frequent term-based text clustering. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 02. doi:10.1145/775047.775110

[70]. He, Q., Li, T., Zhuang, F., & Shi, Z. (2010). Frequent term based peer-to-peer text clustering. 2010 Third International Symposium on Knowledge Acquisition and Modeling. doi:10.1109/kam.2010.5646177

[71]. Corpus linguistics - an introduction. (n.d.). Retrieved from https://www.anglistik.uni-freiburg.de/seminar/abteilungen/sprachwissenschaft/ls_mair/corpus-linguistics

[72]. Y. Chen, et al., “Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment”, Computers in Biology and Medicine, vol. 42, no. 2, pp. 213-221, 2012.

[73]. Depaire, B., Wets, G., & Vanhoof, K. (2008). Traffic accident segmentation by means of latent class clustering. Accident Analysis & Prevention, 40(4), 1257–1266. doi: 10.1016/j.aap.2008.01.007

[74]. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density based algorithm for discovering clusters in large spatial databases with noise," in KDD-96 Proceedings, pp. 226-231, 1996.

[75]. Celik, M., Dadaser-Celik, F., & Dokuz, A. S. (2011). Anomaly detection in temperature data using DBSCAN algorithm. 2011 International Symposium on

Innovations in Intelligent Systems and Applications. doi: 10.1109/inista.2011.5946052

[76]. Saunders, M. N., Lewis, P., & Thornhill, A. (2019). Research methods for business students. New York: Pearson.

[77]. Ausgrid and our customers. (n.d.). Retrieved from https://www.ausgrid.com.au/-/media/Regulatory-Proposal-2019-24-Exec-Summary.pdf.

[78]. Past outages data. (n.d.). Retrieved December 17, 2019, from https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Past-outage-data.

[79]. Dunn, S., Wilkinson, S., Alderson, D., Fowler, H., & Galasso, C. (2018). Fragility Curves for Assessing the Resilience of Electricity Networks Constructed from an Extensive Fault Database. Natural Hazards Review, 19(1), 04017019. doi:10.1061/(asce)nh.1527-6996.0000267

[80]. Simoff, S. J. (n.d.). Form-Semantics-Function – A Framework for Designing Visual Data Representations for Visual Data Mining. Lecture Notes in Computer Science Visual Data Mining, 30–45. doi: 10.1007/978-3-540-71080-6_3

[81]. Customer interruptions and minutes lost: Electricity distribution (RIIO-ED1). (2019, March 20). Retrieved from https://www.ofgem.gov.uk/.

[82]. All energy network charts and indicators: Ofgem. (2019, November 28). Retrieved November 29, 2019, from https://www.ofgem.gov.uk/data-portal/network-indicators.

[83]. RIIO Electricity Distribution annual report 2017-18. (2019, March 8). Retrieved November 29, 2019, from https://www.ofgem.gov.uk/publications-and-updates/riio-electricity-distribution-annual-report-2017-18.

[84]. Local Government Association. (n.d.). Population density, persons per hectare in England. Retrieved from https://lginform.local.gov.uk/reports/lgastandard?mod-

metric=176&mod-area=E92000001&mod-group=AllRegions_England&mod-type=namedComparisonGroup.

[85]. Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70–79. doi: 10.1016/j.neucom.2017.11.077

[86]. Network performance. (2018, November 16). Retrieved from https://www.aer.gov.au/ networks- pipelines/network-performance.

[87]. UK Electricity Networks. (n.d.). Retrieved from https://www.parliament.uk /documents/post/e5.pdf.

[88]. Electricity distribution company performance from 2010 to 2015. (2015, December 16). Retrieved from https://www.ofgem.gov.uk/publications-and-updates/electricity-distribution-company-performance-2010-2015.

[89]. RIIO-ED1 regulatory instructions and guidance: Annex F ... (n.d.). Retrieved December 9, 2019, from https://www.ofgem.gov.uk/ofgem-publications/95354/ annexfinterruptions-pdf.

[90]. Nickolas, S. (2019, December 7). What Does it Mean if the Correlation Coefficient is Positive, Negative, or Zero? Retrieved from https://www.investopedia.com/ask/answers/ 032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp.

