Load models for technical, economic and tariff analysis of medium voltage feeders

Dissertation submitted for the degree of Master of Science in Engineering

Department of Electrical Engineering

University of Cape Town

Prepared and submitted by: **Mr Johannes Lolo Buys**

Supervised by: **Professor Charles Trevor Gaunt**

DECLARATION

I know the meaning of plagiarism and declare that all the work in the document, save for that which is properly acknowledged, is my own. This thesis/dissertation has been submitted to the similarity and originality checking software and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor.

Johannes Lolo Buys

Department of Electrical Engineering

University of Cape Town

South Africa

March 2020

ACKNOWLEDGEMENT

I wish to thank God for making it possible for me to reach this stage and for the strength he gave me to do this work.

I would like to thank my supervisor, Prof C.T. Gaunt, for the support, encouragement, guidance and positive criticism throughout the course of this work, and I acknowledge the opportunity that the University of Cape Town afforded me to undertake this dissertation. I would also like to express my gratitude to Eskom for funding this research and for making the data used as well as the resources required available to me. I wish to also extend my thanks to my mentor, Mr Jack Mathebula, for his excellent mentorship and guidance.

In addition, special thanks go to my manager, Mr Mutenda Tshipala, for his interest in and support of this and created an environment that encourage me to complete it. Finally, wish to thank my family, especially my wife, Malebo, and my children for their patience and support.

Abstract

Load models play an essential role in many studies, including calculating voltage drops and technical losses in distribution systems, for distributed generator (DG) integration planning, and in tariff analysis and design models. The Herman-Beta transform used in the low voltage network modelling studies in South Africa is based on loads modelled as Beta probability density functions. Recently, the transform was extended to make it useful also for probabilistic load flow modelling in medium voltage (MV) networks with non-unity power factor loads and DGs. The electricity supply industry in South Africa has transformed and saw an increased penetration of Independent Power Producers as a result of the government encouraged the renewable independent power procurement programme (REIPPP). There has also been a steady decrease in the costs of procuring power from renewable energy sources, mainly from photovoltaic (PV) systems. South Africa also saw significant tariff increases in the recent past. These have resulted in both new load patterns and uncertainties in the power systems inputs required for network planning and tariff development. Other factors affecting loads and renewable energy output include weather, location and economic factors.

Load models are essential for technical and tariff studies. Long term and short term planning models in both technical and tariff modelling require information about the usage behaviour of customers. Planning cannot be separated from the financial impact and tariffs in general. The literature review indicated that planning has the objective of designing a network for optimal usage, thus minimising the costs and deferring investment where possible. Load patterns have been recognised to represent the usage behaviours of customers better and these behaviours influence the planning parameters. There have been studies by numerous researchers to extract parameters from the load profiles for load flow modelling and simulation purposes. The same challenge exists for South Africa, where there has been progress made on the development of LV models, and the same is not replicated in the MV network space.

The derivation of load models primarily involves the classification of loads, identifying and estimating the parameters of loads, and assigning load profiles to different loads for studies. Customer measurements are an essential input in load model development and load estimation. Identification of parameters is one of the areas where research is ongoing since there is no global consensus on which attributes best describe customer load profiles. In this study, a proposition on how the parameters for technical and tariff analysis models should be defined was made. The use of 24-hour load profiles to classify calendar days into typical days was also suggested. The availability of measurements data made it possible to develop load models for MV and conduct a study on actual customer data. The customers' measurements data, made it possible to identify the parameters and develop load models that could be used for technical and tariff analysis and conduct a pilot study to evaluate the load models. This study proposes a load model that can be used to model typical days and to model customer loads. The load models proposed here uses the k-means clustering algorithm as the basis for classification. The load models enable the classification of loads and assignment of load profiles accordingly.

The results of this study indicated that load parameter models could be extracted from the customer measurements, for technical and tariff studies in distribution networks. It has also been possible to identify and determine the parameters from the load profiles and proposed a process for developing a load model for technical, economic and tariff analysis. The results also indicate that of the five identified parameters, the most significant parameters that affected the clustering results were the load factor, average power and the normalised peak usage parameter when the results of each of the factors were compared on an individual basis. The study also revealed improvements to the clustering results when all the parameters identified in this study were combined and a PCA-based clustering algorithm was used. Finally, the results indicate that the loads in the different economic activity-based classifications do not necessarily have similar shapes although they belong to the same cluster. The modelling process developed in this study may be implemented by utilities for determining load parameter models for MV feeders when measurements are available. The process may also be used to guide future data collection.

Keywords: classification, clustering, load models, load patterns, medium voltage, network planning and operations, sampling, tariffs.

Table of Contents

1. INTRODUCTION 11
1.1. Research context 11
1.2. Motivation for the research 11
1.3. Load models for technical and tariff analysis 13
1.4. Hypothesis 13
1.5. Research questions 13
1.6. Research contribution 14
1.7. Research scope and limitation 14
1.8. Dissertation outline 14
1.9. Concluding remarks 14
2. LITERATURE SURVEY 16
2.1. Introduction 16
2.2. Overview of load models and their objectives 16
2.3. Models for technical and economic analysis 17
2.4. Models for tariff studies 19
2.5. Classification models 24
2.6. Parameters for technical and tariff models 26
2.9. Cluster validation measures or indices 32
2.10. PCA based clustering 34
2.11. Load model validation 35
2.12. Concluding remarks 36
3. DERIVATION OF LOAD PARAMETER MODELS 37
3.1. Introduction 37
3.2. Load modelling principles 37
3.3. Load parameter modelling process 38
3.4. Feeder characterisation and load parameter model derivation 41
3.5. Class concepts and models 46
3.6. Data preparation 48
3.7. Validation of the results 51
3.8. Concluding remarks 52
4. DATA PREPARATION AND PARAMETER ESTIMATION 53
4.1. Data preparation 53
4.2. Data exploration and load parameter estimation 55
4.3. Concluding remarks 75
5. TYPICAL DAY MODELLING 77
5.1. Selecting parameters of a typical day load model 77

5.2. Classifying the days 78
5.3. Allocation of typical days to clusters (Classification) 82
5.4. Concluding remarks 83
6. CUSTOMER CLASSIFICATION LOAD MODEL 84
6.1. Parameter selection 84
6.2. Classifying customers 84
6.3. Validating the results 100
6.4. Implementation consideration of the load models 103
6.6. Concluding remarks 106
7. DISCUSSION OF THE RESULTS 107
8. CONCLUSIONS AND IMPLICATIONS 110
8.1. Answers to the research questions 110
8.2. Unpacking and validating the hypothesis 113
8.3. Implications 114
References 116

LIST OF FIGURES

Figure 1: The context of load modelling 38
Figure 2: Load modelling process using the measurements-based approach 39
Figure 3: Clustering process flow diagram 40
Figure 4: Load parameter model composition 42
Figure 5: typical day profile based on average demand of sample per activity classes 42
Figure 6: Hourly power factors of different classes 43
Figure 7: Stratified sampling process flow diagram 49
Figure 8: Comparison of average monthly energy levels between different economic classes (kWh) 55
Figure 9: An illustration of the normalised daily load profiles for the different SICs from the sample 55
Figure 10: Average daily load profiles of the economic classes as defined in the Eskom database 56
Figure 11: Box and whiskers plots of the customer classes. 58
Figure 12: The distribution of parameters in the agriculture classes. 58
Figure 13: Distribution plots of the parameters of the commercial class 60
Figure 14: Distribution of parameters of the bulk/distributors class 61
Figure 15: Distribution plots of parameters in the industrial class 62
Figure 16: Scatter plot and a density plot of the typical residential class plotted using hourly data for 1 year. 64
Figure 17: Scatter plot and density plot of a commercial class for a year 64
Figure 18: Scatter plot and density plot of the agricultural class for a year 65
Figure 19: Scatter plot of an industrial (Automotive) class indicating weekdays for a year 66
Figure 20: Scatter plot of an industrial feeder (Smelter and mineral processing) indicating weekdays 66
Figure 21: Scatter plot of a mining feeder indicating weekdays for a year 67
Figure 22: Scatter plot of active (kWh) and reactive (kVAr-lag) energy for agriculture and commercial loads 68
Figure 23: Scatter plot of kWh) and kVAr-lag energy for bulk/distributor loads 68
Figure 24: Scatter plot of kWh and kVAr-lag energy for large industrial and mining loads 69
Figure 25: Daily trend cycle analysis of agricultural class 70
Figure 26: Annual trend cycle analysis for Agriculture sector 70
Figure 27: Daily trend cycle analysis for Bulk\Distributors 71
Figure 28: Annual trend cycle analysis for Bulk\Distributors 71
Figure 29: Daily trend cycle analysis for commercial customer class 72
Figure 30: Annual trend cycle analysis for commercial customer class 72
Figure 31: Daily trend cycle analysis for large industrial customer 73
Figure 32: Annual trend cycle analysis for large industrial customer class 73
Figure 33: Daily trend cycle analysis for small industrial customer class 74
Figure 34: Annual trend cycle analysis for small industrial customer class 74
Figure 35: Daily trend cycle analysis plot for mining customer class 75
Figure 36: Annual trend cycle analysis for mining customer class 75
Figure 37: 24 hour load profiles for the different days of the year 77
Figure 38: The silhouette plot and the 3D scatter plots showing 20 clusters 79
Figure 39: Elbow diagram showing the possible optimal number of clusters for typical days. 79
Figure 40: The DBI plot for 20 clusters 80
Figure 41: Silhouette and scatter plots for 2 clusters 80
Figure 42: Silhouettes and scatter plots for 3 clusters – National data set 81
Figure 43: Silhouettes and scatter plots for four clusters 81
Figure 44: Load profiles allocated to the typical day classes 83
Figure 45: Silhouettes coefficients and scatter plots of different pairs of the parameters when 20 clusters were specified. 85
Figure 46: The elbow diagram showing the saturation region(area) 86
Figure 47: Silhouettes coefficients and scatter plots of different pairs of the parameters when 2 clusters were specified. 87
Figure 48: Silhouettes coefficients and scatter plots of different pairs of the parameters for 3 clusters 88
Figure 49: Silhouettes coefficients and scatter plots of different pairs of the parameters when 4 clusters were specified 89
Figure 50: Silhouettes coefficients and scatter plots of different pairs of the parameters when 5 clusters were specified 90

Figure 51: Elbow diagram for the selection of parameters 91
Figure 52: The silhouette scores of different parameters and clusters 92
Figure 53: Box plots showing the distribution of the per-unit values associated with the parameters within each of the five clusters 93
Figure 54: The probability plots for cluster zero for Parameters A to E and the average power. 94
Figure 55: Comparison of the load profiles from the 5 clusters (0 to 4) based on the proposed parameters and the economic classes' profiles 95
Figure 56: Distribution plot of principal components 96
Figure 57: Box plots of the two principal components of the five parameters 96
Figure 58: Average daily profiles of the five resulting clusters 97
Figure 59: Histograms of clusters 0 to 5 98
Figure 60: Load profiles of the clusters within the existing customer classes 99
Figure 61: Results from regression analysis of cluster 0 100
Figure 62: Cluster 1-regression analysis results 101
Figure 63: Cluster 2 regression analysis results 101
Figure 64: Cluster 3 Regression analysis results 102
Figure 65: Cluster 4 regression analysis results 102
Figure 66: Surface plot of proportions of customer classes in the clusters 104
Figure 67: Load models diagram 111

List of tables

Table 1: Summary of sample sizes per economic class 54
Table 2: Summary of the load measurements of the sample 54
Table 3: Symbols to represent different parameters 56
Table 4: Summary statistics of the agriculture class 59
Table 5: Summary statistics of commercial class 60
Table 6: Summary statistics of bulk/distributors 62
Table 7: Summary statistics for the industrial class 63
Table 8: Typical results from the PCA algorithm 78
Table 9: The typical days associated with the clusters 82
Table 10: Calculated averages for the parameter displayed per activity class 84
Table 11: The silhouettes and average DBI scores for different numbers of clusters 91
Table 12: Statistical summaries of the parameters 97
Table 13: Average load of the clusters 99
Table 14: Regression analysis summary for the 5 clusters 103
Table 15: Customer class percentages in each of the clusters 103
Table 16: Limits of parameters for different clusters 105

List of equations

Equation 1 29
Equation 2 32
Equation 3 33
Equation 4 33
Equation 5 33
Equation 6 34
Equation 7 34
Equation 8 34
Equation 9 34
Equation 10 43
Equation 11 43
Equation 12 44
Equation 13 44
Equation 14 45
Equation 15 45
Equation 16 45
Equation 17 46
Equation 18 46
Equation 19 46
Equation 20 47
Equation 21 47
Equation 22 48
Equation 23 50
Equation 24 52
Equation 25 52
Equation 26 52

List of Abbreviations

2D/3D	Two / three dimensions
2D/3D	Two / three dimensions
ADMD	After-diversity maximum demand
CDF	Cumulative distribution function
DBI	Davis Bouldin index
DG	Distributed generation
DR	Demand response
DSM	Demand side management
EPP	Electricity pricing policy
GLF	Geographical Load Forecast
HB	Herman beta
I	Current
IPP	Independent power producer
IRP	Integrated resource plan
kV	Kilo volt
kVA	Kilo Volt-Ampere
kVAr	Kilo Volt-Ampere (reactive)
kW	Kilo watt
LV	Low voltage
MV	Medium voltage
NMD	Notified maximum demand
O&M	Operation and maintenance
P	Power
PC	Principal component
PCA	Principal component analysis
PDF	Probability distribution function
PLF	probabilistic load function
PV	Photovoltaic
RPI	Representative Probability Index
RTP	Real time pricing
SSE	Sum of square errors
SSEG	Small scale embedded generator
SIC	Sector identification code
TDP	Typical day profile
TOU	Time of use
TWP	Typical week profile
Z	Impendence

1. INTRODUCTION

The purpose of this study was to develop a set of coherent load parameter models for both technical and tariff analysis. The load models play an important role in the development of the tools and methodologies used to provide technical and tariff solutions. The tariff analysis and design models for distribution companies rely on accurate representations of the loads in the electric power system. Central to the development of load models are the measurement data, tools and techniques required to transform the measurement data into load models.

This chapter discusses the context and motivation for the study. The research hypothesis is outlined, as are the research questions. The chapter also summarises the anticipated contributions of the study findings from the pilot results and concludes by outlining the scope and limitations of the study.

1.1. Research context

In 2008, Eskom faced a new challenge of not being able to meet the power demand. This highlighted the need for an energy plan that included building new generation capacity to meet the demand. The energy plan then gave birth to the integrated resource plan (IRP, 2011), which was promulgated in 2011. According to this plan, approximately 17.8 GW of renewables and 8.9 GW of other energy generation sources were to be procured from independent power producers (IPPs) (IRP, 2011). In addition, several energy conservation and demand control measures were put in place by the utility to reduce the demand, thus reducing the strain on generation capacity (Cousins, 2009). The decreasing cost of renewable energy sources, especially photovoltaic (PV) systems, encouraged some consumers to install their own PV generation schemes in a bid to reduce the burden of increasing electricity tariffs (hereafter referred to as tariffs). Some customers chose to install energy-efficiency equipment while others curtailed their energy use in a bid to reduce energy costs and improve efficiency.

Tariffs can either cause or defer investments in the distribution network, as well as triggering changes in the electricity usage behaviours of customers. On the other hand, investments are the key drivers of tariffs. Adequately representative estimates of loads can lead to savings in power system planning and investment activities (Xu, 2015). In addition, the tariffs charged for active energy, as well as raising prices during expensive hours and lowering them during inexpensive hours, can be useful in incentivising load shifting or changes in load patterns to lower the system costs for utilities and bring down customer bills (Cousins, 2009). Demand response (DR) methods can also change the load patterns and have been employed by numerous utilities (Eid et al., 2016). The ability to recognise the types of customer loads and to differentiate between them is an important tool in the design of tariffs to incentivise load-shifting and/or other changes in consumption patterns and network planning (Granell et al., 2015, Buys & Gaunt, 2020).

Historically, load research based on extensive surveys and measurements taken over long periods has been at the forefront of understanding customers and classifying them, as well as estimating the load profiles. This was motivated by the lack of metering and database systems. However, this is no longer the case as most utilities have information relating to most of their customers in their databases, especially at the MV feeder level. The availability of measurements data, tools and the improvements of databases have given rise to the possibility of exploiting large volumes of data (Granell et al., 2015) to obtain the data required for any type of load modelling.

1.2. Motivation for the research

The changes in the power system and increasing penetration of DGs are some of the triggers for the need to update the load models. Load models are used by planners, designers and tariff developers to analyse distribution systems and to adapt to the transforming power system as the need arises.

The challenge to utilities lies in understanding the customer usage behaviours taking into consideration the impact of such behaviours on the costs of supplying them (or tariffs) and, conversely, the changes in the usage behaviours that can be encouraged by tariffs (Fisher et al., 2017). Furthermore, utilities must recognise that the tariffs can either encourage or deter new network investments (Cousins, 2009; Granell et al., 2015; Eid et al., 2016).

The convergence of technical and tariff analysis models goes beyond load models as inputs to these studies, as it extends to the application of the results of power flow studies to tariff design. The Electric Utility Cost Allocation Manual (1992) pointed out that the cost information for transmission and distribution networks was obtained from the utility's power flow analysis models when marginal cost methods were used. However, this convergence is seldom explicitly expressed in literature. Many load models were developed around the technical parameters that described the planning and operational characteristics of the power systems but without reference to tariff impacts, with the issue of tariff-related characteristics being addressed in areas of economic research. The convergence of technical and tariff models may be expressed through the derivation of a set of coherent load parameter models for technical and tariff analysis.

The inclusion of the variability of both loads and DG is a challenge, given the asymmetrical variability of the load demand. As a result, deterministic approaches to load modelling could give misleading results (Gaunt et al., 2017). According to ElNozahy et al. (2013), robust load modelling techniques should consider the stochastic nature of system loads. Models based on PLF analysis could be a way to incorporate uncertainties because they consider the stochastic nature of system loads. Knowledge about the load characteristics is needed for PLF analysis of the distribution system, and load models provide the knowledge required for PLF simulations. According to Prusty and Jena (2017), uncertainties may be classified into input and system uncertainties.

Load modelling at MV levels could provide knowledge of the class compositions of various MV and LV loads. Previously, the LV load models used in Eskom were developed from the historical load research projects, which were carried out with data loggers and through door-to-door campaigns (Heunis and Dekenah, 2014). To date, a similar study at MV feeder levels has not been conducted, resulting in a lack of MV load models. To bridge the gap there had been studies that adopted LV models with modification to mimic MV loads. However, the modification of the LV feeder models for use in the MV studies may not give accurate results because the MV feeders are often an aggregate of multiple feeders with different profiles. Similarly, the disaggregation of composite loads to obtain the elementary view may not be achieved unless probabilistic models were used (Gaunt et al., 1999). Nevertheless, the development of load models specifically for MV networks should provide a solution to this problem.

The challenge to utilities lies in understanding the customer usage behaviours, taking into consideration the impact of such behaviours on the costs of supply (or tariffs) and, conversely, the changes in the usage behaviours that can be encouraged by tariffs (Fisher et al., 2017). Furthermore, it was vital that utilities recognise that the tariffs could either encourage or deter new network investments (Cousins, 2009; Granell et al., 2015; Eid et al., 2016).

The convergence of technical and tariff analysis models goes beyond load models as inputs to these studies, as it extends to the application of the results of power flow studies to tariff design. The Electric Utility Cost Allocation Manual (1992) pointed out that the cost information for transmission and distribution networks was obtained from the utility's power flow models when marginal cost methods were used. However, this convergence was seldom explicitly expressed in literature. Many load models were developed around the technical parameters that described the planning and operational characteristics of the power systems, without reference to tariff impacts, while the issue of tariff-related characteristics were being addressed in areas of economic research. The convergence of technical and tariff models may be expressed through the derivation of a set of coherent load parameter models for technical and tariff analysis.

The inclusion of the variability of both loads and DG was a challenge, given the asymmetrical variability of the load demand. As a result, deterministic approaches to load modelling could give misleading results (Gaunt et al., 2017). According to ElNozahy et al. (2013), robust load modelling techniques should consider the stochastic nature of system loads. Models based on PLF analysis were seen as a way in which to incorporate uncertainties since they considered the stochastic nature of system loads. Knowledge about the load characteristics is needed for PLF analysis of the distribution system, and load models provide the knowledge required for PLF simulations. According to Prusty and Jena (2017), uncertainties may be classified into input and system uncertainties.

Load modelling at MV levels may provide knowledge of the class compositions of various MV and LV loads. Previously, the LV load models used in Eskom were developed from the historical load research projects, which were carried out with data loggers and through door-to-door campaigns (Heunis and Dekenah, 2014). To date, a similar study at MV feeder levels has not been conducted, resulting in a lack of MV load models. In an effort to bridge the gap there had been studies that adopted LV models with modification to mimic MV loads. However, the modification of the LV feeder models for use in the MV studies may not give accurate results because the MV feeders were often an aggregate of multiple feeders with different profiles. Similarly, the disaggregation of composite loads in order to obtain the elementary view would not be achieved unless probabilistic models were used (Gaunt et al., 1999). Nevertheless, the development of load models specifically for MV networks should provide a solution to this problem.

This study was motivated by the unavailability of load models for MV feeders in South Africa, the possibility of developing practical load modelling frameworks that use feeder measurements data and the need to test if the existing customer classes were valid in view of the technical and tariff applications given the changes that were taking place in power systems globally. It was anticipated that this would benefit Eskom and utilities without load modelling frameworks for MV technical and tariff applications in place.

1.3. Load models for technical and tariff analysis

Tariff analysis aims to address the tariff design problem by realising several tariff objectives. De Sá Ferreira et al. (2013) identified both aspects of the tariff problem to include the promotion of economic efficiency, revenue adequacy and the provision of incentives related to usage. The input required to provide electricity-related services that were identified by Pérez-Arriaga et al. (2013) to include the characteristics of demand (the need), the costs of supply, and the allocation objectives and mechanisms.

The load flow problem deals with the network planning, network design and operational performance, thereby assessing the voltage drop for given load requirements and power supply availability (Prusty and Jena, 2017; Renmu et al., 2006). The essential inputs for technical analysis may include the demand and the location of the customer, with these inputs being transformed into the voltage drops and/or other thermal transfer parameters that enable the adequacy of the distribution system to be assessed (Herman et al., 2019).

To extract a set of coherent load models for technical and tariff analysis, the parameters for technical studies as well as those of tariff studies need to be identified. Also, uncertainties resulting from the technical and tariff models need to be taken into consideration. The changing load patterns and tariff structures as well as the introduction of DG and energy-efficiency interventions by consumers and utilities have introduced uncertainties into the load parameter models that are vital in the planning and analysis of the power system. These uncertainties mean that the utilities are at risk of losing revenue and of planning inaccurately unless they adapt and begin to incorporate the uncertainties into their load models.

1.4. Hypothesis

Given the availability of measurement data for the MV feeders in most utilities and the need for load models for technical and tariff applications at MV levels, as covered in the preceding sections, the following research hypothesis was adopted for the purposes of this study:

Coherent customer load models suitable for technical, financial and tariff analysis of medium voltage systems can be derived from customers' load measurements and the characteristic parameters.

1.5. Research questions

It was possible to ascertain from previous studies that have been conducted and documented and from experience, the progress that has been made in the development of load models. To test the hypothesis stated above, it was necessary to find comprehensive answers to the following research questions:

a) What are the various load models and what is their purpose?
b) What are the load parameters that are important for technical, financial and tariff analysis models?
c) Is it possible to derive a set of load parameters for technical, financial and tariff analysis from MV feeder measurements data? How can these parameters be derived?
d) How can the clustering algorithms be used to classify loads in load modelling of MV systems?
e) How can the clustering results be tested and validated?
f) How can the load models be tested and validated?

The answers to these questions were sought from the published literature, the development of a process for deriving load parameter models and experimentation using the MV feeder measurements.

1.6. Research contribution

If the hypothesis were proven valid, it was anticipated that the process developed would be of use to electricity utilities in transforming load measurements into models, which could be used for technical and tariff analysis of distribution systems.

1.7. Research scope and limitation

The scope of this research study was limited to the measurements based models on MV feeders in South Africa and owned by Eskom. It was anticipated that the outcome of this study could be used to inform tariff studies and load flow studies. This research was limited to focusing on the correct classification of customer based on the measurements data, and to deriving the load parameters models that can be useful in achieving this.

1.8. Dissertation outline

The dissertation is arranged as follows:

Chapter 2 reviews relevant literature on load models and the way in which they have been applied. Various load models for technical and tariff analysis were reviewed in order to ascertain their relevance in the application to MV loads and what is required in the models.

Chapter 3 presents a theoretical framework and the techniques that were used as the basis of this research. The algorithms are examined in details and the tools used for modelling and analysis discussed.

Chapter 4 discusses the data preparation, namely, the data collection, sampling method and the normalisation technique used. Further chapter 4 provides the results from exploratory data analysis and assessment of the relationship between the active and reactive energy.

Chapter 5 presents the results of the load model for deriving typical day classes.

Chapter 6 presents the results of the customer classification based on a set of parameters for technical and tariff modelling.

Chapter 7 contains a detailed analysis of the results and the load modelling implementation modalities.

Chapter 8 concludes the dissertation by providing summarised answers to the research questions, assessing the validity of the hypothesis and the implications.

1.9. Concluding remarks

This research develops a set of coherent load parameter models needed for technical and tariff analysis. The study was triggered by the development in the power sector, including changes in electricity usage patterns, uncertainties

and the emergence of DG, as well as the unavailability of appropriate load models for MV loads. A research hypothesis and research questions to help test the validity of the hypothesis are formulated. The next chapter reviews the information available and related research questions.

2. LITERATURE SURVEY

The purpose of this chapter is to outline some of the key findings from the studies that have been published on load modelling in order to answer the research questions and to confirm the hypothesis formulated in the previous chapter.

2.1. Introduction

The MV feeder measurements provide a view of the loads connected to the feeder in question and are aggregated to represent the loads as a single profile. The advantages of aggregated measurements compared to a set of individual load measurements include the availability of data, full representativeness, and continuous updates (Gerossier, Barbier and Girard, 2017). Given the developments in the electricity supply sectors globally and the availability of both data and the tools to manipulate it, there is a growing interest in developing load models from feeder measurements.

For the purpose of this study, the terms *parameters* is used to mean any representation, observable, calculated or otherwise quantified attribute, to describe the load pattern or sections thereof that play an important role in the development of load models.

2.2. Overview of load models and their objectives

Load models are seen as a representation of a load and are used for understanding the impact of the load flows on the power systems. Xu (2015) states that "the term load model denotes an analytical, mathematical, equivalent-circuit based, physical-based, component-based or otherwise-established or formulated, representation of a load, which correctly represents the changes in the real and reactive power demands of the modelled load as a function of certain power system parameter (i.e. voltage, frequency) variations".

Load models are an essential input to technical and tariff studies because they provide engineers with information about the system usage by the customers or load characteristics that can be used to simulate various operating states and assess whether the power system is adequate to services customers. Impact assessment report (2016) observed that new structures for power systems were emerging that entailed more actors (DG consumers, change of consumers to prosumers) and changes in relationships.

The introduction of Distributed Generation (DG) in the South African electricity industry, has changed the distribution network structures and has had an impact on the traditional models that were used for analysis. This means that networks require re-examination related to the planning of the distribution systems, the models used for tariffs and the incentive schemes that are aimed at encouraging participation in the optimisation of the network (Impact assessment report, 2016).

According to (Milanovic et al., 2014; Ramírez-Mendiola et al., 2017; Gbadamosi, 2017), load models may be classified under deterministic and probabilistic models. Load models that assume direct causal relationships between the specific inputs and the expected outcomes are considered to be deterministic models, whereas models that make use of statistical methods for the simulated inputs or outputs are referred to as probabilistic or stochastic load models (Chihota and Gaunt, 2018; Gbadamosi, 2017). As the distribution systems evolve and become unpredictable, the need to replace the traditional deterministic models, which were historically successful, with probabilistic models arises (Chihota and Gaunt, 2018; Conti and Raiti, 2007). Load models are an important input to the development of probabilistic load flow models.

Various techniques have been used in developing stochastic load, such as those that were developed through clustering and statistically modelling of the profiles within the clusters (Chihota and Gaunt, 2018; Nassar and Salama, 2015). Other proposed works used clustering techniques to classify loads and attempted to mimic the behaviour of electrical loads through probabilistic modelling of the electrical loads (ElNozahy et al., 2013). Sangrody et al. (2017) used clustering and modelled the daily peak-load using the probabilistic model with PDFs.

Milanovic et al. (2014) cited that for an accurate representation of loads, a detailed analysis of the loads connected downstream of the point of aggregation is essential. Milanovic et al. (2014) further noted that although system loads are classified in general economic activity classes such as residential, commercial and industrial, this might not be sufficient to describe the diverse load composition and structures, and the associated load profiles within each load class. They recommended that general load classes be further divided into sub-classes, for which common load models can be developed.

Load models are used in different applications in technical and tariff studies. Technical models refer to models which are used for solving power flow problems (voltage, losses etc.) while tariff models refer to models which are developed for analysing, designing and setting tariffs.

2.3. Models for technical and economic analysis

Load models are essential for understanding the load variations without the effort of conducting extensive load surveys to make predictions about the customers' behavioural patterns in relation to electric power system usage. According to Gaunt et al. (2011), the distribution networks planning and design engineers require both the load models and algorithms to transform them into voltage drops to enable them to predict the voltage variations and take decisions on network adequacy. Thus, in technical modelling, load models may be transformed into voltage drops, demand and reactive energy amongst other variables for network analysis and capacity planning. Some of the planning models include financial optimization in addition to compliance with voltage drop limits (Herman and Gaunt, 2008). Violation of voltage limits could be costly to the distribution companies. Technical models are often developed to solve multiple objective optimization problems that include the minimisation of the financial impact.

Load models are often classified as either static or dynamic and can be time-variant or invariant single or multi-state (Gbadamosi, 2017). The relatively common method of modelling in power systems involves considering the loads as a combination of constant impedance (Z), current (I), and power elements (P), which are classified as single-state, time-invariant models, that is, energised state models (Milanovic et al., 2014; Gbadamosi, 2017). In most of the steady-state load modelling, the exponential load model and ZIP model are employed to determine active and reactive power and the system voltages (Chihota and Gaunt, 2018; Han et al., 2012; Xu, 2015).

Load models for technical analysis can be used to quantify the uncertainties related to the environmental, social, demographic and economic factors in distribution systems long term and short term or operational planning (Chen et al, 2008). Planning models have the objective of minimising investment, and operational and maintenance costs, and are conducted on the least life cycle cost basis.

2.3.1. Models for network planning

Network planning requires load modes that capture the load variations associated with customer usage behaviour. Load forecasting forms part of the operational planning models and plays a critical role in power system modelling. Thus, load models are key to load forecasting (Kourtis et al., 2011). Load models were essential in providing electrical demand data and load classification as inputs for Eskom's forecasting tool, the geo-based load forecast (GLF) tool (Hashe, 2012; Soni, 2018). The variables of interest in forecasting were found to include the hourly demand, the daily, weekly, monthly and annual peak system demand and energy usage and time of use consumption levels (Kourtis et al., 2011).

Losses modelling and simulation and DG integration planning also forms part of operational planning. Load models have been useful in calculating the technical losses in the distribution system (Sun et al., 1980) and for DG integration planning (Singh et al., 2007). Sun et al. (1980) used typical load profiles as a function of the bus voltage to study the effects of the different load types and the losses on different system components. Sun et al. (1980) used the typical daily load profiles in a load flow procedure for the computation of losses. A noteworthy parameter of their program is the modelling of system imbalances in terms of both bus loading and circuit

configuration using load profiles. Incorporating losses modelling in the planning models has a financial benefit to both the utilities and the customers.

The classical ZIP models are often used in load modelling. There had been findings that suggest that the use of constant ZIP modules in short term operational planning do not yield accurate results. Singh et al. (2007) found that DG planning results from the constant power load model were inaccurate in modelling varying load. Singh et al (2007) further found that load models had an impact on the optimal location and size of DG resources in a distribution system. Furthermore, Ballanti and Ochoa, (2015) found that the constant ZIP classic load models either underestimated or overestimated the network power losses, in both winter and summer seasons. Based on the observations by Singh et al. (2007), Balanti and Ochoa (2015), for short-term operational planning, zip models were inadequate and should rather be used for long term planning where load profiles were to be fixed over a longer-term. The development of probabilistic load flow models is an indication that the static parameter models are no longer adequate for planning.

In long term planning of the distribution systems, annual peak demand data spanning several years is required for load modelling (Black et al., 2018). According to Soni (2018), a planning process is an approach that seeks to investigate the distribution systems adequacy against the forecasted load. Location and economic factors affect long term planning as well as load forecasting. Prusty and Jena, (2017) identified the sources of power system uncertainties as environmental and social factors and demographic and economic factors. The economic factors are linked to operational studies and are said to be responsible for some operational and reliability uncertainties including SAIDI, whereas the demographics are associated with long term planning uncertainty (Prusty and Jena, 2017).

2.3.2. Reliability models

Extreme weather can affect the reliability of the network. A network that is not reliable contributes towards a financial loss by utilities and customers, and thus impact the economy. Black et al. (2018) also expressed this sentiment. According to Black et al., (2018) weather data were used in statistical regression algorithms to estimate the feeder outage and reliability trends based on the utility's reliability indices. Weather conditions were found to be responsible for a large number of customers' outages for extended periods and these significantly affect the SAIDI calculations. Weather data is not often available in the customer measurements databases. The possibilities of modelling parameters linked to weather reliability from customer measurement could benefit many utilities. Outage information is likely to be recorded by distribution companies. Further work is thus recommended in this area.

2.3.3. Probabilistic load flows models

The current trend is to migrate from deterministic load models to risk-based load models. Deterministic models that were traditionally used to determine customer demands in South Africa for network sizing based on after diversity maximum demands (ADMD) are being replaced with the probabilistic models, based on a variation of a beta distribution function, referred to as the Herman Beta (HB) function (Ferguson and Gaunt, 2003). The HB transform algorithm was used to transform residential currents derived from ADMD to voltage drops for network sizing and was demonstrated to be extendable to residential loads where there is limited data (Herman and Gaunt 2008).

The HB transform has been further extended for PLF modelling in MV networks (Chihota and Gaunt, 2018). The absence of load models for MV loads meant that the HB transform had to be modified to adapt it for MV network studies. The modifications that were made included the modelling of DGs as negative load and the addition of the reactive power component. Reactive power had to be considered because MV loads have active and reactive power components, therefore, the assumption of using active power only was not valid. Furthermore, due to data limitations when the model was developed, beta pdf was assumed applicable for MV load modelling on the basis that previous studies have demonstrated that it applied to LV load models (Chihota and Gaunt, 2018). The same

assumption was also considered for generation modelling, thus legitimising the need for load models that are based on the findings of MV feeder studies to validate these assumptions.

A study conducted by Broderick et al. (2014) found that the parameters that are determined by load research undertakings may be summarised as feeder peak load (kW), peak load month/time, feeder minimum load (estimated) (kW), feeder minimum load month/time and relative percentages share of the different economic classes. However as seen in the preceding technical models' section, the contracted power is a deterministic parameter, and the drive is to move towards the probability-based approaches to consider the uncertainties associated with the variations of loads.

Parameters are linked to the variation of loads that are represented as hourly load profiles of active energy. Depending on the desired outcome of the study, the simulation studies could evaluate each time point individually or in groups of hours and the results could also be hourly or in a group of hours. Reactive power is also another important parameter when modelling MV systems as expressed in (Chihota and Gaunt, 2018).

2.4. Models for tariff studies

To understand the task of developing tariff models, the steps leading to the composition of these models should be reviewed. Ortega et al. (2008) identified the essential drivers of load models in relation to tariff analysis to include load demand representation, cost of service and tariff objectives. It has emerged from studies conducted previously that tariff models generally started by classifying customers based on common usage patterns or based on defined parameters derived from the customer measurements(Chicco et al., 2006; Fidalgoa et al., 2012; Ferraro et al., 2016). This was followed by the allocation of representative profiles before assigning tariffs to each class.

As identified by Pérez-Arriaga (2013), this first step may be viewed as quantifying the need. Identifying the need means that the load to be served should be identified. The second step, the cost allocation step, requires a cost of supply study (Pérez-Arriaga, 2013), while the third step refers to the allocation objectives and mechanism, which includes tariff structures that incorporate signals or incentives for usage, as identified by EPP (2008), Cousins (2009) and Pérez-Arriaga (2013), and tariff products such as DR programs (Eid et al.' 2016).

Similar to the technical analysis models, the tariff models require a load model as an input to transform it to either tariffs or tariff components aligned to the modelling objectives. Granell et al., (2015), Impact assessment report (2016) and Eid et al. (2016) are amongst the growing number of researchers who used load models to consider the effects of tariffs on load profiles and network operations and planning. Tariff models include models for cost allocation and tariff design models.

2.4.1. Cost allocation models

In terms of the EPP (2008) distributors are required to conduct a COS study, which is to be submitted and approved by the National Energy Regulator of South Africa (Nersa). This COS study is required every 5 years unless there are major changes with regards to the supply chain. Cost allocation models are required for the cost of supply (COS) studies because they seek to identify and allocate the costs in relation to how the customers utilise energy and the distribution systems, and this means quantifying the need and allocating the costs. This forms part of steps two and three in the task of developing tariffs. Cost allocation models allocate distribution costs to various costs components and customer categories (Eid et al., 2016). Distribution costs comprise network-related costs, which are largely a distribution company's own costs and energy purchase costs, which are often aligned to the costs of production at different hours of the day. According to Feltner (2012), the major cost drivers may be classified into energy, contracted demand and customer-related costs. Network related costs are associated with demand and allocated based on contracted demand. Distribution companies purchase energy at a wholesale price and sell it to customers using retail tariffs. The cost allocation models that are used to allocate network-related costs can be classified into embedded cost models and marginal cost models (CER, 2004).

According to CER (2004), the process of estimating marginal cost involves analysing the possible cost effects of the changes in load. Further, a long-run marginal cost reflects changes in costs when all factors of production can be altered (CER, 2004). A notable requirement from the EPP (2008) is for utilities to consider cost reflectivity when conducting COS studies. According to Cousins (2009), if the prices reflect the costs of providing electricity to customers, then the pricing is considered the cost causality-based or cost-reflective pricing.

Energy costs vary with the amount of energy that the customer uses. Demand costs are linked with the capacity needs of customers or the contracted power, such as maximum demands at particular points in time (Feltner, 2012). Customer-related costs are not linked to customer usage but to services that are rendered such as customer meter reading, account management and administration etc. Therefore, the characteristics that are essential for load models for costing analysis can be summarised as follows:

2.4.1.1. Contracted power (Maximum demand)

In order to obtain a network cost causality function, it is necessary both to determine how the networks are planned and to understand what causes the costs to be incurred (Ortega et al., 2008). Feltner (2012) identified two methodologies that are commonly used for allocating demand-related costs of the distribution systems as the "minimum system" and the zero-intercept. In the minimum system methodology, the standard conductor size is selected and its price is applied to all conductors to determine the minimum system costs. The excess of the conductor costs are allocated as demand-related costs and the minimum system's costs are recovered as the customer-related costs. The zero-intercept methodology recognises a linear relationship between the unit cost of the conductor and its current carrying capability (Feltner, 2012). The maximum demand can be calculated from the current capability of the conductor and vice versa.

Demand is an important parameter for consideration in load models, and various researchers in model development, particularly in classification and load profiling have used it. Fidalgoa et al. (2012) proposed a methodology based on contracted power or notified maximum demand (NMD) and the total energy consumed as indices to create load profiles for LV customers. The profiles that were created were to be used by the trader of energy, taking into account that not all LV customers are metered and billed on a monthly basis, while the traders are billed on a half-hourly basis. Fidalgoa et al. (2012) reported that the method has been adopted for use in Portugal. On the other hand, Panapakidis et al. (2012) proposed categorisation of the customers based on their maximum demand. The model developed by Gaunt et al. (2011) also used ADMDs, which are the contracted demands.

2.4.1.2. Energy related costs

Energy-related costs are costs that are incurred in the procurement of energy and these costs carry signals or incentives for economic usage. The tariff structures that promote economic efficiency with respect to electricity usage should reflect the production costs and encourage consumption patterns that lead to the maximisation of customer welfare (De Sá Ferreira et al., 2013). This means that cost-reflective tariffs should lead to energy costs savings for the customers who are able to respond to the tariff signals (Cousins, 2009; Fischer et al., 2017). A necessary task in load modelling is to determine the parameters that are related to the energy usage behaviours, and also reflect the production costs and can encourage efficient usage of energy. The cost drivers for energy-related costs comprise parameters related to customer usage behaviours. Parameters that are related to the energy consumption behaviour have been explored by various researchers and they include, days of the week or typical days, how customers are classified (economic activity), consumption patterns or time-related consumption, and load factors. Researchers argue that customers utilise energy different for each of these periods. The typical days and customer classes require models that use classification algorithms that are based on load profile parameters to be used. To take into consideration the effect of the time of consumption in cost allocation, the load profiles are segmented in chronological order or time of use related bins.

2.4.1.3. Class costs

Load classes and typical day classes can be viewed as parameters when modelling loads, because of the different consumptions associated with various classes. Where the classes are formed and the load profiles allocated to them fairly represent most of the loads in those clusters, the load profiles of the clusters can be used in extracting the parameters for use in technical and tariff studies, instead of such parameters extracted from individual loads. Costing and tariff studies are reliant on classes for cost allocation and tariff design studies. Various classes and their formations have been explored and documented. Typical days, weeks or months, and customer classes are often part of the technical and tariff analysis.

a) Typical Days

Ferraro et al. (2016) suggested that the day of the week is an important parameter in analysing electrical load since different behaviours can be observed on workdays, pre-holidays and Saturdays and holidays. ElNozahy et al. (2013) supported the day of the week as a parameter of interest in load models and suggested that the behaviour of loads be represented by their daily profiling. Ferraro et al. (2016) found that consumption behaviours are different for the different groups of days. They also concluded that merely clustering based on the consumption of these days might not give the true electrical behaviour on these days, hence the need to use parameters that are linked to daily load profiles. However, Frost et al. (2017) compared the typical daily profiles to typical weekly profiles and found that typical weekly profiles indicated a better temporal resolution as compared to the typical day profile. Figueiredo et al. (2005) proposed a load model that classified days into weekdays and weekends based on experience. It is not clear how the experience was obtained, that is whether there was any study done.

It can be deduced that in costing models, typical days classes load models are useful in determining costs associated with different days of the week as well as reducing the computational burden caused by large measurements data sets and thus they can increase the modelling speed. To obtain typical days a classification procedure that uses the hourly consumption data a multivariate analysis algorithm should be used, to make sure that the vital variability on different days can be captured.

b) Customer class

A customer class is another important input to consider for the cost of supply studies. According to Piscitelli et al. (2019), customers are classified according to their activity such as residential, industrial, commercial etc. and generally, customers belonging to the same class can exhibit different consumption patterns. Classification of consumers and the allocation of load profiles leads to cost-reflective tariffs and accurate load forecasts that are essential for network and tariff planning, improved service quality, proper load management and efficiency strategies (Figueiredo et al., 2003). Similar to determining typical days, to analyse load classes a classification procedure that uses the customer consumption data has to be used.

The electrical loads in South Africa are classified based on economic activity into different categories. These loads have different load factors, often peak at different times of the day and have different usage patterns (EPP, 2008). According to Firestone et al. (2006), classes that are developed from similar and regular consumption patterns enable tariffs that recover the utility's costs and generate some profit from the delivery of electricity. Firestone et al. (2006) further highlight that when there are significant differences in the usage patterns within a class, the pricing signal is weakened, and the tariffs are no longer fair because customers with different peak loads with different impacts on the grid pay the same tariffs. This highlights the importance of accurately identifying the drivers of usage patterns and their relationships to classify customers effectively. In addition, it is necessary to identify the appropriate classification models and/or clustering techniques to assist in estimating the representative profiles for each class.

The Electricity Pricing policy (EPP, 2008) mentions the class compositions and their share of the system demand as follows; domestic (17.2%), agriculture (2.6%), mining (15%), industrial (37.7%), commercial (12.6%), transport (2.6%) and general (12.3%). While there have been significant changes to the power system in South Africa since the publishing of the EPP, it would appear that no studies have been conducted to update the EPP (2008) and, thus, the class compositions stated above have, in all likelihood, changed and need to be updated.

