UNIVERSITY OF CAPE TOWN

MASTERS THESIS

Evaluation of Clustering Techniques for Generating Household Energy Consumption Patterns in a Developing Country

---

Author:
Wiebke TOUSSAINT

Supervisor:
A/Prof. Deshen MOODLEY
Co-supervisor:
Prof. Thomas MEYER

A thesis submitted in fulfilment of the requirements
for the degree of Master of Science

in the

Department of Computer Science

August 11, 2019

Declaration of Authorship

I, Wiebke TOUSSAINT, declare that this thesis titled, "Evaluation of Clustering Techniques for Generating Household Energy Consumption Patterns in a Developing Country" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 11 August 2019

Achilles: Well, my master explained to me that shifting between domains can help. It's like switching your point of view. Things sometimes look complicated from one angle, but simple from another. He gave the example of an orchard, in which from one direction no order is apparent, but from special angles, beautiful regularity emerges. You've reordered the same information by changing your way of looking at it.

Douglas R. Hofstadter *Gödel, Escher, Bach*

UNIVERSITY OF CAPE TOWN

Abstract

Faculty of Science

Department of Computer Science

Master of Science

Evaluation of Clustering Techniques for Generating Household Energy Consumption Patterns in a Developing Country

by Wiebke TOUSSAINT

This work compares and evaluates clustering techniques for generating representative daily load profiles that are characteristic of residential energy consumers in South Africa. The input data captures two decades of metered household consumption, covering 14 945 household years and 3 295 848 daily load patterns of a population with high variability across temporal, geographic, social and economic dimensions. Different algorithms, normalisation and pre-binning techniques are evaluated to determine the best clustering structure. The study shows that normalisation is essential for producing good clusters. Specifically, unit norm produces more usable and more expressive clusters than the zero-one scaler, which is the most common method of normalisation used in the domain. While pre-binning improves clustering results for the dataset, the choice of pre-binning method does not significantly impact the quality of clusters produced. Data representation and especially the inclusion or removal of zero-valued profiles is an important consideration in relation to the pre-binning approach selected. Like several previous studies, the k-means algorithm produces the best results. Introducing a qualitative evaluation framework facilitated the evaluation process and helped identify a top clustering structure that is significantly more useable than those that would have been selected based on quantitative metrics alone. The approach demonstrates how explicitly defined qualitative evaluation measures can aid in selecting a clustering structure that is more likely to have real world application. To our knowledge this is the first work that uses cluster analysis to generate customer archetypes from representative daily load profiles in a highly variable, developing country context.

Acknowledgements

This dissertation owes its quality to the care and devotion of my supervisors Deshen Moodley and Tommie Meyer, who gave me the freedom to explore, the encouragement to seek and the guidance to find. They have made my journey towards completion a better experience than I ever anticipated. I thank the Centre for Artificial Intelligence Research and the Department of Science and Technology for awarding me with a scholarship to deliver the research. Further acknowledgement goes to Marcus Dekenah, Prof. Trevor Gaunt and Dr Schalk Heunis, who willingly shared their decades of experience in domestic load research and household electrification with me.

Contents

List of Figures

List of Tables

List of Abbreviations

AMC	**A**verage **M**onthly **C**onsumption
CCI	**C**ustomer **C**lass **I**ndex
CDI	**C**luster **D**ispersion **I**ndex
CI	**C**ombined **I**ndex
DBI	**D**avies **B**ouldin **I**ndex
DLR	**D**omestic **L**oad **R**esearch
DPET	**D**istribution **P**re-**e**lectrification **T**ool
DTW	**D**ynamic **T**ime **W**arping
GLF	**G**eobased **L**oad **F**orecast
MAPE	**M**ean **A**verage **P**ercentage **E**rror
MdAPE	**M**e**d**ian **A**verage **P**ercentage **E**rror
MdLQ	**M**e**d**ian **L**og Accuracy Ratio
MdSymA	**M**e**d**ian **Sym**metric **A**ccuracy
MIA	**M**ean **I**ndex **A**dequacy
MNLR	**M**ulti**n**omial **L**ogistic **R**egression
PAA	**P**iecewise **A**ggregate **A**pproximation
R	South African **R**and
RDLP	**R**epresentative **D**aily **L**oad **P**rofile
SMI	**S**imilarity **M**atrix **I**ndicator
SOM	**S**elf-**O**rganising **M**ap

List of Symbols

C	current	A
E	energy consumption	kW h
P	power	kW
t	time	hour
$l(t)$	load profile	
$_d$	subscript for an individual day	
h_d	household daily load profile	
$^{(j)}$	superscript for an individual household	
$H^{(j)}$	all daily load profiles of a household	
X	clustering input array	
d	daily demand	
k	set of clusters	
$_x$	subscript for an individual cluster in k	
r	representative daily load profile	
R	set of r	
N	set of load profiles	
n	normalised daily load profile	
c	cumulative normalised daily load profile	

This research is dedicated to all the people in South Africa that transform their homes, communities and our country with electricity, labour, passion and dedication.

Chapter 1

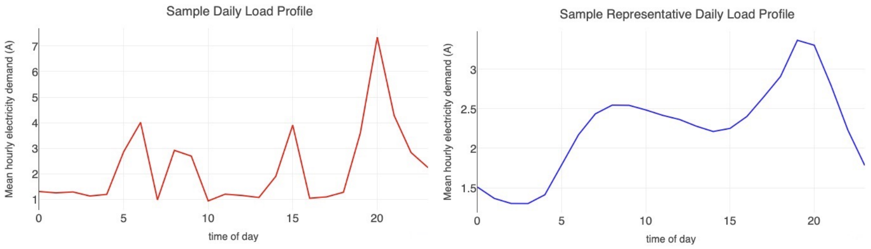
Introduction

1.1 Background and Context

To increase sustainable access to energy, countries need to plan future generation, transmission and distribution capacity of electricity infrastructure to perform grid management and maintenance, as well as grid expansion. Long term energy planning requires insights into the energy consumption behaviour of customers, such as residential households, to build demand forecasts. The promoted shift to renewable energy sources further necessitates localised energy demand models to size mini-grids and assess required generation capacity [39].

The South African NRS Load Research Programme was launched in 1994 to understand how households' social, economic, dwelling and appliance attributes affect their electricity consumption patterns and how this influences the planning and execution of the country's electrification programme, demand side management strategies and long term energy policy [41] [27] [18]. Over a period of two decades from 1994 to 2014, the programme collected metered household energy consumption data and performed an annual household survey. The data collected during the NRS Load Research Programme has been published as the Domestic Electrical Load Study (DELS) dataset [30] [31]. The dataset captures a highly variable population living in rural, informal and urban households, spanning income ranges from unemployed, to pensioners, low and high earning households, stretching over five climatic zones from arid to sub-tropical regions and covering two time zones. It captures both fine grained energy consumption patterns and household attributes.

A daily load profile describes the energy consumption pattern of a household over a 24 hour period (see Figure 1.1). Representative daily load profiles (RDLPs) are indicative of distinct daily energy usage behaviour for different types of households. They are obtained by aggregating a family of daily load profiles that share a common attribute, such as cluster membership or loading condition, e.g. a weekday in winter [15]. With current modelling approaches it is impractical and resource intensive to treat every consumer independently [14]. Customer archetypes (also referred to as customer classes in the literature) are thus developed to represent groupings of energy users that consume energy in a similar manner [4]. RDLPs have been well explored for generating customer archetypes [33] [55] in applications such as tariff development and long term energy modelling [17] [32].



(A) Daily Load Profile (B) Representative Daily Load Profile

FIGURE 1.1: Sample Load Profile Representations

Cluster analysis is an unsupervised machine learning approach that is widely used to create RDLPs [15] [45] [61]. Some challenges with clustering techniques are evaluating algorithm performance and selecting the optimal number of clusters. Algorithms are also highly sensitive to the quality and nature of the input data. Existing clustering results in the energy domain indicate that commonly used metrics are insufficient for guiding the selection of competing algorithms [45], and careful visual evaluation by experts is necessary to overcome this limitation [76] [61]. This further constrains the number of clusters to be small to allow experts to interpret them. In South Africa, economic volatility, income inequality, geographic and social diversity contribute to increased variability of residential energy demand [41], which exacerbates the challenge of creating clusters that are useful for constructing customer archetypes. The focus of this research is to generate, evaluate and select RDLPs for data with large variance, in order to construct customer archetypes for South Africa's highly variable population.

1.2 Problem Statement

Unsupervised machine learning is frequently used for clustering daily load profiles of households to create RDLPs and construct customer archetypes. In developed countries this is an established research domain. Most previous studies focus on the comparison of clustering algorithms and the application of the resultant clustering structure. Limited attention is paid to the effect of data input, data representation and providing a robust evaluation mechanism to select the best set of clusters.