[91]. Hoffman, J. I. (2015). Logistic Regression. Biostatistics for Medical and Biomedical Practitioners, 601–611. doi: 10.1016/b978-0-12-802387-7.00033-0

[92]. Hicks, M. & Malley, C. & Nichols, S. & Anderson, B.. (2003). Comparison of 2D and 3D representations for visualising telecommunication usage. Behaviour and Information Technology. 22. 185-202. 10.1080/0149290310001170 80.

[93]. Important step forward for north of Scotland electricity grid. (n.d.). Retrieved June 22, 2018, from https://www.sse.com/news-and-views/2013/04/important-step-forward-for-north-of-scotland-electricity-grid

[94] Support Vector Machine (SVM). (n.d.). SpringerReference. doi:10.1007/springerreference_106815

[95] R. C. Dubes and A.K. Jain, Algorithms for Clustering Data, Prentice Hall, 1988

[96] C. A. Andresen, B. N. Torsæter, H. Haugdal and K. Uhlen, "Fault Detection and Prediction in Smart Grids," 2018 IEEE 9th International Workshop on Applied

Measurements for Power Systems (AMPS), Bologna, 2018, pp. 1-6, doi: 10.1109/AMPS.2018.8494849.

[97] Joaquim L. Viegas, Susana M. Vieira, Rui Mel´ıcio, "Prediction of events in the smart grid: interruptionsin distribution transformers," 2016 2016 IEEE International Power Electronics and Motion Control Conference (PEMC) doi: 10.1109/EPEPEMC.2016.7752037

APPENDIX

Appendix 1

NaFIRS - National Fault and Interruption Reporting Scheme:

Most DNO's use this to collect information on the number of customers affected by outages and for how long they have been off for.

Each financial year OFGEM gives all DNO's a CI & CML budget, this money is paid upfront and whatever is left over we get to keep and re-invest in our network. We have to find a balance though because too good of performance results in next year's budget being lower and to poor performance means we have to pay the money back.
An Interruption is when supply has been off or not adequate for a total of 3 minutes or longer.

In terms of Fault Reporting, there are five key elements: -

- CI (Customers Interrupted) is a one-off penalty that occurs every time a customer is interrupted
- CML (Customer Minutes Lost) is the duration of outage multiplied by the CI
- SDI (Short Duration Interruption) currently only used for HV figures and captured through CRMS (Control Room Management System)
- RI (Re-Interruptions) customers being off again within 3hours of a permanent restoration or 18hours of a temporary restoration (loop, bunch or generator) don't carry an additional CI penalty
- ONI (Occurrence Not Incentivised) fault work that doesn't require NaFIRS, this can range from anything like changing a cut out to securing a substation and even spiking a cable.

NaFIRS are broken down into different sections.

- DC (Direct Cause)
- MEI (Main Equipment Involved).

DC (Direct Cause)
What has caused this interruption?
These then have different sub-headings

Weather & Environment:-

- 01 Lightning – Lightning striking one of our assets.
- 02 Rain – Rainwater getting into exposed joints or enclosures that would normally be sealed.
- 03 Snow, Sleet and blizzard – As above but with the additional weight of volume

- 04 Ice – Mainly affects trees that can impact our O/H Lines
- 05 Freezing Fog and Frost – As above
- 06 Wind and Gale (excluding Windborne Material) – The wind has blown down the lines
- 07 Solar Heat – Or the Sun, a hot day causing the oil in TX's to overheat and cause failures
- 10 Airborne Deposits (excluding Windborne Material) – Minerals in the air (e.g. sandstorms)
- 14 Condensation – The collection of moisture on our assets
- 15 Corrosion – The metallic elements breaking down over time
- 16 Mechanical Shock or Vibration – Heavy Plant, previous excavation work or even Tremors can affect the physical components of our assets
- 17 Ground Subsidence – The shifting of the ground in which the asset sits can dislodge it from other parts
- 18 Flooding – Excessive water submerging our assets that are not previously protected against such conditions
- 19 Fire not due to Faults – Have we been instructed to, or deemed it necessary to isolate the supplies or has a fire affected our assets
- 20 Growing or Falling Trees (not felled) – Any trees affecting our lines
- 21 Windborne Materials – Anything that is in the air that shouldn't be due to the Wind
- 22 Disruption of Intended Indoor Environment – Anything that would normally be indoors has been exposed to outdoor elements
- 23 Falling Live Tree – The Felling of Trees or Branches that are live onto our lines
- 24 Falling Dead Tree – The Felling of Trees or Branches that are dead onto our lines
- 25 Growing Trees – Trees growing through the lines
- 26 Corrosion due to atmosphere/environment – as 15 except quicker due to location (e.g. waterfronts like Blackpool promenade)