Load research conducted in South Africa (Heunis and Dekenah, 2014) studied and modelled typical profiles for domestic load classes. It was, however, not possible to find documented literature relating to either the estimation of load profiles for non-domestic load classes, MV feeder loads or any such undertaking as part of an ongoing load research project in South Africa. Therefore was necessary to study the different classes using MV measurements to identify different customers or feeder within a class. The classes derived from measurements data may be correlated with those formed using economic activity to determine profile differences within classes and the possible cross-subsidisation within class memberships.

2.4.1.4. Consumption patterns

The consumption patterns of customers vary with time and largely inform TOU tariffs. The cost of producing energy to meet the demand is often the driver of these costs. The costs are often higher during peak and lower in off-peak periods. Herman and Gaunt (2008) and Bobric et al. (2009) identified some of the factors influencing consumption behaviours as follows:

- Weather conditions – the season, the daily temperatures, the wind speed
- Demographic factors – the growth rate of the population, the number of the inhabitants in a certain area, the birth rate etc.
- Economic factors – the gross national product, labour productivity, and economic development rate, the level of life quality and, a very important element, the price of energy.

According to Black et al., (2018) business sectors including the power industry are affected by weather conditions. Weather affects system reliability and it is a key driver of both power supply and demand (Black et al., 2018). Li, et al., (2018) also had the same observation that household demand is influenced by demographic factors, life patterns and location.

2.4.2. Tariff design models

Tariff design models translate the cost of supplying (COS) energy to tariffs. Cost reflective tariff would generally recover the costs as determined in the COS studies. Tariff design is concerned with tariff structures and ratemaking. Tariff structures comprise allocation objectives and mechanism, which includes signals or incentives for usage, as guided by the utilities' policies. Tariffs structure for the recovery of distribution-related costs, which are demand related tariffs, comprise uniform charges for all customers with identical characteristics such as voltage level, electricity consumption, load or generation and incentive-based tariffs (Hinz et al., 2018). Incentive-based tariffs carry a signal related to the siting of electrical plants. Uniform tariffs have the same rate for all customers, independent of their geographical location whereas incentive-based tariffs vary by location (Hinz et al., 2018).

De Sá Ferreira et al. (2013) identified aspects of the electricity-tariffing problem as including the promotion of economic efficiency, revenue adequacy, and the provision of incentives for the adoption of new tariff modalities. Tariff models are intended to provide a solution to the tariff design problem, thus achieving the tariff goals. Energy-based tariffs are meant to recover the costs of energy production and are often time-based. Energy-based tariffs can be either static or dynamic. Static tariffs do not take into consideration the variability of the load with time, while dynamic tariffs are time-varying tariffs that take into account the consumption patterns of customers in relation to time and the costs of serving them (Cousins, 2009; Fischer et al., 2017).

Block-based tariffs are an example of static tariffs meant to be affordable for the vulnerable communities in the residential sectors. The block tariffs are often subsidized either through government direct funding or cross-subsidization sanctioned by regulators.

Time-based tariffs include TOU and special tariffs that are meant to address the short-term system stability requirements. TOU tariffs are designed to reduce peak loading of the power system and/or load shifting from peak to off-peak periods (Cousins, 2009; Layera et al. 2017). The simplest TOU tariffs involve two periods, namely,

the peak and off-peak periods, while the more complex designs comprise peak, standard or intermediate peak and off-peak periods (Cousins, 2009).

Research on dynamic tariffs comprises two major research streams, namely, research that relates to the effectiveness of the dynamic tariffs as a strategy to reduce peak demand through demand-side management (DSM), and research on consumer acceptance of the dynamic tariffs (Layera et al., 2017). The outcomes of the research on dynamic tariffs are often linked to the consumers' behaviour (Fischer et al., 2017), which is influenced by their perception of the complexity of the price (Layera et al., 2017), as well as the price elasticity of the electricity demand (Lijesen, 2007).

Dynamic tariffs such as TOU tariffs can potentially promote economic efficiency. TOU tariffs charged for active energy may be useful in incentivising load shifting or change in load patterns both to lower the system costs for utilities and to bring down customer bills (Cousins, 2009; Fischer et al., 2017). The dynamic tariffs can also be used to drive the behaviours of certain customers to ensure that the network is operated within constraints and, possibly, to defer investments by raising prices during peak hours and lowering them during off-peak hours (Impact assessment report, 2016).

The economic efficiency of tariffs can be distorted when tariffs are not aligned to the costs, that is, when fixed and variable costs are not recovered as fixed and variable tariffs. In Eskom and several other utilities in developing countries, a significant portion of the fixed costs of energy production and the distribution system is recovered through a variable tariff in the form of an energy charge. Picciariello et al. (2015) found that in distribution, energy-based (or volumetric) tariffs carry an inherent risk to the utility in the form of the costs arising from consumption at peak times not being recovered. The risk increases when the nett energy sold contracts and net metering is adopted, as this would mean less recovery of the embedded fixed costs (Ortega et al., 2008). For Distribution companies, contraction in nett energy sold also poses the revenue risk increases. This potential contraction in the nett energy sold is because of increased penetration DG (Ortega et al., 2008) and implementation of energy efficiency schemes and equipment by customers (Cousins, 2009). This highlights the need to design tariffs appropriately, and in alignment with how costs are incurred. The misalignment of costs and tariffs that recover these costs can adversely affect the revenues of utilities.

2.4.2.1. Special tariffs: Demand side management

Special time-based tariff products can be developed in a bid to influence the load patterns. These tariff products include critical peak pricing (EPP, 2008) and real-time pricing (RTP) (Cousins, 2009). In critical peak pricing, TOU tariffs are introduced and these tariffs are characterised by certain periods when the reliability of the power system is threatened and/or during periods where the energy prices are very high. According to Cousins (2009), RTP refers to prices that vary on an hourly or even a sub-hourly basis for some customers. Customers are notified of the tariff rates on a day-ahead or hour-ahead basis. Eskom (2017) found that there were customers who embraced RTP and who were able to use more electricity during the unconstrained periods of the power system as well as to shift the load out of the peak periods, as signalled by the RTP. Therefore, to model RTP, the load demand during the different periods are required. RTP are energy-based, and linked to time or hours of the day.

Demand response (DR) has been a central focus of time-based tariff modelling due to its benefits. DR has been extensively covered in the literature (Eskom, 2017; Ganesan et al., 2019; De Sá Ferreira et al., 2013). According to Eid et al. (2016), DR can lead to the adjustments of the loads to relieve the network capacity constraints and to remain within the technical limitations, therefore reducing the possibility of the power system collapsing. There are different ways in which to activate DR. Eid et al. (2016) summarise these various ways by distinguishing interruptible (direct control) and price-based (indirect control) methods of load modification. Direct methods are contract-based and provide secure flexibility in time and place for the system operator's control. Price-based DR refers to changes in normal consumption patterns or electricity usage by end-use customers in response to changes in the price of electricity over time (Eid et al., 2016). When the price differences between peak and off-peak loads are the main incentive for many DR programmes, the prosumers (and other price-anticipatory participants)

respond by shifting their consumption to cheaper time slots, thus resulting in a new demand profile (Riaz et al., 2017).

2.4.2.2. Customer response models

Customers may or may not respond to tariff signals. Impact assessment report (2016) found that TOU tariffs can drive operational efficiency and defer network investments by influencing the energy usage behaviour of customers who respond to these tariffs. In other words, high peak prices can lead to either peak shaving or peak reduction or load shifting, and consequently, reduce both the investment requirement and the energy losses. Further, Neenan and Eom (2008) found that real-time pricing (RTP) and critical peak pricing (CPP) pilots demonstrate that consumers can and will adjust electricity usage in response to price signals.

Some uncertainties are introduced by time-based tariffs that are related to customer response to tariff signals, and this uncertainty needs to be factored in the tariff design (De Sá Ferreira et al., 2013). Furthermore, De Sá Ferreira et al. (2013) associated the uncertainties of customer response to price elasticities.

Customer response to tariff signals is influenced by various factors that are based on how the energy is used and affordability. A study of the California market by Reiss and White (2002) indicated that approximately 44% of households did not respond to variations in prices. Another finding concerning the Californian market indicated that households that used electric heaters were sensitive to prices and that higher-income households were less sensitive to price variations. Customers that are concerned with environmental values were found to consume relatively low electricity than other customers who are not concerned about the environment (Thorsnes et al., 2012). Inglesi-Lotz and Blignaut (2011) conducted a study that examined the response to electricity price fluctuations in the South African economic sector. They found that the price elasticity in the industrial sector was highly significant and negative, while the rest of the sectors presented insignificant price elasticities.

Although some customers respond to tariff signals, they do not necessarily respond to fluctuations. Studies and models on the price elasticity of demand have been suggested to provide knowledge regarding customer responses to signals. To this effect, De Sá Ferreira et al. (2013) suggested that customer response data should be used to create price-elasticity models that can be incorporated in tariff design models.

2.5. Classification models

In distribution systems, the classification of loads is useful due to both the nature of the distribution networks and a large number of customers (Han et al., 2012, Buys & Gaunt, 2020). While most loads can be uniquely identified and their demand also quantified, the practical way for this identification in load modelling is categorising or classifying them, determining the average class demand and assigning load profiles to the class (Chicco et al., 2006; Tsekouras et al., 2008; Fidalgoa et al., 2012; Ferraro et al., 2016). In most utilities, loads are generally classified based on economic activity. In essence, customer classes can be defined by not only end-user type or economic activity, but the usage size (kW) and voltage (one phase vs. three-phase) (Rauch, 2014).

The classification of loads or feeders is an essential goal of load modelling. However, it requires the definition of the sets of parameters derived from daily load profiles from the utility database (Chicco et al, 2001). The goal of classification influences both the parameters to be selected and the clustering techniques to be used. Some of the goals of classification may be summarised as follows:

- Classification of feeders from a feeder population
- Classifications to determine typical days or typical weeks
- Classifying of customers based on usage behaviour to create customer classes
- Decomposing of composite feeder profiles into elementary profiles

The parameters used in each of these goals are reviewed in the following sections to determine whether they can be used to achieve the objective of the technical and tariff models.

2.5.1. Classification of feeders from a feeder population

There is limited literature on the approaches to differentiate MV feeders. However, Broderick et al. (2014) used clustering methods to uniquely identify classes or types of feeders. They characterised feeders in a way that distinguishes individual feeders from the rest of the feeder population. They used the following variables, namely, nominal voltage level, feeder length, main conductor type, three-phase vs. single-phase feeder length, voltage regulation schemes (load tap changes, feeder regulators, switched capacitor banks), load mix (residential, commercial, industrial), load shape (peak, minimum load, seasonality), existing DG and PV deployment levels (kW), operational practices of the utility and system protection devices. The final parameters were selected after they plotted a correlation map of all these parameters and for those that were correlated, one of them was removed. The k-means clustering algorithm was used and the model referred to as the Cubic Clustering Criterion was used to validate the cluster as well as determine the optimum number of clusters. One of the challenges with multiple parameters when using clustering algorithms is to evaluate which of the parameters are significant. Broderick et al. (2014) found that the voltages and feeder length were the most significant parameters in distinguishing between clusters of feeders and suggested that this finding matches the design criteria for evaluating the feeders.

Broderick et al. (2014) validated their models by selecting six feeders from the data set as validation feeders to assess the results against their data. The choice of the validation feeders was based on the size of the PV, that is, the chosen feeder was to have large utility-scale PV systems.

2.5.2. Classification of customers based on their usage behaviour

Classification of customers means that customers are grouped according to the common parameters or similar load profiles. According to Chicco et al. (2003), customer classification should be related to load profiles. The load profile for the future is not a simple copy of the current, existing load profile; instead, the load profile is slightly modified from day to day and from week to week to reflect changes in consumers' usage behaviour (Bobric et al., 2009; Frost et al., 2017). Classification of customers also requires that certain parameters that are related to contracts and the national policies, for example, economic activity, day of the week, seasons and time-related factors. The factors are used to supervise algorithms that perform classification.

2.5.3. Determining representative days

Determining typical days or grouping similar days is a form of classification where the goal is not for customer classification, but the goal is to study load patterns of different days. Classification methods are a good substitute to the subjective approaches of simply assuming that the consumptions differ based on certain days of the week or weekends. However, Ferraro et al. (2016) did find that consumption behaviours are different for the different groups of days. They also observed that merely grouping based on the consumption levels of these days might not give the true electrical behaviour on these days, hence the need to use clustering algorithms that consider load patterns. In their study, they go on to classify data into three classes, namely, weekdays, holidays and pre-holidays. They stated that they have compared their results with calendar-based observations and reported that the results indicated the accuracy of their clustering. Frost et al. (2017), ElNozahy et al. (2013) and Green et al. (2014) used the k-means clustering algorithm to create typical day profiles by grouping similar days of a year together.

ElNozahy et al. (2013) created typical load profiles by clustering annual hourly profiles, with similar profiles being grouped in the same cluster.

2.5.4. Decompose composite feeder profiles into elementary profiles

Gerossier et al. (2017) proposed a method to decompose aggregated load profiles to their elementary components. They assumed that the demands aggregate different shares of the elementary profiles associated with different customer categories. Gerossier et al. (2017) proposed a method called the augmented Lagrangian method that relied on minimising prediction errors to determine the elementary profiles. Their method requires several feeder

demand curves and a description of customers as input. A clustering technique that involves optimisation is then used to assign each profile to a cluster.

2.6. Parameters for technical and tariff models

The important parameters or load attributes considered for technical and tariff models have been reviewed in the preceding sections and they include typical days, economic activity, consumption patterns (time of consumption), and load factors. It has been established that weather, demographics and economic activity are the major drivers behind the demand patterns. In the South African context, weather conditions are closely linked to seasonality. The defined seasons in load modelling for South Africa are winter and summer. Winter tends to be colder and summer is warmer. Economic factors are viewed in terms of the activity linked to usage, that is, when the economy is favourable, activities in many sectors increase and thus increase in energy usage. Classification of customers is considered central to load modelling. Classification in most utility databases is based on economic activities. Examples of such economic activities include mining, agricultural, commercial and industrial activities amongst others. It can also be established that short term planning models, energy losses models and reliability models require not only the demand or contracted power-related information, but the load patterns, to determine peaks and valleys in the consumption, whereas in long term planning an average demand over a longer period and contracted power can be sufficient. This is also the case with tariff models. For short term planning, tariff models require details about the consumption patterns such as peak and off-peak consumption periods, which are not required for long term planning. The parameters from customer load measurements data can thus be summarised as follows:

2.6.1. Contracted power (Maximum demand)

Demand is one of the parameters that is commonly used. Fidalgoa et al. (2012) proposed a methodology based on contracted power or notified maximum demand (NMD) and the total energy consumed as indices to create load profiles for LV customers. The model developed by Gaunt et al. (2011) also used After Diversity maximum Demands (ADMDs), which are the contracted demands. Demand in the form of annual peak loads has been identified as a key parameter for long term planning and load forecasting (Black et al., 2018, Soni, 2018).

Demand has also been used in various tariff models. To obtain a network cost causality function, it is necessary both to determine how the networks are planned and to understand what causes the costs to be incurred (Ortega et al., 2008). The maximum demand can be calculated from the current capability of the conductor and vice versa.

Various clustering-based classification models also use contracted power as the parameter for distinguishing between different classes as seen in the preceding sections (Chicco et al., 2003; Fidalgoa et al., 2012; ElNozahy et al., 2013; Benítez et al., 2014; Broderick et al., 2014; Green et al., 2014).

2.6.2. Energy consumption

Energy consumption as a parameter can be viewed from two perspectives, namely average consumption and consumption patterns. Average consumption is often mentioned in static models whereas consumption patterns have been explored extensively in time-varying models as well as probabilistic models for both technical and tariff studies. The customer usage behaviours are associated with load profiles, which often represent consumption at different times or intervals. Load profiles have been used in the classification of loads. The drawback of using load profiles for classification is that the class composition often includes loads with different average or maximum energy consumption levels. Load duration curves provide a clearer view of load profile for active and/or reactive power demand of the load during a specified period for different values (Milanovic et al., 2014).

2.6.3. Seasonal demand/consumption

Climate conditions are closely linked to seasonality, and thus changes in usage behaviours of energy by customers. Certain economic activities also increase/decrease depending on the season and this has an impact on the energy

usage patterns. Seasonality has been identified as another attribute of note in load profile variations (Bobric et al., 2009; Broderick et al., 2014). Bobric et al. (2009) proposed a method to analyse the trend cycles and seasonality, based on time series decomposition.

2.6.4. Time of use (TOU) interval consumption

Time factors are some of the important characteristics used to explain load variation in both long term and short term operational planning models, as well as tariff models. In long-term planning annual peak usage has been used (Riaz et al., 2017). Peak and minimum usage characteristics were used by Broderick et al. (2014) classification of loads. The consumption patterns of customers vary with time and largely inform TOU tariffs. The cost of producing energy to meet the demand is often the driver of these costs. TOU tariffs can drive operational efficiency and defer network investments (Impact assessment report 2016). There is extensive literature covering the use of time-based usages for various load modelling, and our aim is not to exhaust them but to highlight that the TOU intervals are key in load modelling.

2.6.5. Load factors

Numerous publications have suggested the use of load factors as the basis of load modelling (EPP, 2008; Milanovic et al., 2014; Sharma and Singh, 2014). Load factors have been identified and considered essential in several tariff models (Sharma and Singh, 2014). According to Sharma and Singh (2014), a load factor is used to determine the difference between average demand and maximum demand and thus a measure of uniformity or variance in energy usage. A good load factor is desired as it is often an indication of a constant consumption rate.

2.6.6. Classes

Customer load classes and typical day classes can be viewed as parameters when modelling loads, because of the different consumptions associated with the different classes. When the classes formed and load profiles allocated to them are representative of all loads, the load profiles can be used in extracting the characteristics parameter for use in technical and tariff studies, instead of such parameters extracted from individual loads. Typical days, weeks or months, and customer classes are often part of the technical and tariff analysis (Figueiredo et al., 2005; Ferraro et al., 2016; Frost et al., 2017). A customer class is another important attribute to consider for the cost of supply studies and has been identified and used for various studies (Figueiredo et al., 2003; Firestone et al. 2006; EPP, 2008; Heunis and Dekenah, 2014; Piscitelli et al., 2019).

2.6.7. Parameter extraction from measurements

According to Milanovic et al. (2014), load identification models can be classified into the following two categories, namely, component-based, where individual electrical components are aggregated to form a single representative load model and the measurement-based approach, which relies on specific measurements from feeders to form load models. It is necessary to investigate how the parameter for technical and tariff analysis models as discussed above can be identified and extracted from the feeder measurement data.

2.6.8. Component-based models

The component-based model relies on prior knowledge of the composition of loads and corresponding device-specific models of its main components (Liang et al., 1998; Gbadamosi, 2017). The component-based approach, which is also referred to as the bottom-up approach, is derived from prior knowledge of the individual load components to build an aggregate load model. The loads are often divided into sectors such as residential, industrial etc. (Gbadamosi, 2017).

A component-based modelling approach was used in the modelling and validation of electrical load profiling in residential buildings (Chuan and Ukil, 2014). Chuan and Ukil (2014) carried out the load profiling of residential

buildings in Singapore to classify the types of buildings and to build the profiles of the building bottom-up from the elementary load components. They suggested that this approach is suitable for smart grid-related applications as it provides information about the usage of the individual components.

The component-based approach remains relevant in load models as it provides information about elementary loads that may be difficult to obtain from aggregated feeder measurements.

2.6.9. Measurement-based models

The measurement-based approach to modelling relies on specific measurements, which are collected at a point in the system representing load consumption (Liang et al., 1998; Maitra et al., 2008; Gbadamosi, 2017). According to Shi and Renmu (2003), the measurement-based load modelling approach has the advantages of directly monitoring the true dynamic load responses and, in addition, it is easy to update the parameters for this modelling approach when the load characteristics change. They further mention that a disadvantage of the approach is the fact that it cannot use load models with large parameter dimensions.

An important conclusion by Milanovic et al. (2014) is that a hybrid of the component-based models and the measurement-based models be used.

The implementation of advanced metering infrastructure (AMI) and improved database systems make it possible for utilities to obtain detailed information regarding the electricity consumption of a large number of customers. However, this benefit for the utilities comes together with a challenge related to the efficient and effective mining of this data (Frost et al., 2017). Nevertheless, it is possible to deal effectively with the challenges mentioned by Shi and Renmu (2003) about large parameter dimensions, and that of Frost et al. (2017) relating to data mining, through the use of dimension reduction or compression techniques, data mining and machine learning algorithms (ElNozahy et al., 2013; Ferraro et al., 2016).

2.7. Load parameters from customer load measurements

The instant power consumption for each consumer are indexed with the activity code and the notified maximum demand or contracted power and this information has been obtained from billing data and used in load modelling to classify and profile customers by various researchers including (Chicco et al., 2003; Fidalgoa et al., 2012; ElNozahy et al., 2013; Benítez et al., 2014.). Activity code, voltages and the contracted power are often specified a-priori as they are contractual (Chicco et al., 2003).

Load profiles, derived from customer load measurements are generally used by researchers to derive characteristic indices or parameters for various studies for load modelling for classification. In classification, Weekdays have been derived by aggregating measurement data into different days of the week, that is, averaging the load diagrams related to the weekday for the whole week or specific weekdays (Chicco et al., 2001, Chicco et al., 2003). However, Ries et al. (2016) represented weekdays by using Wednesday load profiles and non-weekdays represented by Saturday's load profiles. The authors stated that the decision for this is based on the impact that Malta's climate has on energy usage. While it is possible that there could be only 2 possible typical days or groups of days, it is unlikely that selecting a single day's load measurements to represent a group can yield accurate results. Aggregating similar days profiles or days that have common attributes and using their average profiles as the representative of either weekdays and non-weekdays seems to be preferable.

Green et al. (2014) reported that their most accurate simulations from a typical day study they conducted were obtained when they took the variation in average hourly demand that had the same sign as the cluster average estimator. Green et al. (2014) also observed that clustered data presents researchers with a trade-off between the accuracy of results against the ability to conduct Monte Carlo studies.

The consumption patterns related to customer measurements have been studied extensively in load modelling. Chicco et al. (2003) defined a set of indices to represent the usage behaviour of customers. They distinguished

between a priori indices (i.e. contractual and historical data stored in the utility's database) and field indices (extracted from measurement campaigns) which can be extracted from the customer load profiles. Chicco et al. (2001) designed daily load curve indices that included what they termed the non-uniformity coefficient (ratio of minimum to maximum power), the fill-up coefficient (ratio of average over maximum power), the modulation coefficient at peak hours (ratio of the average of peaks and average of the day) and the modulation coefficient at off-peak hours (ratio of the average of off-peaks and average of the day). They found that the use of the shape indicators resulted in a reduction in the resulting number of clusters and reported that this is because these indices level the differences in between consumption patterns. They also found that the contractual parameters are poorly connected to the load patterns and recommended further work to produce global shape indices with which to capture the customers' consumption behaviours.

Similarly, Qiu et al. (2016) and Chicco et al. (2005) defined parameters from the peak and valley periods and determined these indices to be average load rates during peak1, average load rates during peak2, the ratio of total load during peak1 to peak2, the ratio of total load during peak to valley and ratio of average load during peak to valley. They found that applying these parameters to the profile clusters results in an overlap between the indices. In addition, they also discovered that some clusters make it difficult to categorise the customers based on electricity usage patterns because profiles are usually very similar to each other. They proposed an improvement by identifying and removing the edge points in the parameters as a way to solve the problem.

Parameters linked to time-related usage behaviour have also been extracted from daily load profiles through the segmentation of the load profiles into several time segments, which are then used in clustering algorithms to classify customers. Such approaches include dividing daily profiles into segments based on the input parameter and window size and assigning a symbol to each interval (Benítez et al., 2014; Lavin and Klabjan, 2015; Fonseca et al., 2017), selecting the morning slope together with the principal component analysis (PCA) of a typical daily profile (Ferraro, Crisostomi, Tucci and Raugi, 2016), as well as the use of the load factor and cluster loss factor on different days and in particular zones (Sharma and Singh, 2014). The load information used is based on customers' electric energy usage and time of use details, seasonality and demand share, as well as categorisation/classification information (Elkarmi, 2008)

The time-series decomposition models are useful in providing information about the parameters related to trend, seasonality and cycles that may be present in temporal data. The model is used to determine whether there are any cycles or seasonality that should be considered ensuring that seasonal variations are not neglected, particularly when typical days are determined. The load curve represents the power variation in terms of the determinant parameter. If the parameter taken into consideration is the time (t), the curve can be divided into several components that induce the load profile (Bobric et al., 2009):

- T is the trend of the main load variation.
- C is the cyclic component indicating the slow time varying events that tend to reoccur after some defined period. These could be because of economic or political events.
- S is the seasonal component, which represents seasonal fluctuations that usually last for a short period and are the same for every year.
- ε is the random or error component, which results from the randomness inherent in data.

The decomposition model is expressed using the following equation

$$\boldsymbol{Y_t = f(S_t, T, E)} \qquad \textbf{Equation 1}$$

Where Y_t is the represented time-series, S_t is the seasonal component, T is the trend, and E_t is the error component as explained above. The decomposition method determines the span of the trend cycle required.

2.8. Clustering in load modelling

Clustering in load modelling has primarily been useful in the classification of customers based on some defined or computed similarities. The literature distinguishes between clustering and classification. Classification is defined as a supervised learning problem of assigning an object to one of several predefined categories based on the attributes of the object, while clustering refers to a problem of grouping objects based on either distance or similarity (Pandeeswari and Rajeswari, 2015). Clustering frameworks are categorised under classification algorithms in machine learning. Clustering may be performed using supervised, semi-supervised and unsupervised learning methods (Simeone, 2018).

Clustering is defined as objects or observations that are homogenous within the group (Mooi and Sarstedt, 2011). A cluster is essentially a unimodal component within an appropriate, finite mixture model with internal cohesion of within-cluster objects and external isolation of other cluster objects (McNicholas, 2016). Clustering is further defined as an unsupervised data mining technique where similar data are placed into homogeneous groups without prior knowledge of either the subgroups or other information about their composition (Aghabozorgi et al., 2015; Rai and Singh, 2010). From these definitions, it is clear that the goal of classification is to use clustering algorithms to create clusters with internal cohesion of group members and external isolation of the subgroups. Accordingly, the measure of this internal cohesion of cluster members is critical for the success of the technique selected.

Based on the definition of clustering and the frameworks as discussed above, it is necessary to establish if clustering can be used to identify and extract load models parameters of MV loads. This can be done by evaluating some of the commonly used clustering techniques.

Literature mentions several clustering techniques that have been used. These techniques are generally classified as either partitioning or non-partitioning methods (Mooi and Sarstedt, 2011). However, Fraley and Raftery (1998) reported the following three techniques for clustering data, namely, hierarchical methods, relocation methods and model-based methods. Sarda-Espinosa (2017) referred to five techniques, namely, partitioning (or partitioned), hierarchical, density-based, grid-based and model-based methods.

Model-based clustering can also be referred to as probabilistic clustering or a mixture of Gaussian clustering methods. Accordingly, McNicholas (2016) defines model-based clustering as referring to the use of (finite) mixture models to perform clustering. In model-based clustering, each cluster can be mathematically represented by a parametric distribution, such as Gaussian or Poisson (ElNozahy et al., 2013).

Hierarchical methods, partitioning methods and two-step clustering (hybrids) are some of the popular methods of clustering identified by Mooi and Sarstedt (2011). Some of the popular methods used in the classification of customers include the non-partitioning method, known as hierarchical clustering, and a partitioning method known as the k-means clustering.

Other methods which are beginning to gain momentum include the machine learning methods used in artificial intelligence (neural networks, fuzzy systems) and frequency domain approaches (harmonic analysis and wavelets) as reviewed by Chicco et al. (2002). Kohonen self-organising maps have also been deemed an appropriate artificial neural network method, which may be used for classification (Chicco et al., 2002; Verdú et al., 2004).

2.8.1. Non partitioning methods: Hierarchical clustering techniques

Hierarchical clustering techniques are characterised by the nested or tree-like structure, which is referred to as a dendograph, established in the course of the analysis.

According to Broderick et al.(2014), Omran (2010), Jain and Dubes (1988), ElNozahy et al. (2013) and Mooi and Sarstedt (2011), there are two categories of hierarchical clustering namely, agglomerative or divisive clustering. Furthermore, agglomerative hierarchical clustering begins with the creation of a cluster for each element and the similar clusters, are then merged to form a single cluster. This process is repeated until the desired number of clusters have been formed. On the other hand, divisive hierarchical clustering begins with all the elements in a single cluster. The cluster is then divided into sub-clusters based on either the isolation criteria or the maximum

distance between the member objects. Although the analysis of hierarchical clustering (the popular agglomerative method) is straightforward, it is considered computationally expensive.

Broderick et al. (2014) maintain that the two main benefits of hierarchical clustering are flexibility in selecting the number of clusters and the visualisation of the clusters, and the distance between the clusters through a resulting dendrogram. However, the disadvantages include the fact that the choice of cluster variables is often randomly determined. In addition, it is time-consuming and is often decided upon arbitrarily (Broderick et al., 2014; Sarda-Espinosa, 2017).

2.8.2. Partitioning or relocation methods

Partitioning clustering algorithms allocate objects or elements into predetermine clusters iteratively based on the Euclidean distance measures. An average Euclidean distance is calculated between each element and each of the centres of the clusters and the element is allocated based on the shortest distance (Hartigan and Wong, 1978; Fraley and Raftery, 1998; ElNozahy et al., 2013; Al-Wakeel and Wu, 2016). In analysing load profiles to create load classes, load profiles are grouped into various pre-determined clusters such that a cluster contains similar load profiles, and these load profiles are different to those that are in adjacent clusters (Al-Wakeel and Wu, 2016; Al-Wakeel et al., 2017).

The simplest and widely used partitioning technique is k-means clustering. K-means clustering has been reported to be successful in grouping load profiles together by numerous researchers (Buys & Gaunt, 2020). The k-means clustering procedure applied to the pairwise vectors of parameters is described below (Chicco et al., 2003; Figueiredo et al., 2005; Tsekouras et al., 2008; Panapakidis et al., 2012; Buys & Gaunt 2020):

- Let M be a set of consumers (consumer: m = 1,..., M) to be classified and their corresponding load profiles be denoted as $P_h^{(m)}$ and let h = 1 ,…H be the time domain of the load.
- The population data, is denoted as X = $\{x^{(m)}$, m = 1,...,M$\}$ is used to obtain a vector (m) x
- The clustering procedure allocate the customers into K clusters $C^{(k)} \in X$ for k = 1,...,K and $1 \leq K \leq M$.

The above-mentioned researchers acknowledge that the challenge with the k-means clustering algorithm is that the number of clusters has to be pre-determined or estimated before applying the k-means algorithm. Various techniques and measures were used to estimate the optimal number of clusters and these will be discussed in the Cluster adequacy measures' section. However, this challenge can be overcome by performing multiple iterations of clustering, selecting different numbers of clusters and testing their compactness. The classical elbow method, the silhouettes, and DBI are used for this purpose as these methods were shown to be successful in the literature review.

Other clustering techniques are variations of the above group of clustering methods with enhancements. The fuzzy k-clustering method has emerged as a very similar but improved version of the k means clustering wherein the centroids are randomly selected and its results provide the degree to which the members belong to a cluster (Ferraro et al., 2016).

2.8.3. Time-series (TS) clustering models

Time-series clustering is a type of dataset with a sequence of real values and differs from the temporal sequence that comprises a sequence of symbols from a particular alphabet (Aghabozorgi et al., 2015). Time-series decomposition methods have also been used to classify customers. Two aspects identified by Rani and Sikka, (2012) for managing time-series data include methods to determine representation and similarities. Some of the approaches include those that are aimed at dividing the load profiles into sections to estimate time-based

parameters linked to daily profiles (Benítez et al., 2014; Lavin and Klabjan, 2015; Fonseca et al., 2017) and time series decomposition methods discussed in 2.6.3.

It has been established in the preceding sections 2.1 through 2.6 that multiple parameters need to be considered for load models that seek to track the behaviour of customers. This implies that grouping data or customers based on a single parameter may not provide an accurate result for simulation studies. Ferraro et al. (2016) suggested that merely grouping data based on the consumption of the days might not give the true electrical behaviour on these days, hence the need to use clustering methods. Where data is to be grouped based on similarities, clustering appears to have been preferred, as can be deduced from sections 2.4, 2.5 and 2.6.

2.9. Cluster validation measures or indices

The quality of the cluster formation must be assessed to validate both the clusters and the clustering algorithm used. The quality of clusters can be assessed using cluster validity measures. These measures are based on the extent of cohesion or separation. The cluster validation application that uses the same data that was used for the clustering and is known as internal evaluation. If there are separate or different data, other than that used in the formation of the clusters used for the evaluation, this is referred to as external evaluation. Most of the indices estimate and combine the cluster cohesion and the cluster separation and use ratios to determine a cluster quality (Filchenkov et al., 2016).

Filchenkov et al. (2016) reviewed nineteen of the popular cluster adequacy measures and found that there is no universal clustering validity measure and that the cluster validity indices should be chosen for specific problems. They suggested a meta-learning approach to solving the problem of choosing the appropriate cluster adequacy measure. The cluster adequacy measures that are commonly used for vector or profiles cluster adequacy based on internal evaluation were identified by Figueiredo et al. (2005), Figueiredo et al. (2003), Tsekouras et al. (2008) and Chicco et al. (2004). They include the mean square error (SSE_k) that is based on the Euclidean distance of each load profile from its cluster representative profile (, Panapakidis et al., 2012).

Tsekouras et al. (2008), proposed the use of a similarity matrix indicator (SMI). Granell et al (2015) proposed the ratio of the within-cluster sum of squares to the cluster variation, the variance ratio criterion or Calinski-Harabasz Index based on a ratio between the intra-cluster and inter-cluster factors, and the scatter index which compares the distance of data points and centre with the mean of the population. ElNozahy et al. (2013) further calculated the probability of occurrence of each daily demand profile and concluded that the frequency of occurrence of a cluster variable is dependent on the cluster segments relative to the total number of population segments.

The silhouette statistics, the elbow method and the DBI, described below, are the methods that are useful in evaluating clusters based on performance scores (Filchenkov et al., 2016). . However, the literature shows that there is no consensus on global indices that can be applied universally and thus further work in this area is recommended (Figueiredo et al., 2003, Panapakidis et al. 2012, Qiu et al., 2016).

2.9.1. The elbow method

The elbow method looks at the total within-cluster sum of square errors as a function of the number of clusters: One should choose several clusters in such a way that adding another cluster does not improve the total. The number of clusters can be determined as follows:

1. Compute clustering for different values of k, using a clustering algorithm.
2. For each k, calculate the total within-cluster sum of square errors.

$$SSE_k = \sum_{r=1}^{k} D_r$$ **Equation 2**

$$\mathbf{D_r} = \sum_{\mathbf{i}}^{(\mathbf{n_r}-\mathbf{1})} \sum_{\mathbf{j}}^{(\mathbf{n})} (\mathbf{x_i} - \mathbf{x_j})^2$$ **Equation 3**

3. Plot the curve of SSE_k according to the number of clusters k.
4. The location of a bend in the plot is the elbow point and it is an indicator for an optimal number of clusters.

2.9.2. Silhouette statistic

The silhouette is a further cluster validity measure that is commonly used together with k-means clustering. Rousseeuw (1987) introduced the silhouette graphical display method to measure how well an object has been allocated to clusters. Resulting coefficients that equal or are closer to one (1) indicate that the sample is dissimilar to the sample in neighbouring clusters and are desired. Results that are near zero (0) indicate weaker clusters. Negative values indicate that sample objects that have been assigned has not been clustered correctly and there is a possibility of misallocation of sample to clusters. Both the partitions and resulting values are used for constructing the silhouette plot. The silhouette is reported to be effective with k-means clustering.

According to Rousseeuw (1987) the silhouette statistic s(k), measures how well all the objects have been allocated to clusters. The s(k) is calculated as

$$\mathbf{s(k)} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{s_i}$$ **Equation 4**

and where $(-1 \leq s_i \leq 1)$

2.9.3. Davis Bouldin index

Another popular validity measure is the Davies–Bouldin indicator (DBI) which represents the average of the similarity measures of each cluster with the clusters that are significantly similar to it (Davies and Bouldin, 1979). Lower values of the indicator are desirable as they correspond to better clusters (Davis and Bouldin, 1979). Davies and Bouldin (1979) developed, used and tested the DBI with the k means algorithm being used to determine the K cluster. Accordingly, the DBI is suited for use with k-means clustering.

To determine the DBI there is a need to compute the distance measures. The similarity measure is calculated using Equation 5 and Equation 6 (Davis and Bouldin, 1979).

$$\mathbf{R_{ij}} = \frac{\mathbf{X_i} + \mathbf{X_j}}{\mathbf{D_{ij}}}$$ **Equation 5**

Where:

D_{ij} is the distance between the vectors

X_i and X_j are the dispersions of the clusters i and j.

Therefore, the DBI measure of interest is calculated as the average:

$$\mathbf{DBI_i} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{DBI_{ij}} \qquad \textbf{Equation 6}$$

A lower value of DBI_i, is desirable as it indicates the compactness of the clusters.

2.9.4. Statistical summaries

Statistical measures have also been used for evaluating the clusters. In particular, the mean and the standard deviation are used to evaluate clusters (Bobric et al., 2009). Bobric et al. (2009) used these statistical measures to evaluate the membership of clusters, and concluded that these measures were adequate for this purpose. The formulas for the statistics that were used as follows:

Mean: Measure of central tendency,

$$\bar{\mathbf{x}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i} \qquad \textbf{Equation 7}$$

Measure of dispersion of observations from the mean:

Variance,

$$\mathbf{var(x)} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\bar{\mathbf{x}} - \mathbf{x_i})^2 \qquad \textbf{Equation 8}$$

Standard deviation,

$$\boldsymbol{\delta} = \sqrt{\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\bar{\mathbf{x}} - \mathbf{x_i})^2} \qquad \textbf{Equation 9}$$

2.10. PCA based clustering

One of the challenges in relation to models of parameter based clustering noted in the literature relates to data dimensionality and data exploration. Visualisation had been a necessary step in data analysis, because it made is possible to explore and make sense of the dataset (Engel et al., 2011). Dimension reduction represents a solution

to the problems associated with high dimensionality when dealing with vectors (Carreira-Perpinan, 1997). According to Carreira-Perpinana (1997), dimension reduction refers to the search for "a low-dimensional manifold that embeds the high-dimensional data". Silipo (2015) found that principal component analysis (PCA) is a better technique, following a viewed of a number of the dimensionality reduction techniques available and accepted in the data analytics landscape at the time of the study. Zhang and Yang (2016) support this assertion and state that the PCA is able to reduce the vector dimensions of data while maintaining the desired variability that distinguishes the load curves. According Zhang and Young (2016), "PCA is a statistical procedure that orthogonally transforms the original n coordinates of a data set into a new set of m coordinates known as principal components". The PCA is also preferred because of its simple and efficient algorithms (Carreira-Perpinan, 1997). The success of the PCA lies in the principal components accounting for 80 to 99% of the variation for each day. The PCA is suited to normalised data as it is sensitive to the variances of the initial variables (Carreira-Perpinan, 1997).

The goal of PCA based clustering is to ascertain which groups of samples share a similar profile defined by the principal components (PC) variables. Although there are various methods, which may be used, for computing PCA, this study focused on the simplest algorithms used in clustering. This method is based on the covariance method and the singular value decomposition (SVD).

Various software programs have functions that are able to compute PCA. The goal of the PCA algorithm is to transform a given data set X^p into an alternative data set, Y^l, of a smaller dimension.
Where variables p are parameters of a dataset ${x^p}_i \in X^p$. This is performed as follows:

1) Arrange the data to create a vector $x_1 : : : x_n$ with each x_i representing a single grouped observation of the p variables.
    - Set $x_1 : : : x_n$ as row vectors, with p columns.
    - Create a matrix, X, of some dimensions n by p from the single matrices of the row of vectors.
2) Calculate the mean of each dimension and create a vector of means of dimension px1, and j=1 ..p entries.
3) Calculate the distances from the mean
    - Calculate the difference between the mean vector (Xu) and the original matrix (X).
    - Store the results in the matrix A.
4) Calculate the covariance matrix
5) Determine the eigenvectors and eigenvalues
6) Rearrange the eigenvectors and eigenvalues: Sort the columns of the eigenvector matrix and the eigenvalue matrix in order of decreasing eigenvalues.

The final principal components (PCs) assigned to each of the load profiles are used with a clustering algorithm to partition data into different clusters. There are built-in PCA modules in Python (scikit-learn) software tools that have been developed to automate the algorithm described above. The input required for the PCA module is data in a frame format, and the required accuracy or desired number of principal components, depending on the PCA function chosen.

2.11. Load model validation

When a load model is developed, it is necessary to determine whether it is valid and that the model results can be trusted before the model is implemented. Another key aspect of model validation is the generalisation of the proposed model. Models validation in the context of measurements based load modelling requires that various aspects of the model be validated, either in parts or completely where possible. Various researchers have used case studies as validation of their model frameworks. The first important validation happened on the input data that will be used. A joint multivariate evaluation of all parameters is proposed as a richer methodology than traditional and simple univariate analysis methods.

Where primary sources or documents exist to provide a reference, these sources are useful for models validation. Ferraro et al (2016) used the calendar days to validate their results, thereby recording the observation of

consumption on these calendar days and comparing them with their modelling results. They reported that the results indicated the accuracy of their clustering.

Customer classification models are often evaluated against customer profiles that are stored in databases. The class profiles are compared with the known average customer profiles of the related class, and when their error is small, the results are accepted. Often statistical methods of determining the error such as standard deviations, mean square errors, least-square errors and others are employed in this validation process. Customer classification models that are based on clustering algorithms use the cluster validity indices for validating the results. The class profiles are usually derived from the average of the load profiles within the cluster or in some instances the median profile, the most prevalent profile or the median profile. When models are based on measurements data, it is also common to see the data being divided into training data and experimental or test data. In such approaches, the models are fine-tuned using training data. The models are then tested against the experimental data to see if the results of the model are valid and the results acceptable. Piscitelli et al. (2019)) used a training dataset used with the decision trees in their automatic classification model to allocate the new load profiles to existing load classes.

Probabilistic models have been validated through a comparison of the results accuracy of the PLF model against the Monte Carlo simulation results. Monte Carlo simulations can be complex and often take longer time to complete the simulation run and provide the results because it is iterative (Green et al, 2014; Chihota et al, 2018).

2.12. Concluding remarks

The literature reviewed highlighted several salient points pertinent to load models for technical and tariff studies:

- Load models have been developed to
    - Identify customers or customer groups, classify them and estimate the profiles for the group.
    - Estimating the demand and supply for network planning.
    - Identify cost drivers of energy usage by different customers.
    - derive load profile distributions for probabilistic for power flow simulations, and
    - Identify feeders from a feeder population and classify them.
- Static single state models are generally used for long term planning while multi-state models are suitable for short term planning, in both technical and economical, and tariff planning.
- Load profiles reflect the energy usage behaviour of customers over a period and comprise a range of parameters, which may be identified and quantified.
- The load profile is affected by weather conditions, economic factors and demographic factors.
- Tariffs can be used to drive a certain behaviour pattern on the part of the customers and, thus, change the load profile.
- Load models can be seen as a representation of load variations through identifiable characteristics. There is a growing interest in the development of load models given the changes in usage patterns, the introduction of new behind-the-meter technologies, and the emergence of distributed generation.
- Central to the development of load models is the classification of customers (or loads). Clustering algorithms have been used and have been reported to perform successfully in classification when based on defined parameters. The k-means clustering algorithm is widely used due to its simplicity and popularity in the clustering space.
- There is no global consensus concerning the common parameters that should be used in classifying customers or loads. The choice and derivation of parameters are often informed by the goal and the context of modelling.
- Cluster adequacy measures may be used for determining the optimal number of clusters and validating the adequacy of clusters, where the k-means clustering procedure is used.

Clustering adequacy measures do not always provide an adequate, accurate measure of the performance of clustering and, consequently, more than one measure is often used. Statistical analysis techniques have also been used in estimating parameters and validating clusters/class formations.