The optimal number of clusters and best clustering algorithms are difficult to determine from quantitative clustering metrics alone, a challenge that is heightened when data are highly variable. While some studies have shown that qualitative evaluation measures provide a more nuanced ranking of clustering algorithms [23], most studies rely on expert judgement, which is costly, time-consuming and not always available in developing countries, for evaluating their performance.

In South Africa in particular, the DELS data captures daily load profiles and household attributes of a population with high variability along temporal, geographic, social and economic dimensions. Consequently, a large number of RDLPs is anticipated. This is impractical for expert validation, but necessary to produce clusters that represent a large variety of consumers. While consumer behaviour is assumed to correlate strongly with weekday and seasonal trends, this assumption should be flexible for customers that are not constrained by the routine of a standard work week, such as people living in rural or informal settings. Pre-filtering the input data along temporal dimensions as is typically done in developed countries, is thus undesirable when the metered population is highly variable. Pre-binning, which is a two-stage clustering approach, was found to improve clustering results for variable populations in prior work [79], but was only tested on data captured over a short time span.

Residential customer archetypes have been developed by domain experts in South Africa, however, the archetypes are static, explanations for how they were developed are not readily available and regenerating them is resource intensive. Some studies have demonstrated how socio-demographic characteristics of households can be used to classify RDLPs and construct customer archetypes with explicitly defined attributes. To our knowledge, these methods have not been explored in developing countries like South Africa.

1.3 Aim and Objectives

The overall aim of this research is to select a useful clustering structure from South Africa's Domestic Electrical Load Study (DELS) dataset to create residential energy customer archetypes. Unsupervised machine learning algorithms and different combinations of pre-binning and normalisation approaches are investigated to generate clusters that represent daily energy consumption behaviour of households in South Africa. The clusters are used to generate, compare and evaluate representative daily load profiles (RDLPs). The best RDLPs are selected and characterised to construct customer archetypes for a real-world application.

1.3.1 Specific Objectives

The specific objectives of this research are to:

1. Develop a qualitative evaluation framework for selecting the best clustering structure based on a specific application context.
2. Compare the results of three clustering and four normalisation algorithms to create RDLPs from South Africa's DELS dataset.
3. Evaluate the effect that pre-binning load profiles has on clustering results.

4. Apply and evaluate the RDLPs for the purpose of developing customer archetypes for long term energy planning.

1.4 Overview of Research Design

The study is limited to data collected during the South African NRS Load Research Programme between 1994 and 2014. This dataset was collected to inform South Africa's electrification strategy and presents the most detailed view available of residential energy consumption in South Africa. Cluster analysis is explored to create representative daily load profiles (RDLPs) from the dataset and to use these to construct customer archetypes.

The self-organising maps (SOM) and k-means clustering algorithms, as well as a multi-step algorithm that combines SOM and k-means are implemented and evaluated to compare their suitability for creating good clustering structures. In addition, pre-binning by average monthly consumption and by integral k-means are evaluated alongside four normalisation techniques. The compactness and distinctness of the resulting clustering structures are evaluated with a relative index that combines three quantitative metrics commonly used in cluster analysis in the residential energy domain. The clusters are used to create RDLPs.

To distill the characteristics of good RDLPs for creating customer archetypes, the development of a qualitative evaluation framework is initiated with a domain analysis and expert consultations. The insights gathered are used to formulate competency questions, which inform the development of qualitative evaluation measures. The measures are weighted and combined in a cluster scoring matrix that ranks experiments according to their relative performance. From this evaluation the best experiment is selected.

Multinomial logistic regression is used to classify the best clusters, thus characterising the RDLPs by socio-demographic attributes. The relevance of the approach is demonstrated by using the characterised RDLPs to construct customer archetypes that can be used for long term energy planning in South Africa. The archetypes are evaluated against existing expert benchmarks.

1.5 Contributions of the Study

The primary contribution of this study is the development and selection of RDLPs for residential energy consumers, and the application of the RDLPs to construct customer archetypes for long term energy planning in South Africa. Where most existing research has been conducted in developed countries over short time horizons and limited geographic spread, this work tests the performance of clustering techniques used in the residential energy sector within a developing country context over a two decade period and an expansive geographic region. Best-practice approaches are validated by providing a robust comparison of commonly used clustering algorithms, normalisation and pre-binning techniques.

Unlike most previous studies that select the optimal clustering structure based on quantitative clustering metrics alone, this study develops a qualitative evaluation framework that considers existing expert knowledge of the application context. By combining traditional clustering metrics with qualitative evaluation measures, a clustering structure is selected that is more usable, representative and homogeneous than what would have otherwise been the case. This has the advantage of reducing the reliance on subjective expert validation when selecting the best clustering structure.

This work will be of particular interest to energy generation and utilities companies that are seeking to expand and maintain their infrastructure in developing countries.

1.6 Thesis Outline

The remainder of this thesis is structured as follows:

- **Chapter 2 - Literature Review:** This chapter reviews the relevant literature related to cluster analysis, time series clustering, energy load profiles and residential energy customer archetypes.
- **Chapter 3 - Data and Domain Analysis** This chapter provides an overview of the dataset and of decision and evaluation criteria important to domain experts. The qualitative evaluation measures and cluster scoring matrix are presented.
- **Chapter 4 - Clustering Algorithms:** This chapter discusses the algorithms and experiments, presents the quantitative metrics and explains how RDLPs are derived from clusters.
- **Chapter 5 - Results and Analysis:** This chapter presents the clustering results and their interpretation.
- **Chapter 6 - Application:** This chapter demonstrates the development and evaluation of customer archetypes for South Africa from the best cluster set.
- **Chapter 7 - Discussion:** This chapter discusses and evaluates the impact of the work in light of its intended application.
- **Chapter 8 - Conclusion:** This chapter summarises the research and highlights opportunities for future work

Chapter 2

Literature Review

The literature review starts with a broad overview of unsupervised machine learning and clusters analysis. It proceeds to review distinct aspects of clustering time series data. The input data and data representation, clustering algorithms and parameters, and evaluation methodologies of 25 studies that use cluster analysis to construct energy load profiles are then analysed. The chapter concludes by discussing approaches to constructing residential energy customer archetypes.

2.1 Unsupervised Machine Learning

Machine learning is the field of research concerned with programming computers to automate the process of converting observational data into an output that ’learns’ from the input data [67]. The objective of a learning algorithm is typically to perform a task, such as learning to categorise images, detect anomalies, or group similar observations. The input data to the algorithm is called training data and consists of a set of features that are predictors of the output. If the features in the training data are labelled with the output variable, they can be used to guide the learning process - which is then called supervised learning. In contrast, if the training data contains only feature variables with no labels, the learning process is called unsupervised learning.

The goal of unsupervised machine learning is to discover structure in the input data in the absence of training labels that are otherwise used to approximate the error for each observation [38]. Unsupervised machine learning can be used to estimate the distribution of data in the input space, to reduce high dimensional data to lower dimensions for the purpose of visualisation or to discover groups of similar observations in the data, which is called cluster analysis. This research focuses on cluster analysis.

2.2 Cluster Analysis

Cluster analysis is used to abstract the underlying structure in data as a group of individuals or hierarchy of groups without requiring labelled training data [65]. It is widely used in applications ranging from navigating and organising text documents [80] to discovering cancer-related genes [38]. The goal of cluster analysis is to create clusters (or groups) of objects where objects within a cluster a more closely related to each other than to objects that are not in the cluster. Analysing the degree of similarity (or difference) between objects assigned to the same cluster is thus central to creating meaningful clusters.

Clustering techniques assign a single pattern to a single cluster, thus segmenting a finite set of objects into meaningful subsets within a particular problem context [65]. A desirable clustering structure groups similar objects into the same cluster, and separates dissimilar objects into different clusters. The degree of similarity is quantified with a distance measure that computes the difference between pairs of objects within the clustering structure. The similarity of the set of objects is the average of all the distances. An example of a common distance measure is the Euclidean distance between two points. While the clustering process appears obvious, finding the 'correct' clustering structure is ambiguous. Take the four spheres below, which can be clustered in two equally correct ways.

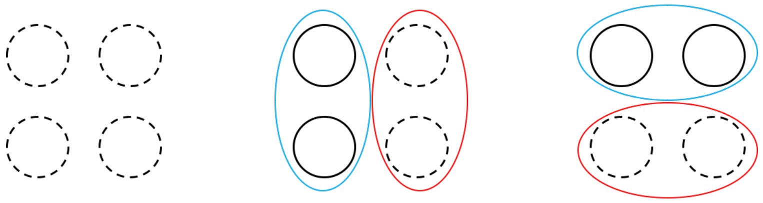


FIGURE 2.1: Example of two equivalent clustering structures for the same set of objects [67]

In real applications there exist different ways of clustering any set of objects into meaningful segments, depending on how the notion of similarity has been defined [67]. The clusters resulting from a set of customers being clustered by age could be very different to clustering the same customers by occupation. In practice clusters can have any shape. However, once a distance measure is applied, it imposes a structure on the data that constrains the shape of the clusters. Euclidean distance, for example, imposes a spherical shape on clusters. True clusters can only be uncovered if the data structure conforms to the distance measure [65].