Birds, Animals and Insects:-

- 30 Birds (including Swans and Geese) – Bird Strikes
- 32 Vermin, Wild Animals and Insects – any non-Human owned creature
- 33 Farm and Domestic Animals – Cattle rubbing or household pets chewing through lines

Third-Party (accidental contact, damage or interference): -

- 39 Wilful Damage, Interference or Theft. (not including Metal Theft) – Anything where a 3rd party has deliberately disrupted our network
- 40 Metal Theft – Disruption to our network solely by the removal of metal from our assets or the need to isolate because of it
- 41 By Cable TV Companies or their contractors – e.g. Virgin media
- 42 By Public Telecommunications Operator or their contractors – e.g. BT or Mobile networks
- 43 By Gas Company or their contractors – any companies undertaking gas works

- 44 By Water/Sewage Company or their contractors – any companies undertaking water/wastewater works
- 45 By Highway Authorities or their contractors – any companies undertaking Highway's work
- 48 Involving Farm Workers or Farm Implements – anything affected by the Farming process or Employees in the Farming industry
- 49 Involving Aircraft or Unmanned Balloons – e.g. aerial vehicles, balloons & corporate drones
- 50 By Private Individuals (excluding 49 and 56) – not a business or company
- 53 Unknown Third Parties – Cleary damaged but no way to prove by whom
- 54 By Local Building Authorities or their Contractors – any government-funded authorities in the construction of buildings
- 55 By Private Developers or their Contractors – any business in the construction of buildings
- 56 Involving Leisure Pursuits – Hot Air Balloons, Kites & Drones controlled through private individuals
- 57 By Other Third Parties – Anything else
- 58 By Cable TV Companies or their contractors – not to be used see 41

Companies: - (This section refers to the DNO and their staff)

- 60 Accidental Contact, Damage or Interference by DNO or their Contractors (incl Live Line Work) – Have we inadvertently caused the disruption
- 61 Switching Error By DNO Personnel – Switching in and affecting customers not previously affected
- 62 Testing or Commissioning Error by DNO Personnel – anytime supplies are lost where they weren't anticipated due to carrying out these works.
- 63 Incorrect or Inadequate System Records, Circuit Labelling or Identification – Where supplies have been removed in error
- 65 Incorrect Application of DNO Equipment – Using equipment where it shouldn't be used
- 66 Faulty Installation or Construction – Any supply problems caused by this
- 67 Load Current above Previous Assessment – Circuits operating on load
- 68 Incorrect Protection Settings or Fuse Rating – Circuits operating due to lower fuse sizes or protection
- 69 Unsuitable Protection Characteristics – The protection can't function normally due to the current characteristics or the circuit
- 70 Inadequate Rupturing or Short Circuit Capacity – The circuit having a greater fault current bypassing the 1st protection
- 71 Deterioration due to Ageing or Wear (excluding corrosion) – Equipment or components that have exceeding the original lifespan specifications

- 72 Fault on DNO Equipment affecting Adjacent Equipment – Equipment failure next to the source of the fault
- 73 Unsuitable Paralleling Conditions – A circuit operation of an in-feed to an open point on a connecting circuit
- 75 Operational or Safety Restriction – The removal of supplies to make something safe or dig back where not enough length is available to effectively isolate
- 76 Extension of Fault Zone due to Fault Switching (including ASC held faults) – Switching out a larger network to isolate the fault
- 77 Inadequate or Faulty Maintenance – a circuit failing after maintenance to improve it
- 81 Switching Error by Contractors – as 61 but by non-DNO employees
- 82 Testing or Commissioning Error by Contractors – as 62 but by non-DNO employees
- 83 Incorrect application or equipment by Contractors – as 65 but by non-DNO employees
- 84 Faulty Installation or Construction by Contractors – as 66 but by non-DNO employees
- 85 Fault on customers network causing operation of Network Protection – IDNO or Commercial customers
- 86 Interruption to remove local generator or restore temp connection (>18hrs) – Removing supplies to restore onto the main network.