3. DERIVATION OF LOAD PARAMETER MODELS

This chapter aims to describe the concepts and the methodology for deriving the load parameter models for technical, financial and tariff analysis. The chapter also covers identified key principles and theoretical aspects, which emerged from the literature review that are essential in the construction of load parameter models.

3.1. Introduction

Loads models represent customer loads for analysis purposes given a large number of loads in a distribution system. Based on the literature review, load models for technical and financial analysis, as well tariff analysis, should comprise a collection of methods, techniques, algorithms and formulae to:

- Quantify specific parameters from load measurements data.
- Group customer load profiles into classes that represent the study periods, create or analyse typical day consumption, and determine and assign the profiles for each of the different typical day classes.
- Group customers into different classes based on specific parameters and estimate the load profiles for each of these classes.
- Allocate new customers or loads to the related classes based on calculated values of the parameters.

To begin developing a load model, the principles that guide its derivation and use should be laid out. In addition, the parameters that define the load models should be identified and justified in terms of the load model objectives.

3.2. Load modelling principles

There are important parameters that planners and tariff designers need to take into account in their practices, for example, demand levels, the time of consumption, and usage patterns, the load factor and the power factor, which are derived from the feeder data. Data scientists have also identified several other sets of parameters but not all of them are relevant for technical and tariff modelling. In other words, a set of coherent parameters for this purpose will exclude any of the parameters that are not directed towards the application in technical, financial and tariff studies.

Figure 1 depicts the context within which the load models for technical, financial and tariff studies were developed. The diagram demonstrates the convergence of technical and tariff models, that is, the customer or feeder measurements data and other inputs. From left to right, Figure 1 shows the inputs that are required, the load modelling based on these inputs, the application and the objectives that the application is intended to realise.

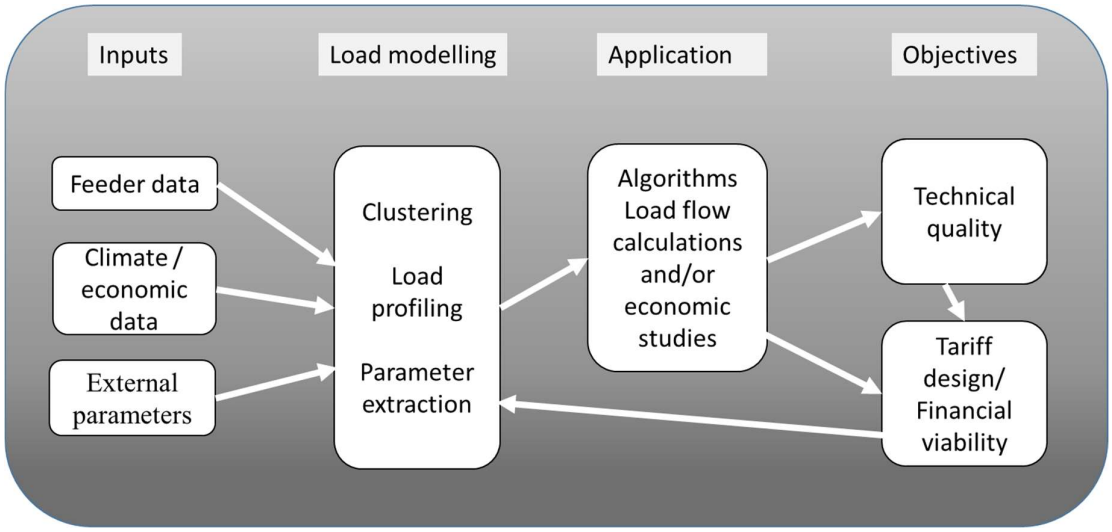


Figure 1: The context of load modelling

The external parameters in Figure 1 refer to the inputs that were mentioned in the literature review as *a priori indices* or *contractual data* (Chicco et al., 2003; Panapakidis et al. 2012). These inputs were of external origin and were estimated using a different data set or contractual rules. The arrows indicate how the inputs may be transformed into load models and how the load models are used in application algorithms for various studies, which in turn provide the desired outputs. Some of the outputs may be required as inputs to other modelling applications, for example, the calculation of the tariffs for energy losses. Load models prepare the data for the various applications, taking into consideration all the variables that are required to provide the outputs. The following principles were proposed for deriving the coherent load models parameters:

a) The load parameter models should be easily identifiable and extractable from the measurement data.
b) The load parameter models should take into account customer demand and the time at which the demand is utilised.
c) The load parameter models should incorporate factors that are related to both weather and economics data that could affect the variability of loads.
d) The load parameter models should enable the utilities to classify customers using simple classification methods.
e) The load parameter models should simplify and reduce the time required to perform technical, financial and tariff application studies without compromising the quality of the solutions.

The study also proposed a modelling process, which is discussed next in Section 3.3. The explicated practice explains feeder characterisation and the theory of classes. Data preparation and statistical analysis would also be discussed.

3.3. Load parameter modelling process

This research is interested in the parameters that are related to both the shape and the levels of the customer load to enable the creation of tariffs that support the optimal usage of the networks. A measurement-based approach was deemed an appropriate approach for this study. The selection of this approach was motivated by the availability of MV feeder measurements. The key steps load modelling process using the measurement-based approach are depicted in Figure 2.

The proposed load modelling process comprises the following steps:

- Data selection and pre-processing. Data is collected from the MV-90 database. Sampling methods are used for the selection of the data while the normalisation technique is used in the pre-processing of the data.
- Data compression (optional step). The PCA dimension reduction techniques was used to create fewer manageable variable. Thereafter the k-means clustering algorithms was applied to group daily profiles into typical days.
- Parameter selection - Identification and selection of parameters from all load profiles.
- Load modelling - Apply a k-means clustering procedure to classify customers and allocate to each class, the representative profiles.
- Testing and validation - The resulting classes and clusters are validated using cluster validity measures and statistical tools.

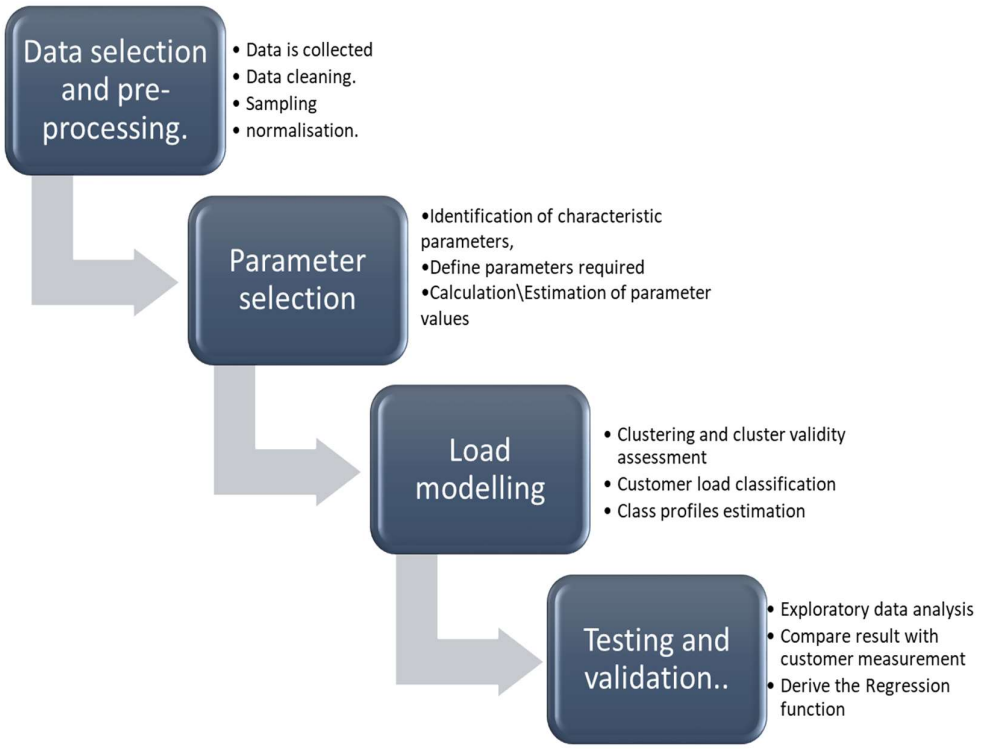


Figure 2: Load modelling process using the measurements-based approach

The detailed load modelling process flow that is proposed to derive the load parameter models is depicted in Figure 3. The proposed load model will lead to the creation of typical days as well as the identification and classification of customers. The typical day results are key for reducing the number of profiles or the volume of input data required for the application studies. The classification model enables the categorisation of customers based on coherent parameters for technical, financial and tariff applications.

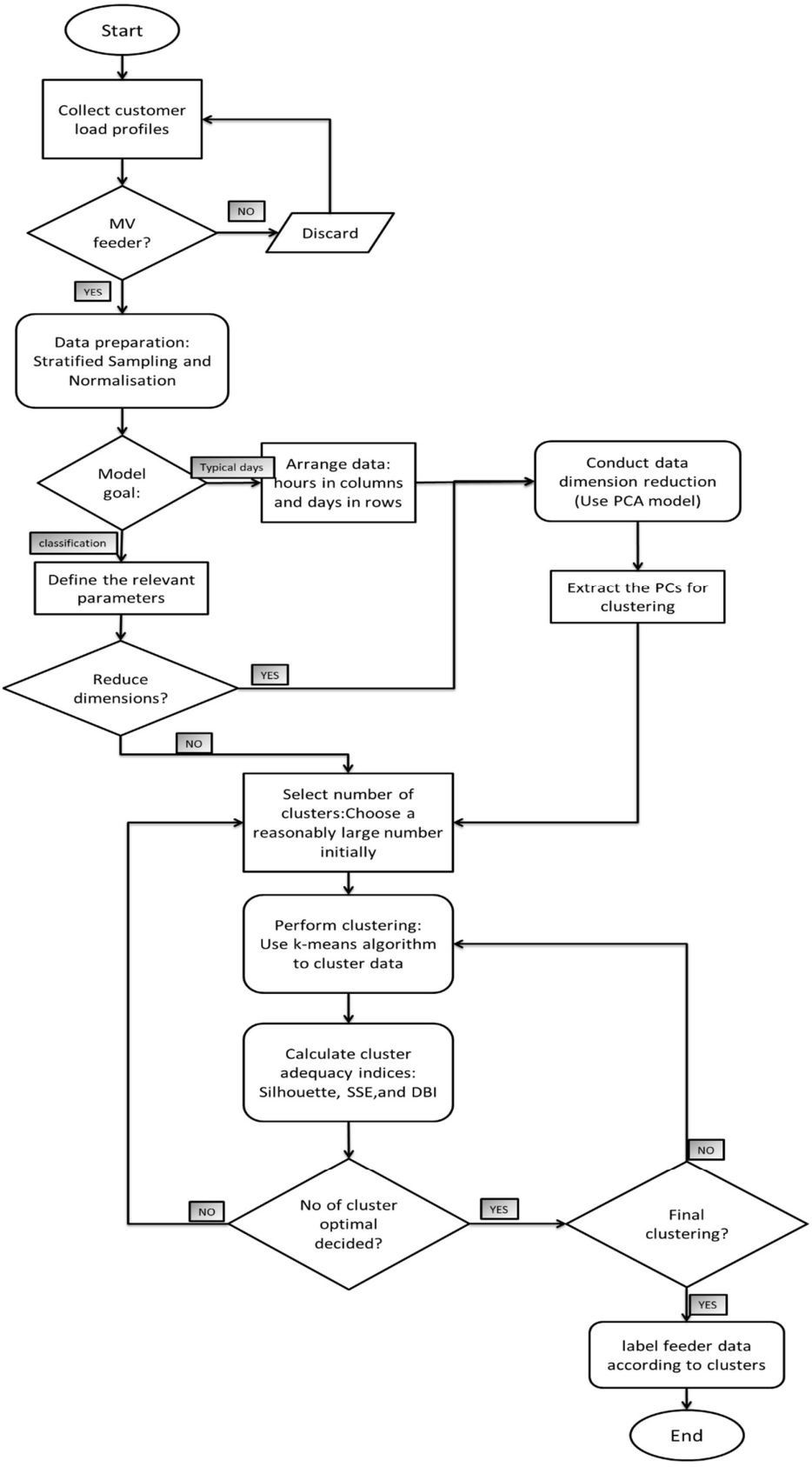


Figure 3: Clustering process flow diagram

The concepts that were found to be relevant to the modelling steps and the theory underpinning the models are discussed in the following sections.

3.4. Feeder characterisation and load parameter model derivation

The literature review suggested numerous parameters for characterising customers. Customer loads have been characterised by the shape of profiles (day, month or year), size (kW/kVA), morning slope of the profile, time-of-use intervals of the daily profiles (Lunchtime, Peak, Standard and Off-Peak), load type, economic activity, "valleys" in the profiles, load factors, seasonality, typical day or week, month, holidays, pre-holidays), and occupancy in buildings. While these parameters could be extracted from measurements data, it was observed from the literature review, that parameters such as economic activity, typical day (week and month), seasonality, holiday (and pre-holiday), and time-of-use needed to be pre-defined.

The assessment of the various attempts to pre-define the load parameters showed that some were subjective while others were based on studies. The proposed approach was intended to identify and study the different parameters, and select those that would enable the creation of the load models for technical, economical and tariff studies. The load profiles of customers may differ based on daily cycles of high and low consumption periods. Similarly, the overall system load profiles may also have their trend cycles. Utilities may use the trend cycle information to define the TOU intervals. The TOU intervals may be fixed for a period and apply to most customers, including those with peak intervals that do not coincide with the system peak. Identifying the parameter that is required for clustering purposes means that the parameter selected should ensure that customers within a cluster demonstrated common behaviours. The advantage of taking into consideration seasonal weather conditions and the time of use interval factors was that it enabled common behaviours that shared common triggers and occurring in a similar context to be identified, thus improving the quality of the clusters. In South Africa, two seasons had been defined for modelling, namely, winter (high demand) and summer (low demand). These had been linked to the weather, which may be considered external because it is not possible to control them. Therefore, based on this information, two propositions were made to distinguish between the exogenous and endogenous parameters to improve the results of clustering and to obtain good (representative) load models for the defined groups of customers.

Proposition 1: *Exogenous parameters are linked to weather, location and economic parameters. These parameters are not derived from the same data used for load model development and are specified a-priori. For this study, these parameters will include seasonality, customer economic activity class and the TOU periods.*

Proposition 2: *Endogenous parameters are parameters that are derived directly from the data that is used for load model development.*

Figure 4 illustrates how the exogenous and endogenous parameters combine to derive the desired load models and to achieve clarity about the clusters formed.

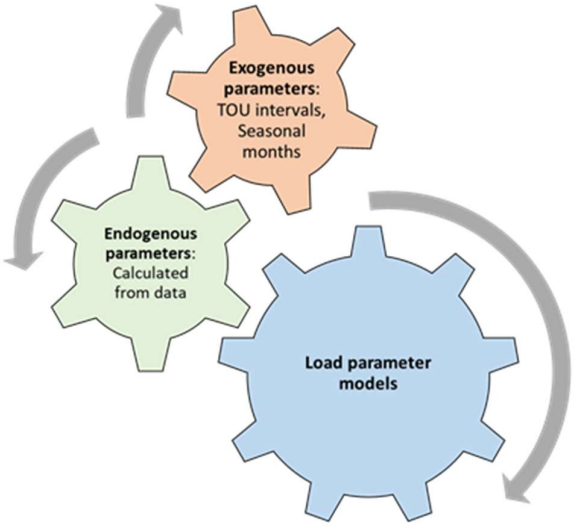


Figure 4: Load parameter model composition

3.4.1. Exogenous parameters

Ballanti and Ochoa (2015) found that the classical constant ZIP load models either underestimated or overestimated the network power losses in both the winter and summer seasons. Accordingly, they concluded that time-varying load models should be used. Exogenous parameters may play a supervisory role in determining the load model parameters because they provide predefined time intervals for which the parameters have to be estimated. A typical system profile is presented in Figure 5 and shows two peaks per day. There are usually two levels of consumption, linked to seasonality. Figure 5 is a typical profile that forms the basis of the TOU tariff structure used in Eskom.

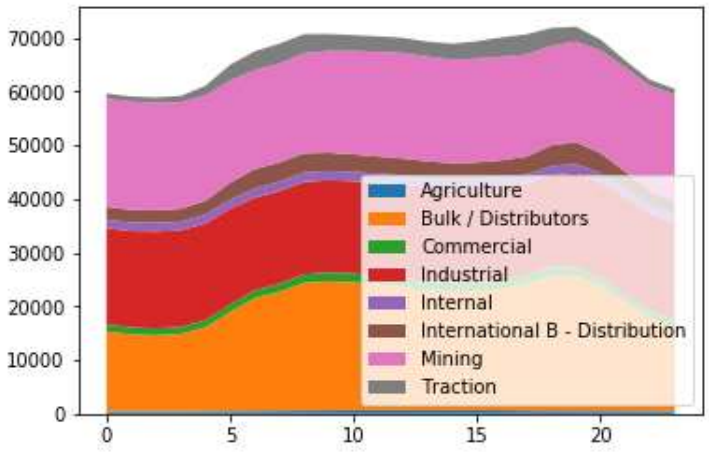


Figure 5: typical day profile based on average demand of sample per activity classes

The profiles in Figure 6 below are showing power factors variations with time. Comparing the power factor plots and the load (active power) profiles, it could be seen how the power factors were tracking the load profiles. Therefore, it was evident from the figure that the power factors varied with time and loading.

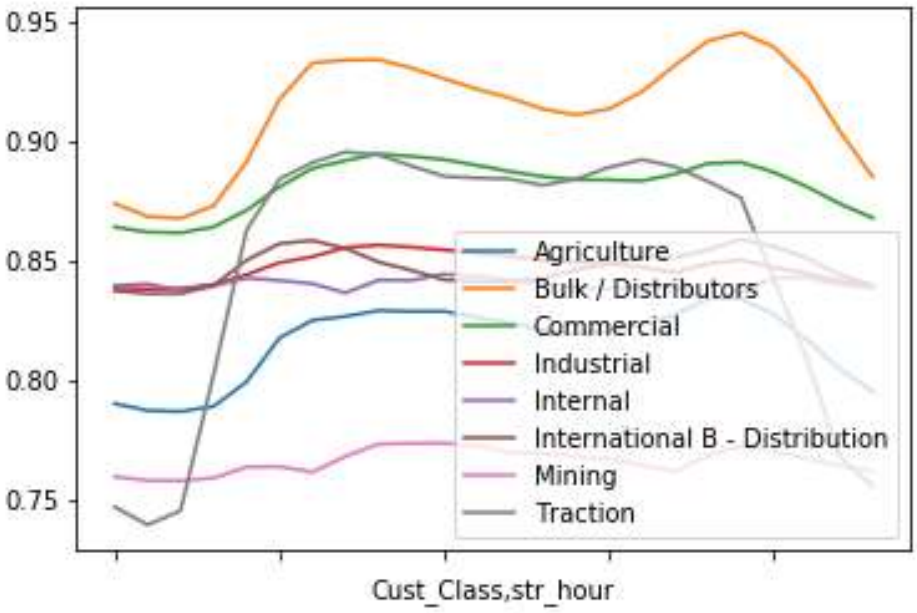


Figure 6: Hourly power factors of different classes

To determine the exogenous parameters, the daily load profiles allocated to different seasons and TOU periods may be partitioned as follows:

a) Seasonality factor – the two seasons used in technical and tariff model development were winter or the high usage season lasting three months and starting in May, and summer or low usage seasons for the remaining nine months.
b) The time of use tariff charged by Eskom comprised three segments, namely, peak, standard and off-peak segments, the combination of which depended on the typical day of the week:
    1) For weekdays (Monday to Friday): The peak segment occurred in the morning between 07:00 and 10:00 and from 18:00 to 20:00. The standard segment was from 06:00 to 07:00, 10:00 to 18:00, and from 20:00 to 22:00. The remaining hours were in the off-peak segment.
    2) Saturday: No-peak segment. The standard segment was from 07:00 to 12:00 and from 18:00 to 20:00, while the remaining hours were in the off-peak segment.
    3) Sundays were charged at off-peak rates.

To apply the time of use concept to data, the load curve for a period H is defined as follows:

$$\mathbf{P} = \{\mathbf{P_h}: \mathbf{h} = \mathbf{1}, \dots, \mathbf{H}\} \qquad \textbf{Equation 10}$$

The load segment is,

$$\mathbf{P_\tau} \subseteq \mathbf{P} \qquad \textbf{Equation 11}$$

for

$$\boldsymbol{\tau} = \mathbf{T}.$$

Where, T is hours in the time of use intervals, peak, standard and off-peak. H is the total number of hours in a day (24). The time series (load profiles) were partitioned into different segments of data that were identified by assigning the symbols P, S, O, and the arithmetic means of each of the segments are determined.

3.4.2. Endogenous parameters

In distribution networks, load models are concerned with the characterisation of feeder demand. In LV networks, it was shown that active power was adequate for analysis and planning studies (Gaunt et al., 1999; Ferguson and Gaunt 2003). However, this was not the case for MV network models due to the significant presence of inductive loads. Accordingly, the power factor in MV feeder models may be taken into account (Chihota and Gaunt, 2018). Several parameters were identified in the literature as suitable for explaining the profile differences and variations. It also emerged from the literature that high load factor customers were preferred over low load factor customers because of their impact on energy production costs and system utilisation. The higher load factor means that there is less variation in terms of load profiles. Figueiredo et al. (2005) found that the load factor and the night-time demand levels were the most relevant parameters in describing the customer's usage parameters. Since these parameters were identified as important for modelling and played different roles, they needed to be considered when deriving the load parameters. The levels of usage within the defined periods may be obtained from measurement data. It is apparent that, for technical and tariff analysis, it is essential to know the demand, time and duration of the demand for each user. Other important parameters that could provide information on usage behaviour and the characteristics of the feeders, such as load factor and power factor, are also important. This information is obtainable from the customer or feeder measurements. The parameters derived from the above principles may therefore be expressed as:

- **Load factor (**LF): Load factors play a significant role in distinguishing customers and are essential for technical models as well as tariff models. Generally, large industrial customers have higher load factors while residential and agricultural customers tend to have lower load factors.

$$\mathbf{LF} = \frac{\mathbf{P_{ave}}}{\mathbf{P_{max}}} \qquad \textbf{Equation 12}$$

- **Average Power factor** (PF**)**: Average daily power factor. The power factor is useful in identifying customers based on the type of load. For example, residential customers tend to have resistive loads while mining and industrial customer tend to have inductive loads. However, distortions in this perception may exist as customers install power factor correction units. Nevertheless, it is still evident that large industrial customers in particular during high consumption periods, tend to draw more reactive energy and thus have slightly lower power factors.

$$\mathbf{PF} = \frac{\mathbf{P_{ave}}}{\mathbf{S_{ave}}} \qquad \textbf{Equation 13}$$

- **Peak usage ($\mathbf{P_p}$)**: The parameter representing peak hours related to the peak average to the daily average. In Chapter 2, it was established that tariffs can either cause or defer distribution network investments and that PLF and TOU forecasting models lead to optimised network plans. Therefore, the usage level during peak time is key to both technical and tariff analysis.

$$\mathbf{P_p} = \frac{\mathbf{P_{ave,p}}}{\mathbf{P_{ave}}} \qquad \textbf{Equation 14}$$

- **Off-peak usage ($\mathbf{P_{opk}}$)**: The parameter representing off-peak hours. Both technical and tariff models would benefit from knowledge of the off-peak usage levels.

$$\mathbf{P_{opk}} = \frac{\mathbf{P_{ave,opk}}}{\mathbf{P_{ave}}} \qquad \textbf{Equation 15}$$

- **Standard usage ($\mathbf{P_{std}}$)**: The parameter representing standard hours,

$$\mathbf{P_{std}} = \frac{\mathbf{P_{ave,std}}}{\mathbf{P_{ave}}} \qquad \textbf{Equation 16}$$

Where:

- P_{ave} , S_{ave} refer to the average active power (kW) and apparent (kVA) respectively.
- $P_{min}(P_{max})$ is the minimum (maximum) power demand of the representative day
- $P_{ave,p}$,$P_{ave,oph}$,$P_{ave,std}$, is the average power demand during daily peak (off-peak) hours.
- Parameter **F**: Normalised average kW demand.

Given this analysis, the research highlighted the need to analyse customer energy demand to understand consumer behaviour from both the load profiles and the effective management of demand in the form of price signals. Therefore, another parameter to consider relates to the average demand.

To allocate the load profiles into meaningful classes using the k-means algorithm, a two-step process for identifying parameters may be followed. These parameters should be arranged into pairwise vectors. The pairwise parameter vectors would be subjected to the k-means clustering algorithm. The first step would be to segment the data according to the TOU hours while the second step would be to calculate parameters from the segmented data, as explained in the next sections.

A) Load profile segmentation:

A daily load curve comprises several TOU periods. Partitioning of the load curve may be achieved by dividing the load profile into TOU segments. TOU segments were found to be essential in tariff and technical analysis. The daily load profiles may be segmented using equations 10 and 11.

Since the tariffs would be modelled for typical days, H may be set to 24. The sets of data belonging to different segments of data may be identified by assigning peak, standard and off-peak to the median of each of the segments. The hours corresponding to the time-of-use (TOU) periods are the segments of interest and the parameters are derived from them.

B) Compute the endogenous parameters:

The parameters may be computed from the measurement data and the segments defined as exogenous parameters, using (4) to (8).

3.5. Class concepts and models

A class contains loads or customers whose parameters may be either similar or in close proximity but which are relatively distant to those of the other clusters. There is a possibility that within the class there may be different subclasses that are linked to parameters that are unique to the sub-class. These include the level or amount of power. Accordingly, a class C from M customers is a cluster K given by:

$$\mathbf{C^{(k)}} \in \mathbf{X} \qquad \textbf{Equation 17}$$

Where X is a sample of all the profiles x represented in the database

$$\mathbf{X} = \{\boldsymbol{x^{(m)}}, \mathbf{m} = \mathbf{1} \ldots \mathbf{M}\} \qquad \textbf{Equation 18}$$

3.5.1. Typical day classes

Classes that represent typical days can be achieved by using a method suggested by ElNozahy et al. (2013) where day variations may be defined as parameters. This involves a two-step process. The following steps may be performed:

1) *Definition and estimation of parameters.*

The exogenous parameters that are necessary for this instance include seasonality. This parameter may be used to sectionalise the data. As defined by Eskom and the municipalities in South Africa, three months were allocated to winter, namely, May June and July, while the remaining months were allocated to the summer season. The data was sectionalised, as per the seasonality parameter, into high (winter) and low (summer) demand. The summer season accounts for 273 days while 92 days were allocated to winter in a single year. Parameters were defined as hourly consumption for a day. This means that for a year, the feeder data would be arranged in 365 x 24 hours. The PCA algorithm may be applied to compress or project the data into a lower-dimensional subspace, that is, from a 24D (parameter) space to 2D or 3D space, where it would be possible to visualise the data and the processing might be less complex. Supposing that each hour represented a single parameter of the specific day, therefore a day can be represented as T projected in H space as illustrated in Equation 19 below:

$$\mathbf{T^H} = \{\mathbf{H}: \mathbf{h} = \mathbf{1}, \mathbf{2} \ldots \mathbf{24}\} \qquad \textbf{Equation 19}$$

Using the PCA algorithm described in chapter two (section 2.9), T can be compressed to two or three-dimensional space R as in Equation 20 :

$$\mathbf{T^R = \{R: h = 2 \text{ or } 3\}} \qquad \textbf{Equation 20}$$

2) Applying the clustering algorithms

To group the calendar days into typical days for a selected year, one of the dimensions should be the date or the corresponding chronological indices. Therefore a typical day may be defined as the representative of cluster K of elements with reduced parameters to R;

$$\mathbf{Typical\ day = D_i = D^{(R)(k)}} \qquad \textbf{Equation 21}$$

which can be found by applying clustering algorithms to the new space data (principal components).
The typical day classes may be determined by applying the k-means clustering algorithm to the reduce vector. As noted in the literature, the k-means algorithm is an unsupervised clustering method. The k-means algorithm also requires that the number of clusters is specified upfront. To determine the number of clusters to be specified in the k-means procedure, various index adequacy measures may be used. The same adequacy measure may be used to evaluate the performance of the clustering algorithm.

The proposed procedure can be outlined as follows:

1) Prepare the data, with hourly records as parameters in columns.
2) Reduce the data dimension using PCA.
3) Perform the first clustering iteration. Apply clustering algorithm with large numbers of iterations.
4) Use adequacy measure (Silhouettes scores, DBI and elbow (SSE)) results to estimate (k) the number of clusters.
5) Analyse the results, (scatter plots with centroids) to make sense of the clusters.
6) Perform k-means clustering with the set number of clusters.
7) Validate the resulting clusters using the same adequacy measures.
8) Label and classify the data into typical days.
9) End of procedure.

When the clusters have been created, the next task is to interpret them and to be able to implement them. These are clusters of days, and the days that are associated with each cluster must be identified accordingly. This may be done by reverting to the original data and labelling each of the dates with the cluster number, where it belongs. After labelling the data, the number of days associated with each cluster should be counted and their relative percentage or ratio to the entire dataset may be determined. The allocation of chronological days to typical days classes may be done based on the probabilities of each day falling into the typical day class.

3.5.2. Customer classification

The classification of loads follows after the classes have been created, that is, feeder or the customer groups are formed based on the parameters. To create classes, a clustering algorithm may be used. Classification refers to the allocation of loads into clusters, using the probability approach for this purpose. Histograms may be used to determine the frequency of a particular feeder in each cluster and allocate such feeder to where its probability of occurrence is higher; that is, where it has the highest possibility of appearing.

3.5.2.1. Clustering using the k-means algorithm

The machine learning methods used in clustering can be summarised into supervised, semi-supervised and unsupervised learning (Simeone, 2018). This study proposed a method based on the k-means clustering algorithm, based on the algorithm's popularity and simplicity as highlighted in the literature review. However, for the type of modelling required in this study, the clustering algorithm may require to be supervised to provide good results. Supervision in clustering may be provided by labelling the data before applying the clustering algorithm. In this study, the suggestion was to use exogenous parameters for labelling and thus provide the desired supervision.

3.5.2.2. Selecting representative profiles

When the load profiles have been allocated into different clusters to form a class, a single load curve that represents class may be determined. The representative profile may be calculated as the average of the load profiles in the cluster or the median profile. Alternatively, the centroid of the cluster may be used as a class representative profile, since it is the mean of the profiles assigned to the cluster. In this study, cluster $C^{(k)}$ contains $N_k$load patterns and has a centroid $c^{(k)}$. The representative load profile $P_h^{(k)}$ is selected to be (Buys & Gaunt, 2020):

$$\mathbf{P_h^{(k)} = \frac{1}{N_k} \sum_{n}^{k} P_h^{(n)}} \qquad \textbf{Equation 22}$$

where h = {1,…H} is the time domain

In short, the procedure for determining the representative profile may be as follows:

1) Identify all members belonging to a cluster (k).
2) Determine and visualise the daily profiles of each member.
3) Calculate the average profiles of all members in the cluster (k).
4) Use this as the class profile.

3.5.3. Cluster adequacy measures

The silhouette statistics, the elbow method and the DBI methods were found to be useful in determining the optimal number of clusters based on clustering performance. The elbow method, the silhouettes statistics and the DBI were also preferred in this study for determining the number of clusters as well as evaluating the performance or adequacy of the clusters.

3.6. Data preparation

The data preparation stage is important when using customer measurements for load modelling because the measurements data may contain unwanted characters, the extreme value resulting from measurement errors, gaps that may affect the model results, and the measurements data set may be too large for the tools and computer systems that are used for modelling. This section focuses on ensuring that the data can be handled and processed using inexpensive computers and that it is ready for load modelling.

3.6.1. Sampling

Dealing with a large number of customers can be computationally expensive (Aamir, 2014; Frost et al., 2017). Data sampling techniques may be used to reduce a large amount of data into manageable sizes while maintaining

the desired representation and variability inherent in the dataset. When applying sampling methods to data, it is important to show that the selected sample is representative of the target population and the study about the demographic and other relevant characteristics that may affect the outcome of the study (Aamir, 2014). The sampling technique that would be suited for this research and that allowed for generalisation should be probability-based or random sampling methods (Acharya et al., 2013). The types of random sampling methods include simple, systematic and stratified random sampling (Aamir, 2014; Acharya et al., 2013). Other probabilistic sampling techniques include cluster sampling and systematic sampling (Barreiro and Albandoz, 2001). According to Barreiro and Albandoz (2001), random sampling may be done when samples are drawn either with or without replacement. In addition, random sampling could ensure that all the objects taken from the population had equal chances of being selected in the sample (Barreiro and Albandoz, 2001).

Stratified sampling was classified under probability sampling methods. According to (Acharya et al., 2013), probability-sampling methods were considered superior to non-probability sampling methods because they did not have an unexplained bias in the samples and the data variability could be explained. In stratified random sampling, the data may be divided into various subgroups called strata (Buys & Gaunt, 2020). Each member object of the stratum may share some common parameters and may be sampled separately. Strata were recommended where the population comprised different sections, which varied in size or type (Kitchenham and Pfleeger, 2002). According to Barreiro and Albandoz (2001), stratified sampling provided better results than random sampling when there were more differences between the strata but the strata were more homogeneous internally. According to Acharya et al. (2013), three may be used to distribute the size of the sample between the strata, namely:

- Allocation in proportion to the size of each stratum, which means determining the ratio of the stratum size to the population.
- Proportional to the defined variability parameters in each stratum. For example, it is known that customers are grouped into activity sectors, sector identification codes (SIC), and thus the sample should maintain that proportion.
- Equal proportions to the strata. This applies where the intention was to promote the smaller strata and where there was no concern over the precision of the bigger ones. Eskom customers are categorised according to classes such as mining, bulk/distributors, commercial and industrial (Buys & Gaunt, 2020).

In each customer class, there are several SICs to further distinguish and categorise the customers. It is important to ensure representation from all the SIC codes to capture all the customers that are represented without the distraction of the dominant loads. Accordingly, the customers may be separated according to the SIC codes (Buys & Gaunt, 2020). A simple random sampling technique, which ensures the same probability of being drawn for each of the members of the stratum, is used to draw the samples (Buys & Gaunt, 2020). The process flow of the stratified sampling procedure is summarised in Figure 7 below.

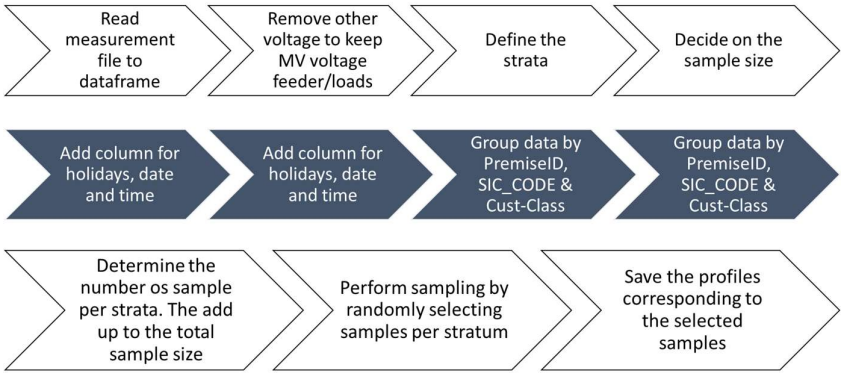


Figure 7: Stratified sampling process flow diagram

The arrows indicate the process steps, with the process flow following the arrows starting from the top row. The steps are automated using python software. The approach used involved calculating the sample based on the proportion of the total consumption of the stratum. The procedure may be summarised as follows:

1) Filter data for MV loads only (6.6kV or 11kV to 33kV).
2) Select the stratum (Activity class e.g. agriculture).
3) For each stratum:
    a. Calculate the total power consumption as the sum of all the loads within the stratum (Activity class).
    b. Calculate the proportion of the accounts per SIC for the stratum.
    c. Multiply this proportion with the required sample size.
4) Go to the next stratum.
5) Indicate how many samples per SIC.

3.6.2. Normalisation

Data normalisation was used in research and modelling to minimise the dominance of large values, which were likely to affect the outcomes of the model. Muralidharan (2014) defined normalisation as the process of minimising the effects of systematic sources of variation. According Muralidharan (2014) normalisation was useful for classification algorithms based on distance measures. The normalisation techniques include min-max normalisation, normalisation by decimal scaling and z-score normalisation or standardisation (Muralidharan, 2014). One of several commonly used normalisation techniques is referred to as min-max standardisation, and it yields values in the range of [0, 1] (Muralidharan, 2014; Buys & Gaunt, 2020). The procedure was chosen based on its advantage of providing data that was free from both outliers and bias. Equation 23 was used for min-max normalisation. The min-max procedure was deemed suitable for the purposes of this study.

Normalisation formula (Muralidharan, 2014, Buys & Gaunt, 2020):

$$\mathbf{P_j} = \frac{\mathbf{P_i} - \mathbf{P_{min}}}{(\mathbf{P_{max}} - \mathbf{P_{min}})} \qquad \textbf{Equation 23}$$

Where:

- Normalised data point is (P_j) ,
- (P_{min}) is the minimum value of power over a defined period, and
- (P_{max}) is the maximum (peak) power over the period.

This normalisation procedure has the advantage of providing a data set that is free from the effects of outliers and missing data (Buys & Gaunt, 2020).

3.6.3. Data compression – PCA method

The minimum number of the parameters that have been identified for our load model is six, applied in the customer classification model. Various data compression or dimensionality reduction methods have been studied (Silipo, 2015). The PCA method was one such method and was used with clustering algorithms. Based on its simplicity and popularity, PCA was the dimension reduction algorithm chosen for this study. The PCA was explained in the literature review in Chapter 2.

3.6.4. Software and tools used in modelling

The procedure and algorithms used in this research were applied iteratively for a large number of customers and thus the procedures were automated. Python open-source software, including the associated scientific tools, was used to develop a code for both automation and the processing of the models. The analysis of the results was carried out primarily using the Python matplotlib library. Some of the analysis results were obtained using the Orange analytic program, which comes with Python packages. Microsoft Excel was also used for parts of the analysis.

3.7. Validation of the results

As indicated in the literature review chapter, statistical techniques have also been used for the validation of both experiments and results. This study also used statistical tools and techniques to validate the parameters. The criteria used for the validation of the parameters were based on the concepts of analysis of variance (ANOVA) widely used in exploratory data analysis. The load model parameters were interpreted and expressed in a form of a mathematical function for representing each class in relation to the estimated consumption value. In addition, linear regression statistics were also used to further evaluate the effect and significance of each parameter in the relationship or function that defined each class.

The ANOVA statistics that were used to ascertain statistical significance included the mean, maximum, minimum, standard deviation and p-value. A smaller p-value (less than 0.05) signifies that the parameter is significant in determining the outcome of the function representing the class. To validate the results, the load models may be compared with the customer profiles as provided from the measurements database. The comparison is based on the summary statistics of the original customer class profiles and the profiles from the results. Further, regression models are useful in assessing the validity of the variables in relation to the results. The suggestion was to use the regression models to validate the load model results.

3.7.1. Exploratory data analysis

The necessary step in utilizing data for deriving the pdfs or CDFs, and making statistical inferences, is to explore and understand the data, identify the categorical and numerical variables, the inherent relationship between the variables that describe it and the underlying distributions of the variables of interest. In addition, challenges may arise concerning outliers, which can be detected early through exploratory data analysis. Outliers tend to skew the data and it is therefore important that they are identified and removed where applicable. It is thus necessary to explore the data to detect whether there are outliers and underlying patterns.

3.7.2. Linear regression models

A multiple-linear regression model may be used to assess the validity of the classes derived using clustering algorithms. Regression models have been widely studied and have been found to be useful in describing the relationships that may exist between the independent variables and the dependent variable. The validity of the regression models may be determined by assessing the value of the coefficient of determination (R^2) and its adjusted equivalent (adjusted R^2). If the R is close to one, this indicates that the function is explaining the variation of the parameters in the represented class. The coefficients represent both the extent and the direction of the influence of the variable on the outcome of the function.

The benefit of fitting a regression function to the model is that the coefficients are determined based on the least square error estimation, thus indicating that the coefficients are the best estimates. Improved accuracy may be obtained when the intercept, which is a numerical constant, was eliminated, as it carried no meaning in the context of multiple variables that are present in each class. Logically, it is not possible that all members of a class can have zero consumption and still get a value as an average of the class consumption. The basic equation for linear regression is:

$$\mathbf{y = a + b_1 x_1 \ldots b_n x_n + \varepsilon} \qquad \textbf{Equation 24}$$

Where a is the intercept, b_i is the coefficient of the variables (parameters) x_i and ε is the error term.

Without the intercept, Equation 24 becomes

$$\mathbf{y = b_1 x_1 \ldots b_n x_n + \varepsilon} \qquad \textbf{Equation 25}$$

The coefficient is determines using the following:

$$\mathbf{b_i = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}} \qquad \textbf{Equation 26}$$

$\bar{x}$ and $\bar{y}$ are the means of the data sets x and y.

3.8. Concluding remarks

This chapter discussed the theoretical development related to the process of load modelling using measurements. The process discussed had four steps or phases, namely, Data selection and –pre-processing, parameter extraction, load modelling, and model validation. The theoretical basis for identifying parameters led to the proposal of the exogenous and endogenous parameters for the study. The exogenous and endogenous parameters may be combined to provide the desired load models. The use of clustering algorithms was also suggested to develop customer classes that were required in load modelling. Further, the proposal was also to use the exogenous parameters to provide supervision to the clustering algorithm and thus improve the results. Also suggested was the use of the PCA algorithm for reducing data dimensions as a way of including all the desired parameters of the load profiles, because these multiple parameters pose a dimensionality problem when employing clustering algorithms. K-means clustering was recognised as simple yet effective in classification models, and it forms part of many load models. A supervised k-means clustering algorithm was also proposed for the classification of loads.

Various adequacy measures may be used to guide the selection of the number of clusters, as well as to validate the clusters themselves. Three basic measures that can provide adequate clustering results have been suggested namely, the elbow method, the silhouette statistic and the DBI. These measures may be used in determining the optimal number of clusters as well as validating the clustering results. These measures have been recognised for their simplicity and effectiveness when combined.

The model results may be validated using statistical measures and plots. The use of ANOVA statistics and regression models were suggested for both numerical and graphical analysis of the results. Regressions models have been suggested for validating the model results and assigning loads to the different classes.

4. DATA PREPARATION AND PARAMETER ESTIMATION

The process discussed in Chapter 3 showed that load modelling might be done for creating typical days and customer classification. This chapter explains how, following the steps illustrated in Figure 3, the data was being handled and prepared for both the typical days and customer classification modelling. The results of each step are presented. The chapter also presents the results of the exploratory data analysis, which was conducted on the measurement data. The purpose of the preliminary data analysis was to understand the data used and identify if there were any anomalies. Statistical analysis tools were used to analyse the measurement data. The data were classified according to the economic classes defined in the database and the statistical analysis was performed on each class.

4.1. Data preparation

When dealing with large volumes of data, data preparation is the first step to improve the speed of processing, remove any outliers and ensure that the data is in the required format. The next sections discuss how the data was prepared for analysis and use in the modelling process.

4.1.1. Data used

The data used was received from the Eskom database. The data used for the study was Eskom's customers measurements data recorded between 2017 and 2018. The measurements included other details such as the customer account details, including the location, sector and economic class of the customers, in addition to the half-hourly kW and kVAr (lead and lag) import and export data. This means that, for a single year, there are approximately 17520 records for each customer. The data was acquired in text file format and was processed using Python programming language. Python, which is an open-source programming language, was also used throughout the study and for all the experiments and computations.

4.1.2. Data preparation and sampling

The data that is used for the study has been used for billing customer. Therefore, the presence of extreme values did not necessarily imply that there were errors in the data, but that there could be customers with larger or smaller demand levels, especially when viewing the data from an economic activity group or other categorisation that did not necessarily consider the size of the load. Essentially, the outliers could imply that the class compositions as defined in the database needed improvements and/or contained missing data and errors that were later corrected. The results of this study are intended to indicate whether there is a need to deal with such extreme values. To prepare the data, the accounts/or meters with zero consumption for all periods were removed. The annual load factors of all meter points were also calculated and where the load factors were less than 5%, the points were removed. This was to eliminate the data distortions resulting from customers moving premises or changing accounts.