Given the importance of distance measures, much attention has been devoted to their development in the literature. While many novel measures claim superiority over previous approaches, [47] demonstrates that these claims are made in the absence of evaluation against standard tests. Under rigorous evaluation, they perform significantly worse than Euclidean distance. Ultimately, the distance measure is context dependent and depends on both data attributes and the clustering objective.

2.2.1 Clustering Algorithms

Different types of clustering algorithms are used for cluster analysis, and can produce different sets of clusters for the same set of objects. Traditionally, algorithms perform either hierarchical or partitional clustering. Self-Organising Maps (SOM) present a third, neural network-based approach. These three types of algorithms are discussed below.

Hierarchical Clustering

As the name suggests, hierarchical clustering algorithms produce a hierarchy of clusters, where clusters on each level are formed by merging clusters from one level below [38]. Hierarchical clustering methods differ based on the strategy applied for creating the clusters, as well as the distance measure that defines the degree of similarity between clusters. A clustering strategy is either agglomerative or divisive. An agglomerative, hierarchical clustering is a bottom-up strategy that starts by placing each object in its own cluster and merges individual clusters one-by-one to form increasingly larger clusters. A divisive strategy reverses this process and fragments one large cluster containing all objects into ever smaller clusters in a top-down manner. Agglomerative strategies are more common in the literature than divisive strategies.

Different agglomerative clustering algorithms measure the similarity between clusters in different ways. Single linkage defines the similarity of two clusters to be the distance between the most similar pair of objects in the clusters. Complete linkage does the opposite and takes the maximum distance between two objects in the clusters. Average linkage clustering computes the average distance between objects in both clusters.

For all agglomerative clustering algorithms a stopping criterion must be defined that determines when the merging of clusters ends. The clustering structure can be visualised as a dendogram, which is a graphical representation of complete cluster merging if no stopping criterion is applied. Dendograms are interpretable, which has contributed to the popularity of hierarchical clustering [38]. With several hundred patterns dendograms however become impractical [65].

Partitional Clustering

Partitional clustering algorithms partition (or cluster) input observations into a set of clusters, such that objects in a cluster are more similar to each other than to objets in different clusters [65]. The theoretical solution to this is to select a clustering criterion and to evaluate it for all possible clustering structures that produce the desired number of clusters. The preferred structure is then the one that optimises the criterion.

Practically, this approach would result in an insurmountable number of possible partitions. A more feasible approach is to only evaluate a small subset of partitions that have a high potential of containing the optimal partition. Such strategies specify an initial partition and then change the object assignment in an iterative manner, so that the clustering criterion improves with each iteration. Two challenges with this approach are that algorithms can easily converge to suboptimal local minima, and that the initial number of clusters must be pre-specified.

The k-means algorithm is one of the most common partitional algorithms that works in this iterative fashion [38]. It uses the squared Euclidean distance as clustering criterion to evaluate the similarity between objects and cluster centres. K-means is initialised with random cluster centres. Each object in the data is assigned to the closest cluster centre, and the cluster centres are updated with the value obtained by averaging all objects assigned to that cluster. This processes is repeated until the algorithm converges. Two similar algorithms are k-medoids and k-median. K-medoids works like k-means, but requires the cluster centroids to be members of the dataset. K-medians does not square the distance between an object and its closest centroid [67].

Self-Organising Maps

The Self-Organising Map (SOM) is an artificial neural network based on principles of competitive learning [50]. The principle of competitive learning can be understood as follows. Assume an input vector $x(t)$ with values observed over t time steps, and a set of variable reference vectors $m_i(t)$ with $m_i(0)$ randomly initialised. Imagine simultaneously comparing $x(t)$ with each $m_i(t)$ at time t, and updating the best matching $m_i(t)$ to better match $x(t)$. If the comparison is based on a distance measure, it can be updated in a similar manner so that it is reduced at the index of the best matching reference vector m_i. The different reference vectors thus become 'tuned' to domains in the input variable x.

A simple form of the SOM algorithm consists of a network of cells positioned on a two-dimensional rectangular grid. The observations of the data input are mapped onto this grid by finding the cell that is the closest match to each input x_i in Euclidean distance. That cell and all its neighbours are then updated to move towards x_i. The result of the update is that the cells move closer to the data while the relationship between the cells remains smooth [38].

Amongst other applications, SOM can be used for pattern recognition, dimensionality reduction and noise reduction. Vesanto et al. illustrate the benefit of using SOM in combination with traditional clustering algorithms such as k-means and hierarchical clustering in a two-stage clustering process [75]. Using SOM as an abstraction layer has the advantage of reducing the computational cost of clustering large datasets, due to the smaller set of input objects. The SOM can also be used to provide a rough visual representation of the clusters.

2.2.2 Cluster Evaluation

Validation is the most challenging part of cluster analysis. Clustering algorithms have a tendency of creating clusters in data even when no natural clusters exist [65]. Clustering results must thus be validated with caution. Compactness and distinctness are the two main properties that determine validity of individual clusters. In a compact cluster, members lie close to each other and the cluster centroid. A distinct cluster is separated from its neighbouring clusters. A valid cluster is unusually compact and unusually distinct. The challenge in cluster evaluation lies in establishing a reference against which 'unusual' has meaning, that has a theoretical foundation and that makes intuitive sense [65].

The major objectives of cluster evaluation are to assess clustering tendency of the data, to determine the number of clusters and to evaluate cluster quality [44]. Finding the 'right' number of clusters can be challenging if this number is unknown. The elbow method, which plots the values of a clustering index against the number of clusters, is a popular visual way of establishing the number of clusters [44]. It too has its challenges, as square error, for example, is sensitive to sample size and dimensionality [65].

Clustering Metrics

Validity metrics use external, internal or relative validity indexes that measure how well a clustering structure represents the true structure of the data. External indexes evaluate a clustering structure against prior known class labels in a supervised evaluation process. External indexes are only useful if class labels are available. Internal indexes measure the effect between the clustering structure and the data, using only the data. Establishing internal indexes is challenging and relative indexes can be employed with less difficulty [65]. A relative index quantifies which clustering structure best matches the data when compared against competing structures. Internal indexes can be used as relative indexes. The Silhouette Index and Davies Bouldin Index are frequently used to evaluate clustering structures and are discussed in greater detail.

The Silhouette Index is an internal index that can also be used as a relative index. For an individual pattern p in the dataset

$$silhouette(p) = \frac{distinctness(p) - compactness(p)}{max\{distinctness(p), compactness(p)\}} \tag{2.1}$$

Compactness is the average distance between p and all other patterns in the same cluster. Distinctness is the average distance between p and all remaining patterns that are not in the same cluster. The Silhouette Index has a value between -1 and 1. Patterns in a good cluster will have a small compactness value, a large distinctness value and consequently the Silhouette Index will be approaching 1. If a pattern lies further from patterns in the same cluster than from patterns in other clusters, the Silhouette Index will be negative. The Silhouette Index of a cluster and dataset can be calculated by averaging the Silhouette Indexes of all member patterns [44].

The Davies Bouldin Index (DBI) was designed to compute the system wide average of the similarity of each cluster with its most similar cluster. It can be used to evaluate clusters when no prior knowledge is available of the data structure. For two clusters it is calculated as the ratio of the sum of cluster dispersions, and the distance between the two cluster centroids.

$$DBI(i,j) = \frac{dispersion(i) + dispersion(j)}{distance(i,j)} \tag{2.2}$$

Cluster dispersion can be calculated using different measures. A simple method for computing it is as the average distance between the centroid of a cluster and each pattern in the cluster. The DBI for the dataset is obtained by averaging the similarity measure of each cluster and its most similar cluster, $DBI(i,j)_{max}$, for all clusters. A small DBI value indicates that cluster dispersions are small and distances between clusters are large, which is desirable. When plotting the DBI against the number of clusters, the optimal number of clusters can be visually identified. It is possible for the DBI to have several local minima [20].

Regardless of the index used, in practical applications no single index is likely to provide consistent results across algorithms [8]. A suggested strategy for overcoming this shortfall is to compare very different clustering algorithms, vary their parameters and collect ranked indexes for all experiments. Consistent results are an indication that a meaningful structure may exist in the data.

2.2.3 Summary

This section discusses cluster analysis as an unsupervised machine learning approach that is useful for finding groups in a dataset when no labelled training observations are available. Three common types of clustering techniques used in cluster analysis are hierarchical clustering, partitional clustering and Self-Organising Maps (SOM).