Generating Companies:-

- 87 Local Generation Failure – DNO owned generator stopping on long term repairs
- 88 DNO Equipment Affected by National Grid Personnel or Equipment – any of our assets being impacted by National Grid
- 89 DNO Equipment Affected by Private Generator or Authorised Operator – 3rd party hired generators that fail with an operational contract

Unclassified or Unknown:-

- 90 Faulty Manufacturing, Design, Assembly or Materials – anything that doesn't do what it's supposed to after installation
- 97 No-Fault Found – supplies are restored from a back feed, but the fault hasn't been located
- 98 Causes Unclassified in this Table – An Identified cause that can't be allocated to a listed cause
- 99 Cause Unknown – Any fault where the cause isn't certain
- A1 Transient Fault No Repair – The operation of a fuse in the normal feeding point
- A2 Premature Insulation Failure – The breakdown of insulation sooner than recommended (possibly need proving for this code to be used)

MEI (Main Equipment Involved).

This is what part of the network has been affected by the fault and is broken down into four categories, Overhead, Underground Main, Underground Service & Switchgear/Fusegear/Link-box/Cut-out.

Overhead:-

- 01 Overhead Main Bare Conductors – Un-insulated Main
- 02 Overhead Main Insulated Conductors – Insulated Main
- 03 Overhead Main Aerial Bundled Conductor – ABC main
- 09 Overhead Main Mixed – any combination of the above
- 11 Overhead Service (Metered) Bare Conductors – Un-insulated Service
- 12 Overhead Service (Metered) Insulated Conductors – Insulated Service
- 13 Overhead Service (Metered) Mixed – any combination
- 14 Overhead Service (Metered) Concentric – Concentric Service
- 15 Overhead Service (Metered) Duplex/Triplex – Duplex/Triplex Service
- 16 Overhead Service (Metered) Aerial Bundled Conductor – ABC Service
- 19 Overhead Service (Metered) Other – anything else
- 20 Surface Wiring Main – Mural Main
- 30 Surface Wiring Service – Mural Service

Underground Main:-

- 41 Underground Main PLCS (armoured or unarmoured) – Main PLCs
- 42 Underground Main CONSAC – CONSAC Main
- 44 Underground Main Waveform – Waveform Main
- 45 Underground Main Districable – Districable Main
- 49 Underground Main Mixed or unclassified – Mixed or Unclassified Main

Underground Service:-

- 51 Underground Service (Metered) PLCS – PLCS Service
- 52 Underground Service (Metered) Plastic Insulated Concentric Types – Plastic Service
- 53 Underground Service (Metered) CONSAC – CONSAC service
- 54 Underground Service (Metered) Waveform – Waveform Service
- 55 Underground Service (Metered) Districable – Districable Service
- 59 Underground Service (Metered) Mixed or Unclassified – Mixed or Unclassified Service

Switchgear/Fusegear/Link-Box/Cut-Out :-

- 61 Switchgear/Fusegear (excl. service cut-outs) Circuit Breaker – Breaker on the circuit
- 62 Switchgear/Fusegear (excl. service cut-outs) Pole Mounted Isolator – isolation point on the pole
- 63 Switchgear/Fusegear (excl. service cut-outs) S/S fuse board or pillar – LV board for TX
- 64 Switchgear/Fusegear (excl. service cut-outs) Pole Mounted fuse gear – PMT

- 65 Switchgear/Fusegear (excl. service cut-outs) Street Feeder Pillar – FP's
- 66 Switchgear/Fusegear (excl. service cut-outs) Multi-Service Pillar or Turrett – Flats/Apartments
- 67 Switchgear/Fusegear (excl. service cut-outs) Link Box – LB's
- 68 Switchgear/Fusegear (excl. service cut-outs) fused wall box – Wall Boxes
- 69 Switchgear/Fusegear (excl. service cut-outs) other – anything else
- 72 Cut Outs (Metered) Overhead – COOH's default to this
- 73 Cut Outs (Metered) Underground – COUG's default to this
- 82 Un-Metered Service Overhead – Overhead Service to supply without a Meter
- 86 Un-Metered Service Underground – Underground Service to supply without a Meter
- 90 Other – anything else

Component
The next section is Component and is directly linked to the MEI. It describes the equipment that has faulted. Dependant upon your MEI depends upon the outcome of your component as there are multiple options for the same input.