To determine the sample size, a decision was made that:

- Only MV measurement should be contained in the sample.
- Each stratum should contain at least 1000 samples.
- Each economic class must be represented in the stratum.
- Each SIC must be represented in the stratum.

The sampling was done for all the nine provinces in the country, which were all sampled individually and later combined to create a national sample. The resulting national sample comprised 783000 customer accounts. These records covered all the 160 SICs in the country. Table 1 presents the number of samples per customer class or sector. The PREMISE ID column shows the count of customer accounts in the population

(database) while the SIC column shows the number of SICs that were accounted for in the sample. The load measurement from the sample are summarised in Table 2 below.

Table 1: Summary of sample sizes per economic class

Cust_Class	PREMISE_ID	SIC
Agriculture	814384	22
Bulk / Distributors	1160011	10
Commercial	2158357	65
Industrial	692160	49
Internal	78840	1
Mining	595305	16

Table 2: Summary of the load measurements of the sample

Class	Average (kW)	Maximum (kW)
Agriculture	819.67	34 365.63
Bulk / Distributors	21 098.51	514 129.81
Commercial	1 928.30	213 218.25
Industrial	20 743.66	1 105 262.05
Internal	4 105.72	18 129.09
Mining	23 691.51	669 833.75
Traction	3 729.96	34 999.24

4.1.3. Normalisation

The sample contained different types and sizes of loads. Larger loads dominate smaller loads in terms of consumption levels (see Figure 8). Figure 8 shows the average energy consumptions of different customer classes per month for a year. It is clear from the graph that the average energy consumption for different classes in the database is significantly different. This can have an impact on the load model results. The bulk/distributor dominated all the other economic classes. The dominant class will skew the results and therefore the desired parameters from some of the classes could be missed. The load profile of each customer and class is important for developing robust models. Therefore, the data needed to be normalised to the extent that the profiles and variations of each load class could be adequately captured.

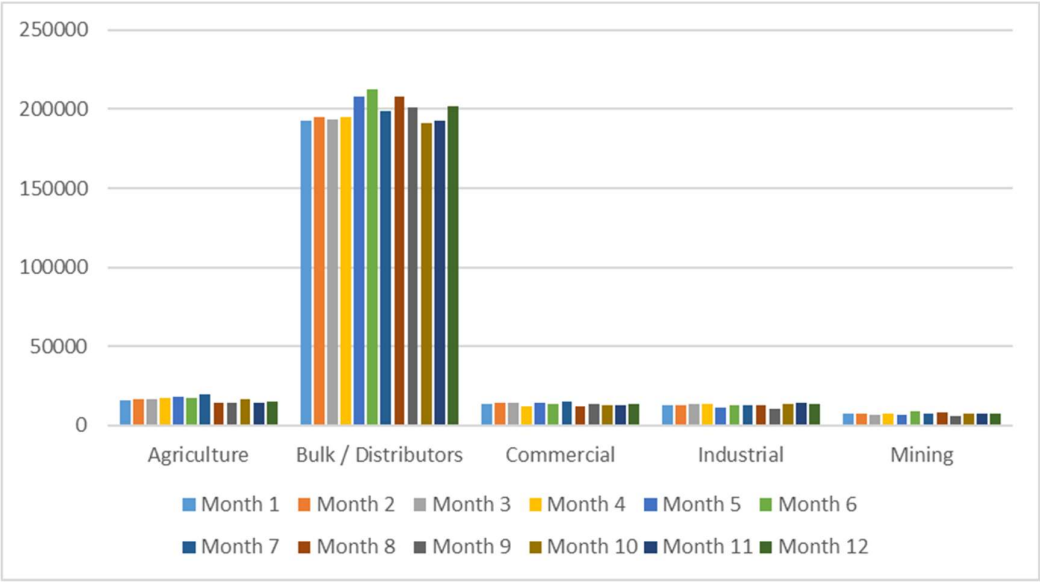


Figure 8: Comparison of average monthly energy levels between different economic classes (kWh)

The normalised sample profiles, summarised per SIC code are depicted in Figure 9. The purpose of this picture is to illustrate the different shapes and that these shapes are visible when the data is normalised. Due to the picture size limit all the SIC codes could not appear in Figure 9.

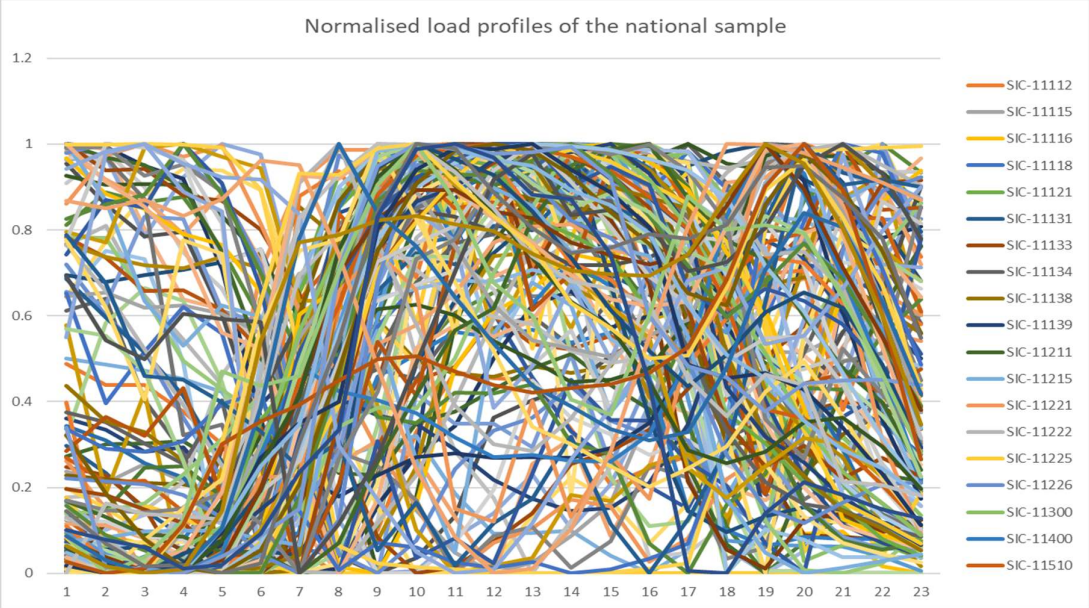


Figure 9: An illustration of the normalised daily load profiles for the different SICs from the sample

Normalising the profiles eliminated the dominance of classes so that the focus could be on the shapes of the curves and other distinguishing usage–based parameters to identify and classify loads. The consumption levels of customer demands w still key in the load models, but they are limited in distinguishing customers because loads that are equal in size (kW or kVA) may have different profiles. Load variations are largely a result of energy usage behaviours of customers. Further exploration of the measurement data may be done using statistical methods. Generally, data exploration is key to data preparation, and establishing the bases of testing and validating the results. In particular, the results were validated by comparing the resulting customer class distribution and profiles to those in the database.

4.2. Data exploration and load parameter estimation

The tools for exploratory data analysis used in this study included the box and whiskers plots, the frequency/distribution plots and the summary statistics. The economic classes as used in the Eskom database were evaluated using each of the EDA tools and the parameters identified in chapter 3. These parameters are the load factors, power factors, normalised peak usage parameter, normalised standard usage parameter and normalised off-peak usage parameter. The evaluation results were key in determining whether the existing economic classes comprised loads or customers that had similar energy usage patterns.

4.2.1. Endogenous parameters in economic classes

Statistical techniques discussed in chapter 3 were used to estimate the values of endogenous parameters for the load model. Since the number of loads used in the study was large, the economic activity classes defined in the Eskom database were used to illustrate how the statistical tools could be used to estimate the values for the identified parameters. The process started by analysing the loads, before using statistical plots and summaries to decide on the estimated values to use.

Figure 10 presents the average load profiles of the customer classes as they were defined and used in Eskom at the time of the study. Each customer class was represented using a different colour. Some profiles appeared to be similar. Both the agriculture and the commercial customer classes had similar shapes for the majority of the hours and peak at about midday. The noticeable difference between the agriculture and the commercial profiles was between 8:00 p.m. and 10 p.m. Eskom's own consumption was represented by the internal class and followed a similar pattern to that of the bulk/distributor class, with a peak in the morning and another in the evening.

The industrial class indicated a significantly low consumption during the day. The industrial class was known to have a flatter profile than the other classes. This new profile indicated the presence of alternative power supplies that are available during the daytime. Logically, this could mean that the industrial customers have installed PV systems that supply them during the daytime, and as the sun disappeared in the evenings, they switched back to Eskom mains. A notable decrease in consumption during peak periods was observed for the mining class. This could be an indication that the mining class was responding to the TOU tariff signals, which were expensive during peak hours. These trends signalled the deviations of the load profiles and customer classes from the historic models and thus a shift in the load models that has yet to be captured. There is an opportunity for future research dedicated to unpacking these trends and understanding their drivers.

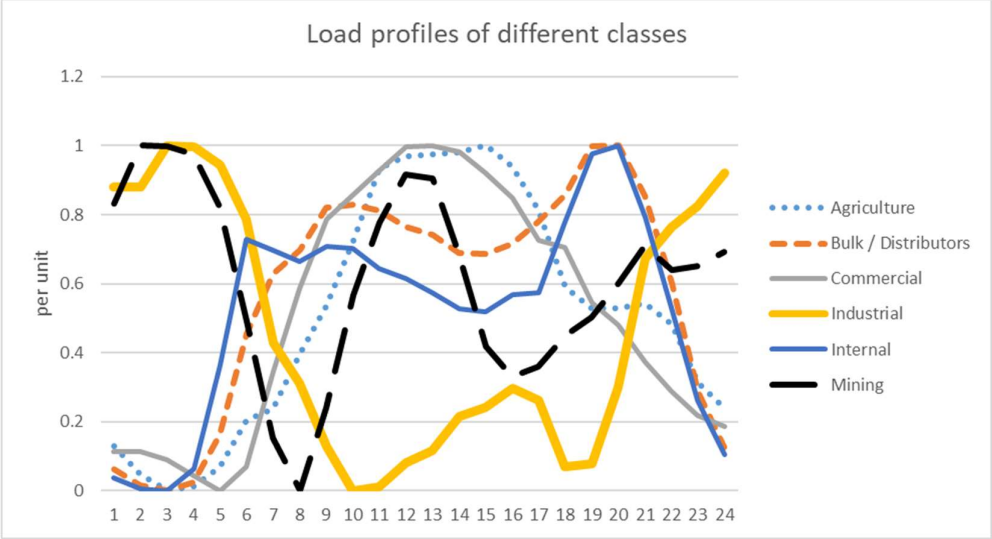


Figure 10: Average daily load profiles of the economic classes as defined in the Eskom database

Figure 11 depicts the box and whiskers plots of the economic classes. The alphabets LF, PF, P_UF, S_UF, O_UF and Pav_UF were used to represent the parameters for visibility in the different plots and the tables used for the analysis. The parameters and associated symbols are shown in Table 3 below.

Table 3: Symbols to represent different parameters

Symbol	Parameter
LF	Load factor
PF	Power factor
P_UF	Normalised Peak usage
S_UF	Normalised Standard usage
O_UF	Normalised Off-Peak usage
Pav_UF	Normalised Average Power

The box plots in Figure 11 illustrated the spread of each of the parameters, and the presence of outliers and the shape of the data, without making any assumption on the distributions of these parameters. The analysis was based

on the hourly data. There are instances where, in a single day, some values ranged from zero to maximum. An example is during peak time where the usage values for some of the usage parameters such as standard and off-peak would be zero. Therefore, instead of recognising zeroes as outliers, they were left in the data but excluded in the calculations so that the results were not distorted. It can be seen that there were also outliers associated with some of the parameters. An analysis of each of the parameters follows.

A) LF parameter

The load factor (parameter LF) had fewer outliers, shown above the top whisker for higher values, none for the lower values and, as indicated by the red median line and the long upper tail (whisker), the dataset was skewed towards the right. This was the case for all the economic classes in the data sample. The median line was closer to the 75th percentile. The use of the median or mean as the estimate for the load factors parameter would be representing mostly the load in the upper percentile of the classes. The spread of the load factors values was also large relative to the other parameters, as the indicated size of the "box" in the box plot.

B) PF parameter

Power factor (parameter PF) showed the presence of a significant number of outliers on both the low and the high values for all the classes except for the mining class.

C) P_UF parameter

The normalised Peak usage (parameter P_UF) revealed the presence of outliers, and the median line was slightly below the 50th percentile. As indicated by the sizes of the box (quartiles), the dispersions were different for each class.

D) S_UF parameter

The normalised Standard usage (parameter S_UF) showed no presence of outliers while the dispersion appeared to be uniform for all the classes. The median line for parameter S_UF was closer to the 50th percentile, except for the bulk/distributor class, where it was closer to the 75th percentile.

E) O_UF parameter

The normalised Off-Peak usage (parameter O_UF) was dominated by extreme values below the 25th percentile. The dispersion of this parameter was very low and only the bottom tail was visible for all classes. The bulk/distributor and the commercial classes show the smallest of the quartile box with more outliers.

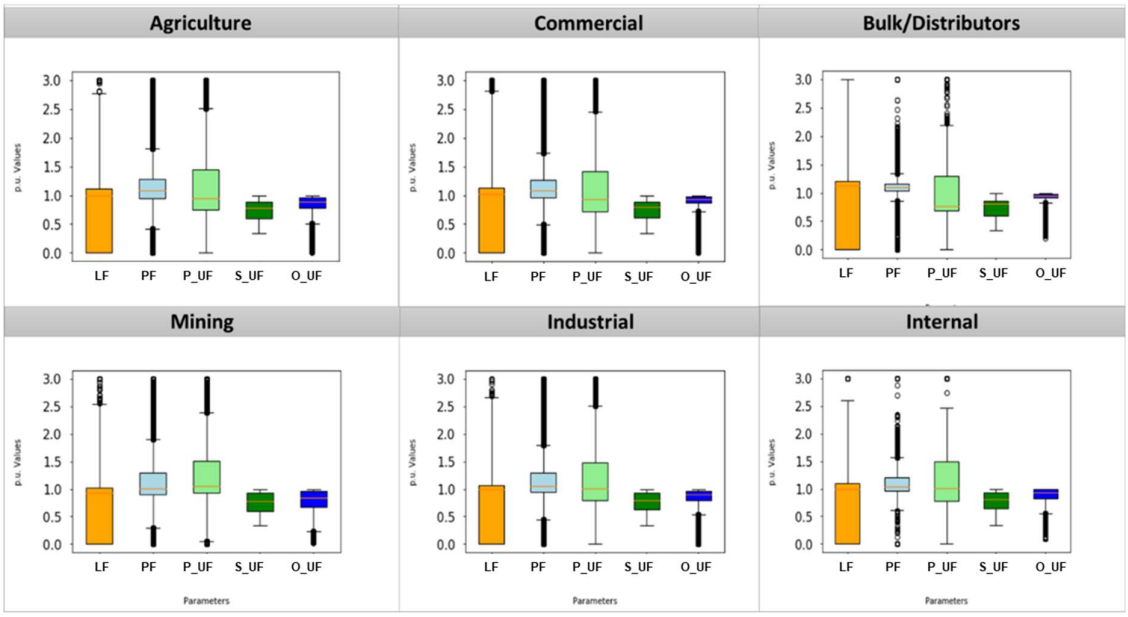


Figure 11: Box and whiskers plots of the customer classes.

The results of further analysis on various customer classes are presented below in the form of histograms/distribution plots and numerical statistics.

4.2.1.1. Agricultural Class

Figure 12 depicts the distribution plots of the different parameters in the agriculture class. The normalised parameters were calculated and the frequency of their occurrence in each of the existing classes and assessed in the form of distribution plots. As indicated in Figure 12, the y-axis of each of the distribution plots indicated the frequency as a count of the normalised values of the parameters, while the x-axis is the associated per-unit value of the parameter. There was a zero for each parameter, and this was because the parameters were largely time-based, that is, there would be parameters that would be zero, when others are greater than zero. An example is during peak, where the standard and off-peak values must be zero.

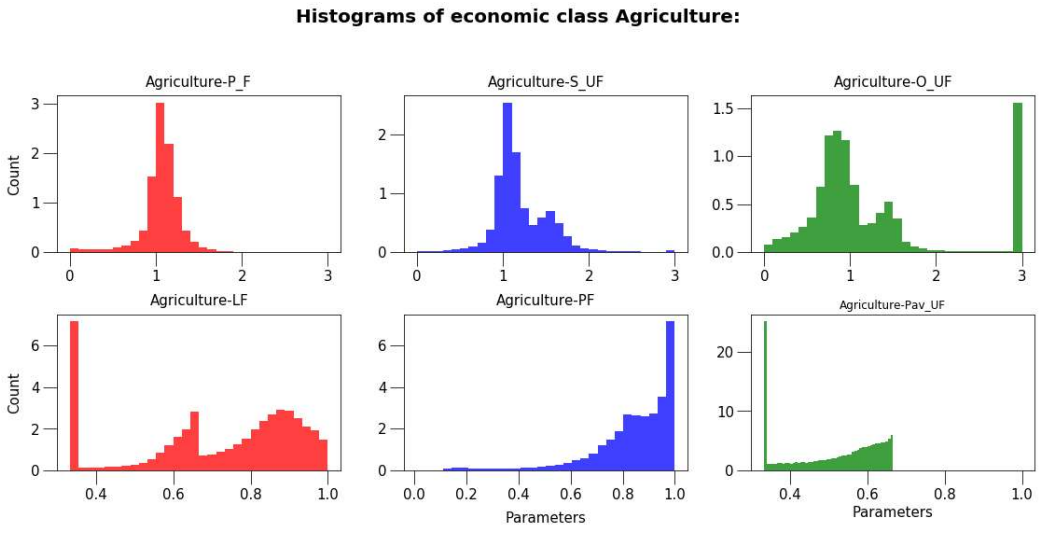


Figure 12: The distribution of parameters in the agriculture classes.

There were similarities in the distribution plots of all the classes relating to all the parameters. The plots of the Agricultural, commercial, Bulk/Distributors, and industrial mining are depicted in the same order in Figure 12, Figure 13, Figure 14, and Figure 15. The plots were not necessarily identical for the different customer classes but they had similar shapes. An exception was Figure 14 the bulk/distributors distribution plot of parameters P_UF, S_UF and O_UF, which were narrowly exhibiting a low dispersion. The following observations an interpretation can be made from the figures except for Figure 14:

- There were a significant number of loads with load factors ranging between 0.6 and 0.9.
- A large number of loads in this class had low load factors of about 0.3.
- The power factors for most agricultural customers were between 0.8 and 1.
- There are customers whose load factors were close to one.
- The very low power factors could indicate the presence of outliers or faulty conditions.
- Parameter P_UF, S_UF and O_UF indicate the relative utilisation of energy during different time of the day. Parameter (peak usage) had a high count of zero or numbers close to zero load, and this was expected as the peak hours were very few in South Africa relative to the standard and off-peak times.
- The off-peak was generally marked with high consumption because it was cheaper to utilise electricity during this periods. The spike in off-peak usage (O_UF) could be an indication of the high usage associated with this period.

The numerical statistics of the agriculture class are summarised in Table 4. As depicted in Figure 12 and Table 4. The parameters P_UF, S_UF and O_UF appeared to be symmetric around the means, whereas LF and PF were skewed to the left. A summary of the statistics of these parameters is presented in Table 2 below and confirms the observations from the histograms. The beta distribution may be a possible fit. As explained in the literature review, previous studies have indicated that the loads in South Africa followed a beta distribution. From the results it is also valid to assume that the MV loads may be represented using a beta distribution. Therefore, this means the adoption of the Herman-Beta transform could be extended to MV network analysis as suggested by (Chihota and Gaunt, 2018). It is still necessary to ascertain this by conducting a thorough investigation into the MV loads. The results of this study will be pivotal in providing the required load models for this type of investigation.

Table 4: Summary statistics of the agriculture class

	P_UF	S_UF	O_UF	LF	PF
Count	135451.0	135451.0	135451.0	135451.0	135451.0
Mean	0.728	1.017	1.255	0.713	0.839
Std	0.535	0.53	0.83	0.21	0.168
Min	0.0	0.0	0.0	0.333	0.012
25%	0.0	0.941	0.75	0.597	0.779
50%	0.999	1.078	0.952	0.777	0.879
75%	1.121	1.288	1.455	0.886	0.963
Max	3.0	3.0	3.0	1.0	1.0

4.2.1.2. Commercial Class

Observations similar to those of agriculture class were made in the commercial class. Accordingly, it was possible to posit that the assumptions about the distribution of parameters about the agriculture class could also apply to

the commercial class. The commercial class distributions are presented in Figure 13 and the summary statistics are shown in Table 5.

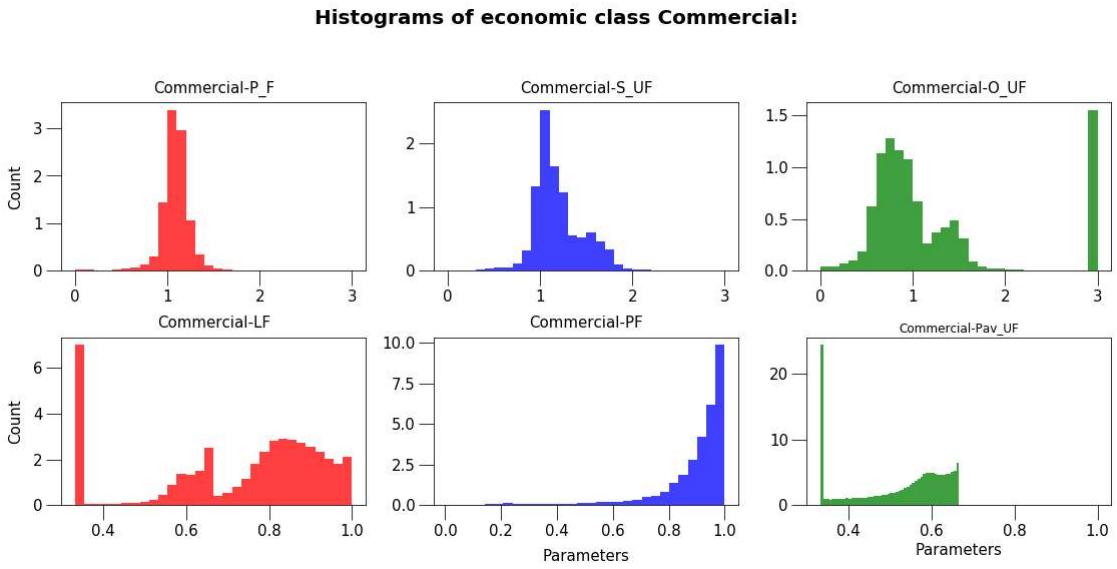


Figure 13: Distribution plots of the parameters of the commercial class

Table 5: Summary statistics of commercial class

	P_UF	S_UF	O_UF	LF	PF
Count	411776.0	411776.0	411776.0	411776.0	411776.0
Mean	0.743	1.012	1.245	0.727	0.897
Std	0.525	0.503	0.814	0.207	0.134
Min	0.0	0.0	0.0	0.333	0.01
25%	0.0	0.962	0.723	0.612	0.873
50%	1.017	1.083	0.936	0.797	0.939
75%	1.124	1.272	1.417	0.885	0.978
Max	3.0	3.0	3.0	1.0	1.0

4.2.1.3. Bulk/Distributors

The histograms of the different parameters of the bulk/distributors class are presented in Figure 14 and the corresponding summary statistics are shown in

Table 6. Similar to the agriculture class it can be observed that the normal distribution of the bulk/distributors class may not be a good fit for all the parameters, as the summary statistics indicated both that the underlying distribution is not symmetric and also that the mean is lower than the median. While similar conclusions in respect of some of the probability distributions can be drawn, it is also to be noted that the other parameters exhibit two distributions or peaks in a single plot, that is, they may be bimodal or there could be loads that should not form part of this class. This provides an opportunity for exploring this classification further.

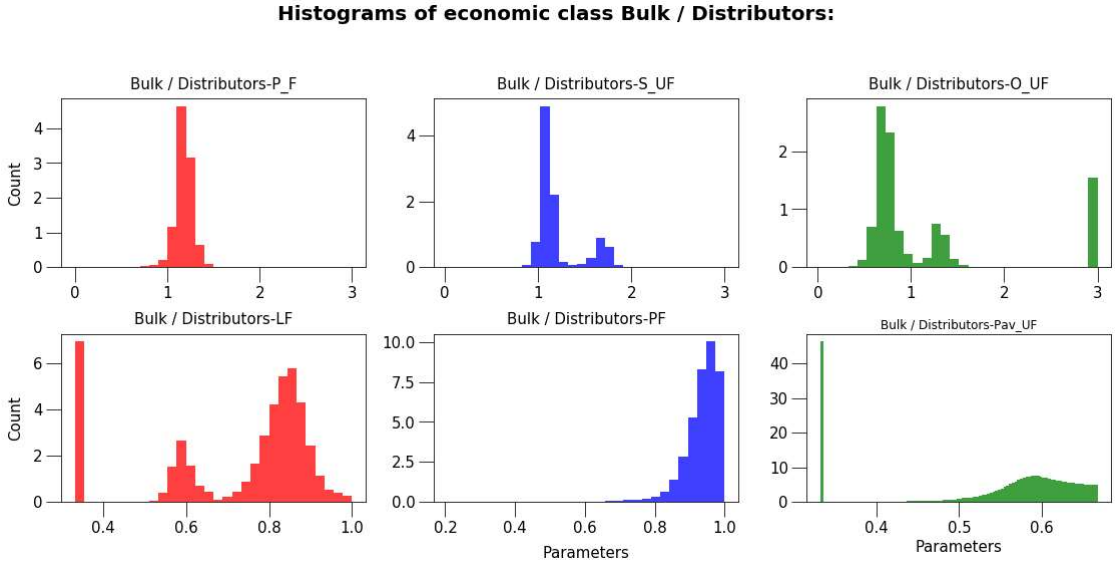


Figure 14: Distribution of parameters of the bulk/distributors class

Table 6: Summary statistics of bulk/distributors

	P_UF	S_UF	O_UF	LF	PF
Count	166591.0	166591.0	166591.0	166591.0	166591.0
Mean	0.806	1.021	1.173	0.722	0.934
Std	0.553	0.49	0.813	0.194	0.057
Min	0.0	0.0	0.0	0.333	0.208
25%	0.0	1.034	0.693	0.597	0.913
50%	1.133	1.094	0.772	0.811	0.945
75%	1.205	1.155	1.296	0.858	0.97
Max	3.0	3.0	3.0	1.0	1.0

4.2.1.4. Industrial and mining classes

All the results of the industrial and mining classes are similar. However, an interesting observation in relation to these classes had to do with the average power distribution, which was flatter but skewed to the left. The flatness w an indication of a relatively constant load that had a high factor load. The distribution plots of the industrial class are presented in Figure 15. Since the mining distributions are almost identical to those of the industrial class they are not shown separately in this analysis.

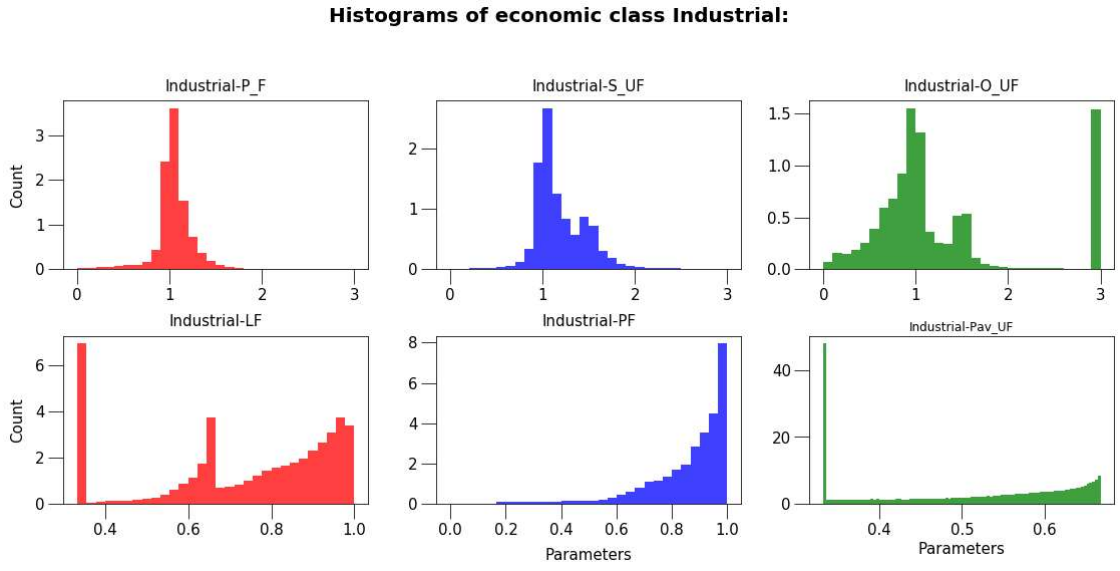


Figure 15: Distribution plots of parameters in the industrial class

The statistics for the industrial class are presented in Table 7. The observations made in respect of the other classes stated above also apply to this class, that is, the underlying distribution may be a beta distribution. However, further studies are required to infer and draw conclusions on the distribution of the industrial class.

Table 7: Summary statistics for the industrial class

	P_UF	S_UF	O_UF	LF	PF
Count	143227.0	143227.0	143227.0	143227.0	143227.0
Mean	0.719	1.003	1.277	0.736	0.858
Std	0.515	0.501	0.813	0.218	0.159
Min	0.0	0.0	0.0	0.333	0.0
25%	0.0	0.954	0.802	0.623	0.799
50%	0.988	1.053	1.002	0.797	0.909
75%	1.073	1.292	1.486	0.925	0.97
Max	3.0	3.0	3.0	1.0	1.0

4.2.2. Exogenous parameter: Seasonality and trend cycle

This section presents an analysis of the customer profiles from different customer classes to justify the proposal to consider the exogenous factors. Scatter plots were used to plot each consumption point and the (time of day in hours). Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another. In this section, scatter plots are drawn to show how much is the energy usage level affected by time of day. These plots demonstrated the shape of the profiles and the differences in the levels of energy usage. Each class was analysed as follows:

4.2.2.1. Residential class

Figure 16 illustrates on the left-hand side are the scatter plot of residential demands against time of day (in hours) for all the days in a year and the corresponding density plots are plotted on the right-hand side. In the residential classes, as depicted in Figure 16 below, there were differences in the level of consumption for the different TOU intervals. During peak time periods the consumption levels were relatively higher, as indicated by the red dots, while off-peak is represented by the blue dots, the opposite was observed during the off-peak periods. The green dots and the green density plot represent consumption in standard time. The defined TOU periods were not aligned to all the hours of high consumption. The red dots started at 17:00 and ended by 20:00 with higher energy usage evident from 19:00 to 20:00. Nevertheless, these did capture the greater part of high consumption. Accordingly, it is evident that these time segments constituted the necessary parameters for the load models. These observations supported the relevance of the proposition related to the exogenous parameters made in chapter 3.

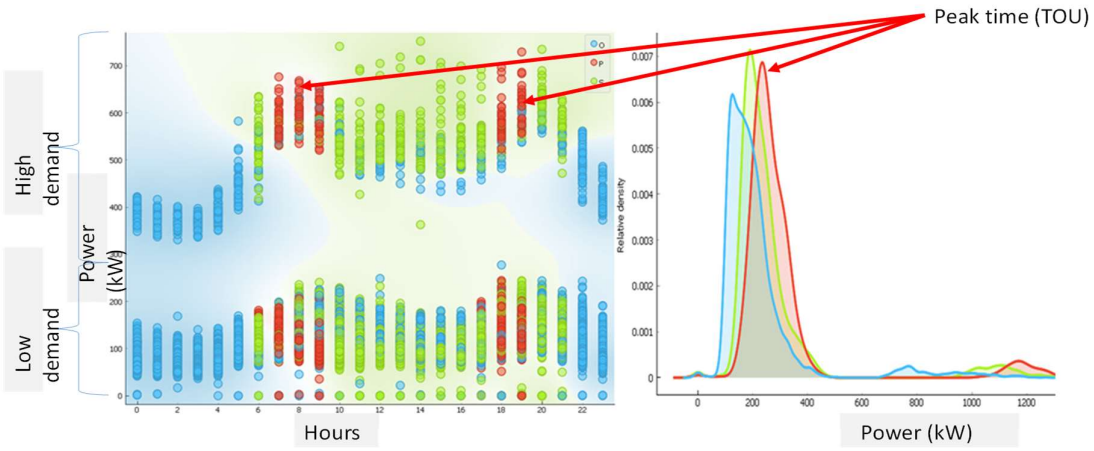


Figure 16: Scatter plots and density plots of the typical residential class plotted using hourly data for 1 year to indicate how energy is affected by time of day.

4.2.2.2. Commercial class

Figure 17 presents the profiles of a commercial customer plotted using the measurements of a randomly selected commercial customer. The profiles indicated that the commercial class customers reached their peak consumption around midday. The operating times of the majority of office-based commercial type-loads were usually from seven o'clock in the morning. As the office staff arrived at work, the consumption increased as air-conditioners, lighting and other office equipment were being switched on. Accordingly, any peak-usage control mechanism may have a minimal impact on these customers. These customers were also likely to benefit from power generated by solar-based renewable sources, such as Solar PV. In the countries where there are liberal electricity trading markets and real-time prices, these customers were likely going to pay an even smaller tariff on average as most of the supply from renewable energy sources, PV in particular, would occur when the system demand was relatively low. This could be an incentive for them to install PV plants. The histogram also indicated the standard period as being ahead of the peak and off-peak periods, thus signalling high consumption levels during the standard TOU periods. The seasonal differences were non-existent in respect of this economic class.

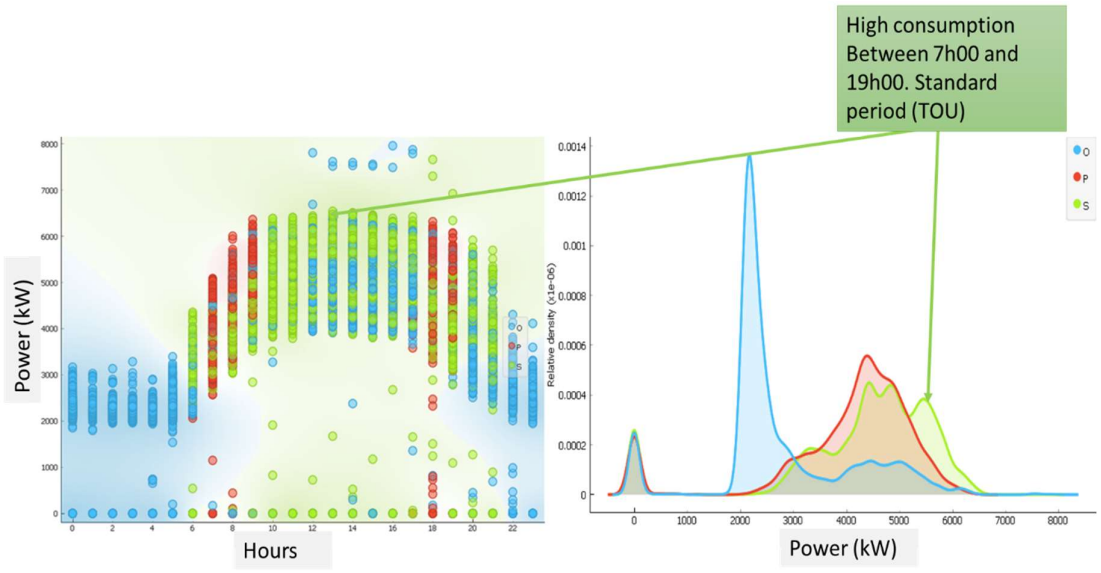


Figure 17: Scatter plot and density plot of a commercial class for a year

4.2.2.3. *Agricultural class*

Like the commercial class, the agriculture class also peaked during the daytime – see Figure 18. It is worth noting that the energy usage of this customer class was higher during the morning peak and lower during the evening peak. Generally, as was reflected by the higher morning peak, farmers tended to start working early and finish early. The histograms of the TOU on the right-hand side of Figure 18 indicated that the peak and the standard consumptions being higher. The larger density was during the off-peak periods, followed by the peak periods. The farmer also tended to use water pumps during off-peak and standard times. The customers in this class would not look at an increase in the standard and morning-peak tariffs favourably as this would significantly increase their costs.

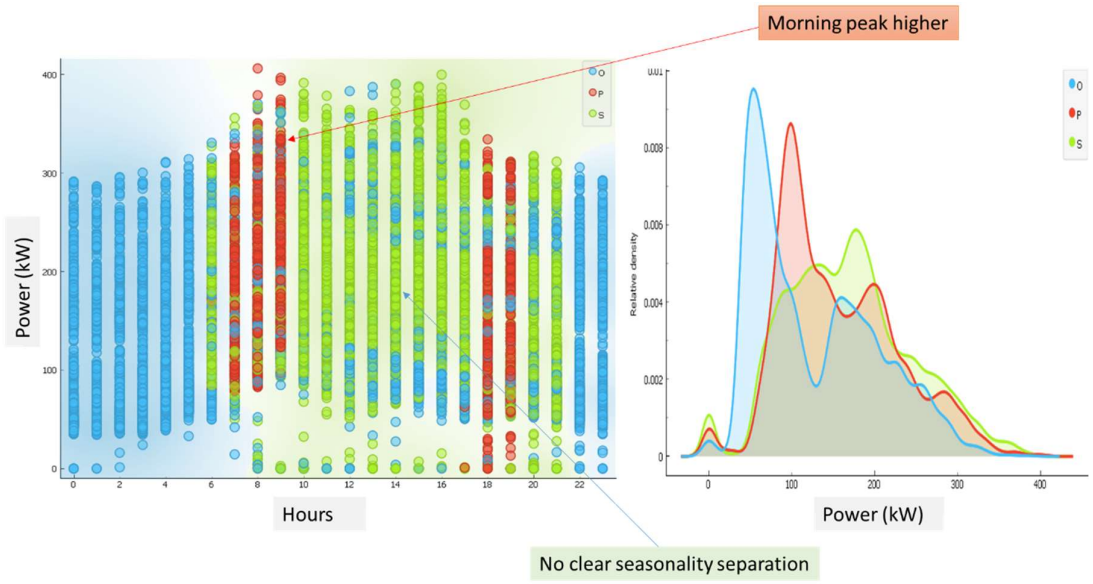


Figure 18: Scatter plot and density plot of the agricultural class for a year

4.2.2.4. *Industrial (Automotive and other manufacturing)*

The industrial sector tended to use more energy on weekdays than over the weekend. This is shown in Figure 19, which depicts a scatter plot of the industrial class that is dominated by the automotive and manufacturing industries. The scatter plot was colour coded to distinguish between the various days of the week. The blue (dark) colour represented weekdays (Monday to Friday) and the lighter colours weekends. There was low energy usage during the morning peak although this increased during the daytime (standard hours).

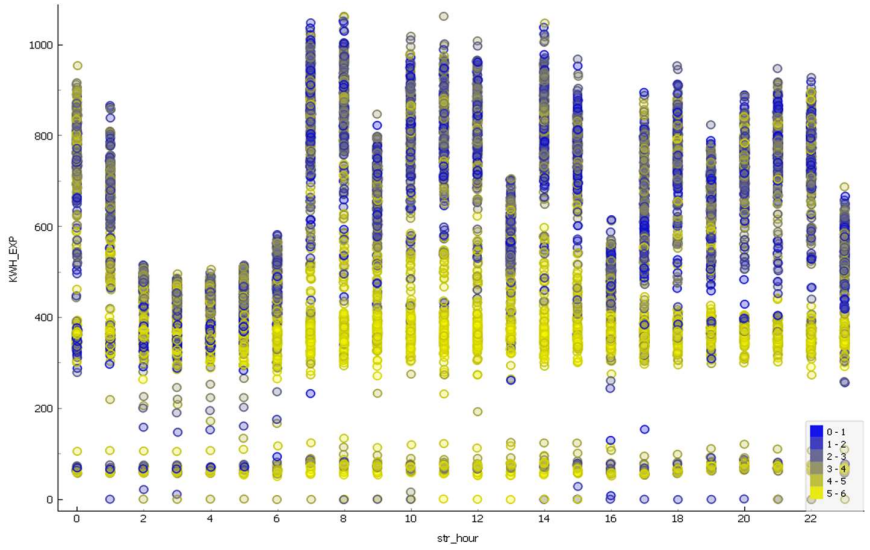


Figure 19: Scatter plot of an industrial (Automotive) class indicating weekdays for a year

4.2.2.5. Industrial (Smelters and mineral processing)

Figure 20 depicts the scatter plot of another group of industrial plant (smelters and mineral processing) with different patterns. Two profiles were visible in the plot, indicating the presence of at least two activity classes. In this case, the activity types are mining, with profiles similar to those of pure mining loads, and smelter and mineral processing plants that demonstrated a flatter profile.

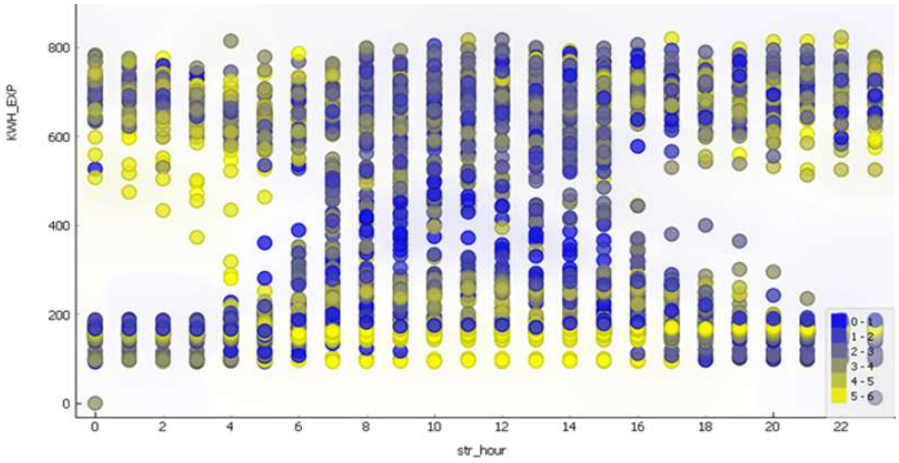


Figure 20: Scatter plot of an industrial feeder (Smelter and mineral processing) indicating weekdays

4.2.2.6. Mining

Figure 19 presents the profile taken from a feeder that supplies the mining loads. It was observed that that the mining sector appeared to be operating mainly between 07:00 and 19:00. It was also evident that there were fewer activities during weekends as indicated by yellow dots.

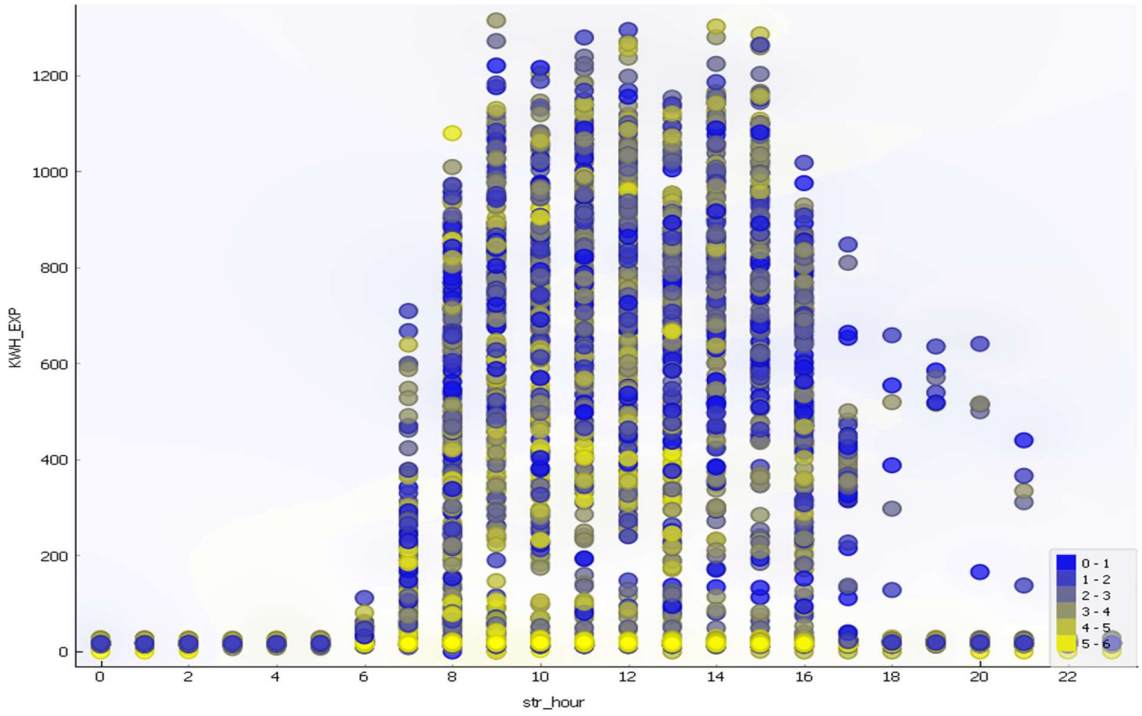


Figure 21: Scatter plot of a mining feeder indicating weekdays for a year

4.2.3. kW and kVAr relationship of loads in different economic activity classes

The analysis of the kW and kVAr relationship may help to further distinguish between loads or customers. For example, some customers (with larger induction loads) tended to have poorer load factors when their load increased. The Reactive energy charges raised by utilities were developed from the basis of recognising that during high load periods, the power factors decreased and hence the charges were raised. Often a threshold was specified and customers were encouraged to operate above it. Eskom's threshold was 0.96 and other utilities used 0.95, and this applied only during peak and standard periods.