K-means is introduced as a popular partitional clustering algorithm and an intuitive interpretation of the SOM is provided. Two clustering indexes used for evaluating clustering structures are reviewed. Previous studies highlight that cluster evaluation is a challenging task. A comparison of multiple indexes is thus recommended to evaluate whether results are consistent.

2.3 Cluster Analysis of Time Series Data

Time series clustering is a special type of clustering that operates on temporal sequences [1]. In time series clustering the objects can be characterised as patterns, and the objective of a clustering algorithm is to find the partition of the dataset such that the distance between patterns in the same cluster is minimised. Applications are concerned with seeking to discover frequently occurring patterns, identifying anomalous patterns or generating representative patterns [34] in a large variety of domains [1]. Time series clustering is challenging in that time series are high dimensional, with high feature correlation and often large amounts of noise [47]. They represent observations of processes that change over time [54] and tend to be large in size, which increases clustering run times [1].

Most time series clustering algorithms that have been developed either adapt existing algorithms used for clustering static data, or convert the time series data into a static form so that static algorithms can be used directly [54]. Algorithms of the first kind are referred to as shape-based or raw-data-based approaches, as they are employed on raw time series data. The algorithms applied are the same as those used for clustering static data, but the distance measures are modified to accommodate time series. Algorithms of the second kind are either feature-based or model-based, deriving feature vectors and model parameters respectively from the raw time series data before applying a conventional clustering algorithm [1][54].

Critical factors that determine time series clustering success are the selection of an appropriate distance measure, choices around data representation, feature representation and input data size, algorithm selection and the selection of suitable evaluation measures [26][47][54]. The suitability of a particular clustering algorithm is highly dependant on both the data and problem context, which may require invariance towards particular distortions such as warping, phase, offset or amplitude scaling [5].

2.3.1 Distance Measures

Maximising the similarity between time series patterns is the key objective of the clustering process. A similarity function applies a distance measure to calculate the distance between two time series. Distance measures that compare time series on a one-to-one basis are known as lock-step measures and include the ubiquitous

Euclidean distance, also known as L_2 norm, and its variants. Elastic measures such as Dynamic Time Warping (DTW) allow for one-to-many and one-to-none comparisons [26]. Other approaches include feature-based and model-based approaches [66].

The size of the dataset has a significant effect on the classification accuracy and speed of elastic distance measures [26]. Comparing DTW with Euclidean distance, Ding et al. found that as the dataset size increases, so does the classification accuracy of Euclidean distance. For datasets with several thousand objects there is no statistically significant difference in accuracy between the two measures. In terms of speed, the speed of DTW increases and approaches that of Euclidean distance as the size of the dataset increases.

Most comparative studies focus on the evaluation of distance measures for time series classification, rather than clustering. The choice of distance measure depends on the size of the dataset, time series characteristics such as the regularity of the sampling interval and known distortions, time series length, data representation and the objective of the clustering problem [1][5]. Euclidean distance, like other lock-step measures, fixes the mapping between points and the measure is unable to handle time shifts and out-of-phase patterns [26]. For complex patterns and invariance towards particular distortions, better distance measures have been proposed [5]. If distortions and phase shifts are not a concern, sampling intervals are regular, time series length is fixed and the training dataset is large, Euclidean distance is still considered a competitive distance measure that is easy to implement, parameter free and has linear complexity [26].

For a unique application of clustering building energy patterns, [43] show that Euclidean distance obtains the best, general clusters. The work however is based on a single, small dataset and does not present evidence that the results are generalisable.

2.3.2 Data Representation

Considering the central role of distance measures in clustering algorithms, data normalisation is an essential preprocessing step to ensure that amplitude and offset distortions do not dominate the differences in pattern shapes captured by the distance measure [5]. Without normalisation, time series similarity is considered meaningless [47]. Different approaches to normalisation include z-normalisation, standardisation, scaling to a range of zero-one and unit normalisation.

Beyond normalisation, time series data representation techniques are concerned with reducing noise and reducing data size to a manageable size in order to increase computing speed and decrease storage requirements, while maintaining the local and global features of the original data [78]. Outliers, missing data and erroneous observations make time series data inherently noisy, which presents a challenge to clustering algorithms that are typically sensitive to the quality and representation of input data [44]. Choosing an appropriate data representation can help reduce noise.

Due to the high dimensionality of time series data, processing and storing it in its raw format is expensive [26]. Dimensionality reduction is used to transform time series either by reducing the data into a lower dimensional space (for example from 5 minute intervals to hourly intervals) or by feature extraction (for example extracting the time of peak occurrence and the peak value). Dimensionality reduction has the benefit of reducing memory requirements and the time required for computing similarity functions [1].

There are two high level categories of data representation. Data adaptive methods choose transformation parameters based on the data properties, while non-data adaptive methods fix parameters, irrespective of the features of the data [49]. Piecewise Aggregate Approximation (PAA) [48] is a popular non-data adaptive method that can be understood intuitively as dividing a time series into n equal-length segments, and representing each segment by the mean of the data values in that segment. Based on a thorough comparative study, Ding et al. suggest that the choice of data representation technique is application dependant and that there is no single best, generic approach [26].

2.3.3 Summary

This section reviews approaches to cluster analysis for time series data, which often has high dimensions and is large in size. The choice of distance measure is critical when clustering time series. Data and context related attributes must be taken into account when selecting the distance measure. When the dataset is large, intervals are regular and time series length is fixed, Euclidean distance remains a competitive distance measure that is fast to execute. Further considerations when clustering time series data are normalisation and dimensionality reduction techniques. Piecewise Aggregate Approximation is often used to average time series over longer time windows.

2.4 Clustering Energy Load Profiles

In the energy domain cluster analysis is used extensively to segment energy consumers for targeted energy efficiency campaigns [2], pricing [13], energy forecasts [53] and small scale renewable generation [79]. Depending on the application and the data available, clustering techniques are either applied to socio-demographic attributes collected through household surveys [39], or to metered energy consumption time series data [45]. This review is limited to the application of clustering techniques to metered energy consumption data, as the resulting clusters can be used for generating representative daily load profiles [51].

This section reviews 25 studies from the past two decades that cluster load profiles of energy consumers for the purpose of generating representative daily load profiles and customer archetypes. Details of the studies and related works are discussed, and their applicability within the South African context is considered. The section starts with a summary of the studies under consideration, followed by a discussion on the representation of load profile data in the energy domain. An overview of algorithms used for clustering energy consumers is provided. The section concludes with a review of the evaluation measures used to determine the best clusters.

2.4.1 Overview of Reviewed Studies

The 25 reviewed studies are summarised in Table 2.1. They are analysed in relation to their input data and data representation, the clustering algorithms and parameters, and the evaluation methodologies, as these have a significant impact on achieving good clustering results. The abbreviations used for algorithms, distance measures and clustering indexes are listed in tables 2.2, 2.3 and 2.4 respectively.

2.4.2 Data Input and Representation

Time-varying energy consumption patterns are called load profiles or load curves. A daily load profile captures the average load drawn from the electrical grid over a metered interval as a current (A), power (kW) or energy (kWh) value. If a daily load profile averages consumer behaviour for a particular loading condition, such as a particular year, season, month and daytype, it is called a representative daily load profile (RDLP).

For residential electricity consumers, behavioural patterns vary throughout the day, on different days of the week and in different months according to consumers' preferences. Given the diurnal nature of household energy consumption behaviour, the standard representation of a residential load profile is a t dimensional vector spanning over a 24 hour period from 00:00 to 23:59. It is common to have separate RDLPs for summer and winter months [74]. A further distinction is usually drawn between weekday and non-weekday profiles [41][59][29].

Load Profile Feature Extraction

Fine-grained metered observations are frequently reduced using Piecewise Aggregate Approximation with 15, 30 or 60 minute windows, to produce input vectors of 96, 48 or 24 dimensions respectively [63][77][19]. Other data reduction methodologies extract features such as total demand, peak demand and number of peaks [11][13], or apply dimensionality reduction using Principal Component Analysis [28] or Self-Organising Maps [62]. Xu et al. represent daily load profiles as a normalised vector that sums consumption over time, to capture load shape as well as consumption levels [79]. Granell et al. investigate the impact of temporal resolution on clustering algorithms in the residential energy domain [37]. The study finds that cluster quality is best at a resolution of 8 or 15 minutes. For the k-means algorithm performance is robust in a band of temporal resolutions between 4 to 60 minutes.