Overhead:-

- 0 Conductor – the cable
- 1 Jumper or Dropper – this part of the circuit
- Pole incl. Stays and Steelwork - anything about the pole and it's the structure
- Insulator – this part of the circuit
- Lead-in – the connection to the cut out from a mural service
- Jointed Termination Compression – Compression joint
- Jointed Termination Line Tap – Line Tap joint
- Jointed Termination other – any other joint
- Attachment to Building – Bolts, Screws, Fixings
- Other – anything else
- X No Component identified – no one piece determined as the origin of the fault

Underground:-

- 0 Cable excluding joints and terminations – the cable
- 1 Mains Joint (CNE to SNE cable) – Combined Neutral Earth Cable to Separate Neutral Earth Cable
- Mains Joint – equal value cores to cores
- Service Joint – the last joint before the customers cut out
- Heat Shrink termination Pole Mounted – Termination to overhead
- Heat Shrink termination Other – other Termination
- Non Heat Shrink termination Pole Mounted – Termination to overhead
- Non Heat Shrink termination Other – other Termination
- X No Component identified - no one piece determined as the origin of the fault

Switchgear:-

- 0 Link – Solid Link
- 1 Fuse – Fuse
- 2 Fuse Carrier – Housing for fuse
- 3 Main Contacts – LV board Jaws
- 4 Busbars – Conductive Bar
- 5 Busbar supports – what holds the bars in place
- 6 Trip Mechanism & coils - tripping device
- 7 Enclosure – the frame (e..g actually the link box)
- 8 Connections – how it connects to the network
- 9 Other – anything else
- X No Component identified - no one piece determined as the origin of the fault

Other:- only accessible through MEI 90

- 0 Other Link – Solid Link
- 1 Other Fuse – Fuse
- 2 Other Fuse Carrier – Housing for fuse
- 3 Other Main Contacts – LV board Jaws
- 4 Other Busbars – Conductive Bar
- 5 Other Busbar supports – what holds the bars in place
- 6 Other Trip Mechanism & coils - tripping device
- 7 Other Enclosure – the frame (e..g actually the link box)
- 8 Other Connections – how it connects to the network
- 9 Other Other – anything else
- X Other No Component identified - no one piece determined as the origin of the fault

AB (Area Board)

The final part is the AB (Area Board) boxes; these are to determine the cause of damage, whether a repair has been done, what material is involved and how it is billed. Only underground faults need AB 1, 3, 4 & 5. Overhead & Switchgear only require AB 5

AB 1:- how has the fault occurred

- 0 No Damage – nothing has contributed to the cause
- 1 Mechanical Excavator – Diggers
- 2 Other Machines – Anything else
- 3 Hand tools – Spades, Breakers, etc
- 4 Jointing Operations – anything involving jointing
- 5 Displacement – the equipment has been moved causing damage
- 6 Penetration – piercing devices through the ground
- 7 Corrosion due to previous damage – to be used with DC 15 or old damages
- 9 Unknown/unclassified – anything else

AB 3: - how the fault was dealt with

- 0 Fault not Cleared – still on the system but supplies restored (follow up required)

- 1 Fault Cleared/Repaired – completed repair work
- 2 Transient Fault – a fuse replacement in the normal feeding point of an underground network (used only with DC A1)

AB 4:- The material type of conductor involved

- 0 Aluminium Live – Live core made of Aluminium
- 1 Copper Live – Live core made of Copper
- 2 Mixed Live – any combination
- 3 Aluminium Neutral – to be used on Neutral faults
- 4 Copper Neutral – to be used on Neutral faults
- 5 CONSAC / Waveform – to be used with MEI's 42, 44, 53 & 54