The measurements on active and reactive power for both import and export lag and lead were obtained for MV loads. Figure 22 shows the plots from the measurements of agricultural and commercial loads. Loads were represented as exports on the database and therefore the kWh_Exp field from the data was used. The reactive energy for the lagging power factor was represented by the kVAr_Lag_Exp for loads. There was a positive correlation observed between the active and reactive energy for the loads in the agricultural class and those in the commercial class. This means that as the load increased, the reactive power also increased. It can also be seen from Figure 6 that power factors varied with the time of the day. This implies that when the load increased, which often happened during certain times of the day, a different power factor would result. This observation was aligned to the tariff regimes that had reactive energy charges that were linked to a time of use intervals, in particular, the high usage periods such as standard and peak time.

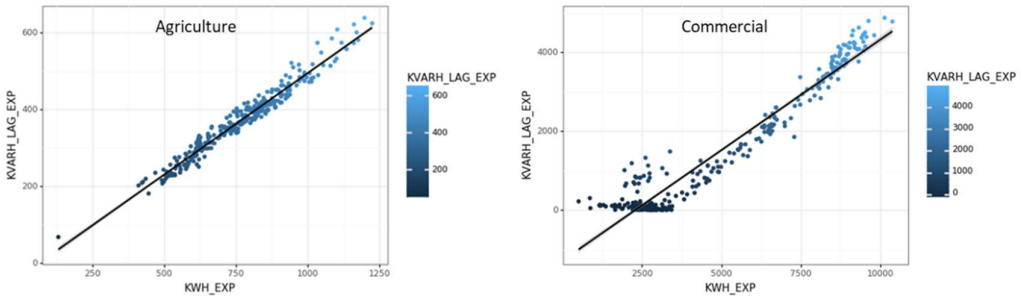


Figure 22: Scatter plot of active (kWh) and reactive (kVAr-lag) energy for agriculture and commercial loads

The measurement of bulk/distributors revealed some negative correlation between the active and reactive energy. Although the correlation was not linear, the pattern was evident from Figure 23. The bulk/distributors comprised predominantly residential and small commercial and industrial loads. The relationship between the active and reactive energy revealed that the power factor improved as the load increased.

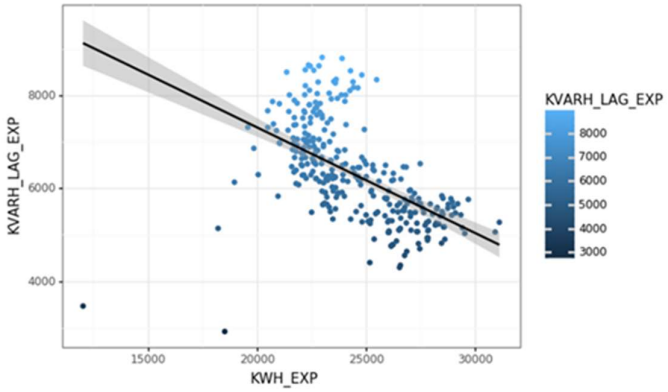


Figure 23: Scatter plot of kWh) and kVAr-lag energy for bulk/distributor loads

A slightly weaker correlation was also observed between the active and reactive power of bulk/distributors loads. As the load become larger, the power factor appears to improve. A logical explanation for this relationship is because the load generally increased during peak periods, where residential customers were generally using more energy. This was confirmed by the scatter plot in Figure 16.

A positive correlation was also observed between the active and reactive powers of the large industrial and mining loads – see Figure 24. It is noticeable that as the load became larger, the power factor was poorer. The mining power factor (on the right-hand side) was worse than the industrial loads at higher consumption levels. There may be a potential benefit for both the industrial and mining customers if they could improve their power factors. The mining loads often comprised large induction motors for mine winders. Since these large induction motors start and stop, as the mine winders ascend and descend, the power factors vary during these processes. There are often variable speed drives and soft starters that are installed to limit the input current during the start and stop the process, but the power factors are still impacted as induction motors draw reactive energy during the process. Industrial furnaces in the industrial sector tend to consume the largest of the loads. The furnaces could also be the reason for the lower power factor during high loading periods.

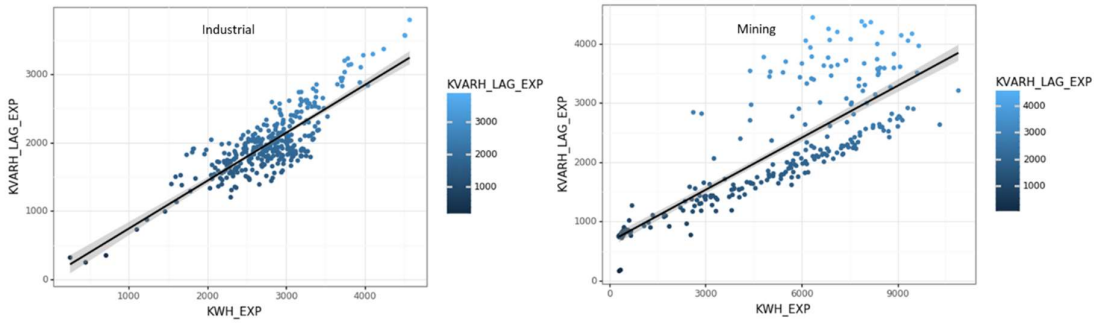


Figure 24: Scatter plot of kWh and kVAr-lag energy for large industrial and mining loads

4.2.4. Time series decomposition: Trend cycle analysis

Two directions underlie the direction of the development of the decomposition method. First, it is the recognition that there may be a correlation between the variables, and that correlation must be eliminated. Second, the elements of economic activity can be separated such that the cycle could be isolated from the seasonal pattern and other random events. In this section, the time-series of the customer data was decomposed to determine the trend cycle for each economic-activity class in the sample data. This is to ensure that the trend cycles that represent all the customers were capture. The model could be an additive or a multiplicative decomposition method depending on whether there is an increasing \decreasing trend or not, with the latter calling for the additive model. To get the trend a seven Moving Average (7MA) smoother was used. This was informed by the knowledge that the data had 7 days weekly cycles. The smoothness or absence of patterns in the residual plots indicated that all patterns could have been removed. For monthly trend analysis, a 12-month centred moving average (12 MMA) was used. Additive decomposition was appropriate for the data used because the trend was constant. For all the plots in this section, the top graph depicts the load measurement volumes, followed by trend, seasonal and residual plots respectively. The x-axes of all the following plots indicate the period, for the annual trend the axis represents day number (1 to 365) and for the annual analysis it represents the month number (1 to 12).

4.2.4.1. Agriculture class

Agriculture showed stationary trend data around the mean. The trend cycle seemed to be repeating and was constant as seen on the trend plot of Figure 25. The monthly patterns were also explored since most of the seasonal parameters could only be seen at the level of the monthly analysis. The resulting plots are shown in Figure 26. The results indicated that there were repeating cycles, a seen on the trend plot. The results also showed an increasing trend between the second month and the fourth month but remained constant from then onwards. Agricultural sector load tended to increase in line with the seasonal changes as it got colder and there was less rainfall. This meant that there were three possible cycles or seasonal patterns annually. The lack of pattern in the residual plot signified that the trend cycles were well captured, that none of the underlying patterns in the data was missed.

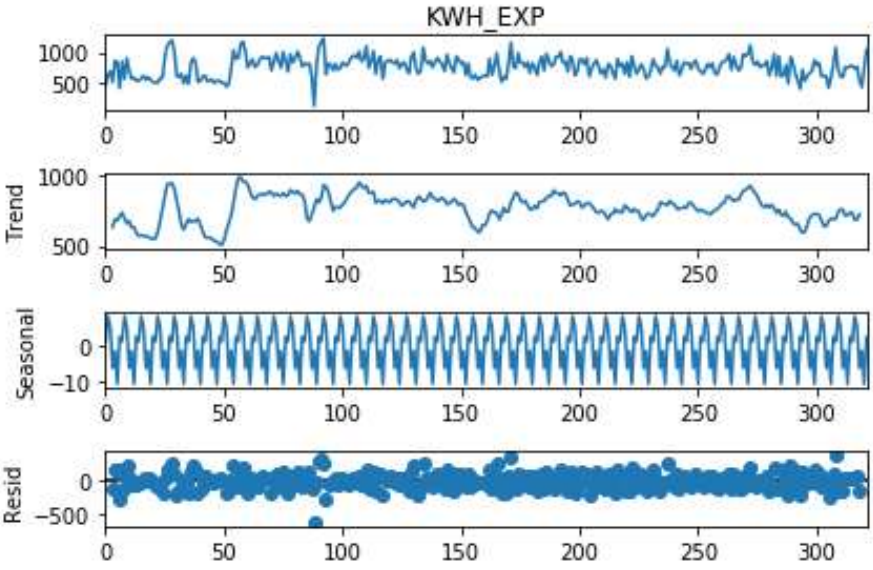


Figure 25: Daily trend cycle analysis of agricultural class

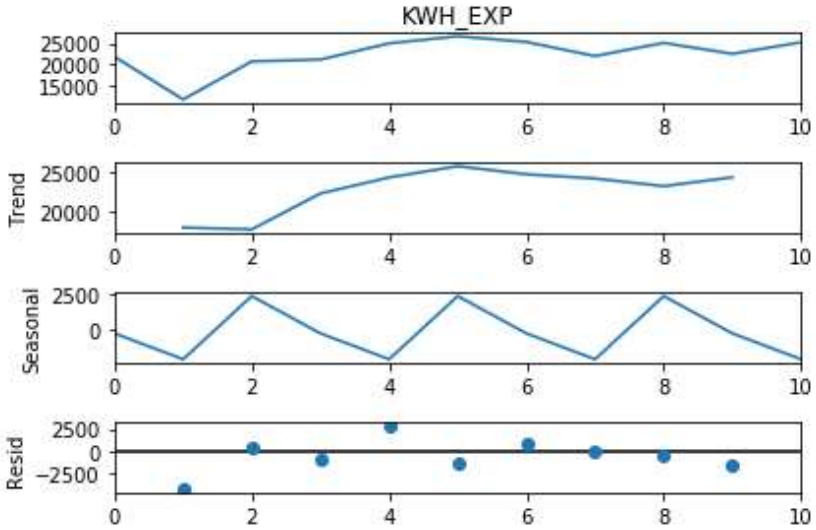


Figure 26: Annual trend cycle analysis for Agriculture sector

4.2.4.2. Bulk\Redistributors

According to the results, the Bulk\Redistributors responded to winter and summer seasonality because this class contained almost all the load mix. Residential customers profiles tended to dominate on the Bulk supplies, while distributors would have varying percentages of mixed customers depending on the area. Some distributors, especially small metros comprised predominantly residential customers, and this increased their response to seasonal trends. This can be seen by an upwards trend in the period of colder months and a downward trend in the months when the temperatures increase. The Bulk\Distributor class also indicated that there may be three seasonal cycles as opposed to the known two (winter and summer) as seen in Figure 27 and in more pronounced plots in Figure 28.

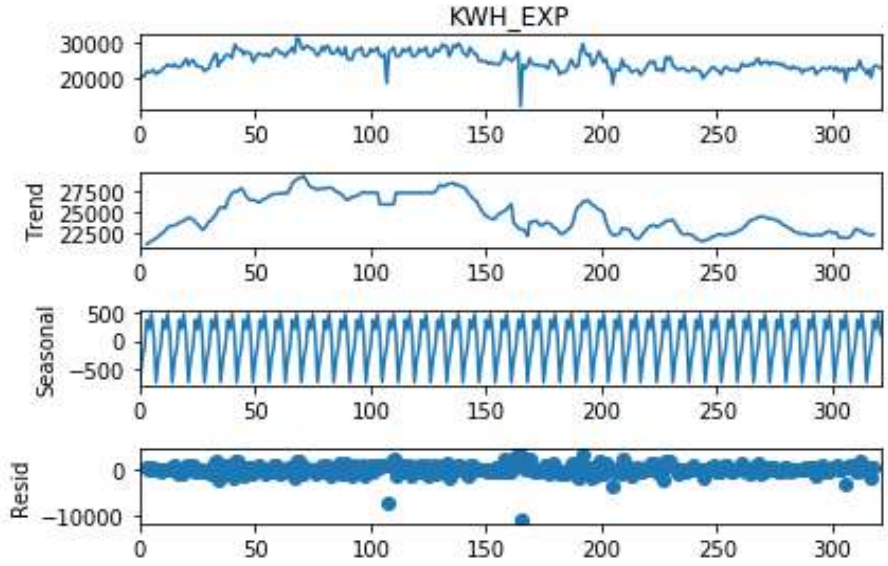


Figure 27: Daily trend cycle analysis for Bulk\Distributors

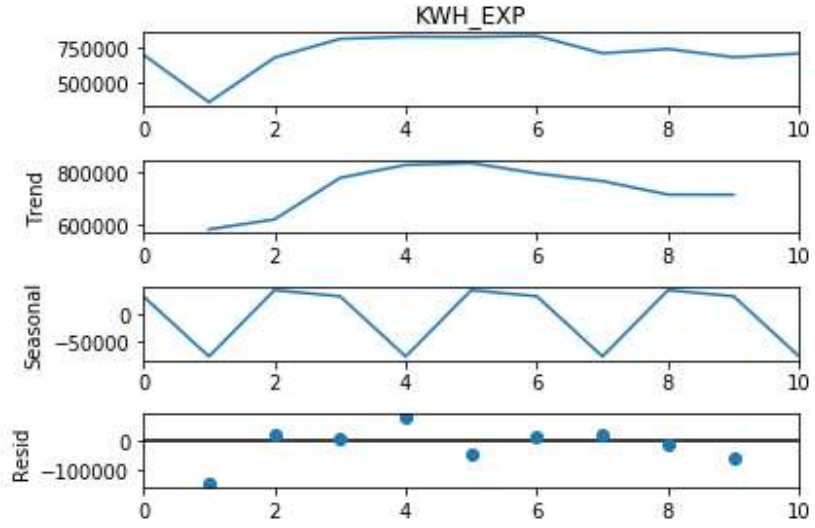


Figure 28: Annual trend cycle analysis for Bulk\Distributors

4.2.4.3. Commercial loads

Commercial also responded to winter summer seasonality. This can be seen by an upwards trend in the winter months and a downward trend in the months when the temperatures increased. There was also a possibility of a weekly cycle from the results. The commercial class results indicated a negative trend towards the end of the year (from months 5 to 12).

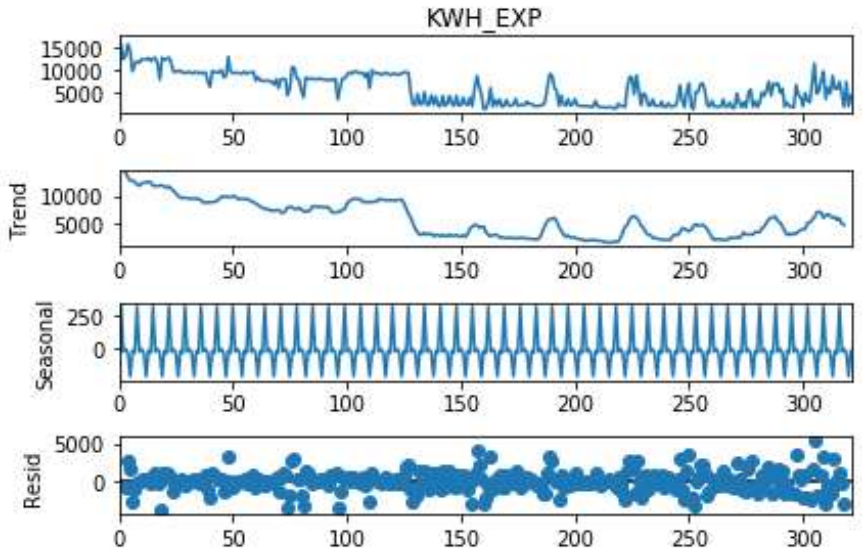


Figure 29: Daily trend cycle analysis for commercial customer class

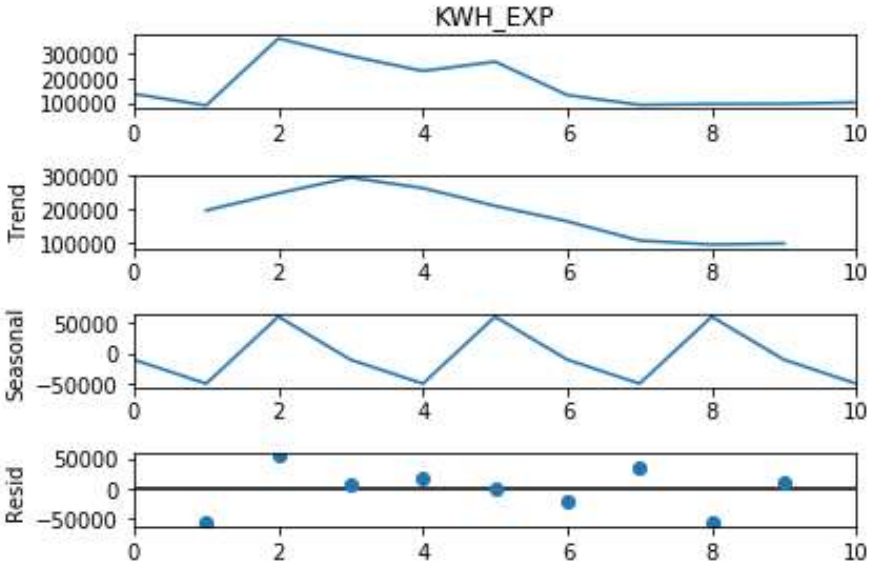


Figure 30: Annual trend cycle analysis for commercial customer class

4.2.4.4. Large Industrial loads

The large Industrial class had stationary trend data around the mean. From Figure 31 it can be deduced that the larger industrial loads were stable and did not respond to daily seasonal changes. There may have been a presence of seasonality on a weekday. The annual trend analysis indicated that there were three seasonal cycles annually as shown in Figure 32.

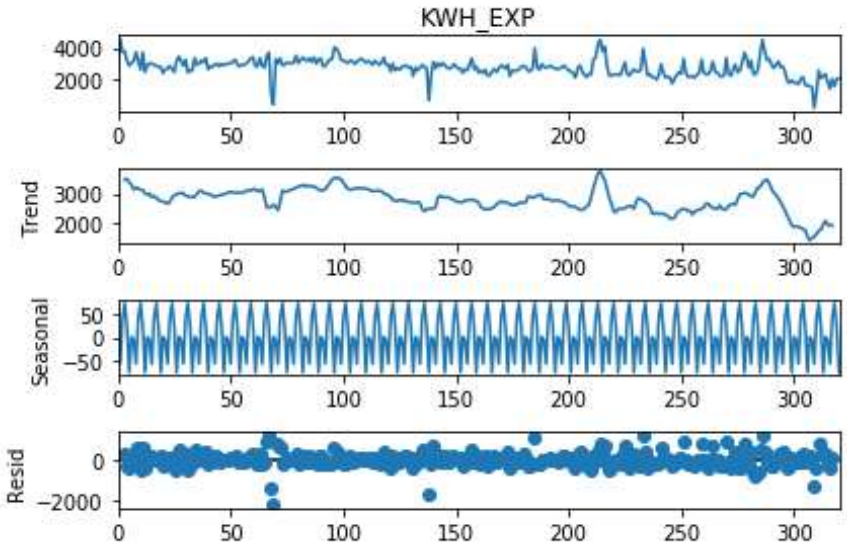


Figure 31: Daily trend cycle analysis for large industrial customer

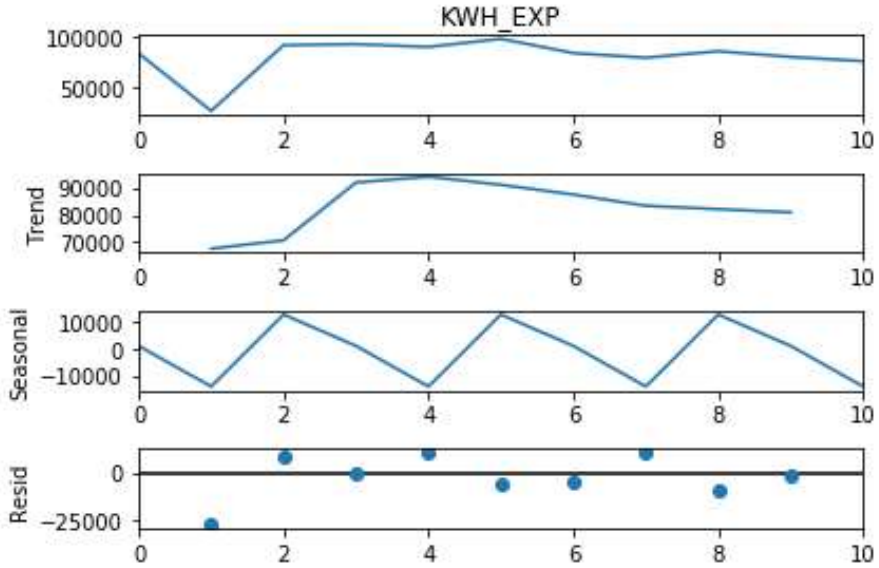


Figure 32: Annual trend cycle analysis for large industrial customer class

4.2.4.5. Small Industrial

The analysis results showed that Small Industrial responded to winter and summer seasonality. This can be seen by an upwards trend in winter months and a downward trend in the months when the temperatures increased as shown in Figure 33. This trend w more pronounced in the annual trend cycle plots in Figure 34. It is also worth noting that there were also three seasons annually suggested by the results from the monthly trend cycle analysis.

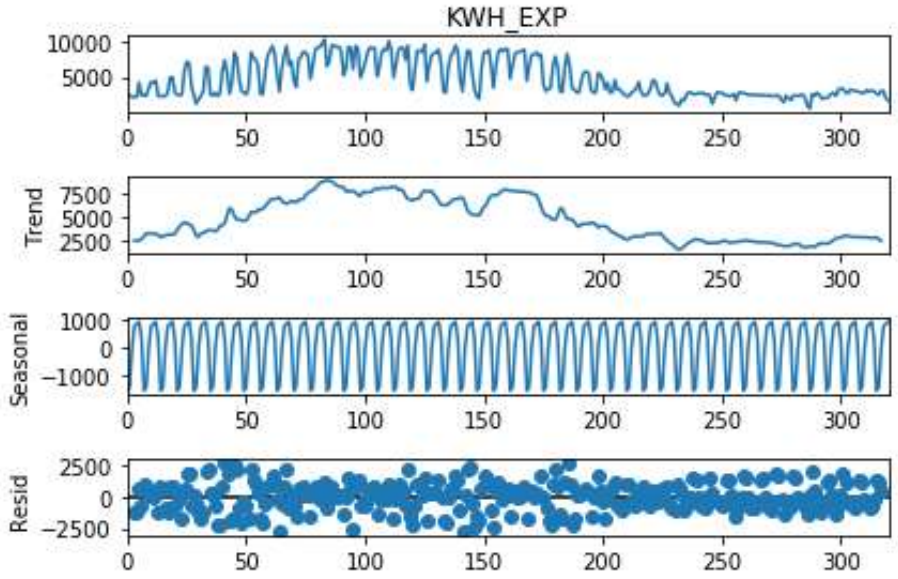


Figure 33: Daily trend cycle analysis for small industrial customer class

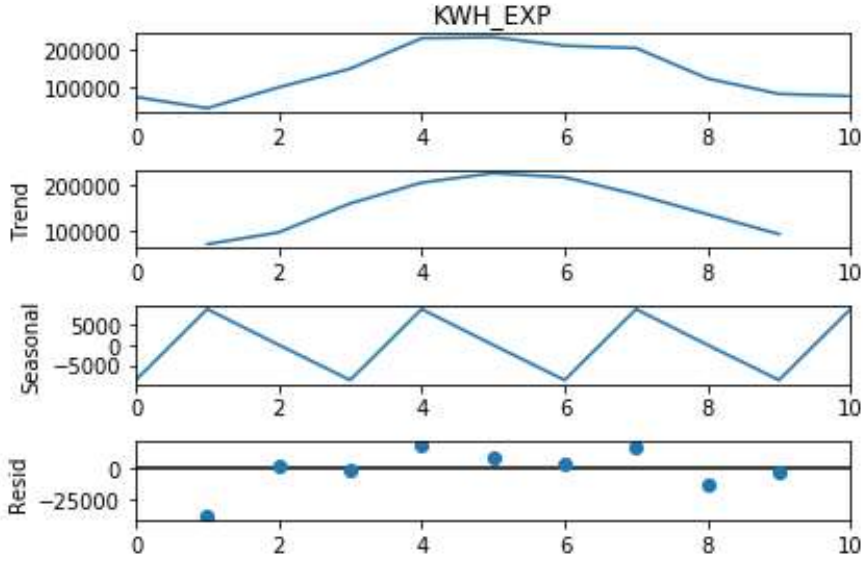


Figure 34: Annual trend cycle analysis for small industrial customer class

4.2.4.6. Mining sector

Mining had stationary trend data, although the results showed that activities were picking up at the beginning of the year and slowing down towards year-end. Seasonal patterns existed in the mining sector data. No patterns were seen on residuals, which meant that the remainder was random noise. See Figure 35 and Figure 36.

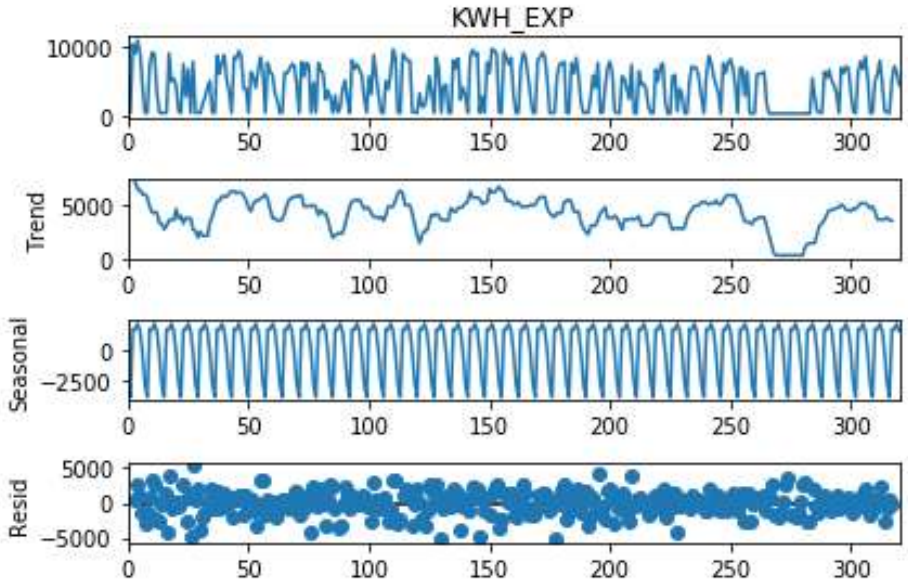


Figure 35: Daily trend cycle analysis plot for mining customer class

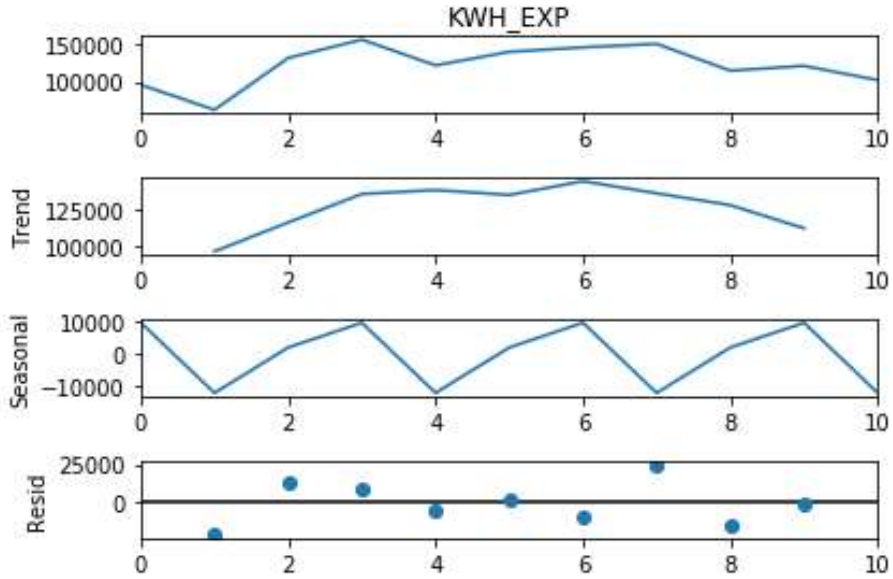


Figure 36: Annual trend cycle analysis for mining customer class

4.3. Concluding remarks

Stratified sampling simplifies and improves the modelling speed. However, it is important to understand the structure of the measurements database for the successful application of this sampling procedure to define the strata well. Statistical tools are useful in the exploratory analysis of customer measurements data. Statistical tools such as box plots and distribution plots as well as numerical summaries may be used on the data to detect outliers and to determine the best estimators of the parameters. Statistical tools were used on the load measurement to estimate the values of the endogenous parameters proposed for load modelling in this study. The box plots were able to demonstrate where the median was in relation to the minimum, maximum and other quartiles. In addition, the box plot also indicated the possibility of the data being skewed and that there were outliers in the data.

Different customer classes have different load profiles. EDA analysis results indicated that the various classes of customers exhibited different consumption behaviours linked to both the time of consumption and seasonality, but that not all classes nor all customers in a class responded to seasonal changes in the same way. It was also established that most customer classes respond to seasonal cycles. The analysis of trend cycles suggested three seasonal cycles, however, only two seasons are defined in the context of South Africa. The seasons are still relevant looking at how the consumption tends to increase towards the winter months and reduce thereafter.

5. TYPICAL DAY MODELLING

This chapter discusses the results of the typical day load model. The goal of the typical day load model is to represent the annual data in terms of typical day averages. This modelling is used mainly in modelling where there is a need to recognise that the energy usage varies for different days of the year and that some of the days may follow the same pattern. Modelling each of the 365 days individually can be a computationally expensive and time-consuming task and therefore creating typical day classes can be a useful solution.

5.1. Selecting parameters of a typical day load model

The interest in the load model for typical days was on grouping similar days based on the profiles. Therefore, the chronological load profiles of half-hourly energy data were used. In a typical day model, the interest was on the variations of loads with time in a twenty-four day. Therefore, each of the hours was be assumed to be a parameter, and this resulted in 24 parameters for each day. The next step was to arrange data with hours in columns and days in rows. Arranging data in this way resulted in a data frame of 365 rows x 24 columns. Figure 37 shows the normalised daily profiles for a year plotted in a three-dimensional (3D) graph. The x-axis represents the hours of the day (sum of Hour_0 to Sum of Hour 23), the y-axis indicates the per-unit normalised values and the z-axis shows different dates. To draw any plot of the 365 x 24 data frame is not possible with the current systems, because they were mainly limited to 3D. The aim was to visualise the data frame to be able to group the days to the different clusters that contain similar weekdays.

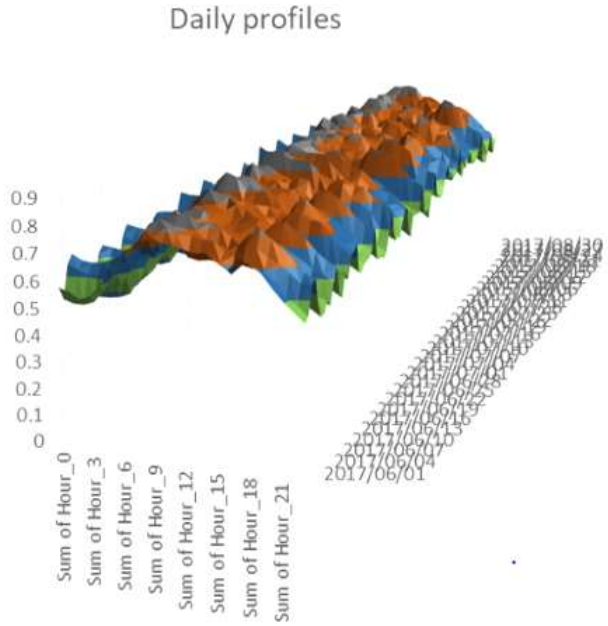


Figure 37: 24 hour load profiles for the different days of the year

As mentioned in chapters two and three, PCA was found to be useful in projecting multi-dimensional vectors into equivalent two or three-dimensional space. In this case, the interest is in compressing the 24 dimensions for ease of visualising and clustering. The PCA algorithm was used based on the reasons provided in chapter 3 and the results are the principal components (PC) of each of the days. Table 7 shows a snap view of the typical results from the PCA algorithm. For each day, the results are the PC values that represent the variability of that day. The results of PCA were obtained and analysed by province. The results may be summarised as follows: for the Western Cape, the first three PCs accounted for 92% of the data variability, 90% for North West and Limpopo, 88% for Gauteng, 80% for Free State, Eastern Cape and Northern Cape respectively while, for Mpumalanga, two PCs accounted for 60% only. Three PCs achieved a representation of over 75% for 24-hour profiles nationally.

Table 8: Typical results from the PCA algorithm

Date	PC1	PC2	PC3
2017/04/01	0.046749199	-0.477178438	0.023398248
2017/04/02	0.759487482	-0.219967282	0.004175609
2017/04/03	-0.475730766	0.189751754	-0.320280039
2017/04/04	-0.236674037	-0.022485317	-0.012705263
2017/04/05	-0.267055686	-0.046469197	-0.127159048
2017/04/06	0.048031862	-0.181050952	0.004411806
2017/04/07	-0.48848711	0.07793519	-0.16018456
…	…	…	…
2018/03/31	0.932949357	-0.186691705	0.043460656

5.2. Classifying the days

The PCA provided unique parameter values for each day. Using these parameter values, the days could be classified into different classes to create typical days. K means clustering algorithm was used for classifying the days as explained in chapter 3.

5.2.1. K-means clustering results

Since three PCs adequately described the daily variations, three-dimensional scatter plots were used to visualise and apply in the clustering algorithm. For the first iteration, the number of clusters was set to be twenty. As stated in chapter 3, the first iteration is used to inform the number of clusters to choose and initialise the procedure, since the k-means algorithm requires this number to be determined a-priori.

5.2.1.1. Iteration 1: 20 Clusters

The load measurement data were separated into three months of winter and nine months of summer seasons. The national results were of interest to this study and, thus, this chapter presents the typical day result of the national sample. The number of clusters was set to the highest possible number, in this case, twenty (20). The k-means algorithm was then used to cluster the loads based on the principal components into 20 clusters. After obtaining the results showing which days belonged to which clusters, the measurements were labelled accordingly and the days were classified. The results from the algorithm were also used to calculate the different accuracy measures and the results are explained next.

A) Silhouette score

Figure 38 presents the silhouette score diagram and the scatter plots for 20 clusters with each cluster being represented with a different colour. The PCs are represented on the scatter plot on the right-hand side of Figure 38 as PC1, PC2 and PC3 on the three axes of the diagram. The silhouette plot is on the left-hand side of the figure, while the x-axis of this plot indicate the score and the y-axis represents the number of clusters. The numbering of the clusters is zero-based. It is possible to make the following observations from these plots:

a) If the number of clusters was large, the clusters would overlap, and could not be identified.
b) The silhouette score was very low, approximately 0.3, thus these low silhouette scores signified that the boundaries between the clusters were not clear.
c) As was indicated by the negative silhouettes scores observed in clusters 0, 1, 5 and 6, some of the customers (profiles) may have been allocated to incorrect clusters, as the number of clusters increased beyond the optimal number.

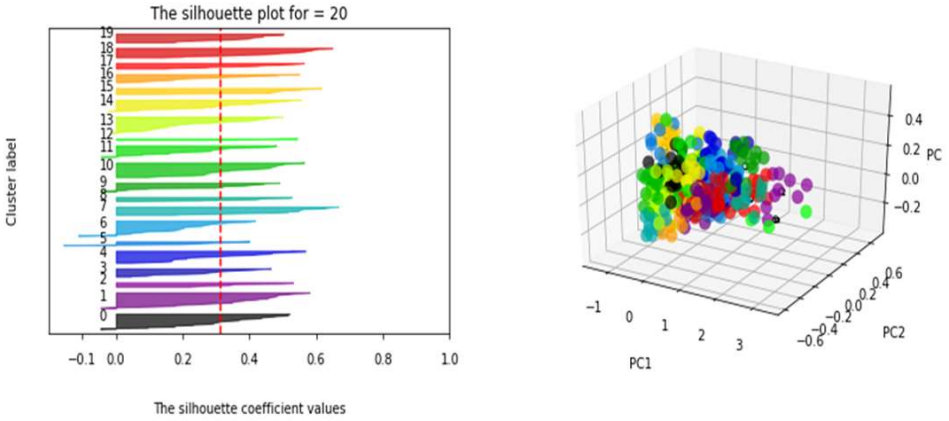


Figure 38: The silhouette plot and the 3D scatter plots showing 20 clusters

B) The elbow diagram

The results of the elbow diagram for the national sample highlighted that the number of clusters required for the typical day model should be on or below seven. This was indicated by the elbow saturation points in Figure 39. These were the points at which the increase in the number of clusters had a relatively minor impact on the SSE (see the shaded block). The saturation point started at three. Discretion, based on experience and the results from other cluster adequacy measures, was necessary to ensure that it was possible to select clusters optimally.

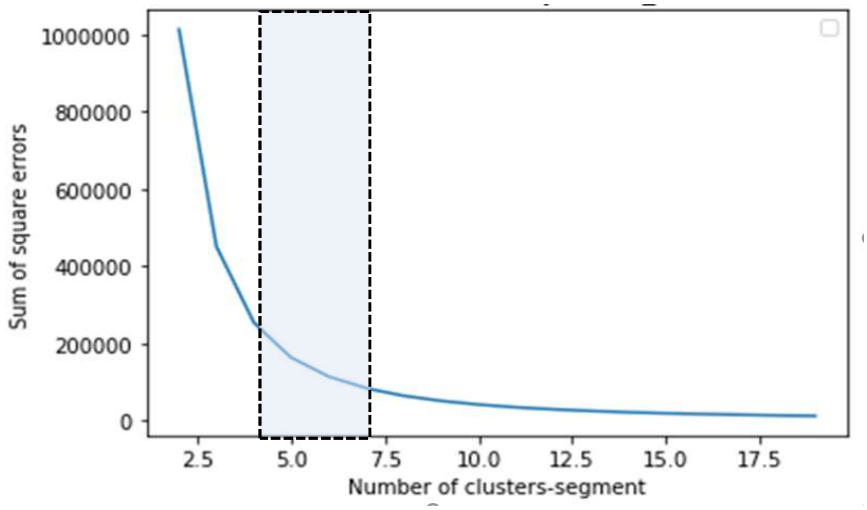


Figure 39: Elbow diagram showing the possible optimal number of clusters for typical days.

C) DBI results

The DBI results are presented in Figure 40. The DBI was the lowest for two clusters with the number increasing as the number of clusters increased, thus indicating a loss of compactness in the clusters as the number of clusters increased. The clusters were still compact in the shaded area of the DBI plot as the DBI scores were still lower, although they increase rapidly as the clusters increase. The next lower DBI was for three clusters.

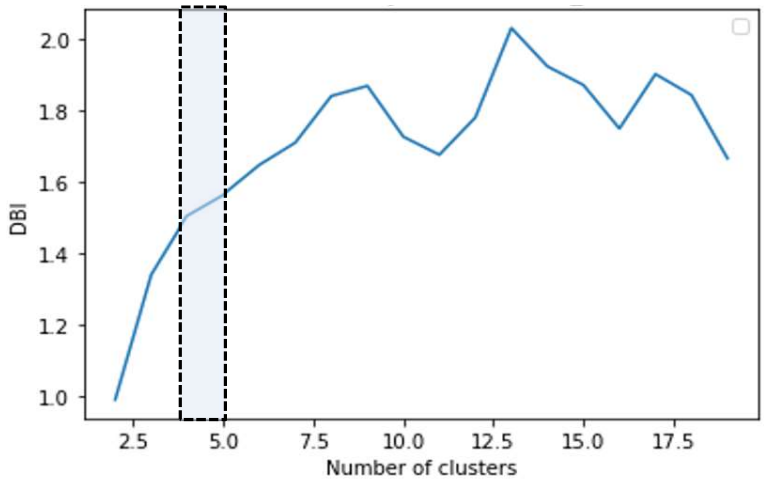


Figure 40: The DBI plot for 20 clusters

5.2.1.2. Iteration 2: Two clusters

It was necessary to evaluate the clustering results for the 20 iterations to see if these numbers of clusters are sensible. For iteration two (2) and the other iterations that follow (three and four) the results of the silhouettes and scatter plots only are presented. The results for iteration two, the number of clusters were pre-set to two as shown in Figure 41 below. The elbow diagram and the DBI plot reflected the lower scores for the two clusters already as seen in Figure 39 and Figure 40 above.

A) The silhouette score

Figure 41 depicts the silhouettes scores, which are indicated by the dotted line on the silhouette plot. As can be seen from the scatter plot on the right-hand side above, it is possible to separate the two clusters. The silhouettes score is above 0.5. Referring back to the elbow diagram in Figure 39, the results indicate that there should be more clusters as the saturation area starts only when there are three clusters. Therefore, it would appear that the two clusters are not adequate from this perspective. Personal judgement and experience are necessary for deciding whether the two clusters should be considered for the results.

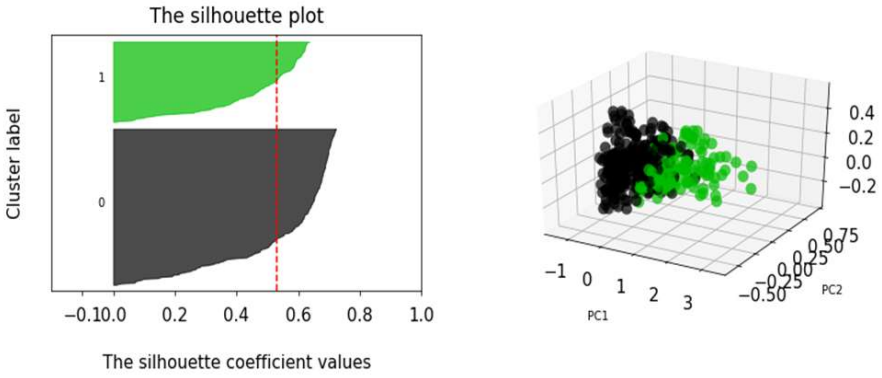


Figure 41: Silhouette and scatter plots for 2 clusters

5.2.1.3. Iteration 3: Three clusters

The silhouettes and scatter plots for clustering results where the number of clusters is set to there (3) are presented in Figure 42. The silhouettes plot showed that three clusters demonstrated no element of overlapping because all

values are above zero. The scatter plot showed the three clusters using colour codes that match the clusters in the silhouettes plots. The red line in the plot indicated a silhouettes score of approximately 0.5.