Authors	**Ref**	**Input patterns**	**Customers**	**Dataset**	**Notes**
Batrinu (2005)	[6]	234	234 non-residential	unspecified	IRC developed by authors, builds on Chicco (2003)
Bidoki (2010a)	[9]	127	127 non-residential	unspecified	performance objective dependent
Cao (2013)	[11]		4225 households	Irish CER dataset	k-means pre-binning
Chelmis (2015)	[12]		115 buildings	USC campus microgrid data	EVI developed by authors
Chicco (2002a)	[13]	471	471 non-residential	Romanian national electricity distribution company	
Chicco (2003)	[16]		234 non-residential	unspecified	suggests potential for pre-binning
Chicco (2006)	[17]		234 non-residential	unspecified	
Dang-Ha (2017)	[19]	3 090	3090 households	Hvaler dataset	
Dent (2014)	[24]	180	180 households	NESEMP	clusters & segments by peak time flexibility
Dent (2014a)	[25]	204	204 households	NESEMP	assesses variability of energy demand
duToit (2016)	[28]		11kV & 22kV feeders	Eskom	shows that PCA dim reduction & NUBS centroids improve results & run time
Figueiredo (2005)	[33]		165 small consumers	Portuguese Distribution Company	
Jin (2016)	[46]	32 611 421	residential	unspecified	
Jin (2017)	[45]	104 673	325 households	unspecified	clustering for preprocessing & data reduction to segment customers
Kwac (2013)	[52]		85 households	PG&E	investigates consistency of consumption
Kwac (2014)	[51]	44 949 750	123 150 households	PG&E	AKM developed by authors
McLoughlin (2015)	[55]		3941 households	Irish CER dataset	
Panapakidis (2018)	[60]	365	1 small industrial user	unspecified	develops a cluster algorithm selection framework
Ramos (2012)	[61]		208 non-residential	Portuguese Distribution Company	
Rasanen (2010)	[62]	3989	3989 small consumers	unspecified	
Rhodes (2014)	[63]	103	103 households	Pecan Street Project	
Teeraratkul (2018)	[70]	23 254	1057 households	Opower Corporation (Oracle)	
Tsekouras (2007a)	[73]		94 non-residential	Greek Public Power Cooperation	applies two-stage clustering
Viegas (2015)	[77]		1972 households	Irish CER dataset	
Xu (2017)	[79]	19 070	residential	Pecan Street Project	best results when applying two-stage clustering

TABLE 2.1: Overview of literature on clustering energy consumers

		Data representation				Data variability	
Authors	**Ref**	**RDLP aggregation**	**Dimensions**	**Interval**	**Norm**	**Time range**	**Spatial cover**
Batrinu (2005)	[6]	weekday (spring)	96	15min	[0,1]		
Bidoki (2010a)	[9]	annual	96	15min	[0,1]	1 year	
Cao (2013)	[11]	weekday	48, 18	30min, features	[0,1]	4 weeks	Ireland
Chelmis (2015)	[12]	none	96	15min		spring semester Monday	University of Southern California, US
Chicco (2002a)	[13]	weekday (winter)	4	features	[0,1]	3 weeks	
Chicco (2003)	[16]	weekday (spring)	96	15min	[0,1]		
Chicco (2006)	[17]	weekday (spring)	2 - 6	features			
Dang-Ha (2017)	[19]	summer, winter, weekday, weekend	96	60min	[0,1]	1 year	Hvaler, Norway
Dent (2014)	[24]	weekday evening peak	2 - 8	features	[0,1]		North East Scottland
Dent (2014a)	[25]	spring weekday evening peak	48, 42	5min, motifs	[0,1]	3 months	North East Scottland
duToit (2016)	[28]	none	48, 8	30min, features	standardise	2 summer months x 8 years	
Figueiredo (2005)	[33]	summer, winter, weekday, weekend	96	15min	[0,1]		
Jin (2016)	[46]	none	24	60min	de-minning	1 year	California, US
Jin (2017)	[45]	none	24	60min	de-minning	1 year	California, US
Kwac (2013)	[52]	none	96, 24	15min, 60min	unit norm	3 summer/ autumn months	City in San Francisco Bay Area, US
Kwac (2014)	[51]	none	24	60min	unit norm	13 months	California, US
McLoughlin (2015)	[55]	none	24	60min		6 months	
Panapakidis (2018)	[60]	none	24	60min	[0,1] custom	1 year	
Ramos (2012)	[61]	weekday	96	15min	[0,1]	6 months	
Rasanen (2010)	[62]		489	features		1 year	Northern Savo, Finland
Rhodes (2014)	[63]	summer, autumn, winter spring	24	60min	[0,1]	1 year	Austin, US
Teeraratkul (2018)	[70]	none	24	60min	unit norm	22 days	
Tsekouras (2007a)	[73]		96	15min	[0,1] custom	10 months	
Viegas (2015)	[77]	summer, autumn, winter spring	48	30min		18 months	
Xu (2017)	[79]	none	96	15min	unit norm	1 month	18 cities, US

Authors	**Ref**	**Distance measure**	**Clustering Algorithms** * is best performing	**Cluster range**	**Evaluation**
Batrinu (2005)	[6]	Euclidean	*IRC, HC, k-means, fuzzy k-means, MFTL	2-15	ScatI, VRC, MIA, CDI
Bidoki (2010a)	[9]	Euclidean	k-means, *WFAKM, MFTL, SOM, HC		MIA, CDI
Cao (2013)	[11]	Manhatten, Euclidean, *correlation, cos	HC, *k-means, SOM+k-means	5- *14	peak overlap, Hamming distance, Wiener filter
Chelmis (2015)	[12]	Euclidean, Hausdorff	k-means, HC, k-medoids, Voronoi decomposition	inconclusive	DI, EVI, VRC
Chicco (2002a)	[13]	w-Euclidean	FTL	7, 9	MIA, CDI
Chicco (2003)	[16]	Euclidean, w-Euclidean	k-means, *MFTL, SOM, HC(ward), HC(avg), fuzzy k-means	10-20; inconclusive	MIA, SMI, CDI, DBI
Chicco (2006)	[17]	Euclidean, w-Euclidean	k-means, *MFTL, SOM, HC(comp), HC(ward), *HC(avg), fuzzy k-means	10-30; inconclusive	CDI, DBI, MDI, ScatI
Dang-Ha (2017)	[19]	Euclidean, cos, Minkowski	k-means, SKM, *SOM, HC (ward), HC(avg), HC(single)	2-50; inconclusive	CDI, DBI, MDI, MIA
Dent (2014)	[24]		k-means	2-10	CDI, DBI, MIA, SMI, BH
Dent (2014a)	[25]	Euclidean	k-means, fuzzy c-means, SOM, HC, Random Forests	8	MIA, CDI
duToit (2016)	[28]	Euclidean^2, DTW, PCC, cos	k-means	*5	DBI, SilhI
Figueiredo (2005)	[33]		SOM+k-means	6-9	MIA
Jin (2016)	[46]		AKM+HC		DBI
Jin (2017)	[45]	Chebyshev, Euclidean	*k-means, *kmedoids, *AKM, *HC(ward), HC(avg), HC(comp), GMM, DBSCAN	10-500	CDI, DBI, MIA, SilhI, SMI, VRSE
Kwac (2013)	[52]		AKM+HC		threshold
Kwac (2014)	[51]		AKM+HC		threshold, entropy
McLoughlin (2015)	[55]		k-means, kmedoid, *SOM	2-16	DBI
Panapakidis (2018)	[60]		k-means, MKM(1), MKM(2), various HC, fuzzy c-means, SOM, others	2-30	CDI, DBI, MIA, SMI, ScatI, VRC, others
Ramos (2012)	[61]	Euclidean	*k-means, HC (ward), HC (avg), HC(comp), HC(norm cut)	2-30	DBI, DI, SilhI, J, others
Rasanen (2010)	[62]	Euclidean	*SOM+k-means, SOM+HC(comp)	2-30	DBI, IA
Rhodes (2014)	[63]	Euclidean	k-means	*2	
Teeraratkul (2018)	[70]	Euclidean, *DTW	k-means, *kmedoids, E&M		WCS, WB, WCBCR
Tsekouras (2007a)	[73]	Euclidean	k-means, AVQ, fuzzy k-means, HC	5-25	CDI, MIA, SMI, DBI, WCBCR, J
Viegas (2015)	[77]		k-means	*2-7	DBI, DI, SilhI
Xu (2017)	[79]	Euclidean	*k-means, AKM+HC, SAX k-means	3 (stage 1), 4 (stage 2)	peak overlap, consumption error, entropy, WCSS

Load Profile Normalisation

Most studies normalise input data by scaling vectors to a min-max scaler so that patterns retain their shape but are scaled to a zero-one range [19][63][9]. This approach is very sensitive to outliers and appears to be an unvalidated domain preference. Jin et al. propose de-minning as a more robust form of normalisation, but do not offer a quantitative comparison against other approaches [46]. De-minning has the drawback that it only considers profile shape. Considering the importance of normalisation, it is surprising that some studies do not provide any details about the normalisation technique at all. The selection of normalisation algorithms is mostly unsubstantiated. No studies with a rigorous comparison of different normalisation approaches were found.

Time Range and Spatial Cover

Geographically and temporally most studies cover a single location and a maximum time period of 18 months. Typically studies first derive RDLPs for individual customers at specific loading conditions and then cluster the RDLPs, which significantly reduces the number of input patterns. Some studies, such as Jin et al. and Kwac et al. cluster all daily load profiles and derive a set of consumption patterns, described by the cluster centroids, that represents distinct daily energy usage behaviour for different types of consumers [45][51].

Considerations for Developing Countries

Only limited studies have been done in developing countries. Some assumptions around data representation and cleaning should be reconsidered when clustering energy consumers in this context. The decision to remove very low consuming households made by Kwac et al. and Cao et al. [51][11] excludes consumers in rural or informal households. While individual household consumption of these groups is low, they present a significant percentage of households in the NRS Load Research dataset and are key stakeholders in the context of energy access. Moreover, their low consumption base presents an opportunity for high growth which has important implications for utilities.

2.4.3 Clustering Algorithms

Over twenty different algorithms are used for clustering daily load profiles. Many studies claim superior performance of one algorithm over the next. Algorithm abbreviations, their frequency counts and the number of studies that indicate the algorithm as one of the best are listed in Table 2.2. Best performing algorithms have been denoted with a * in Table 2.1.

Abbreviation	**Algorithm**	**Frequency count**	**Best ***
	k-means	19	4
AKM	Adaptive k-means	1	1
	fuzzy k-means	4	
MKM	Modified k-means	1	
	SAX k-means	1	
SKM	Spherical k-means	1	
WFAKM	Weighted Fuzzy Averages k-means	1	1
	fuzzy c-means	2	
	kmedoids	4	2
AVQ	Adaptive Vector Quantisation	1	
	DBSCAN	1	
E&M	undefined	1	
FTL	Follow-The-Leader	1	
MFTL	Modified Follow-The-Leader	4	2
GMM	Gaussian Mixture Model	1	
HC	Hierarchical Clustering	12	2
IRC	Iterative Refinement Clustering	1	1
	Random Forests	1	
SOM	Self-Organising Maps	7	2
	Voronoi decomposition	1	
	SOM+k-means	3	1
	AKM+HC	3	
	SOM+HC	1	

TABLE 2.2: Abbreviations and usage frequency of algorithms

The majority of studies compares more than one algorithm, with only nine studies implementing a single algorithm. Four of these nine studies implement k-means [24][28][63][77], three implement a multi-step algorithm that combines adaptive k-means with hierarchical clustering (AKM+HC) [46][52][51], one study combines SOM with k-means (SOM+k-means) [33] and one study implements a follow-the-leader (FTL) algorithm [13].

Of the 16 studies that compare more than one algorithm, one quarter does not indicate which algorithm performs best and one quarter indicates that k-means performs best or amongst the best algorithms [11][45][61][79]. SOM [19][55], k-medoids [45][70] and modified follow-the-leader (MTFL) [16][17] are each identified as the best algorithm in two studies. Finally, weighted fuzzy averages k-means (WFAKM) [9], adaptive k-means (AKM) [45], SOM + k-means [62], hierarchical clustering (HC) with Ward [45] and average [17] linkage criteria and iterative refinement clustering (IRC) [6] are each identified as the best or amongst the best clustering algorithms in a single study.

K-means is implemented in 19 out of the 25 studies, making it both the most popular algorithm and the algorithm that is most likely to produce the best results. In general, most studies perform no benchmarking and insufficient comparative evaluations. Results across studies are thus contradictory, inconsistent and inconclusive.

Euclidean distance is used most frequently as distance measure, as shown in Table 2.3. A third of studies compare distance measures, though only two conclusively propose a best measure.

Abbreviation	Distance measure	Frequency count
	Chebyshev	1
	correlation	1
cos	cosine	3
DTW	Dynamic Time Warping	2
	Euclidean	15
Euclidean^2	Euclidean squared	1
w-Euclidean	weighted Euclidean	3
	Hausdorff	1
	Manhatten	1
	Minkowski	1
PCC	Pearson Correlation Coefficient	1
	unspecified	8

TABLE 2.3: Abbreviations and usage frequency of distance measures

Clustering with Pre-binning

Pre-binning, often referred to as two-stage clustering, is suggested by [16] and implemented in [11], [79] and [73]. The results and effectiveness of pre-binning as suggested by Tsekouras et al. are unclear, in part because the input data and data representation have not been documented. Xu et al. have found that using a two-stage approach that first clusters by overall consumption and then by load shape produces better results than clustering by load shape only. The influence of different types of pre-binning has not been investigated.

Limitations of Existing Clustering Approaches

Most studies are primarily concerned with the comparison of different clustering algorithms, and neglect to investigate the effects of data representation and parameter selection. The impact of the input dataset on clustering algorithms is largely unacknowledged, with one third of studies in Table 2.1 omitting to specify the data source. Almost half the studies do not explicitly state the number of patterns in the input dataset and over half the studies compare clustering algorithms on very small datasets consisting of less than 500 input patterns. Very few studies explore the effect of the distance measure on clustering results, with a third of studies omitting to specify the distance measure. Euclidean distance is used as default.

These observations are similar to those made by Miller et al. in their review of clustering approaches of non-residential buildings, which describes challenges around research reproducibility and ambiguity of algorithm applicability [57]. This is a wider problem in the data mining community that has been detailed by [47] more than a decade ago.

2.4.4 Evaluation Measures

The common evaluation measures and their frequency of usage are listed in Table 2.4. The Davies Bouldin Index (DBI), Cluster Dispersion Index (CDI) and Mean Index Adequacy (MIA) are used most frequently, with the Similarity Matrix Indicator (SMI) and Silhouette Index having a couple of use cases. Evaluation of clustering results remains a challenge [45], which some authors try to overcome by proposing metrics of their own. Insufficient testing and evaluation of measures such as the Energy Variance Index presented in [12] however means that their reliability is uncertain and these metrics are not often adopted by other studies. MIA, which is proposed in [13] is an exception and has been adopted by many subsequent studies.

Abbreviation	**Clustering index**	**Frequency count**
BH	Ball & Hall	1
CDI	Cluster Dispersion Index	11
DBI	Davies Boulding Index	13
DI	Dunn Index	3
EVI	Energy Variance Index	1
IA	Index of Agreement	1
J	Mean square error or error function	2
MDI	Modified Dunn Index	2
MIA	Mean Index Adequacy	11
ScatI	Scatter Index	3
SilhI	Silhouette Index	4
SMI	Similarity Matrix Indicator	5
VRC	Variance Ratio Criterion / Calinski Harabasz Index	3
VRSE	Violation rate of Root Squared Error threshold	1
WCS	Within Cluster Sum	1
WCSS	Within Cluster Sum of Squares	1
WB	between cluster variation	1
WCBCR	WC/WB	2

TABLE 2.4: Abbreviations and usage frequency of clustering metrics

Jin et al. find that standard performance metrics pose a trade-off between compactness and distinctness for cluster selection [45]. Dang-Ha et al. conclude that the standard evaluation measures are unreliable due to bias towards isolating outliers and insufficient penalisation of large, noisy clusters [19]. The study further suggests that automatic selection of clusters should be done with caution and clusters should be judged manually within the context of the application. Dent et al. find that a single metric on its own is insufficient to adequately represent cluster performance and suggest a combination of measures to ensure optimal cluster selection [24].

Determining the Optimal Number of Clusters

Where stated, the optimal number of clusters is indicated with a * in Table 2.1. With the exception of [45] and [19], cluster ranges are constrained to small numbers of less than 30 clusters, both to ease expert interpretation and to produce clusters that correspond with existing user groups. It is unclear whether the possibility of local minima has been taken into consideration in determining the cluster ranges. Only few studies conclusively suggest the optimal number of clusters.

External, Qualitative Evaluation Methods

Sarle et al. define cluster evaluation as a statistical problem [65]. In real world applications it is however equally important that clusters are interpretable, useable and satisfy constraints deemed important to users [44]. These qualitative constraints are challenging to evaluate and many clustering applications rely on visual examination and expert judgement to assess cluster validity [76] [61] [45]. Relying on expert input however introduces subjectivity into the evaluation process that makes the quality control of clustering difficult and burdens users [44].

Beyond being compact and distinct, the ability of a cluster to be representative of its individual members and to meet the requirements of a specific context are important. Creating load profiles for new user groups, for example, requires that clusters are linked with the energy company's databases and that they exhibit proximity to the number of classes that are currently used by the energy company [62]. This influences the optimal number of clusters.