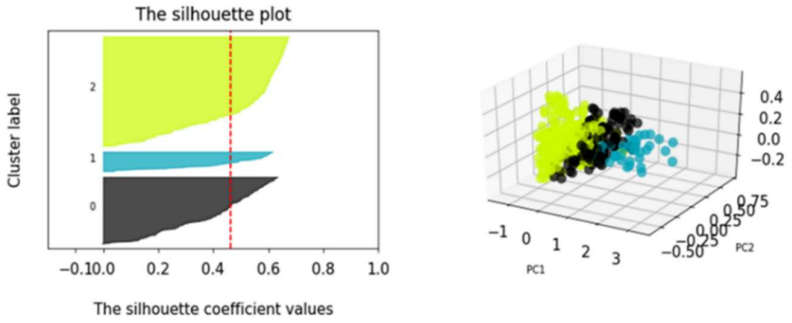


Figure 42: Silhouettes and scatter plots for 3 clusters – National data set

The silhouette score of the three clusters is lower than that of the two clusters in the previous iteration. It is clear that, as the number of clusters increases, the silhouette score decreases. Although the clusters were extremely close to one another, there was, nevertheless, still separation as was indicated by the absence of non-negative scores in the silhouette plot. The scatter plot indicated that the boundaries were very close and likely to overlap if more clusters were to be added.

5.2.1.4. Iteration 4: Four clusters

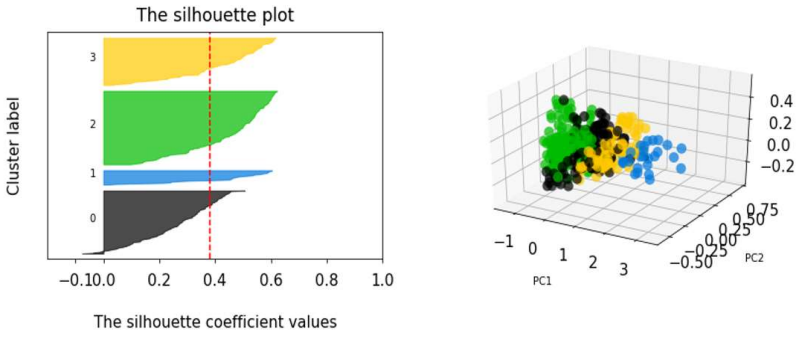


Figure 43: Silhouettes and scatter plots for four clusters

When the number of clusters was increased beyond three, the problem of allocating profiles (days) to wrong clusters arose. Figure 43 presents the silhouette and scatter plots when the number of clusters was increased to four. It can be observed that there was a degree of overlap, with cluster members (load profiles) being far from the clusters that they ought to be in. The silhouette plot in Figure 43 shows that cluster zero went to negative, thus indicating that some of its members were in the incorrect clusters. This problem increased as the number of clusters were increased further, indicating that the results would be invalid beyond three clusters. Therefore, it was not necessary to evaluate further increases in the number of clusters as the overlapping and incorrect allocation of days to incorrect clusters would worsen.

The silhouette score was the highest at two, three and four. The average silhouette score is 0.62 for two clusters, 0.59 for three clusters and then 0.57 and 0.55 for four and five clusters respectively. This indicated that clusters with internal cohesion should be approximately three or four if the elbow point between three and seven was considered and the DBI was still low at this point. The results from the adequacy measures indicated that a lower number of clusters would be sufficient to represent the days for the whole year. While the silhouettes and the DBI

point to two clusters, the elbow pointed to three or more. As stated above, going beyond three results in less optimal clusters and, therefore the conclusion was that three clusters should be used for modelling.

5.3. Allocation of typical days to clusters (Classification)

The results above indicated that the optimal number of clusters was three. We, therefore, performed the second iteration of clustering and labelled the individual days accordingly. That is, all the days that fall into cluster 0, were provided with a zero label, those in cluster 1, were given one label etc. Thereafter the weekdays associated with particular clusters were identified before grouping them, thereby allocating the typical days. Selecting three clusters meant that there were three profiles only to represent the whole year. Thus, there was a need to establish which of the weekdays in a year were associated with each of the clusters so that the cluster profiles could be allocated to these days. Table 8 shows the results of allocating the weekdays to clusters. The table depicts results for both the winter and summer seasons, represented by the letters w and s respectively. In the weekdays' column, number 1 represented Mondays and 7 represented Sundays. The percentages represented the proportions of the weekdays in each of the clusters.

Table 9: The typical days associated with the clusters

		Cluster		
Seasons	**Weekdays**	0	1	2
s	**1**	55%	21%	24%
s	**2**	56%	8%	36%
s	**3**	56%	8%	36%
s	**4**	51%	5%	44%
s	**5**	44%	13%	44%
s	**6**	65%	35%	0%
s	**7**	13%	87%	0%
w	**1**	21%	7%	71%
w	**2**	8%	0%	92%
w	**3**	8%	0%	92%
w	**4**	8%	0%	92%
w	**5**	15%	0%	85%
w	**6**	46%	23%	31%
w	**7**	62%	31%	8%

The results in Table 8 indicated the following based on the relative percentage levels of each day associated with the clusters:

- Summer weekdays, namely, Mondays through to Saturdays, were associated with cluster 0 because their proportions in cluster 0 were higher for this cluster.
- The Sundays in summer were associated with cluster 1.
- The Fridays in summer were also in cluster 2.
- Cluster 2 did not seem to be dominated by any of the days in summer.
- In winter there appeared to be a shift in this arrangement with Mondays to Fridays being strongly linked to cluster 2,
- The Saturdays and Sundays in winter were aligned to cluster 0.
- Cluster 1 appeared to be irrelevant for winter.

The results showed different load profiles for various days and that there was a seasonal impact. This analysis was confirmed by the profiles allocated for each typical day class as shown in Figure 44. The winter profiles for the cluster that were relevant as established in the table above, were winter clusters 0 and 2, indicated by line plots w0 and w2. The summer profiles were represented by s0 and s1 for clusters 0 and 1. The clusters that were not dominated by any profile were considered irrelevant and were excluded from the final selection of clusters

.

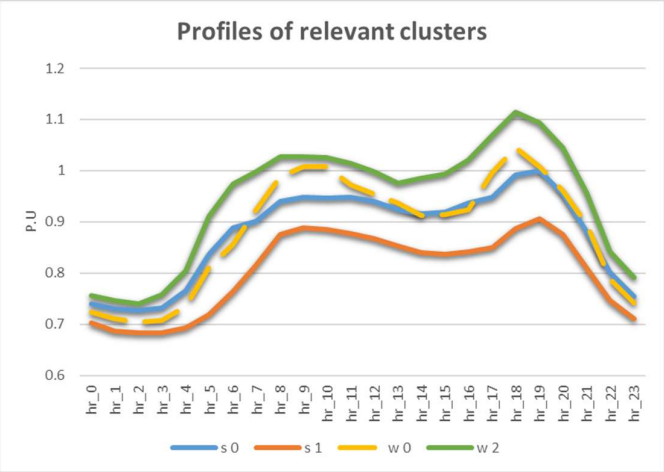


Figure 44: Load profiles allocated to the typical day classes

The typical day cluster presented in Figure 44 indicated that there was only one (1) profile for winter and three (3) for summer. The plot also shows that clusters 1 and 2 in summer are almost equivalent with minor differences only.

From Table 8 it can be suggested that the cluster profile s0 is to be associated with summer weekdays (Monday to Friday), and s1 is summer weekends (Saturday and Sundays). Similarly, for the winter weekdays, w0 profile can be used whereas for the weekends w2 was more appropriate. Associating the load profiles to the weekdays and weekends has been a practice for many utilities. The model result provided a scientific way of determining the load profiles associated with these days. The profiles selected this way are more accurate than those selected through averaging all weekday profiles and all weekend profiles, or those selected only through experience.

5.4. Concluding remarks

The results discussed above illustrated the possibility of modelling typical days using measurement data. Typical day load models reduce the burden of modelling each day of the year separately and improve the modelling speed. The results also indicated that certain days demonstrated similar patterns and, therefore, it was not necessary to model each day on its own. In addition, the results suggested that the weekends were not the same as weekdays and, thus, the distinction made in the existing Eskom models, which differentiate these days, may be valid. Essentially, the year may be represented by four different profiles.

Although the results may be expected to differ with various utilities and selected test periods, they do, nevertheless, provide a sound foundation for reducing the number of daily profiles used in load modelling and, hence, improving the computational speed. Power systems simulation software such as Powerworld and Digsilent require a test horizon to be determined when modelling power flows and related sensitivities for time-varying loads. The load profiles for the corresponding time horizon are inputted to the simulator. This test horizon in modelling and simulations can be reduced from 8760 time points, often used for a single year to 96 representative time points without compromising the quality of the results since these 96 time points are representative of the actual profiles in both summer and winter season. Thus, this is an important benefit of the typical day load model.

The typical day load model can be implemented for different customer classes. However, most studies are not concerned with typical day's classes for individual customers or categories of customers but require the typical day classes for all customers. For example, instead of having different typical days for each customer class, which is likely to complicate the model, it is better to have an informed assumption for all loads being modelled. The results of this model provide the basis for this assumption.

6. CUSTOMER CLASSIFICATION LOAD MODEL

The results presented above are useful in providing a study with a reduced horizon for modelling or power flow simulations. In other words, instead of simulations that span a year, it is possible to use the profiles of typical days. However, it is still necessary to classify the customers based on usage patterns. This section presents the results of the clustering process, which created the customer classes and allocated the profiles to each class. The classification of MV customers or loads was performed using k-means clustering algorithms. The application steps, as summarised in Chapter 3, were followed. This chapter presents the results of each of the key steps in the process for classifying customers. These steps include parameter selection, clustering, classification and determining the representative profile.

6.1. Parameter selection

The parameters (LF, PF, P_UF, S_UF, and O_UF) identified created a dimensionality problem. The concern was not only to find the representative or equivalent variable for the parameters but also to evaluate each parameter in terms of the extent to which it was able to contribute to classifying the customers based on their load patterns. The parameters and their estimated values are depicted in Table 9. The parameters were calculated for each of the customers in the sample. For ease of display, the averages of each parameter are displayed per activity class in Table 6, instead of showing the values for each load. As shown in Table 3, LF is the load factor, indicating that the agricultural loads had a load factor (LF) of 67%, an average power factor (PF) of 0.85 and normalised usage factors for peak (P_UF), standard (S_UF) and off-peak (O_UF) parameters of 0.59, 0.61 and 0.51 respectively.

Table 10: Calculated averages for the parameter displayed per activity class

CLASS	LF	PF	P_UF	S_UF	O_UF
Agriculture	0.67	0.85	0.59	0.61	0.51
Bulk / Distributors	0.71	0.95	0.78	0.72	0.55
Commercial	0.69	0.90	0.64	0.66	0.53
Industrial	0.75	0.84	0.63	0.67	0.58
Internal	0.70	0.97	0.74	0.71	0.60
Mining	0.70	0.80	0.53	0.58	0.50
Grand Total	**0.70**	**0.88**	**0.65**	**0.66**	**0.53**

6.2. Classifying customers

The results are evaluated for two distinct cases. In the first case, a 2D (pairwise) based algorithm was used and each parameter (LF and PF) was paired with the parameter Pav_UF (the normalised average power). The rationale behind creating pairwise entries using parameter F was that this parameter formed the basis of all the parameters with all other factors being subject to it. In the second case, all the parameters were used, represented by their PCs, while pairwise entries are also created for the application in the k-means clustering algorithm. The PCs were determined using the PCA algorithm.

6.2.1. Case 1: Parameters A-E paired with parameter F

The pairs that formed entries into the model are (LF; Pav_UF), (PF; Pav_UF), (P_UF; Pav_UF), (S_UF; Pav_UF), and (O_UF; Pav_UF). The description of what each symbol mean is in Table 3, in chapter 3.

6.2.1.1. Iteration 1: Twenty (20) clusters

As in Chapter 5, 20 clusters were chosen for the first iteration. It was expected that the resulting number of clusters would be below 20 and, hence, the selection of 20 for the first iteration. The following cluster validity measures were analysed after using the k-means algorithm.

A) The silhouette score

Figure 45 and Figure 46 below present the results of 20 clusters in the form of silhouette plots on the left, and scatter plots showing the clusters on the right. Different colours and the circled numbers distinguish the clusters. The colours and numbers of the scatter plots correspond with the colours and the numbers of the silhouette plots. Clusters without clear separation overlapped on the scatter plots while their corresponding silhouettes demonstrated a negative tail.

Cluster separation became a challenge when there were too many clusters. This could be seen in the scatter plots on the right-hand sides of Figure 45 and Figure 46 where the clusters were overlapping, especially where the data density was higher. The overlapping was also confirmed by the silhouette diagrams on the left-hand side of these figures.

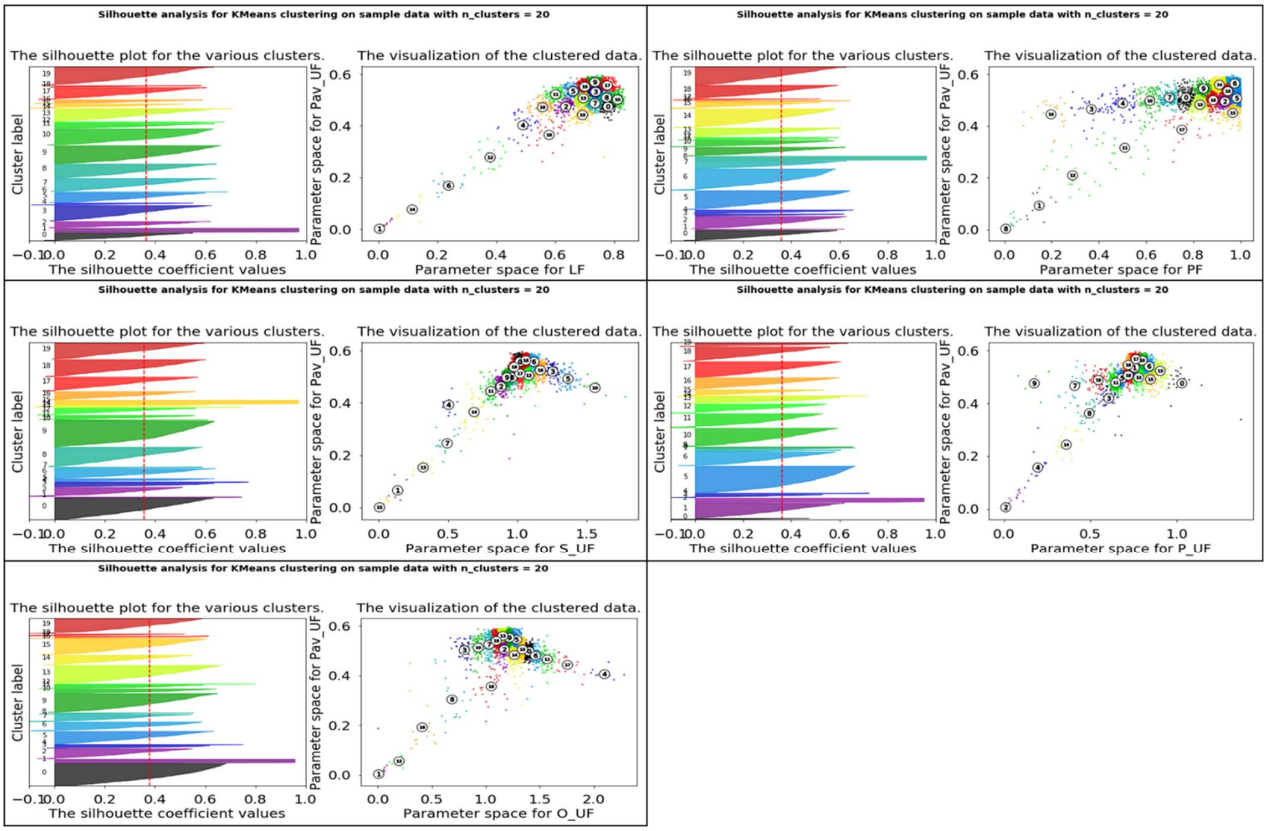


Figure 45: Silhouettes coefficients and scatter plots of different pairs of the parameters when 20 clusters were specified.

The results showed that the higher number of clusters, in this case, was not optimal for all parameters. The average silhouette score of each of the parameters was approximately 0.3, which is relatively low. A higher value, closer to 1, is desirable for the separated clusters. The silhouette plots also showed negative values for some of the clusters, indicating that some of the customers were allocated to the incorrect clusters.

B) Elbow results

The results of the elbow diagram on the national sample revealed that the number of clusters should be between four and seven. This was indicated by the elbow saturation points (Figure 46), that is, where the increase in the number of clusters had a relatively small impact on the SSE, as shown by the shaded block. The saturation point started at four. It was important for analysts to use discretion, based on experience and the results from other adequacy measures, to ensure the optimal selection of the clusters.

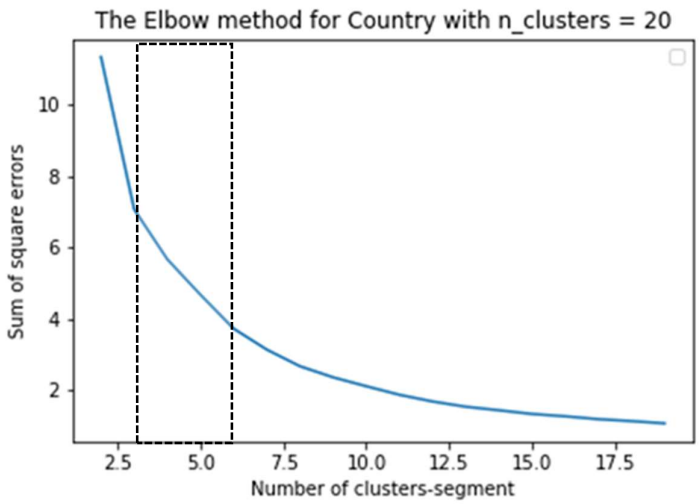


Figure 46: The elbow diagram showing the saturation region(area)

Again, the DBI scores became consistently lower as the number of clusters increased. Accordingly, fewer clusters were evaluated iteratively with a cluster being added to ascertain an optimal number of clusters.

6.2.1.2. Iteration 2: Two clusters

The scatter plot showing two clusters is depicted in Figure 47. The numbering of the clusters is zero-based, indicating that the counting started from zero. Iteration has two clusters, which is where the silhouette was the highest and the DBI the lowest. Ideally, this should be the recommended number of clusters. However, according to the elbow diagram, this is also where the SSE is larger, thus implying that this choice may not yield parameters that could assist in predicting the next class of customers.

The scatter plots of all these parameters indicated that there could be more clusters and, therefore, it is necessary to incorporate additional clusters.

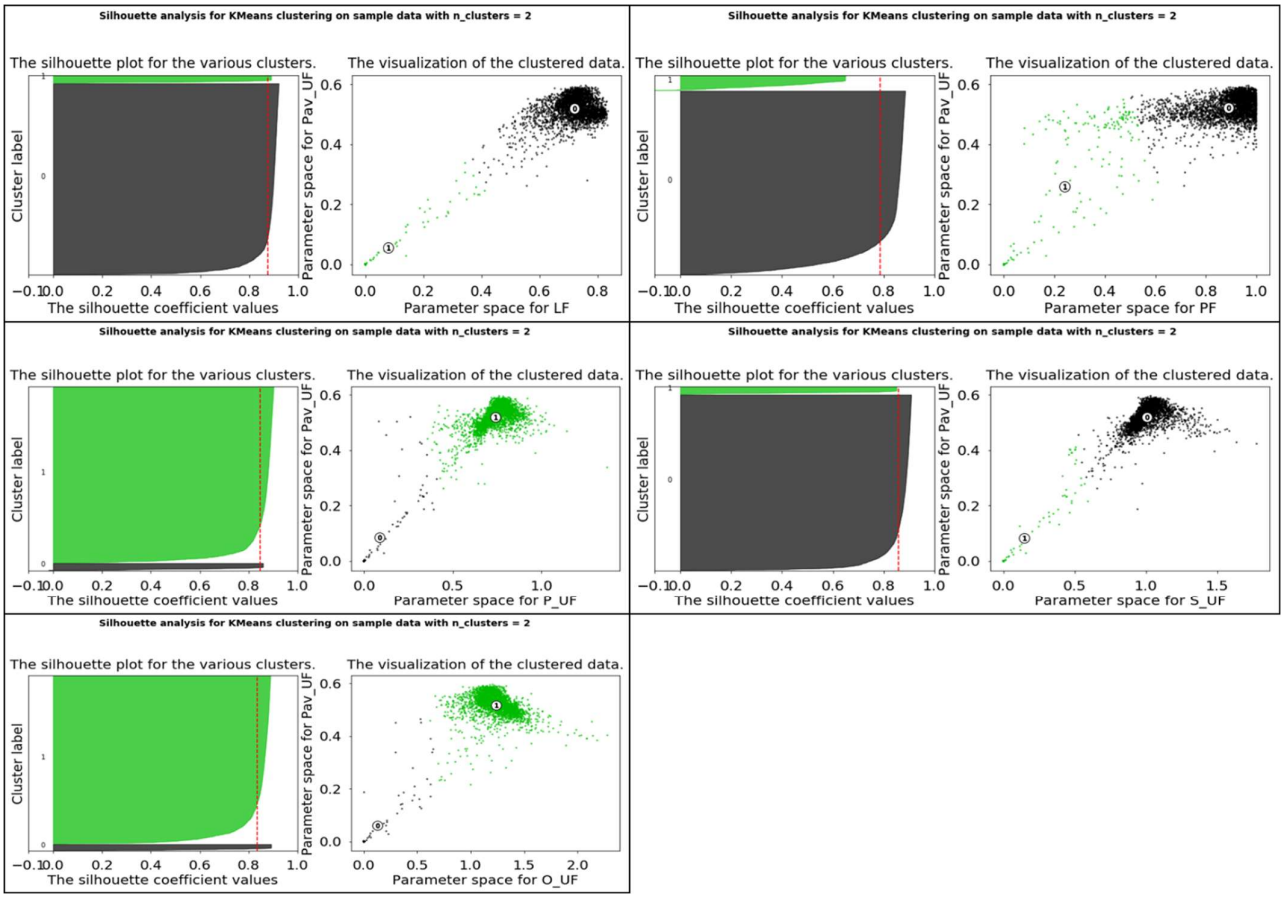


Figure 47: Silhouettes coefficients and scatter plots of different pairs of the parameters when 2 clusters were specified.

6.2.1.3. Iteration 3: Three clusters

The scatter plot showing three clusters is evaluated in this section. This is the next highest point in the silhouette. Although it is still above the arbitrary threshold of 0.5, the DBI is not necessarily the next lowest average in line. Parameters LF and PF present some unique findings when three clusters were assumed.

Figure 48 depicts the silhouettes and scatter plots of parameters LF and PF in three clusters, with parameter space of LF at the top of the figure. The results of parameter LF showed some incorrectly allocated memberships in cluster zero as indicated by a negative value in this cluster. There was a possibility that adding more clusters would result in more loads being misrepresented. The opposite was also envisaged to be possible.

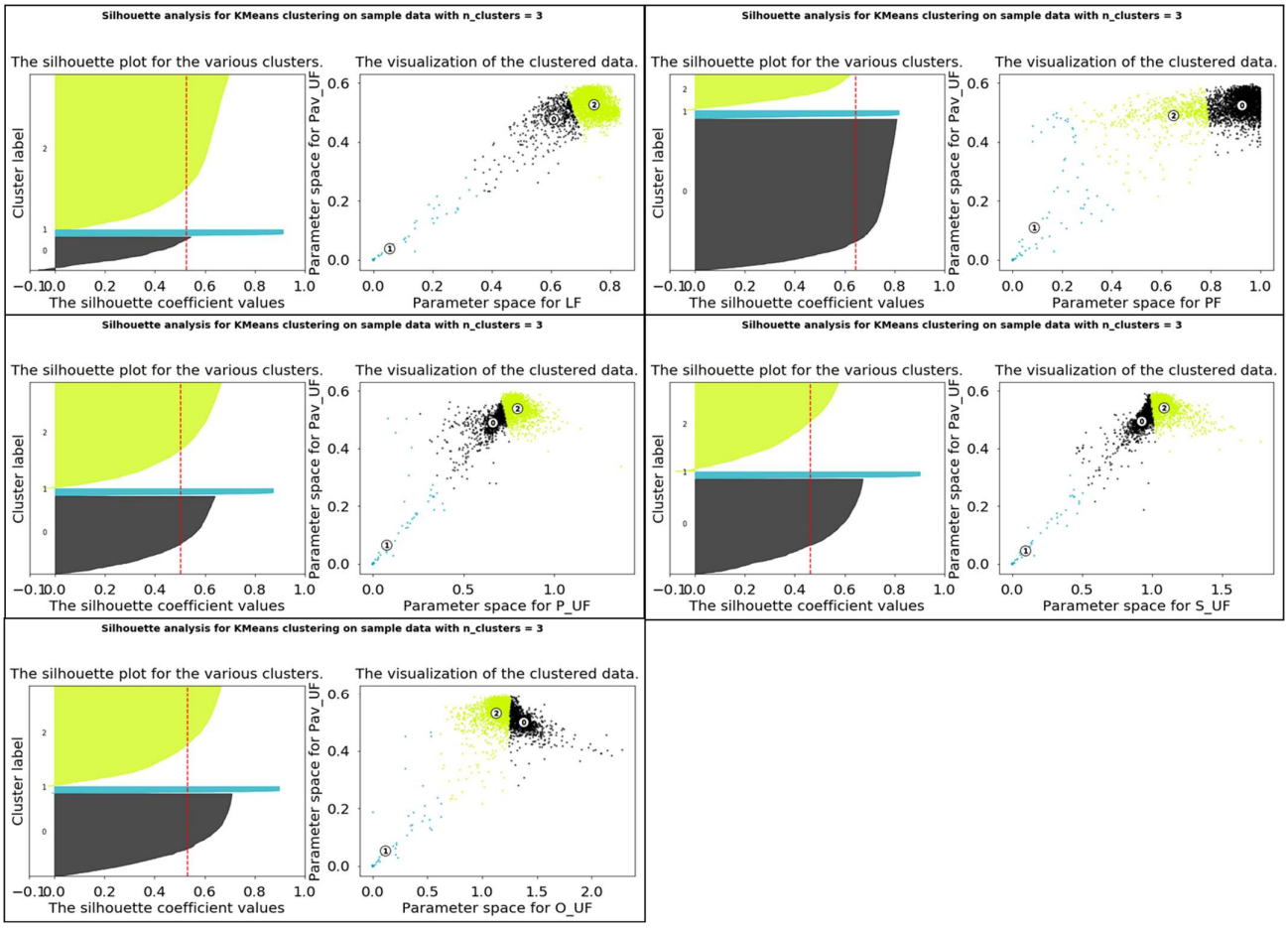


Figure 48: Silhouettes coefficients and scatter plots of different pairs of the parameters for 3 clusters

6.2.1.4. Iteration 4: Four clusters

The silhouette and scatter plots showing four clusters are shown in Figure 49. This is where the silhouette was the highest and the DBI was the lowest. Ideally, this could be the recommended number of clusters.

It was clear from Figure 49 that some of the loads were at the boundaries and that beyond this point, most of the loads would possibly end up being allocated incorrectly. This was evident from the smaller negative values of the silhouettes. The scatter plots showed that some of the cluster compositions were adequate although there could still be room for additional clusters, especially when looking at the parameter spaces of LF, PF and P_UF where there was space between the cluster at the bottom left-hand corner of the scatter plot and the rest of the clusters. In addition, the negative values were also very small (less than 0.1), thus indicating that there was still some room for additional clusters.

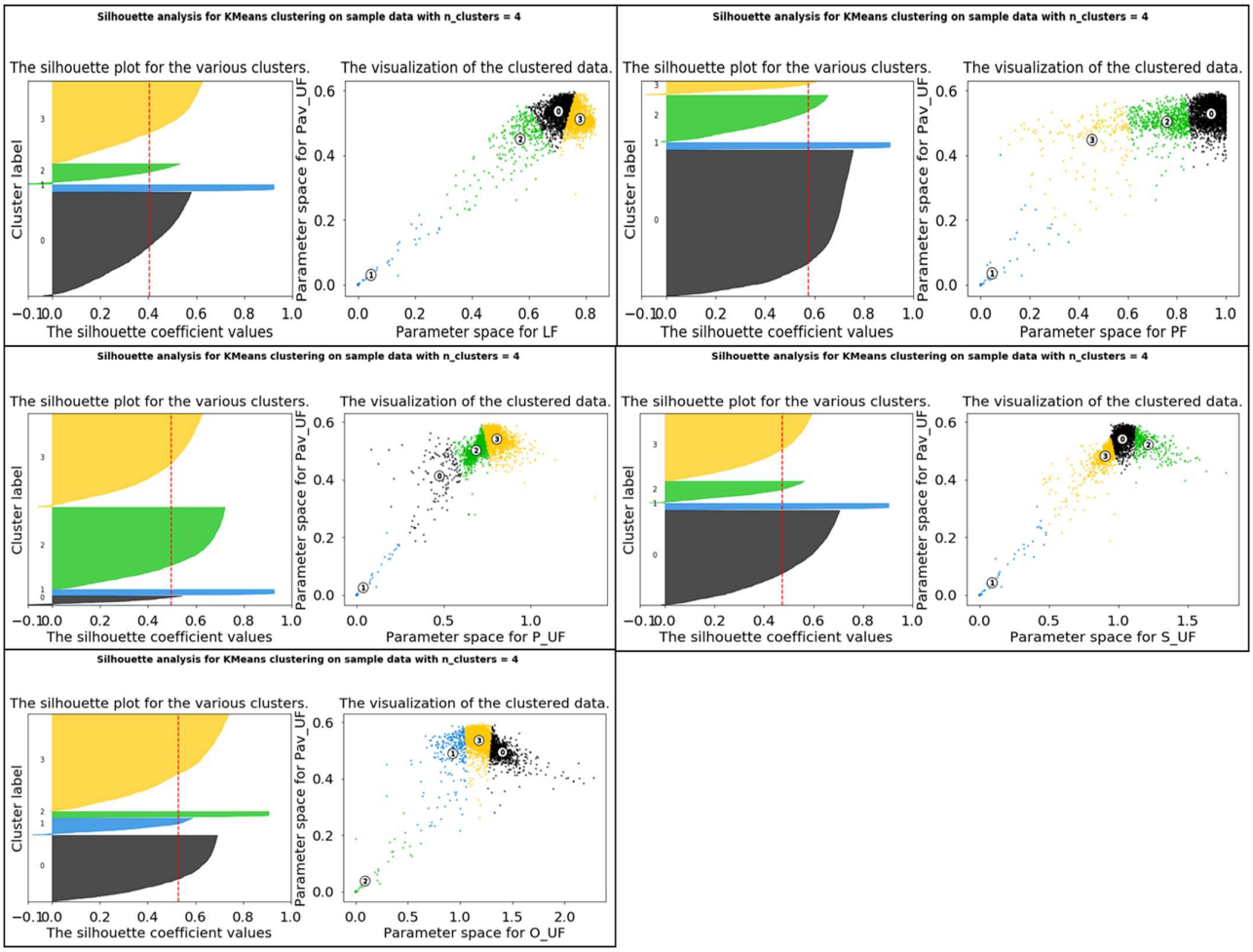


Figure 49: Silhouettes coefficients and scatter plots of different pairs of the parameters when 4 clusters were specified

6.2.1.5. Iteration 5: Five clusters

The scatter plot showing five clusters is presented in Figure 50. This is where, as stated earlier; the silhouette was at the threshold.

As indicated in Figure 50 the parameters LF and P_UF represented both the worst and the best score if the choice of clusters was five. The remaining parameters were somewhere in between the worst and best scores. Based on the threshold, the evaluation was halted at five clusters, which was considered the optimal number of clusters.

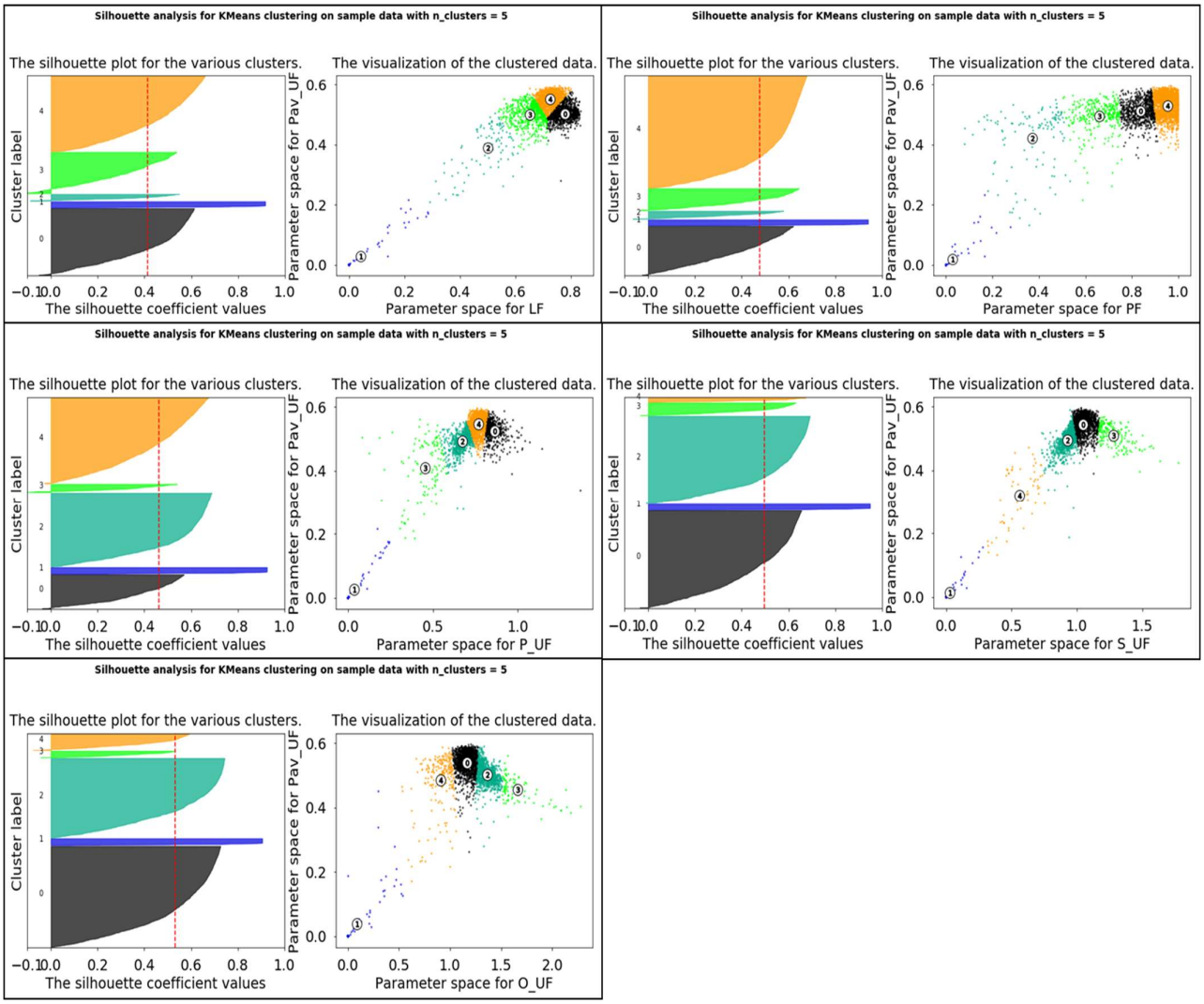


Figure 50: Silhouettes coefficients and scatter plots of different pairs of the parameters when 5 clusters were specified

6.2.1.6. Results of the cluster adequacy measures

The elbow diagram below indicated that it was possible to consider approximately five or more clusters. Accordingly, the different numbers of clusters were evaluated and a choice was made based on the silhouette plot, and resulted in five clusters chosen.

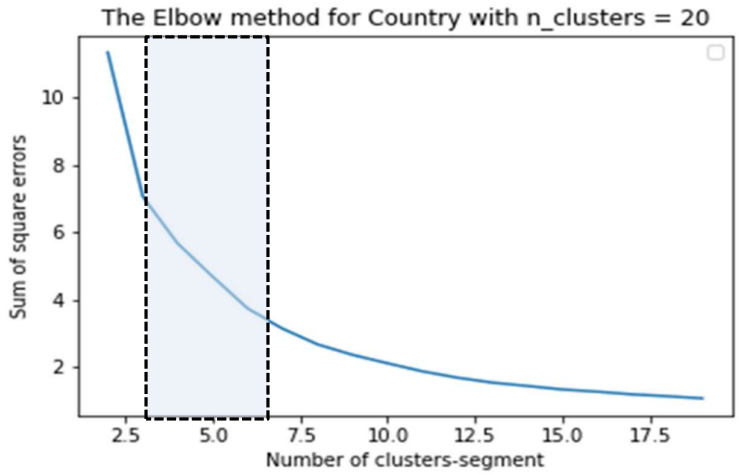


Figure 51: Elbow diagram for the selection of parameters

Table 11: The silhouettes and average DBI scores for different numbers of clusters

	Silhouette scores					Average DBI scores				
Number of clusters	**LF-Pav_UF**	**PF-Pav_UF**	**P_UF-Pav_UF**	**S_UF-Pav_UF**	**O_UF-Pav_UF**	**LF-Pav_UF**	**PF-Pav_UF**	**P_UF-Pav_UF**	**S_UF-Pav_UF**	**O_UF-Pav_UF**
For n_clusters =2	0.88	0.79	0.85	0.86	0.84	0.62	0.99	0.66	0.67	0.62
For n_clusters =3	0.53	0.64	0.50	0.46	0.53	3.91	2.24	4.08	4.52	3.52
For n_clusters =4	0.40	0.57	0.50	0.47	0.53	5.81	2.71	3.64	3.99	3.05
For n_clusters =5	0.41	0.48	0.46	0.50	0.53	5.03	3.58	3.99	3.50	2.69
For n_clusters =6	0.41	0.41	0.47	0.47	0.47	4.30	4.16	3.33	3.58	3.27
For n_clusters =7	0.41	0.38	0.40	0.47	0.47	3.67	5.05	4.10	3.31	2.89
For n_clusters =8	0.40	0.38	0.39	0.41	0.45	3.91	4.35	3.81	4.26	2.76
For n_clusters =9	0.40	0.38	0.38	0.41	0.44	3.50	3.86	3.62	3.98	2.83
For n_clusters =10	0.40	0.37	0.40	0.41	0.40	3.22	3.72	3.27	3.92	3.33
For n_clusters =11	0.38	0.37	0.39	0.38	0.39	3.67	3.36	3.17	3.89	3.11
For n_clusters =12	0.36	0.38	0.40	0.38	0.39	3.63	3.06	2.88	3.58	3.57

It was expected that if a silhouette score threshold of 0.5 was used, according to the lines drawn, the number of clusters based on parameter LF would be three, on PF and P_UF there would be four, while, for S_UF and O_UF, there would be five clusters. The elbow point, at which the decay begins to saturate, is indicated in the diagram depicted in Figure 40. A saturation area was subjectively created for the evaluation of clusters and made a selection based on the cluster that performed the best in this region. Since the silhouette and average DBI scores were pointing to up to five clusters, two clusters were evaluated first, then three, four and five, as indicated in Table 11. The silhouette plots and the scatter diagram showing the clusters were used for this exercise. Figure 52 gives a graphical view of the results in table 7. It is clear from the plots that additional clusters, beyond five, were not significantly affecting the silhouette scores.

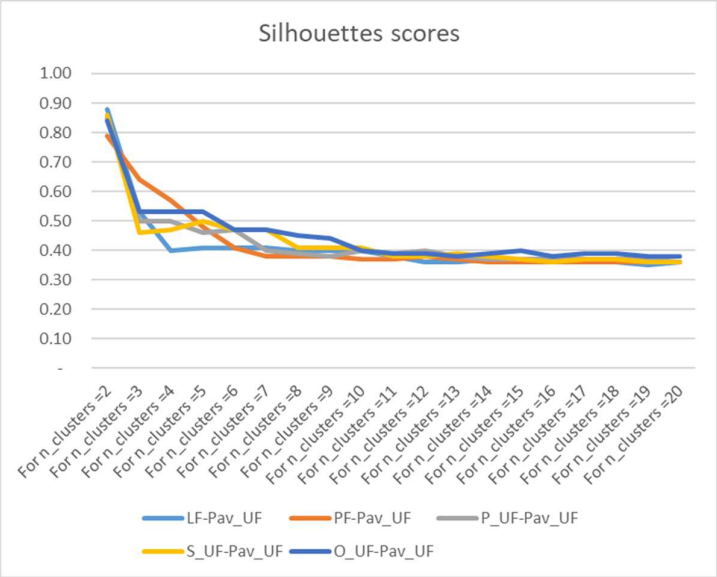


Figure 52: The silhouette scores of different parameters and clusters

6.2.1.7. Interpretation of the results

The scatter plots provided a view of the relationship and correlation between the various parameters. The parameters, which demonstrated a positively strong correlation, gave similar results and, thus, suggesting that it was possible to eliminate some of the combinations.

The resulting classes showed improvements in the grouping of the customers when comparing with the economic activity classes. The number of outliers was reduced to insignificant levels while the class averages were more representative as they are around the 50th percentiles of the customer measurements distribution as

can be seen in Figure 53

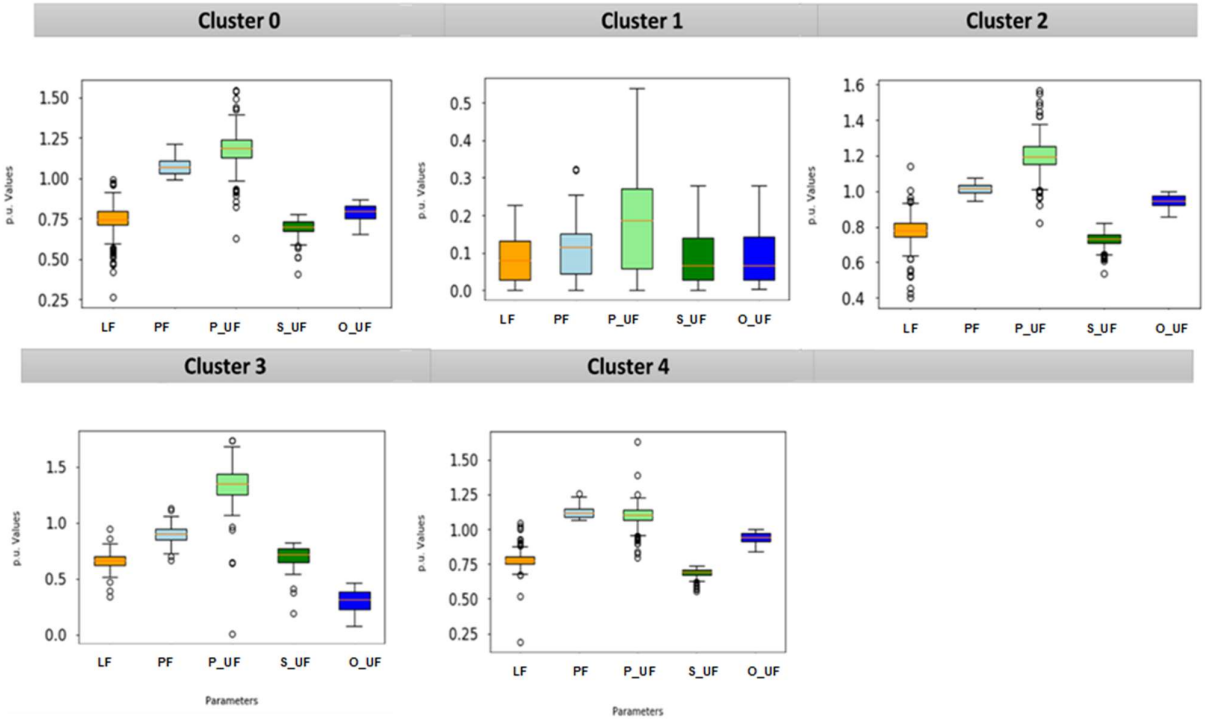


Figure 53: Box plots showing the distribution of the per-unit values associated with the parameters within each of the five clusters

The histograms for parameters LF, PF, P_UF, S_UF and O_UF and the average power on parameter Pav_UF are depicted in

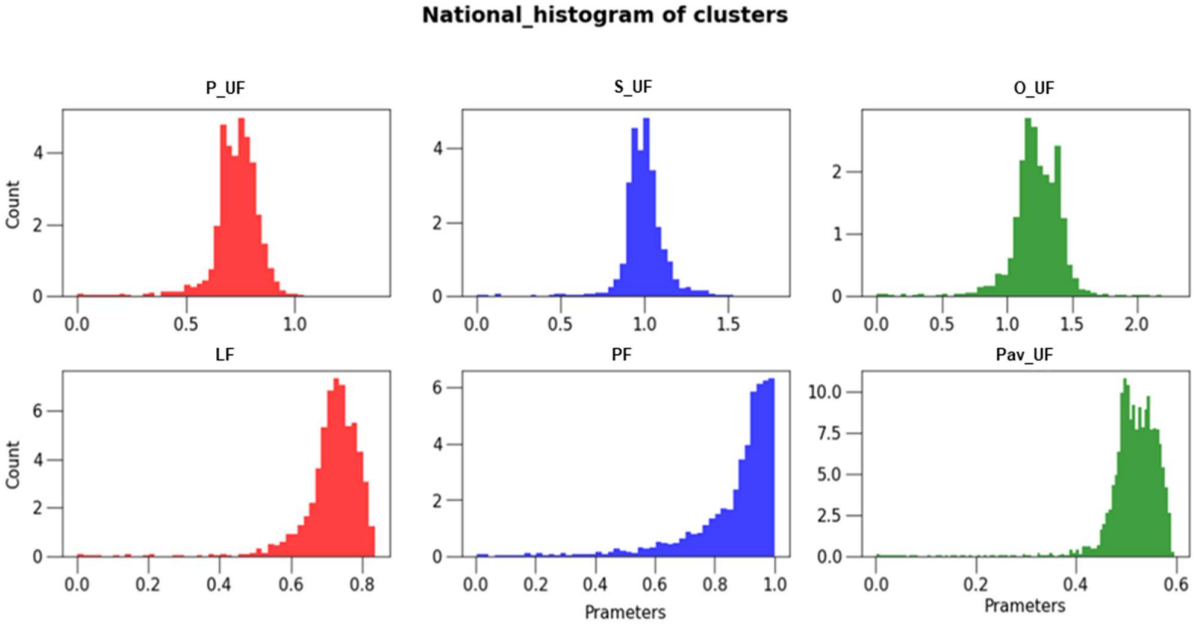


Figure 54.