Dent suggests that quantifying qualitative measures that support expert judgement can be helpful to successfully segment customers for practical use [23]. Taking inspiration from customer segmentation in the marketing sector, Dent proposes that clusters should be substantial, accessible, differentiable, actionable, stable, familiar, parsimonious, relevant, compact and compatible. Dent suggests a composite clustering measure to objectively evaluate some of these attributes that would typically rely on subjective judgement. The composite measure is used to assess the relative effectiveness of different clustering structures and was found to compliment traditional quantitative measures. While this approach shows promise to improve the evaluation process, explicitly defined qualitative evaluation metrics are not common in the literature.

Cluster Entropy as External Evaluation Measure

Entropy is widely used to calculate information gain to facilitate attribute selection of decision tree algorithms [44]. It has been introduced by Zhao et al. as an external evaluation criteria for partitional document clustering to evaluate how class labels are distributed across clusters [80]. Entropy is calculated as

$$entropy = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{2.3}$$

where m is the number of classes in the dataset and $p_i = {n^i}_c / n_c$ is the fraction of patterns in cluster c assigned to class i. The entropy of a clustering structure is then calculated as the sum of individual cluster entropies weighted by the cluster size [80]. This approach is extended by Rosenberg et al., who introduce V-measure as an external entropy-based clustering criteria to evaluate cluster homogeneity and completeness [64]. Homogeneity requires that a single class is assigned to a single cluster, which implies that for perfect homogeneity entropy is zero.

Entropy has been used as an external measure to capture customer variability in the energy domain. Teeraratkul et al. employ entropy to evaluate clustering algorithms. Lower average entropy indicates less variability in cluster assignment to customers, and thus presents a preferred algorithm that produces more suitable clusters [70]. Xu et al. and Kwac et al. use entropy to characterise customer lifestyles and the stability of household energy consumption behaviour [51] [79].

2.4.5 Considerations for Clustering Residential Customers

While cluster analysis typically yields good results for consumers in the industrial and commercial sectors, the variable nature of individual households makes it more challenging to interpret clustering results in the residential sector [69]. To use the clusters to construct RDLPs, it is necessary to cluster all individual daily load profiles rather than clustering aggregated RDLPs. Pre-binning appears to be a promising technique to apply to highly variable residential consumers. The k-means and SOM algorithms deliver good clustering results in most studies. The input dataset and algorithm parameters must be taken into consideration when selecting a clustering technique. Comparative approaches should evaluate data representation and algorithm parameters, not only the type of clustering algorithm used. Evaluation of clusters is challenging, and qualitative measures have shown promise in easing the evaluation process. Entropy in particular has been used in several studies to evaluate the variability of household energy use.

2.5 Residential Energy Customer Archetypes

This study uses the term customer archetype to refer to a set of RDLPs that describes the characteristic energy consumption behaviour of customers with particular attributes. In the literature the terms profile classes, customer classes and customer segments are also used to describe customer archetypes [7]. While most studies associate customer archetypes directly with RDLPs generated by aggregating the load profiles of cluster members, a limited number of studies in the residential domain uses socio-demographic attributes of households to characterise clusters, and constructs customer archetypes based on these attributes. This work focuses on the latter. Existing work that develops residential customer archetypes by characterising clustering results and attributes that affect household energy consumption are reviewed in this section. The review also introduces other approaches that have been used by experts in South Africa to construct customer archetypes.

2.5.1 Application of Cluster Analysis to Develop Customer Archetypes

Customer archetypes can be derived from RDLPs by classifying load profiles according to socio-demographic characteristics [63] [55] [76]. The approach followed by Rhodes et al. [63] and Viegas et al. [76] applies context filtering to the daily load profiles of households to establish RDLPs under specific loading conditions, such as four seasons, for each household. The RDLPs are then clustered into optimal groups using k-means. A probit regression analysis is done on the resultant cluster centroids to discover significant relations with explanatory variables obtained from survey data, thus deriving customer archetypes.

An alternative methodology is developed by McLoughlin et al. [55]. For each day over a six month period the daily hourly load profiles of 3941 Irish households are clustered separately. RDLPs (referred to as profile classes in the study) are calculated by averaging the consumption of member households of each cluster for every day. The RDLP used by every customer on every day is captured in a Customer Class Index (CCI). The statistical mode of the CCI is assigned to each customer to obtain its most frequently occurring profile. Finally, multinomial logistic regression is used to classify the CCI (and thus RDLPs) by socio-demographic and appliance variables. The regression model is used to determine the likelihood that a customer with specific socio-demographic characteristics uses a particular RDLP. Most socio-demographic and appliance variables were found to influence RDLPs either positively or negatively and the methodology is promising for developing customer archetypes.

This work extends the approach of [55] by clustering daily load profiles over a continuous period of two decades, rather than a single day. Instead of classifying the statistical mode of the CCI, this work characterises each RDLP based on the frequency of assignment to individual households. The resultant characterised RDLPs can then be used to customise customer archetypes according to attributes defined by users.

2.5.2 Attributes of Customer Archetypes

Energy demand models rely on the development of customer archetypes whose energy demand can be adapted to different geographic scales such as a city, province or nation, to adequately predict future energy demand [39]. While the patterns and nuances of individual household consumption are poorly understood, the factors affecting consumption have been well researched in developed and developing countries, though literature from developed countries is dominant. The factors considered in the development of customer archetypes depend on the objectives of the model in which they are used. In energy efficiency and demand side management interventions, behavioural attributes are dominant. Network planning and new electrification models focus on physical and socio-demographic attributes, which can be approximated in the absence of a historic baseline. Data availability, model complexity and research paradigm restrict the attributes that are incorporated in models.

In developed countries the factors that are understood to influence household energy consumption are broad and include building features, socio-demographic and economic indicators, occupant behaviour, occupancy, weather, time, indoor environmental conditions, appliance ownership, attitude and culture [3] [36] [68].

While most of these factors hold true in developing countries, additional attributes must be considered. In South Africa research has shown that the significance of factors that influence energy consumption varies considerably depending on the economic status of a household [41]. Disposable income, time since electrification, availability and cost of alternative fuels, circuit breaker size, appliance availability and complementary infrastructure such as piped water affect the energy consumption of a household. For example, piped water access to a home is a precondition for hot water geysers, which typically consume the most power in residential homes. Tradition and energy cost are other factors that can inhibit low-income communities from adopting electricity as primary source of fuel after electrification [18]. Many rural households continue to cook on wood rather than use electricity after electrification. This may be due to cultural preferences, such as the smokey taste of wood-fired food, or due to the availability of free fire wood. Once electrified, the accumulation of energy consuming appliances costs money and is thus an incremental process.

2.5.3 Expert-derived Customer Archetypes in South Africa

Energy access, household electrification, consumer adoption of new energy services and security of supply shape the need for developing residential customer archetypes in developing countries. The Geo-based Load Forecast (GLF) standard [10] and the Distribution Pre Electrification Tool (DPET) [41] [40] emerged out of an expert working group of South Africa's NRS Load Research Programme to gain a better understanding of energy consumers in South Africa. To date the DPET model is regarded as the best available model on domestic energy consumer classes in South Africa. The GLF has been adopted in the master planning methodology of South Africa's energy utility, Eskom.

The GLF and DPET models define RDLPs for a set of residential customer archetypes characterised by temporal, geographic, dwelling, appliance and socio-demographic attributes that experts consider to be indicators of energy consumption. Experts have gained insights into the attributes of customer archetypes and their corresponding energy consumption behaviour by analysing metered consumption data with statistical techniques and analytical methods such as k-means clustering [56], GAMs [21], and linear regression models [42]. Due to resource constraints the DPET model was last updated in 2011.

2.6 Summary

This chapter reviews the potential and limitations of cluster analysis for clustering time series data. Specifically, the application of clustering techniques in the energy domain as applied to the task of generating representative daily load profiles is discussed.

The data representation, clustering algorithms and evaluation approaches of 25 previous studies are reviewed. Most of the studies are based in developed countries and focus on comparing clustering algorithms. While many different algorithms have been tested, k-means and SOM generally perform well. Pre-binning has shown promising results in studies where it was applied. Most studies focus on comparing clustering algorithms and place little emphasis on the size and representation of the input dataset. Many datasets are small, with less than 500 patterns. Zero-one normalisation is prominent and appears to be based on domain preference, rather than robust comparisons. 60 minute time intervals are used most frequently and have been found adequate with the k-means algorithm.

Next, commonly used cluster evaluation metrics are reviewed. There is no accepted best metric, though DBI, CDI, MIA, SMI and the Silhouette Index are used frequently. Several studies recommend comparing results against a combination of measures. The majority of studies finds evaluation to be a challenge and only few studies are able to offer a definitive recommendation on the optimal number of clusters to be selected. Qualitative approaches and external measures such as entropy have improved the evaluation process if they have been used to augment traditional quantitative metrics.