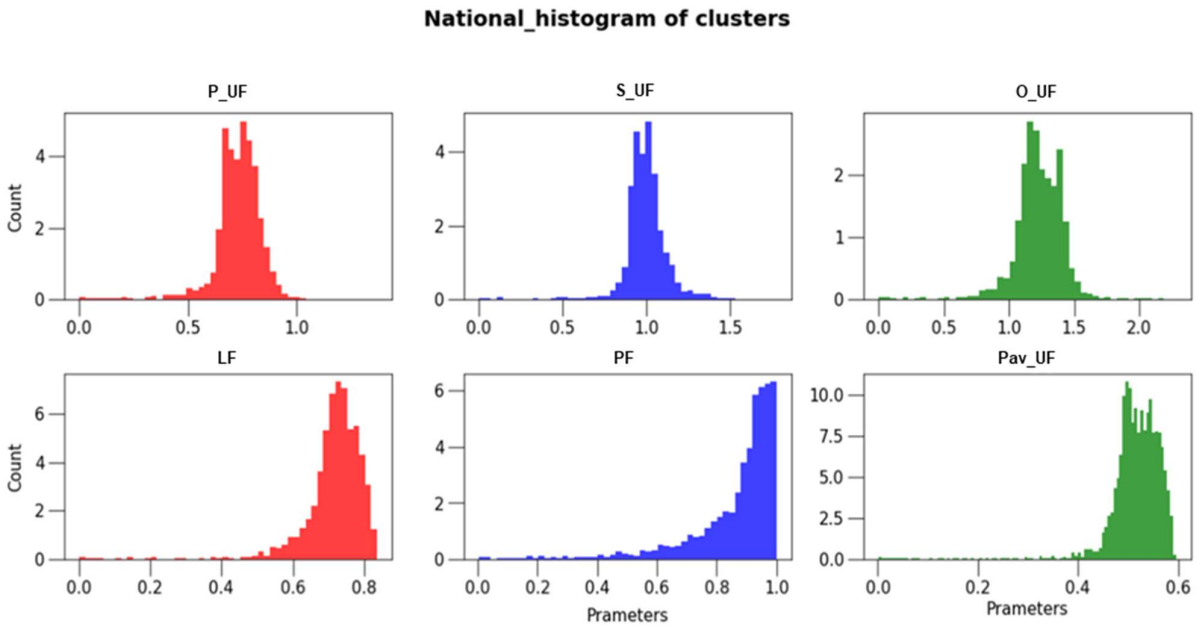


Figure 54: The probability plots for cluster zero for Parameters and the average power.

It can be seen from Figure 54 that the distributions were clearer and there were very few extreme values as compared to the economic classes' box plots. There was homogeneity in the clusters' distributions, which had not been seen in either the economic or the exogenous customer classes evaluated in Chapter 4. None of the distribution plots showed two or more peaks. The results also indicated that more customers were represented in the clusters than in the economical classes as defined in the Eskom database. This is also applied to the other four clusters whose graphs are not shown.

6.2.1.8. Allocation of profiles to different clusters

The profiles were allocated using the proposed method stated in chapter 3. The normalised averages of each cluster profile for all of the parameters combined are presented in Figure 55. These averages were plotted using the normalised data to ensure that they were all reflected in the same scale. Min-max normalisation was used. There were distinguishable patterns that showed that the clusters had different shapes.

The resulting class profiles were also compared with the economic activity based class profiles. As depicted in Figure 55, reading from top-left to right, the cluster developed using each of the parameters LF, PF, P_UF, S_UF and O_UF plotted against time. These were the average daily profiles of the various clusters. The last plot on the bottom right (Pav_UF) was based on the existing economic classes. The profiles of each class were found to be significantly different for each of the parameters/parameters. Classes formed using parameters LF and PF had similar load profiles. The profiles of the clusters based on parameters P_UF, S_UF and O_UF were unique, with each showing at least four significantly different shapes. Essentially the results were found to be different for each of the different parameters that were identified, for all the clusters. In addition, none of the parameters produced class profiles that were similar to those of the economic classes.

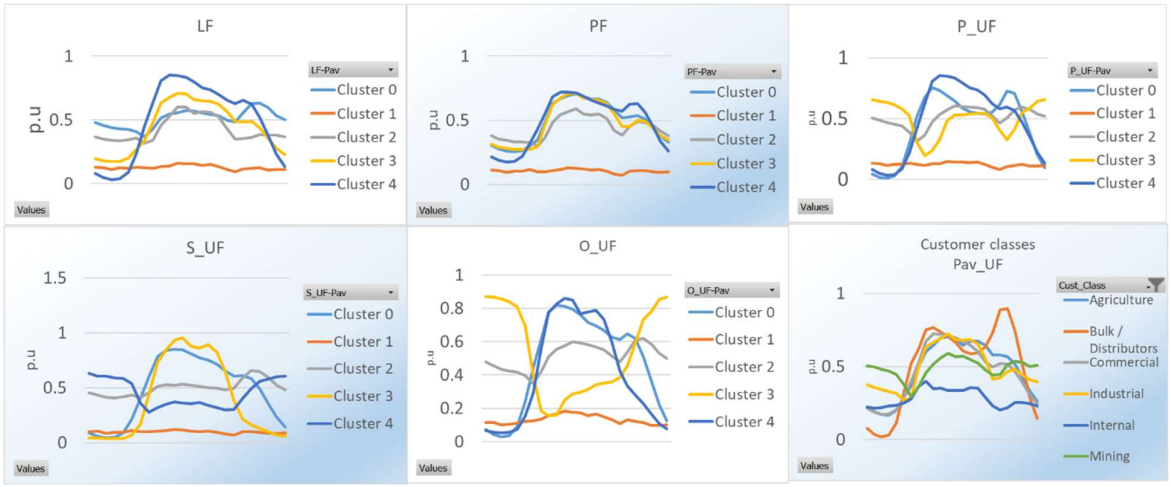


Figure 55: Comparison of the load profiles from the 5 clusters (0 to 4) based on the proposed parameters and the economic classes' profiles

A single parameter may not be adequate in the classification of load customers, and as indicated in the results, some of the parameters had more influence on the shape than a level while others had more impact on the level. This could be seen by observing S_UF and O_UF, which are related to the standard and off-peak usage parameters. Generally, there was high consumption during this period due to the lower cost of energy. About 20% of energy was utilised during peak, and the remainder was shared between standard and off-peak. However, it was expected that even on the normalised or per-unit basis their values would be higher than those of the peak usage. Therefore, the peak, standard and off-peak usage parameters may be linked to the usage level at different times of use intervals. This implies that technical and tariff models built only on parameters P_UF, S_UF or O_UF would not provide accurate results.

The classification using the power factors (PF) and load factors (LF) provided clusters with profiles that were similar in shape but differ in size. The load factors on the other hand appear to cater for both the shape and the level. This meant that the results could be improved if these five parameters were combined to perform the task of classification. This was investigated in case 2, which follows next.

6.2.2. Case 2: Use of PCA-based clustering algorithm to consider the identified parameters together

The next step in the analysis of the parameter was to consider using all five parameters in the same clustering algorithm. The significance of this consideration is that it would be possible to create classes that were based, not on a single parameter, but all the identified parameters LF, PF, P_UF, S_UF and O_UF. However, the problem of multi-dimensionality arose because there were five parameters. To be able to use k-means clustering, the dimensions were reduced using PCA.

6.2.2.1. Selecting an optimal number of clusters

The five endogenous parameters could not be projected in a scatter plot, which generally formed that basis for visualising the data groups in clustering; therefore, these parameters were compressed using the PCA algorithm to two PCs. The two PCs were used in the k-means clustering algorithm to create clusters based on the distance measures between the various PCs. The PCs were also evaluated to ensure that they accounted for the most significant variability associated with the parameters. Using some of the Python tools Scicit learn, a PCA algorithm has been developed and can be called a library. In the PCA library, the method for evaluating the performance of the PCA in terms of the variability that it accounts for is provided. The PCA algorithm was also used to determine the extent to which the PCs would represent the parameters. The PCs used accounted for 88.5%

of the variability of these parameters. The number of clusters used for this case was five, and the choice of five was informed by the results from 6.1 above. The decision to use the results from 6.1 is based on the need to compare the results from the same basis for the two study cases. However independent evaluation of the number of clusters formed using the PC indicated that the number of five is acceptable, although the cluster adequacy indices were mostly pointing towards four clusters the distributions resulting from the five clusters are shown in Figure 56. The PCs did not overlap but some of the values were very close to each other indicating that they were on the boundary.

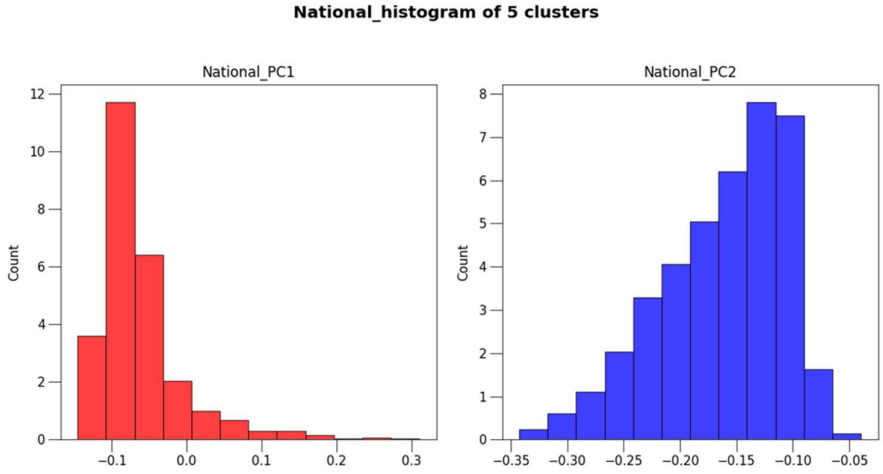


Figure 56: Distribution plot of principal components

Figure 57 shows the values that make up each of the two PC in a box plot. From the figure, it can seen that the PC are distinct and do not overlap. This is desired for the classification model because the resulting classes will have clear boundaries if there are no overlaps between the parameter that are chosen to represent the loads.

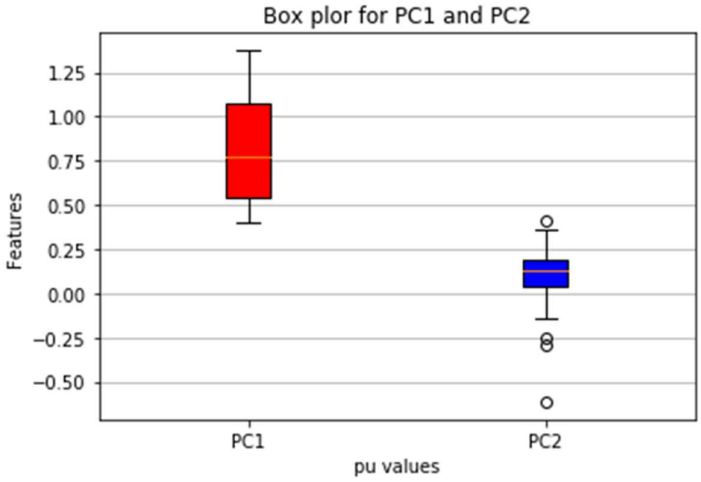


Figure 57: Box plots of the two principal components of the five parameters

6.2.2.2. Analysis and interpretation of the results

When extracting the parameters for a specific sub-interval (or time), or other parameters, statistical analysis of the results is often used and this type of analysis provides a basis for estimating parameters. In particular, the statistical summaries include the means and standard deviations of each cluster at a particular time, while they also provide information to estimate the parameters and to show how the data is distributed. The statistical summaries are

presented in Table 12. The smallest standard deviation was desirable as it indicated how the data points represented varied about the mean. From the table, it was observed that clusters one and four had almost equal arithmetic means. There was a possibility of overlapping loads between these two clusters. In Figure 9 the distribution of the resulting cluster is provided.

Table 12: Statistical summaries of the parameters

Statistic	*Cluster 0*	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4*
Mean	0.464	0.870634	0.192406	0.775775	0.876769
Median	0.469	0.865098	0.158259	0.814823	0.912783
Standard Deviation	0.140	0.081979	0.127621	0.161316	0.111083
Sample Variance	0.020	0.006721	0.016287	0.026023	0.01234
Kurtosis	- 0.035	0.515446	0.809634	-0.58485	-0.21335
Skewness	- 0.089	-0.61202	0.902717	-0.58837	-0.90251
Range	0.713	0.36555	0.526369	0.641273	0.427262
Minimum	0.116	0.63445	0.008009	0.358726	0.572738
Maximum	0.829	1	0.534378	0.999999	1

The difference in the shapes of these distributions, that is, positive or negative skewness and asymmetrical graph, indicates that a distribution that can accommodate this difference in the shape would be the preferred choice. The beta distribution is one such distribution. It is also possible to calibrate the parameters for each cluster.

6.2.2.3. Allocation of profiles to different clusters

When clusters have been determined, the next step in the load models is to assign load profiles to the different clusters. The method proposed in chapter three for allocating profiles to the clusters was used. The method essentially calculated the average profiles for the loads in the cluster use that as a class or cluster profile. The profiles of the resulting clusters are presented in Figure 58. The profiles of clusters 0, 2 and 4 were found to be very close to each other in the daytime between 07:00 and 14:00, but began to differ in the evening although their shapes were similar. These observations were in line with the observation made by Figueiredo et al. (2005) that the nighttime demand levels may be a significant differentiator of classes. The profiles also showed the morning peak occurring outside the prescribed hours of between 06:00 and 08:00; instead, they could be occurring between 08:00 and 10:00. The drivers of these demands included loads that peaked during the daytime. Essentially, this meant that the commercial loads and mining activities, as shown in the profiles in Chapter 5, could be the significant contributors to the peak.

A noteworthy observation was that the statistical mean values indicated that clusters one and four were close to each other, whereas the profiles of these clusters are not the same in shape and size. Therefore, this implies that the similarity cannot be drawn only from observation or simple averages but the parameters of the loads.

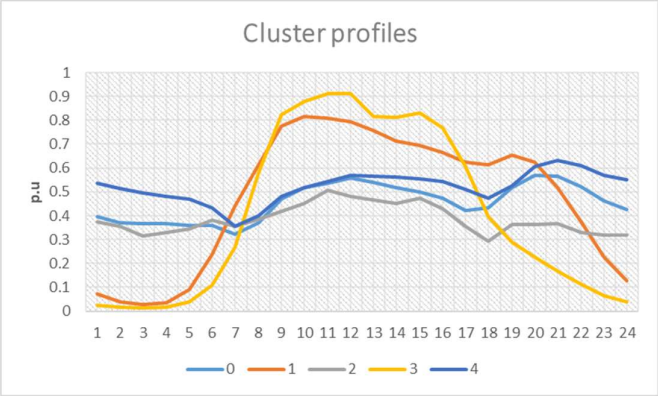


Figure 58: Average daily profiles of the five resulting clusters

Histograms provided a graphical analysis of parameter values and when used in conjunction with summary statistics, could make it possible to estimate the best pdf. The pdfs could be used as input for technical analysis. Histograms for clusters 0 to 5 are presented in Figure 59. The following deduction can be made from the results:

- There was no evidence of extreme values in any of the resulting clusters.
- The cluster had different distributions, indicating that they were dissimilar or distinguishable.
- There were overlaps of the PC values on each of the clusters, and this indicated a possibility of overlaps in clusters.

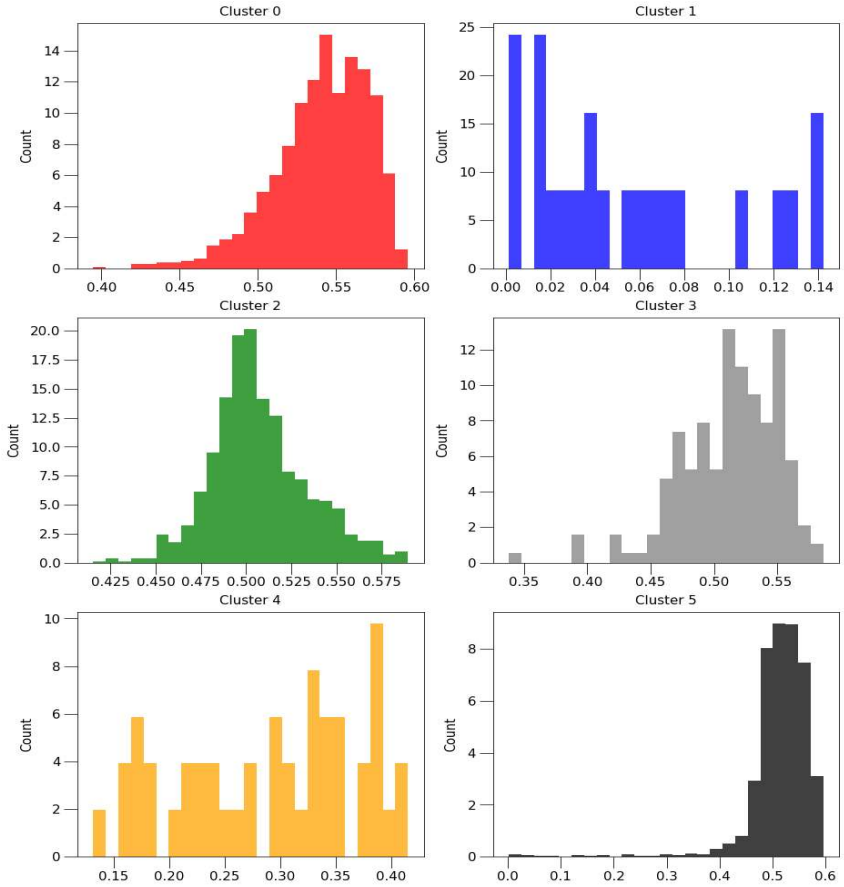


Figure 59: Histograms of clusters 0 to 5

The average customer load for the different clusters is depicted in Table 13 below. The tables also show the economic zones in rows while the clusters from the model results are in columns. The table illustrated that customers in the agriculture class had an average consumption of 316 kW and comprised loads the were distributed in the different clusters as follows; the customers with the largest average consumption were allocated to cluster 4, and those with low consumption in cluster 2. This means that if there is a need to further classify the loads in the agricultural class, the profiles of cluster 4 could be allocated to the customer consuming at this level. The same interpretation may be extended to other classes in the table to further allocate appropriate profiles to customers within these different classes. The interpretation provided above is important when allocating costs in tariff models, and when there is a need to represent loads more accurately while maintaining the reference to economic activity class, in load flow studies. The values were shown to two decimal places, therefore values lower than 0.00 were shown as 0.00 and where there were no values this was indicated by a space in Table 13. The loads with the highest average power were in cluster 4 and those with the lowest average power were allocated to cluster 2 for all economic activity classes except Internal. The averages of each class related to the clusters provided an opportunity to classify new customers, to the different clusters and based on the knowledge of their economic activity.

Table 13: Average load of the clusters

Class\Cluster (Aver:	0	1	2	3	4	Average
Agriculture	213.59	334.23	5.62	193.71	837.72	316.97
Bulk / Distributors	296.64	17 803.52	0.00	1 720.18	25 198.04	9 003.68
Commercial	94.42	924.24	22.59	271.04	2 284.57	719.37
Industrial	123.69	630.77	0.00	197.39	25 552.43	5 300.86
Internal	8.32	883.11	223.87		4 584.81	1 425.03
Mining	439.06	718.86	113.52	234.91	27 996.22	5 900.51

The results in Table 13 implied that it was possible to further classify the economic activity based-classes into sub-classes and allocate profiles to these subclasses. Based on the results, the subclass profiles for each of the economic activities would be as depicted in Figure 60. For each class, the corresponding sub-class was indicated by the name of the class and the cluster number, that is, for commercial class and similarly for the other classes, the sub-classes are: commercial - 1, commercial - 2, commercial - 3, commercial - 4 and commercial – 5 within which to categorise our loads.

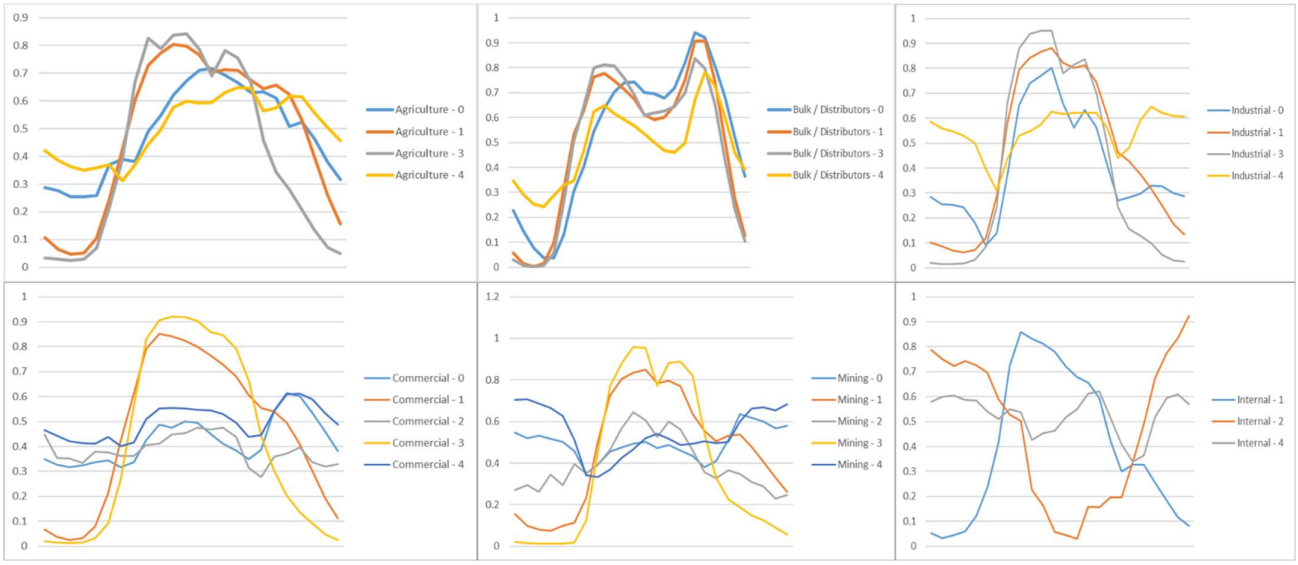
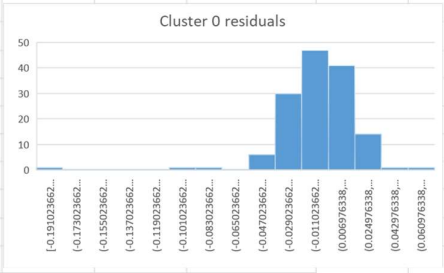


Figure 60: Load profiles of the clusters within the existing customer classes

6.3. Validating the results

An important step in modelling is validating the results. While the clustering result themselves were validated using cluster adequacy measures, the model also needed validation and that could be achieved through the evaluation of the results and comparison of those results with reality, that is, the actual data and/or some predefined benchmarks. The cluster validity tools used to define the number of clusters, such as the silhouettes, DBI and the elbow methods, were used to determine the number of clusters. Using these scores as a guide provided a reasonable level of validation that is carried out upfront since the methods that were used utilised the same data. The model validation could be achieved by evaluating the performance of the clusters using statistical tools and comparing the load profiles from the different classes as well as the use of intuition in establishing whether the profiles were different for each class and that there were unique linear relationships that describe the model. Regression analysis and analysis of variance (ANOVA) were conducted to determine the validity of each cluster. The data was then arranged per customer class with each parameter being the independent variable, and the average consumption of the customer being the dependent variable. The regression model built in Excel was used. The results of the regression analysis of cluster 0 are presented in Figure 61.

SUMMARY OUTPUT



Regression Statistics	
Multiple R	0.998146
R Square	0.996295
Adjusted R Square	0.988942
Standard Error	0.027647
Observations	143

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	28.36685	5.673369	7422.381	7.6452E-165
Residual	138	0.105482	0.000764		
Total	143	28.47233			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
P_UF	0.228686	0.036918	6.194354	6.31E-09	0.15568714	0.301685	0.155687	0.301685
S_UF	0.161934	0.022297	7.2625	2.53E-11	0.117845244	0.206022	0.117845	0.206022
O_UF	0.061344	0.014226	4.312031	3.06E-05	0.033214371	0.089474	0.033214	0.089474
LF	0.098055	0.043025	2.279003	0.024202	0.012980735	0.183129	0.012981	0.183129
PF	0.032832	0.012286	2.672361	0.008439	0.008539223	0.057124	0.008539	0.057124

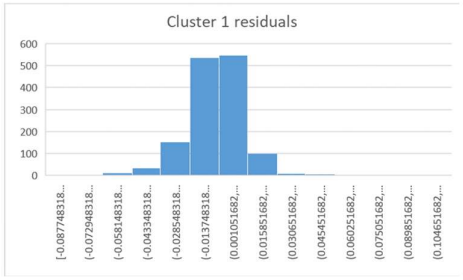
Figure 61: Results from regression analysis of cluster 0

As indicated by the adjusted R-square, which was closer to one, the results showed that the class was well explained by the parameters. Using the widely used p-value threshold of 0.005, the only insignificant variables for this class were parameters LF and PF as their p-values were greater than the threshold. A histogram of the residuals was also plotted and shown in the top-right corner of Figure 61. The plot showed that the residuals were concentrated around 0.007 and 0.011, which were sufficiently small to indicate the acceptable performance of the model.

The results for cluster 1 are depicted in Figure 62. The results were found to be similar to those for cluster 0 with the one difference being that the only insignificant parameter for cluster 1 was PF.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999598
R Square	0.999197
Adjusted R Square	0.99848
Standard Error	0.015378
Observations	1405



ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	5	411.8214	82.36427	348292	0
Residual	1400	0.331073	0.000236		
Total	1405	412.1524			

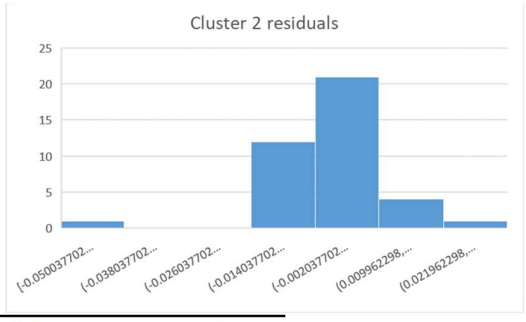
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
P_UF	0.055156	0.006582	8.37972	1.27E-16	0.042245	0.068068	0.042245	0.068068
S_UF	0.165193	0.004981	33.16522	1.8E-178	0.155422	0.174964	0.155422	0.174964
O_UF	-0.35561	0.009566	-37.1742	5.4E-211	-0.37437	-0.33684	-0.37437	-0.33684
LF	1.042681	0.018749	55.61271	0	1.005902	1.07946	1.005902	1.07946
PF	-0.00226	0.005652	-0.3993	0.68973	-0.01334	0.00883	-0.01334	0.00883

Figure 62: Cluster 1-regression analysis results

The coefficients of parameters PF and O_UF were negative, thus indicating that they had an opposite effect on the results. This implied that by increasing these specific parameters the average kW decreased. For cluster 2 shown in Figure 63, the results indicated that when all the coefficients, except PF and O_UF, were increased the average power would also increase, while increasing PF and O_UF, would reduce the average power. This means that for this cluster as the power factor increases and usage in off-peak is increased, the output decreases. This is implied that in this cluster when the load was higher, the power factors became poorer.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99704251
R Square	0.99409377
Adjusted R Square	0.96398715
Standard Error	0.01251175
Observations	39



ANOVA

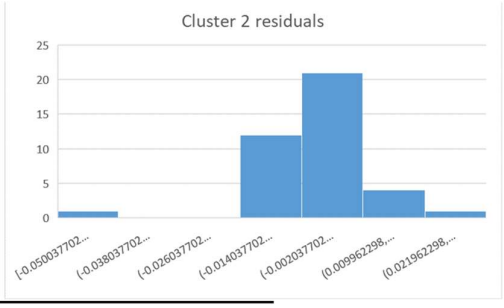
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	5	0.895843	0.179169	1144.526	5.76E-36
Residual	34	0.005322	0.000157		
Total	39	0.901165			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
C	0.23369316	0.036406	6.419063	2.47E-07	0.159707	0.307679	0.159707	0.307679
D	0.11061135	0.022338	4.951695	1.98E-05	0.065215	0.156008	0.065215	0.156008
E	0.10666842	0.030604	3.485442	0.001375	0.044474	0.168863	0.044474	0.168863
A	0.11007449	0.076124	1.44599	0.157336	-0.04463	0.264777	-0.04463	0.264777
B	-0.0329254	0.026712	-1.23259	0.226184	-0.08721	0.021361	-0.08721	0.021361

Figure 63: Cluster 2 regression analysis results

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99704251
R Square	0.99409377
Adjusted R Square	0.96398715
Standard Error	0.01251175
Observations	39



ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	5	0.895843	0.179169	1144.526	5.76E-36
Residual	34	0.005322	0.000157		
Total	39	0.901165			

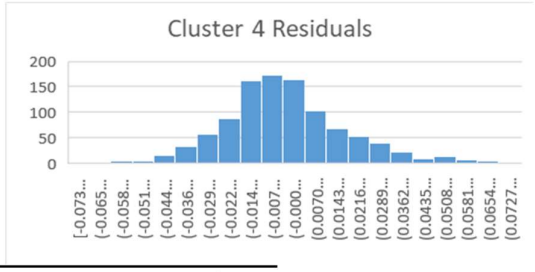
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
P_UF	0.23369316	0.036406	6.419063	2.47E-07	0.159707	0.307679	0.159707	0.307679
S_UF	0.11061135	0.022338	4.951695	1.98E-05	0.065215	0.156008	0.065215	0.156008
O_UF	0.10666842	0.030604	3.485442	0.001375	0.044474	0.168863	0.044474	0.168863
LF	0.11007449	0.076124	1.44599	0.157336	-0.04463	0.264777	-0.04463	0.264777
PF	-0.0329254	0.026712	-1.23259	0.226184	-0.08721	0.021361	-0.08721	0.021361

Figure 64: Cluster 3 Regression analysis results

The results presented Figure 65 indicated that all of the parameters were significant and that the adjusted r-square was the highest at 0.997. The model or regression equation for cluster 4 was found to be a good fit.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999133
R Square	0.998266
Adjusted R Square	0.997272
Standard Error	0.020885
Observations	1018



ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	5	254.3382	50.86764	116622.6	0
Residual	1013	0.441844	0.000436		
Total	1018	254.7801			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
P_UF	0.121203	0.012428	9.752066	1.55E-21	0.096814	0.145591	0.096814	0.145591
S_UF	0.244988	0.008472	28.91823	1.5E-134	0.228364	0.261612	0.228364	0.261612
O_UF	0.057815	0.004039	14.31467	1.82E-42	0.04989	0.065741	0.04989	0.065741
LF	0.122778	0.01599	7.678487	3.78E-14	0.091401	0.154156	0.091401	0.154156
PF	0.019663	0.006496	3.027142	0.002531	0.006917	0.03241	0.006917	0.03241

Figure 65: Cluster 4 regression analysis results

Table 14 presents the summary of the regression results. As it can be seen from the table, indicated by the smaller p-values there was a significantly stronger relationship between the parameters and the average demand of each cluster. These parameters also explained the variability or the average power as indicated by the strength of the coefficient of determination (R-square). It is therefore possible to conclude that these clusters were valid for the dataset used.

Table 14: Regression analysis summary for the 5 clusters

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
R Square	0.9963	0.9992	0.9941	0.9992	0.9983
Adjusted R Square	0.9889	0.9985	0.9640	0.9941	0.9973
Standard Error	0.0276	0.0154	0.0125	0.0146	0.0209
P_UF p-value	0.0000	0.0000	0.0000	0.0000	0.0000
S_UF p-value	0.0000	0.0000	0.0000	0.0000	0.0000
O_UF p-value	0.0000	0.0000	0.0014	0.0000	0.0000
LF p-value	0.0242	-	0.1573	0.0000	0.0242
PF p-value	0.0084	0.6897	0.2262	0.0882	0.0084
P_UF	0.2287	0.0552	0.2337	0.0581	0.1212
S_UF	0.1619	0.1652	0.1106	0.0669	0.2450
O_UF	0.0613	- 0.3556	0.1067	- 0.1282	0.0578
LF	0.0981	1.0427	0.1101	0.8244	0.1228
PF	0.0328	- 0.0023	- 0.0329	0.0115	0.0197

6.4. Implementation consideration of the load models

The implementation of load models in any study requires that the models be related to the existing loads. Therefore, an assessment of the results in the context of existing loads was necessary.

6.4.1. The implication of the results on the existing customer classes

To implement the results it was necessary to revert to the Eskom database, where the representation of the customer within the clusters could be evaluated. Table 15 depicts the number of customer classes from the Eskom database together with the probabilities of each class belonging to each of the clusters. It was possible to draw an inference concerning the dominating activities within a cluster to which the customer class belongs. In addition, there was an observable overlap of customer classes. This indicated that the economic classes contained loads that belonged to the neighbouring classes. This was also evident in the non-homogenous histograms that were plotted for the customer classes in Chapter 6.

Table 15: Customer class percentages in each of the clusters

Customer_classes	0	1	2	3	4	Grand Total
Agriculture	6.19%	47.16%	1.29%	9.28%	36.08%	100.00%
Bulk / Distributors	0.43%	91.36%	0.22%	1.73%	6.26%	100.00%
Commercial	4.51%	54.86%	1.62%	5.79%	33.22%	100.00%
Industrial	3.67%	28.85%	0.24%	9.54%	57.70%	100.00%
Internal	11.54%	38.46%	7.69%	0.00%	42.31%	100.00%
Mining	13.03%	9.92%	2.83%	13.88%	60.34%	100.00%
Grand Total	**5.08%**	**50.23%**	**1.35%**	**7.11%**	**36.22%**	**100.00%**

The results in Table 15 can be shown visually on a surface plot in Figure 66, where the classes are on the x-axis, the allocation percentages on the y-axis, and the clusters on the z-axis.

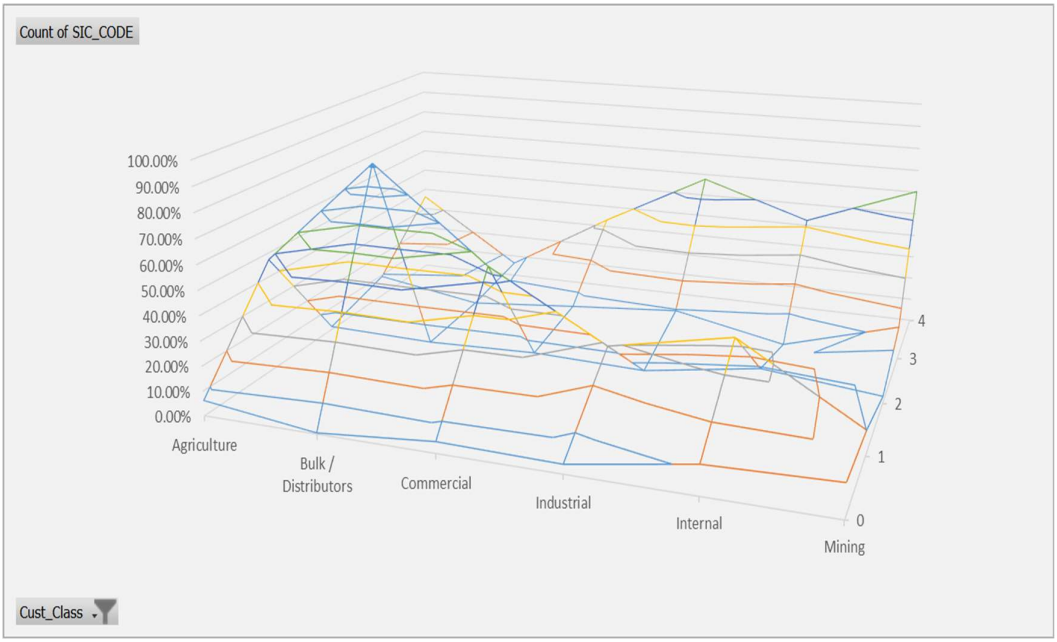


Figure 66: Surface plot of proportions of customer classes in the clusters

The five clusters were not necessarily the same as the predefined economic classes, although there is a possibility that some clusters would have been dominated by a particular class demands. The results in the table indicated that the agriculture class demonstrated a 47% presence to cluster 1, 36% to cluster 4 and the smaller remaining percentages were in the other cluster, thus implying that tariff options could be built around the profiles of clusters 1 and 4 to ensure customers are offered tariff options that suit them. It was found that 91% of the bulk/distributor class accounted for cluster 1 and this is 50.23% of all the clusters. This indicated that this class had a significant influence on the system profile. It appeared that the mining and the industrial classes seemed to dominate cluster 4, whereas the commercial was found mostly in cluster 1.

6.5. Application of the results to new customers

To classify new customers, the values of the parameters LF, PF, P_UF, S_UF and O_UF can be calculated and allocated to the appropriate cluster. The upper and lower limits of each of the parameters indicate the boundaries within which the class members should lie. These limits are presented in Table 16. The coefficient columns contain the coefficients from the linear regression model and represent coefficients of the parameters, while the lower and upper 95% are the upper and lower limits that define the range of the values at a 95% confidence interval.

Table 16: Limits of parameters for different clusters

	Parameter	Coefficients	Lower 95%	Upper 95%
Cluster 0	P_UF	0.229	0.156	0.302
	S_UF	0.162	0.118	0.206
	O_UF	0.061	0.033	0.089
	LF	0.098	0.013	0.183
	PF	0.033	0.009	0.057
Cluster 1				
	P_UF	0.055	0.042	0.068
	S_UF	0.165	0.155	0.175
	O_UF	-0.356	-0.374	-0.337
	LF	1.043	1.006	1.079
	PF	-0.002	-0.013	0.009
Cluster 2				
	P_UF	0.234	0.160	0.308
	S_UF	0.111	0.065	0.156
	O_UF	0.107	0.044	0.169
	LF	0.110	-0.045	0.265
	PF	-0.033	-0.087	0.021
Cluster 3				
	P_UF	0.058	0.039	0.077
	S_UF	0.067	0.057	0.077
	O_UF	-0.128	-0.157	-0.099
	LF	0.824	0.760	0.889
	PF	0.012	-0.002	0.025
Cluster 4				
	P_UF	0.121	0.097	0.146
	S_UF	0.245	0.228	0.262
	O_UF	0.058	0.050	0.066
	LF	0.123	0.091	0.154
	PF	0.020	0.007	0.032

The general equation of the linear regression models has been provided in chapter 3 and may be used as a basis for classifying customers using the information in Table 16. The classification of customers may be based on the average power calculated using the customer's parameters LF, PF, P_UF, S_UF and O_UF. The procedure is as follows:

a) If the customer has profile (projections of hourly usage):
   - *Use the profile to calculate parameter LF, PF, P_UF, S_UF and O_UF, using Equation 11 to Equation 15.*
   - *Determine/Estimate the profiles of the customer based on their economic activity class.*
   - *Allocate the customer to the different cluster based on the values of the calculated parameter , and the boundaries in Table 16*
   - *Allocate a cluster profiles to the customer*
b) Else (if no profiles data but estimated demand)
   - *Use the average demand to find the matching demand in Table 13*
   - *Determine/Estimate the profiles of the customer based on the cluster corresponding to the demand.*
   - *Allocate the customer to the different cluster based on the values of the calculated parameter LF, PF, P_UF, S_UF and O_UF, and the boundaries in Table 16*
   - *Allocate a cluster profiles to the customer*
c) End.

6.5.1. Combining the typical day and customer classification models

When performing load flow studies, there is often a need to know the loads and generators, as well as their associated profiles for simulation intervals or typical days. Power system simulation tools, such as Digsilent and Powerworld, allow for time series inputs. The results of this study indicated that utilities could use the measurements data and the process used to determine/identify their customers and define a simulation span. The typical days for summer and winter could be used for the simulation period, while the classification results could provide information on how the utility's customers or loads are clustered and which of their load profiles to use.

6.6. Concluding remarks

As suggested in the literature review chapter, several parameters can be taken from measurement data. Using customer measurement data it was possible to create load models that can be used to classify customers, and assigning load profiles to customer classes. This was achieved by identifying and combining the endogenous and exogenous parameters. Therefore, the coherent load parameter models for tariff and technical analysis models should comprise the endogenous and exogenous parameters.

This chapter presented the results from the pilot study conducted on measurements data of customers connected at MV. Two cases were studied, the first using a pairwise parameter-based load model, where the parameters were arranged in pairs with normalised average power (Pav_UF) and used to classify customers. In the second case, the five parameters that were derived from the measurements were considered. The PCA dimension reduction algorithm was used to project these five parameters into a 2 D space and used in the k-means clustering algorithm to classify customers. The results shown in Table 13 indicated that the current load classification based on economic activity alone was not adequate for technical and tariff analysis. These economic activity based classes could be further classified.

Regression analysis is one of the statistical tools available for the analysis of the classification results. Regression analysis made it possible to recognise the significance of each parameter in relation to the cluster function that each parameter represented. While clustering algorithms are not restricted by the dimension of the data, the need to visualise the clusters remains a challenge. Nevertheless, dimension reduction techniques provided the options of projecting high dimensional data to lower dimensional space to enable data to be visualised. However, the principal components themselves had no meaning other than that they represented the variability inherent in the data. This meant that their usefulness ended when the classes were created. The cluster members or loads that belong to a particular cluster were identified by their corresponding PC values. That is, since the PCA would produce the PC for each load profile as identified by the parameters, this meant that each load profile would have a PC. To assign loads to the clusters, the clustering algorithm was using the PC to calculate the smallest distances between the average PCs of each of the clusters and PCs of the loads.

A procedure to classify new customers has been developed. The procedure accommodates customers who have the load profiles as well as those without the load profile but have the load information (demand size). In both cases, a profile is required to calculate the parameters. Since the parameters are calculated from the load profiles, profiles need to be assigned to the customer who does not have it and this assignment is based on the customer's activity as well as the size of load using Table 15. Thereafter parameters may be calculated and as a final step these calculated parameters are compared to the values in Table 16 to verify the cluster in which the customer belongs.

7. DISCUSSION OF THE RESULTS

The studies related to load modelling are ongoing as the power sectors continue to evolve. However, particularly in the South African context, in the past the load modelling studies were focussing mostly on the LV loads and, as a result, there has been significant progress made around the load models for LV networks. Based on the available literature, there appeared to be no equivalent efforts in the development of load models for MV loads. Nevertheless, it was anticipated that this study would provide a process for MV load modelling and that the study results would address this need. The customer composition of MV networks could include several loads and embedded generators connected to MV and/or LV systems. Therefore, simply aggregating load models at the component level in an attempt to provide an MV system view would not provide accurate results. Load models needed to be developed to reflect the load variations in the MV systems using customer measurements.

It was understood from the literature review that when modelling MV loads, both the active and reactive powers were essential for consideration. The analysis of the relationship between the active and reactive power from the MV customer measurements in chapter four (4) supported the literature observations in this regard because it showed how the reactive power varied with loading as well as time for all the customer classes. While it was possible to ignore the reactive component for LV loads, this did not apply to MV loads because the MV lines are longer and there were often large industrial and highly inductive mining loads, connected to the MV networks. Reactive power increases the thermal loading of feeders resulting in increased voltage drops at receiving end busses. The growing number of DG that connect to the MV network also warranted the need for inclusion of the reactive component in load models because these DG are likely to provide reactive power support when needed and at times at a fee as ancillary services or system operator charges. It was recognised by most utilities globally that large loads increased the reactive power (lag), in particular during high load conditions. These utilities are charged for excess reactive power based on various thresholds of power factors. The South African Megaflex tariff has the reactive power charge applied to reactive powers corresponding to power factors below 0.96. Most Megaflex customers are connected to the HV or MV feeders. In Figure 6 it was possible to see how power factors vary with time, and in particular, they seemed to vary with loading, when compared with Figure 5. Therefore, the load model parameters for MV loads should include either reactive energy or the power factor. The models developed in this study included the power factor represented by PF.

Two load models have been developed in this study, namely, the typical day load model and the customer classification model. The process described in chapter 3 produced both load models, which are key for technical, financial and tariff analysis of MV feeders. The importance of developing appropriate load models was to ensure adequate planning of distribution systems, encourage optimal usage of the networks, and design fair and cost-reflective tariffs that will ensure the financial viability of utilities. It emerged from the literature review that classification of customers was central to the load modelling process and it was based on the shared attributes of the load patterns. As pointed in the literature review, these attributes or parameters could be extracted from measurement data in various ways, ranging from data mining techniques to statistical and mathematical models that rely on predefined and/or calculated values. It was commonplace, based on the literature review, that load-modelling frameworks comprised three main stages, namely, parameter identification and derivation, classification, and the profile assignment stages. The extraction of load parameter models for technical and tariff analysis requires consideration of both the predefined or exogenous parameters and the measured (endogenous) parameters linked to the modelling requirements.

In chapter four, the customer measurement data were analysed to establish the basis for deriving the load model parameters. The analysis revealed that customers used energy differently and such usage differences were related to the time of day as well as the seasonality. Scatter plots of the typical loads for each economic class were analysed and the results confined that there was a time factor to consider. There were periods of high as well as low consumption in a typical day. Therefore defining the exogenous parameters that linked time of use in line with the time of use periods used in Eskom was a valid approach.

Using the time-series decomposition method and the results in chapter 4 it was found that there were possibly three seasonal cycles in a year. This finding contradicts the assumptions used in tariff and load modelling in South

Africa, Eskom in particular, where only two seasonal cycles were considered. The results did not necessarily define seasonal patterns based on the levels of demand but the cyclic patterns. This means that defining seasons using these results may not necessarily be the best approach where the cost drivers are related to the demand levels and as well as for network planning where demand levels are the main input for adequacy assessment.

There had been interests in distinguishing loads in terms of weekdays, weekends, pre-holidays, holidays etc. The interest in distinguishing loads this way was mostly driven by the tariff modelling. The tariffs that were based on time-of-use were often different on these different days. In South Africa, there is no peak period for Saturday and Sunday. Based on the interests observed from the literature reviewed, it can be concluded that the load models should also be able to classify loads in terms of weekdays and weekends in line with electricity consumer tariffs defined by most of the South African utilities. The model results in chapter five revealed that indeed demand patterns of the loads' sample were not the same for the different days. The results indicated that there were four distinct load profiles to distinguish the loads in terms of the days. In South Africa at the time of the study, the typical day classes that were defined for tariff purposes were weekdays (Mondays to Fridays), Saturdays and Sundays and all had a different time of use bins (periods). The limitation of the proposed model, which is an opportunity for further exploration, is that it has not separated the holidays and pre-holidays from the data. However, the finding that the weekdays in summer were not the same as the weekdays in winter and similarly the summer weekend were different to winter weekends was of significance.

An ideal load model is a model, which comprises all the parameters that describe the variability of the load. The challenge is on deciding which of the parameter should be used to describe a particular load pattern of a customer. The absence of a global consensus on the parameters for use in load models was highlighted in the literature. It emerged from the results that each parameter explained the variability of load patterns in different ways. To make sense of the parameters to be selected, there is a need to understand how each parameter relates to the time as well as consumption variations of customers. In chapter three, the endogenous parameters were identified and their relevance were described in the context of load model development. Using the process in chapter three, the data was prepared. From the results in chapter six, it was observed that the customer classes resulting from clustering using single or individual parameters differed materially, thus indicating that considering any one of them to the exclusion of others was not optimal.

Considering all of the identified parameters also had its challenges. The major challenge was related to the data dimensions. As the data dimensions increases, it becomes difficult to visualise it and to apply the clustering algorithms. To overcome this challenge, the literature suggested dimension reductions techniques and the PCA technique emerged as a preferred method. The PCA algorithm was used as described in chapter 3, for dimension reduction in the typical day load model as well as the customer classification model. In a customer classification model, all five identified parameters (LF, PF, P_UF, S_UF and O_UF) were reduced to two PCs and these PC were used in a clustering algorithm.

Five clusters resulted from our study and this led to an improvement of the results or quality of the clusters. The validity of the clusters was evident from the analysis based on the number of outliers that were present in the economic classes and almost none in the resulting clusters (see Figure 11 and Figure 53). This implied that customer classes in their current form in the Eskom database were not representative of the majority of the customers within these classes. The study results showed that the current economic classes could be further subdivided, as they comprised different clusters of loads as determined by the load model in chapter 6. This fact was established from the summary of the classification results shown in Table 15 and the load profiles of the sub-classes of each customer class in Figure 60. If the economic activities were not considered, the classification would lean towards fewer classes that correspond to the number of clusters. Fewer classes are ideal for simulation studies, in light of the amount of data that may be required for studies that span over several years. However, the clusters must represent all loads with a high degree of confidence, and this could be slightly challenging considering that, there are new technologies and also the need to reduce cost by customers introduces more variability in data.

The k-means clustering algorithm was found to be an effective and easy tool to implement for classifying MV loads when the parameters were well defined. The silhouettes, elbow and DBI were effective in selecting the optimal number of clusters and validating the results. The decision about the optimal number of clusters was not always obvious and required a broader analysis of the adequacy of the clusters formed. The visual judgement also had a role to play in the final classification of loads. It was also found that cluster adequacy measures on their own could not provide an answer to the question of "how many clusters?" and therefore, subjectivity and judgement based on prior knowledge of the expected results also played an important role in the selection of the parameters. The fact that there was human judgement necessary for determining the number of optimal clusters it is a challenge to develop models that can automatically decide on the number of clusters particularly when using k-means clustering. Further work may be necessary for enhancing cluster adequacy models to limit human judgement.

It was also possible to estimate the parameters as well as the class profiles statistically. The summary statistics and the histograms of these parameters indicated some underlying distribution of these parameters with a possibility of a beta pdf since some parameters were skewed while others were symmetrical around their means. This may be better established by future studies to determine the distribution of MV loads. The results of such studies would be beneficial to PLF models in MV networks, identifying whether load flows due to customer reclassification could result in voltage violations, overloading or improved losses.

A linear regression model can be a useful tool for estimating the parameters. The use of the multivariate, linear regression model has provided an alternative way of estimating parameters with the upper and the lower limits. The regression model may be used to guide the boundaries within which the loads lie. Customers' response, as represented by a change in load profiles, may lead to the reclassification of other customers and affect the load flows. Such an analysis is also necessary and forms part of the technical analysis linked to this work.

Scholars and utilities alike may adopt the load models that were proposed in this study when the load models for MV system studies relating to technical, financial and tariff analysis are desired. Since the data is becoming more available for many utilities, the opportunities of using data to improve load models emerge. However, the computational requirement for processing large data is often a challenge unless these utilities are willing to invest large amounts of money in purchasing the required computer systems. Another challenge relates to the ability to mine and engineer the data in a way that it could be readily available for use in load modelling. The research on alternative methods revealed that sampling techniques could be employed to assist the utilities in overcoming the challenges of dealing with big data. Stratified sampling, which was chosen and used in this study for the data preparation stage, has been preferred for most applications because it was probability-based and could include all known sectors in the strata. Therefore, the samples from stratified sampling techniques were representative. The use of sampling made it possible to create a sample that was used for the pilot study.

8. CONCLUSIONS AND IMPLICATIONS

The aim of this chapter is to use the results of the research to assess the validity of the hypothesis formulated in Chapter 1, and to consider the implications. The hypothesis was stated explicitly and research questions to guide the research were identified. The answers to the research questions are detailed beneath the questions.

8.1. Answers to the research questions

What are the various load models and what is their purpose?

There is sufficient information in the literature to explain load models in detail. Load models may be defined as a way or form of representing a load and its variations. This form could be a mathematical, components based or physical representation. Figure 1 presented a diagram illustrating the context of load modelling and its importance in related technical and tariff application studies. Thus, load models helped to explain various aspects of customers based on their consumption data. It was possible to assess their consumption in winter and summer and on different days of the year. In addition, customers could also be classified for tariff and technical studies.

The various load models reviewed in published literature may be summarised in terms of:

1. Load models for technical analysis, which could be divided into short-term models long-term load models that account for the uncertainties related to the climate, demographic and economic factors.
2. Planning models have the objective of minimising costs, and may be conducted on the least life cycle cost basis and therefore may be key for financial analysis of the utilities.
3. Models for tariff analysis, which included the cost allocation models that attempted to identify cost drivers from the customers' energy usage patterns and the tariff design models that translated the allocated costs to tariffs, could provide policy signals that include signals to drive certain behaviours of customers for efficient use of energy.
4. Classification models aiming to categorise loads may be summarised as:
    a. Classification of feeders from a feeder population
    b. Classifications to determine typical days or typical weeks
    c. Classifying of customers based on usage behaviour to create customer classes
    d. Decomposing of composite feeder profiles into elementary profiles

The load models can be summarised using Figure 67 below.

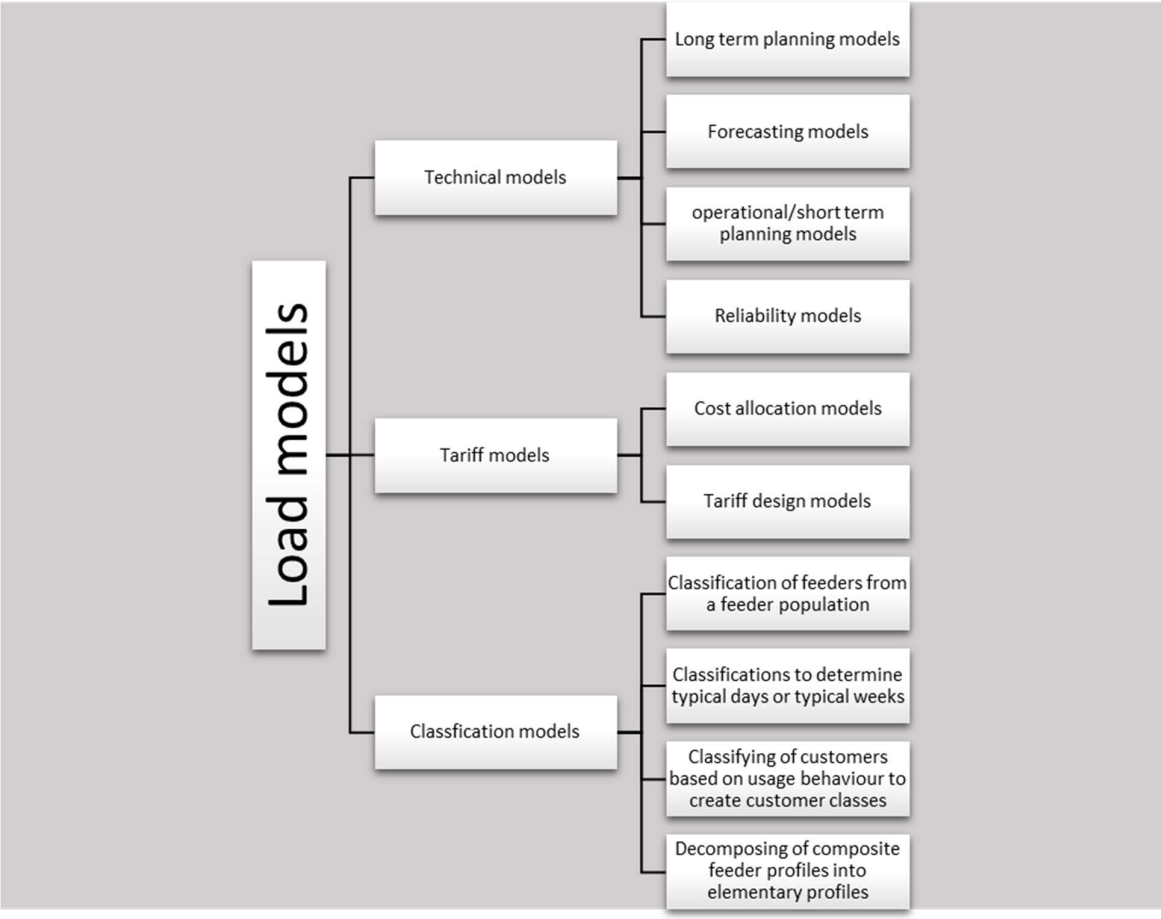


Figure 67: Load models diagram

The purpose of loads models can be summarised as:

- Load models have been developed to
    a. Identify customers or customer groups, classify them and estimate the profiles for the group.
    b. Estimating the demand and supply for network planning.
    c. Identify cost drivers in relation to energy usage by different customers.
    d. Derive load profile distributions for probabilistic for power flow simulations and,
    e. Identify feeders from a feeder population and classify them.

What are the load parameters that are important for technical, financial and tariff analysis models?

It emerged from the literature reviewed that the ability to recognise types of customers and to differentiate between them was an important tool in the design of tariffs to incentivise load-shifting, thus effecting changes in consumption patterns. Therefore, the study was to identify the parameters from the customer load profiles to be able to conduct studies that would make it possible to:

a. plan and design networks that can adequately supply customers,
b. manage the operation of the networks,
c. Design cost reflective tariff models that will ensure optimal usage of the networks as well as ensure financial viability of the utilities.

To identify these parameters, the data were analysed in the context of how the consumption patterns of customers changed during different times of the day and in relation to the defined time of use periods and seasonality. This analysis has led to two propositions to aid in defining the parameters. The first was defining exogenous parameters, which were related to the time of use periods, Peak, Standard and off-peak and seasonality, which is winter and summer. The exogenous parameters are contractual for most users. Some smaller power users may not be on the time-of-use tariffs but their charges may be seasonally differentiated, which includes customers who are on prepaid metering based tariffs. The electricity supply agreements that are entered into with various customers specify the tariff that the customers have chosen and the tariff will have different rates for different time of use periods and the two seasons, winter and summer.

The literature related several attributes that could be used to explain and distinguish loads and these were:

- Contracted power (Maximum demand)
- Active Energy consumption
- Daily/weekly/monthly demand
- Seasonal demand/consumption
- Time-of-use consumption/demand
- Time-series (profiles)
- Daily Load profile segments/sections
- Customer economic activity
- Trend cycles
- Load Factors
- Reactive energy/power
- Lunch time demand
- Morning slope (of the daily profile)
- Night-time consumption
- Principal components(co-variances of data point)
- Undefined/unknown parameters

The undefined or unknown parameters result from the use of machine learning and statistical pattern recognition models such as artificial neural network models.

The periods and seasonality were defined before calculating any other parameter and were used to segment the data. Essentially the exogenous parameter enabled the segmentation of the data set in terms of seasonality and time-of-use periods, and thus providing the desired labels for supervising our clustering algorithm.

c) Is it possible to derive a set of load parameters for technical, financial and tariff analysis from MV feeder measurements data? How can these parameters be derived?

The parameter linked to these time-of-use periods were proposed to be defined as endogenous parameters because they were calculated as the average power for each of the segments. The steps to derive these parameters were described as follows:

Exogenous parameters

1) Data was filtered to select the voltage zone feeder only (MV feeders).
2) Stratified sampling was performed on the data, based on the use of SIC.
3) Seasonality, as defined in South Africa and by Eskom (winter and summer), was used to divide the data for the purposes of clustering and the selection of typical days.
4) Time of use intervals, as defined by Eskom, were used to segment load profiles in order to extract the time-of-use based parameters.

Endogenous parameters

The following endogenous parameters were determined, namely, Parameters LF, PF, P_UF, S_UF and O_UF. These parameters were selected because they formed the basis of technical and tariff application studies. The calculation of these parameters from the measurement data was performed. These parameters were then used together with clustering algorithms to identify the load profiles that belong together and to classify them, thus providing a set of coherent load model parameters as stated above.

d) How can the clustering algorithms be used to classify loads in load modelling of MV systems?

Clustering frameworks have been reported to be successful in the grouping of loads and feeders into different classes by various researchers. Clustering is defined as an unsupervised data mining technique whereby similar data is placed in related or homogeneous groups without advanced knowledge of the definitions of the groups. K-means clustering was used in this study to cluster days, to create representative days and to cluster the load profiles associated with each customer to classify the customers (Aghabozorgi et al., 2015). However, clustering can also be supervised. In the case of this study, supervision was provided with exogenous factors. Supervision was conducted in such a way that the endogenous parameters were defined within the context or boundaries set by the exogenous parameters. Load profiles were determined for each selected class and presented in the results.

e) How can the clustering results and the load models be tested and validated?

Silhouettes SSE (elbow) and Davis Bouldin indicators (DBI) can be used to validate the compactness and validity of the clusters. Amongst many reviewed cluster adequacy measures the three used in this study had been recommended for use with k-means clustering algorithm by various researches in the literature. The use of these validity indices was vital in this study. These indices were used to estimate the number of clusters and to validate the already clustered data. It should be noted that by using these indices to determine the number of clusters, cluster validity is confirmed and therefore the resulting clusters were validated.

It was shown in Chapters 5, 6 and 7 that statistical analysis is an important tool in analysing data. This study also used ANOVA, histograms, statistical summaries and regression models to validate the resulting classification, load models. The validity of the load models was validated using regressions statistics. The use of the p-value threshold of 0.005 helped to establish the significance of the identified parameters in describing the clusters. The histograms of the residuals were also plotted for all the clusters. As indicated in Chapter 6, these residuals were sufficiently small to indicate the sound performance of the model. Statistical summaries indicated the averages and standard deviations of each cluster. The results in Figure 60 indicate that the load models suggested for customer classification were accurate and considered the existence of a variety of profiles within the economic activity based customer classes.

8.2. Unpacking and validating the hypothesis

The research hypothesis stated that: ***Coherent customer load models suitable for technical, financial and tariff analysis of medium voltage systems can be derived from customers' load measurements and the characteristic parameters.***

The research hypothesis was validated through the identification of the exogenous and endogenous parameters and the proposition made on distinguishing between them. A process for load model development has been proposed. Two load models have been proposed and experimented with within this study, using the customer measurements data. The load models parameters have been justified for coherence based on their relevance in technical and tariff studies, which could be key for the financial sustainability of utilities. The parameters were identified and were named LF, PF, P_UF, S_UF and O_UF, which represented the normalised peak, standard (shoulder peak) demands, off-peak demand, load factor, power factor and the normalised average power, was assigned the symbol Pav_UF. The first model is referred to as the typical day model, where the typical day profiles for each season were derived. The second results provided different customer classes and profiles for each class. In addition, a process for developing load models was identified. The results of the statistical analysis also indicated that the classes modelled were valid.

8.3. Implications

Load research has been ongoing in South Africa since 1994. This load research has resulted in the development of various models as well as the recognition that LV loads follow a beta distribution in South Africa. Subsequently, there had been development in PLF modelling of LV systems, where an HB transform was adopted which resulted in improvements in simulation time and which also outperformed the Monte-Carlo simulation while maintaining the same accuracy of results. National standards in South Africa were updated with the models and these models were adopted for LV modelling nationally. The recent developments of the original HB algorithm into the Extended HBE transform for PLF modelling of MV feeders by Chihota and Gaunt, (2018) has highlighted the need for realistic MV load models. The load models for MV feeders developed in the research of this thesis may provide realistic inputs into any technical and tariff studies conducted on MV feeders.

Many utilities have adequate database systems to store large data sets, and thus the measurement data is available for desired studies. The research results on customer classification models indicate that it is possible to recognise patterns from load profiles using limited parameters and this indicates that an acceptably small number of load model parameters can probably be derived from larger data sets. The significance of this finding is that it is not necessary to look for complex models with an excessive number of parameters to define the characteristics of loads in a way that they can be classified in clusters. It was expected that the load models will be useful in technical analysis and for financial analysis since the models described the characteristic behaviour of the customers in each cluster, and the parameters of the load models were relevant to the technical and financial operations. Therefore, it was expected that the development of load models generally similar to this research might be useful to most medium and large utilities with MV customers.

In a scenario where behind-the-meter demand management and energy storage were widespread, smart metering technologies would generally be used. These smart meters may provide the desired information about how the customers consumed and managed their energy. Customer categorization could still be desired largely, but machine-learning techniques might be necessary to segregate customers accurately. Probability-based sampling techniques such as stratified sampling could be applied with the strata defined from the results of the machine-learning categorization. At the utility level, there would be an increase in uncertainty and it would need to be taken into consideration when the load model parameters were determined. This means that simple estimations of parameters may be replaced by stochastic approaches. For example, the Peak usage factor (P_UF) could be modelled statistically, and an empirical distribution assigned to it. Thereafter a Monte-Carlo simulation could be performed or Herman Beta transform used to estimate the P_UF value to use. The same approach could be done also for other parameters of interest, before using them in the load modelling procedure.

In the context of increasing electricity tariffs, unreliable networks, and poor performing generation fleets, as is the case with many developing countries, the behind the meter technologies offer customers greater costs and reliability benefits. The behind-the-meter technologies such as alternative power sources, battery storage, and load management devices are also pivotal in lowering the marginal cost of electricity. The lower costs of electricity may in turn benefit the customers. However, the role of the utility's systems in providing a backup for power supply, reliability and efficiency, will remain important in the future. The traditional distribution systems would possibly be replaced by smart grids and communication between the utility's devices and the customers' devices would be important. It may also be economical for customers to remain connected to the grid to buy electricity at a lower cost, considering the decreasing marginal cost of electricity and that there would be a possibility that there could be surplus energy during certain times of the day, which could benefit the customers. Load modelling at the utility level would still be necessary to ensure that smart networks were adequately designed and operated. To remain financially viable, most utilities could adopt flexible and dynamic pricing structures that would benefit customers if they bought energy during certain periods in a day, while still paying the utility for some of the services. These types of tariff structures would require predictive models to forecast the prices and such models would need robust load models at the utility level.

References

Aamir, O., 2014. Sample size estimation and sampling techniques for selecting a representative sample. *Journal of Health Specialities, 2(4)*, pp. 142-147.

Acharya, A., Prakash, A., Saxena, P. and Nigam, A., 2013. Sampling: Why and how of it?. *Indian Journal of Medical Specialities, 4(2),* pp. 330-333.

Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.Y., 2015. Time-series clustering–a decade review. *Information Systems, 53*, pp.16-38.

Al-Wakeel, A., and Wu, J., 2016. K-means based cluster analysis of residential smart meter measurements. *Energy Procedia, 88*,pp 754-760.

Al-Wakeel, A., Wu, J. and Jenkins, N., 2017. k-means based load estimation of domestic smart meter. *Applied Energy*, *194*, pp. 333–342.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J.M. and Perona, I., 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition, 46(1),* pp.243-256..

Ballanti, A. and Ochoa, L.F., 2015, September. Assessing the effects of load models on MV network losses. *In 2015 Australasian Universities Power Engineering Conference (AUPEC)* (pp. 1-6). IEEE.

Barreiro, P. L. and Albandoz, J. P., 2001. Population and sample: Sampling techniques. *Management Mathematics for European Schools*, pp. 1-18.

Energy, E.C.D.G., 2016. Impact assessment study on downstream flexibility, price flexibility, demand response & smart metering. *European Commission Publications*.

Benítez, I., Quijano, A., Díez, J.L. and Delgado, I., 2014. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. *International Journal of Electrical Power & Energy Systems, 55*, pp.437-448.

Black, J., Hoffman, A., Hong, T., Roberts, J. and Wang, P., 2018. Weather data for energy analytics: from modeling outages and reliability indices to simulating distributed photovoltaic fleets. *IEEE Power and Energy Magazine, 16(3)*, pp.43-53.

Bobric, E. C., Cartina, G. and Grigoraş, G., 2009. Clustering Techniques in Load Profile Analysis for Distribution Stations. *Advances in Electrical and Computer Engineering vol 9 number 1*, pp. 63-66.

Broderick, R.J., Williams, J.R. and Munoz-Ramos, K., 2014. Clustering method and representative feeder selection for the California Solar Initiative. *Sandia National Laboratories report SAND2014-1443: Albuquerque, New Mexico, February 2014*.

Buys, L. and Gaunt, C.T., 2020, January. Load models for technical and tariff analysis of medium voltage feeders. In *2020 International SAUPEC/RobMech/PRASA Conference* (pp. 1-6). IEEE.

Milanovic, J.V., Djokic, S., Matevosyan, J., Resende, F.O., Korunovic, L.M., Dong, Z.Y. and Yamashita, K., 2012. Modelling and aggregation of loads in flexible power networks–scope and status of the work of CIGRE WG C4. 605. *IFAC Proceedings Volumes, 45(21),* pp.405-410.

Carreira-Perpinán, M.A., 1997. A review of dimension reduction techniques. *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, *9*, pp.1-69..

Chen, P., Chen, Z. and Bak-Jensen, B., 2008, April. Probabilistic load flow: A review. In *2008 Third International Conference on Electric Utility Deregulation and Restructuring and Power Technologies* (pp. 1586-1591). IEEE.

Chicco, G., Napoli, R., Postolache, P., Scutariu, M. and Toader, C., 2001,. Electric energy customer characterisation for developing dedicated market strategies. *In 2001 IEEE Power Tech Proceedings, Porto (Cat. No. 01EX502)* (Vol. 1, p. 6).

Chicco, G., Napoli, R. and Piglione, F., 2006. Comparisons among clustering techniques for electricity customer classification. *IEEE Transactions on Power Systems, 21(2*), pp. 933-940..

Chicco, G., Napoli, R., Piglione, F., Postolache, P., Scutariu, M. and Toader, C., 2002, September. A review of concepts and techniques for emergent customer categorisation. In *TELMARK Discussion Forum European Electricity Markets, London*.

Chicco, G., Napoli, R., Piglione, F., Postolache, P., Scutariu, M. and Toader, C., 2004. Load pattern-based classification of electricity customers. *IEEE Transactions on Power Systems, 19(2*), pp.1232-1239..

Chicco, G., Napoli, R., Postolache, P., Scutariu, M. and Toader, C., 2003. Customer characterization options for improving the tariff offer. *IEEE Transactions on Power Systems, 18(1),* pp.381-387.

Chihota, M.J. and Gaunt, C.T., 2018, June. Transform for probabilistic voltage computation on distribution feeders with distributed generation. *In 2018 Power Systems Computation Conference (PSCC)* (pp. 1-7).

Chuan, L., and Ukil, A. (2014). Modeling and validation of electrical load profiling in residential buildings in Singapore. *IEEE Transactions on Power System*s, *30(5)*, 2800-2809.

Conti, S. and Raiti, S., 2007. Probabilistic Load Flow for Distribution Networks with Photovoltaic Generators Part 1: Theoretical Concepts and Models. *D.I.E.E.S. - Università degli Studi di Catania*, pp. 132-136.

Cousins, T., 2009. Using time of use (TOU) tariffs in industrial, commercial and residential applications effectively. *TLC Engineering Solutions*..

Davies, D.L. and Bouldin, D.W., 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), pp.224-227.

De Sá Ferreira, R., Barroso, L.A., Lino, P.R., Carvalho, M.M. and Valenzuela, P., 2013. Time-of-use tariff design under uncertainty in price-elasticities of electricity demand: A stochastic optimization approach. *IEEE Transactions on Smart Grid*, *4*(4), pp.2285-2295.

Eid, C., Koliou, E., Valles, M., Reneses, J. and Hakvoort, R., 2016. Time-based pricing and electricity demand response: Existing barriers and next steps. *Utilities Policy*, 40, pp. 15-25..

Elkarmi, F. 2008. Load research as a tool in electric power system planning, operation, and control—The case of Jordan. *Energy Policy*, 36(5), pp. 1757-1763.

ElNozahy, M.S., Salama, M.M.A. and Seethapathy, R., 2013, July. A probabilistic load modelling approach using clustering algorithms. In *2013 IEEE Power & Energy Society General Meeting* (pp. 1-5). IEEE.

Engel, D., Hüttenberger, L. and Hamann, B., 2012. A survey of dimension reduction methods for high-dimensional data analysis and visualization. In *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

EPP, 2008. *South African Electricity supply industry: Electricity pricing policy*. Department of Minerals and Energy .Pretoria.

Eskom, 2017. *Strategic direction and tariff design principles for Eskom's tariffs*. Eskom. Johannesburg.

Feltner, L., 2012. *Overview of electric cost of service studies*, The Prime Group LLC. http://www.theprimegroupllc.com/COSS_Overview.php

Ferguson, I.A. and Gaunt, C.T., 2003. LV network sizing in electrification projects-Replacing a deterministic method with a statistical method. *In 17th International Conference on Electricity Distribution (Cired)* (No. 68, pp. 1-6).

Ferraro, P., Crisostomi, E., Tucci, M. and Raugi, M., 2016. Comparison and clustering analysis of the daily electrical load in eight European countries. *Electric Power Systems Research, 141*, p. 114–123.

Fidalgoa, J. N., Matosb, M. A. and Ribeiroc, L., 2012. A new clustering approach to derive typical load diagrams based on billing data. *Electric Power Systems Research, 82*, p. 27– 33.

Figueiredo, V., Duarte, F.J., Rodrigues, F., Vale, Z. and Gouveia, J., 2003. Electric energy customer characterization by clustering. In *Proc. ISAP* (pp. 1-6).

Figueiredo, V., Rodrigues, F., Vale, Z. and Gouveia, J. B., 2005. An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions On Power Systems, 20(2),* pp. 596-602.

Filchenkov, A., Muravyov S. and Parfenov, V., 2016, Towards cluster validity index evaluation and selection. *IEEE Artificial Intelligence and Natural Language Conference (AINL),* pp. 1-8.

Firestone, R., Magnus Maribu, K. and Marnay, C., 2006. The value of distributed generation under different tariff structures.

Fischer, D., Stephen, B., Flunk, A., Kreifels, N., Lindberg, K.B., Wille-Haussmann, B. and Owens, E.H., 2016. Modeling the effects of variable tariffs on domestic electric load profiles by use of occupant behavior submodels. *IEEE Transactions on Smart Grid, 8(6),* pp.2685-2693.

Fonseca, J. A., Millera, C. and Schlueter, A., 2017. Unsupervised load shape clustering for urban building performance assessment. *Energy Procedia 122*, pp. 229–234.

Fraley, C. and Raftery, A.E., 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal*, *41*(8), pp.578-588.

Frost, A.E., Azaza, M., Li, H. and Wallin, F., 2017. Patterns and temporal resolution in commercial and industrial typical load profiles. Energy Procedia, 105, pp. 2684-2689.

Ganesan, K., Saraiva, T. and Bessa, R., 2019. On the Use of Causality Inference in Designing Tariff to Implement More Effective Behavioral Demand Response Programs. Energies, pp. 1-20.

Gaunt, C.T., Herman, R., Dekenah, M., Sellick, R.L. and Heunis, S.W., 1999, June. Data collection, load modelling and probabilistic analysis for LV domestic electrification. In *International Conference on Electricity Distribution (CIRED)* (pp. 1-6).

Herman, R., Kadada, H. and Gaunt, C.T., 2019. Design Parameters for LV Feeders to Meet Regulatory Limits of Voltage Magnitude.

Gaunt, C. T., Herman, R., Namanya, E., and Chihota, J. 2017. Voltage modelling of LV feeders with dispersed generation: Probabilistic analytical approach using Beta PDF. *Electric Power Systems Research, 143*, pp. 25-31.

Gbadamosi, A., 2017. Dynamic Load Modelling in Real Time Digital Simulator (RTDS).

Gerossier, A., Barbier, T. and Girard, R., 2017. A novel method for decomposing electricity feeder load into elementary profiles from customer information. *Applied Energy, 203*, pp.752-760.

Granell, R., Axon, C. J. and Wallom, D. C., 2015. Clustering disaggregated load profiles using a Dirichlet process mixture model. *Energy Conversion and Management 92*, pp. 507-516.

Green, R., Staffell, I. and Vasilakos, N., 2014. Divide and Conquer? k-Means Clustering of Demand Data Allows Rapid and Accurate Simulations of the British Electricity System. *IEEE Transactions on engineering management, 61(2),* pp. 251-259.

Han, S.W., Kim, J.H., Lee, B.J., Song, H.C., Kim, H.R., Shin, J.H. and Kim, T.K., 2012. Measurement-based static load modeling using the PMU data installed on the university load. *Journal of Electrical Engineering and Technology*, *7*(5), pp.653-658..

Hashe, S., 2012. Geo-based load forecast standard. Eskom Distribution. Johannesburg

Herman, R. and Gaunt, C., 2008. A practical probabilistic design procedure for LV residential distribution systems. *IEEE Trans. Power Delivery, 23,* pp. 2247-2254.

Heunis, S. and Dekenah, M., 2014, April. A load profile prediction model for residential consumers in South Africa. *In IEEE Twenty-Second Domestic Use of Energy* (pp. 1-6).

Hinz, F., Schmidt, M. and Möst, D., 2018. Regional distribution effects of different electricity network tariff designs with a distributed generation structure: The case of Germany. *Energy Policy, 113*, pp.97-111.

Inglesi-Lotz, R. and Blignaut, J., 2011. Estimating the price elasticity of demand for electricity by sector in South Africa. South African Journal of Economic and Management Sciences, pp. 1-7.

Jain, A.K. and Dubes, R.C., 1988. *Algorithms for clustering data*. Prentice-Hall, Inc..

Kitchenham, B. and Pfleeger, S., 2002. Principles of survey research Part 5: Populations and samples. *Software Engineering Notes, 27(5)*, pp. 17-20.

Kourtis, G., Hadjipaschalis, I and Poullikkas, A. , 2011, An overview of load demand and price forecasting methodologies , *2(1)*, pp. 123-150.

Kryszczuk, K. and Hurley, P., 2010, April. Estimation of the number of clusters using multiple clustering validity indices. In *International workshop on multiple classifier systems* (pp. 114-123). Springer, Berlin, Heidelberg.

Lavin, A. and Klabjan, D., 2015. Clustering time-series energy data from smart meters. *Energy Efficiency, 8(4),* pp. 681-689.

Li, F., Li, R., Zhang, Z., Dale, M., Tolley, D. and Ahokangas, P., 2018. Big data analytics for flexible energy sharing: Accelerating a low-carbon future.*IEEE power and energy magazine, 16(3),* pp.35-42.

Layera, P., Feurerb, S. and Jochemc, P., 2017. Perceived price complexity of dynamic energy tariffs: An investigation of antecedents and consequences. *Energy Policy, 106*, p. 244–254.

Liang, Y., Nwankpa, C.O., Fischl, R., DeVito, A. and Readinger, S.C., 1998. Dynamic reactive load model. *IEEE Transactions on Power Systems, 13(4),* pp.1365-1372.

Lijesen, M., 2007. The real-time price elasticity of electricity. *Energy Economics*, pp. 249-258.

Linvill, C., Shenot, J. and Lazar, J., 2013. Designing distributed generation tariffs well. *Montpelier, VT: Regulatory Assistance Project*.

Maitra, A., Gaikwad, A., Pourbeik, P. and Brooks, D., 2008, July. Load model parameter derivation using an automated algorithm and measured data. In *2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century* (pp. 1-7). IEEE.

McNicholas, P. D., 2016. Model-based clustering. *Journal of Classification, 33*, pp. 331-373.

Mooi, E. and Sarstedt, M., A.,2011 Concise Guide to Market Research The Process, Data, and Methods Using IBM SPSS Statistics.

Moravej, Z. and Akhlaghi, A., 2013. A novel approach based on cuckoo search for DG allocation in distribution network. *Electrical Power and Energy Systems*, *44*, p. 672–679.

Muralidharan, K., 2014. A note on transformation, standardization and normalization. *International Journal of Operations and Quantitative Management, IX(1 and 2),* pp. 116-122.

Nassar, M.E. and Salama, M.A., 2015, May. A novel probabilistic load model and probabilistic power flow. In *2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)* (pp. 881-886). IEEE.

Neenan, B. and Eom, J., 2008. Price Elasticity of Demand for Electricity: A Primer and Synthesis Intern.

Omran, W., 2010. Performance analysis of grid-connected photovoltaic systems.

Ortega, M. R., Pérez-Arriaga, J., Abbad, J. R. and Gónzalez, J. P., 2008. Distribution network tariffs: a closed question?. *Energy Policy*, *36(5),* p. 1712–1725.

Panapakidis, I.P., Alexiadis, M.C. and Papagiannis, G.K., 2012, May. Electricity customer characterization based on different representative load curves. In *2012 9th International Conference on the European Energy Market* (pp. 1-8). IEEE.

Pandeeswari, L. and Rajeswari, K., 2015. K-Means Clustering and Naive Bayes Classifier For Categorization Of Diabetes Patients. International Journal of Innovative Science, Engineering and Technology, 2, pp. 179-185.

Payasi, R. P., Singh, A. K. and Singh, D., 2011. Review of distributed generation planning: objectives, constraints, and algorithms. *International Journal of Engineering, Science and Technology, 3(3),* pp. 133-153.

Pérez-Arriaga, I.J., 2013. Challenges in power sector regulation. In *Regulation of the Power Sector* (pp. 647-678). Springer, London.

Picciariello, A., Reneses, J., Frias, P. and Söder, L., 2015. Distributed generation and distribution pricing: Why do we need new tariff design methodologies*?. Electric Power Systems Research,* pp. 370–376.

Picciariello, A., Vergara, C., Reneses, J., Frías, P. and Söder, L., 2015. Electricity distribution tariffs and distributed generation: Quantifying cross-subsidies from consumers to prosumers. *Utilities Policy, 37*, pp. 23-33.

Piscitelli, M.S., Brandi, S. and Capozzoli, A., 2019. Recognition and classification of typical load profiles in buildings with non-intrusive learning approach. *Applied Energy, 255*, p.113727.

Prusty, B. and Jena, D., 2017. A critical review on probabilistic load flow studies in uncertainty constrained power systems with photovoltaic generation and a new approach. *Renewable and Sustainable Energy Reviews, 69,* pp. 1286–1302.

Qiu, W., Zhai, F., Bao, Z., Li, B., Yang, Q. and Cao, Y., 2016, August. Clustering approach and characteristic indices for load profiles of customers using data from AMI. In *2016 China International Conference on Electricity Distribution (CICED)* (pp. 1-5). IEEE.

Rai, P. and Singh, S., 2010. A survey of clustering techniques. *International Journal of Computer Applications*, *7*(12), pp.1-5.

Rauch, J. N., 2014. *Cost of Service Study and Rate Design*, NARUC –USAID presentation. Viewed on 31 October 2019. https://pubs.naruc.org/pub.cfm?id=5388D962-2354-D714-51A8-F5FD79C756F5

Ramírez-Mendiola, J. L., Grünewald, P. and Eyre, N., 2017. The diversity of residential electricity demand – A comparative analysis of metered and simulated data. *Energy and Buildings*, 151, p. 121–131.

Rani, S. and Sikka, G., 2012. Recent Techniques of Clustering of Time Series Data: A Survey. *International Journal of Computer Applications (0975 – 8887),* 52(15), pp. 1-9.

Reiss, P., White, M., 2005. Household Electricity Demand, Revisited. *The Review of Economic Studies. Vol. 72*: pp. 853-883

Renmu, H., Jin, M. and Hill, D. J., 2006. Composite load modeling via measurement approach. *IEEE Transactions on Power Systems, 21(2),* May, pp. 663-672.

Riaz, S., Marzooghi, H., Verbič, G., Chapman, A.C. and Hill, D.J., 2017. Generic demand model considering the impact of prosumers for future grid scenario analysis. *IEEE Transactions on Smart Grid, 10(1),* pp. 819-829.

Ries, J., Gaudard, L. and Romerio, F., 2016. Interconnecting an isolated electricity system to the European market: The case of Malta. *Utilities Policy, 40*, pp.1-14.

Rousseeuw, P., 1987. Sihouettes: a graphical aid to the interpretation and validation of cluster analysis.. *Journal of Computational and Applied Mathematics, 20*, pp. 53-65.

Sangrody, H., Zhou, N. and Qiao, X., 2017, July. Probabilistic models for daily peak loads at distribution feeders. In *2017 IEEE Power and Energy Society General Meeting* (pp. 1-5). IEEE.

Sardá-Espinosa, A., 2017. Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette*, *12*, p.41.

Schneider, K. P., Fuller, J. C. and Chassin, D. P., 2011. Multi-state load models for distribution system analysis. *IEEE transactions on power systems, 26(4), November*, pp. 2425-2433.

Sharma, D.D. and Singh, S.N., 2014. Electrical load profile analysis and peak load assessment using clustering technique. In *2014 IEEE PES General Meeting| Conference and Exposition* (pp. 1-5). IEEE.

Shi, J. H. and Renmu, L., 2003. Measurement-based load modeling - model structure. *IEEE Bologna PowerTech Conference*, pp. 1-5.

Silipo,R.,2015. KD Nuggets News. Viewed on 31 October 2019. https://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html

Simeone, O., 2018. A very brief introduction to machine learning with applications to communication systems. *IEEE Transactions on Cognitive Communications and Networking*, *4*(4), pp.648-664.

Singh D., Misra R. K. and D. Singh, 2007, Effect of load models in distributed generation planning. *IEEE Transactions on Power Systems, 22(4),* pp. 2204-2212,

Soni, M., 2018. *Assessment of geographical based load forecast approach in distribution planning* (Master's thesis, Faculty of Engineering and the Built Environment).

Sun, D.I.H., Abe, S., Shoults, R.R., Chen, M.S., Eichenberger, P.A.E.P. and Farris, D., 1980. Calculation of energy losses in a distribution system. *IEEE transactions on power apparatus and systems*, (4), pp.1347-1356.

Thorndike, R.L., 1953. Who belongs in the family?. *Psychometrika*, *18*(4), pp.267-276.

Thorsnes, P., Williams, J. and Lawson, R., 2012. Consumer responses to time varying prices for electricity. *Energy Policy, 49*, p. 552–561.

Tsekouras, G.J., Kotoulas, P.B., Tsirekis, C.D., Dialynas, E.N. and Hatziargyriou, N.D., 2008. A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. *Electric Power Systems Research*, *78*(9), pp.1494-1510.

Verdu, S.V., García, M.O., Franco, F.J.G., Encinas, N., Marin, A.G., Molina, A. and Lazaro, E.G., 2004, October. Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters. In *IEEE PES Power Systems Conference and Exposition, 2004.* (pp. 899-906). IEEE.

Xu, Y., 2015.*Probabilistic estimation and prediction of the dynamic response of the demand at bulk supply points*. The University of Manchester (United Kingdom).

Zhang, T. and Yang, B., 2016, November. Big data dimension reduction using PCA. In 2016 *IEEE International Conference on Smart Cloud (SmartCloud)* (pp. 152-157). IEEE.