Finally, studies that leverage regression analysis and socio-demographic data to characterise clusters of load profiles to construct customer archetypes are reviewed. The factors that affect household energy consumption are explored, and the attributes that are particularly relevant in the South African context are described. The chapter concludes by introducing the GLF standard and DPET tool, which contain benchmark customer archetypes used in South Africa.

Chapter 3

Data and Domain Analysis

This chapter provides an overview of the dataset and of decision and evaluation criteria important to domain experts. The chapter starts by introducing the South African Domestic Electrical Load Study (DELS) dataset and details the input data representation. A list of competency questions and qualitative clustering measures that can assess the competency questions are then developed. Finally, the cluster scoring matrix against which top experiments will be evaluated is presented.

3.1 The South African Domestic Electrical Load Study (DELS) Dataset

The NRS Load Research Programme is a collaborative research effort between South Africa's power utility, Eskom, municipalities and research institutions to gain a better understanding of South Africa's residential energy demand sector. From 1994 to 2014 The DELS dataset was collected as part of the programme. It presents the most comprehensive source of information on residential energy consumption in the country. For each participating household a data logger was connected to the household power supply. Current, voltage, real and absolute power readings were collected at five minute intervals. At the end of each study year the collected data was cleaned and load observations marked as valid or invalid to identify sensor failures. This study only uses current readings and excludes all invalid readings from the analysis.

Each study year typically runs from May to May. A front-door survey was performed once during the course of the year to collect socio-demographic, dwelling and appliance information. We use year references to correspond with the household survey year. The dataset treats each categorical survey value as a separate feature, thus capturing 180 mostly sparse features per household annually. Most households were observed between 3 and 5 years. Long term trends of household consumption are not important for the purpose of this study, as we consider each daily load profile as an independent input pattern. Households are thus treated as having separate identities

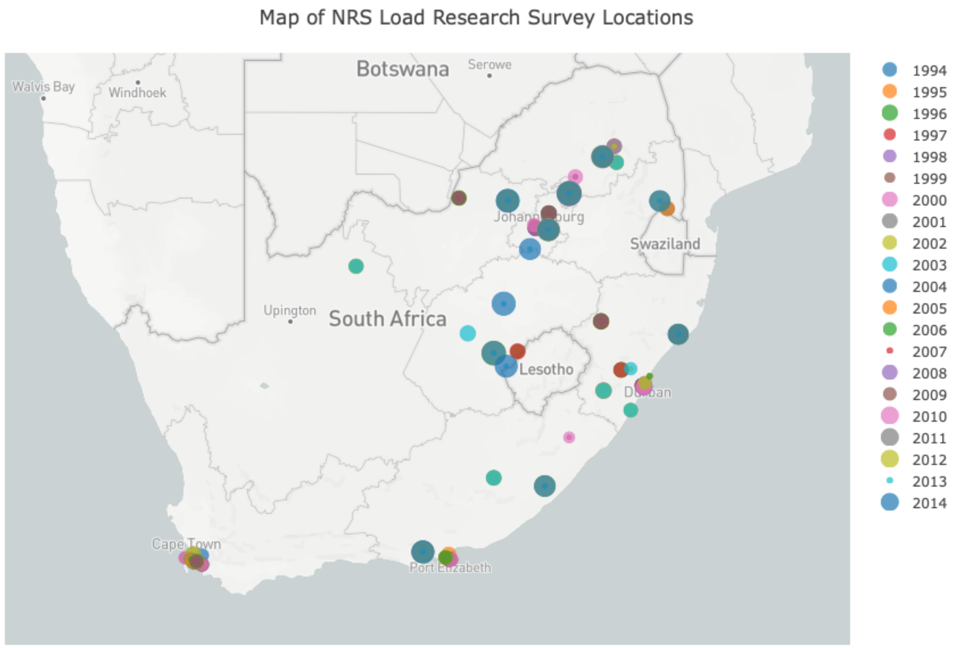


FIGURE 3.1: Map of survey locations

for each year of observation. The dataset contains current observations for a total of 14 945 household identities, from here on referred to as households.

Granular residential energy consumption patterns are inherently noisy, as they record the highly variable daily activities and behaviour of households [69]. This is also true for our dataset. Previous research found the dataset to exhibit seasonal and weekday trends that vary based on appliance ownership, geographic location and other household attributes [40] [22]. For newly electrified households the data also exhibits a demand growth trend that stabilises after 15 years when appliance acquisition is assumed to be saturated [41] .

3.1.1 Description of Sample Population

The DELS dataset captures South Africa's diverse population across the country's geography, covering five climatic zones and rural, informal and urban settlements (see the map in Figure 3.1). Over the twenty year period 8656 household surveys were collected (58% of metered households). No surveys were captured in 2007 and 2013. A harmonised version of the survey data produced during this research has been published as the Domestic Electrical Load Survey - Key Variables (DELSKV) dataset [72].

The sample population is representative of South Africa's unequal society. The majority of households have a low income of less than R5000 (about $340) per month. A fraction of households earns up to 50 times that amount (see Figure 3.2). Figure 3.3 shows a similar distribution for dwelling size. Most households occupy dwellings

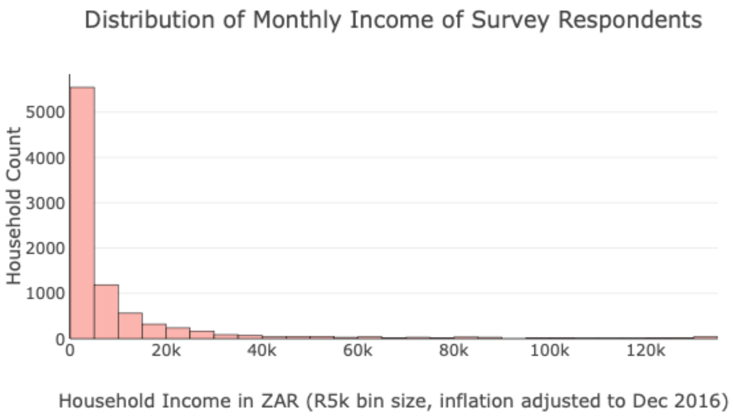


FIGURE 3.2: Distribution of survey responses for monthly income of survey responses

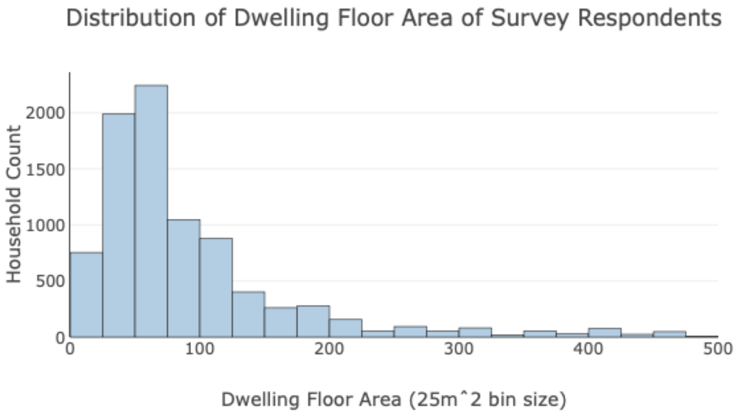


FIGURE 3.3: Distribution of dwelling floor area of survey responses

between $25m^2$ and $100m^2$. It is important to include the entire spectrum of households in our analysis, as the small number of affluent households are disproportionally large energy consumers.

As can be seen from Figure 3.4, the dataset covers a large number of newly electrified households. Figure 3.5 shows the proportional use of wall materials, roof materials and water access points of survey respondents. Less than half the surveyed households have access to piped water in the home and less than a quarter of households live in dwellings with brick walls. More than half the households have a corrugated iron or zinc roof - a construction material that is particularly popular in rural and informal settlements due to its availability and low cost.

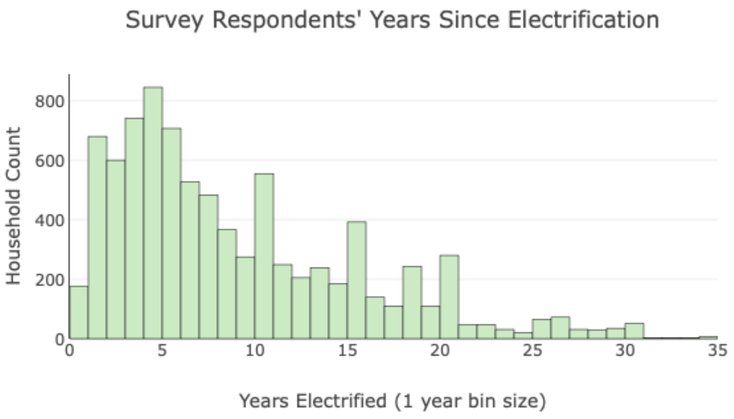


FIGURE 3.4: Distribution of years electrified of survey responses

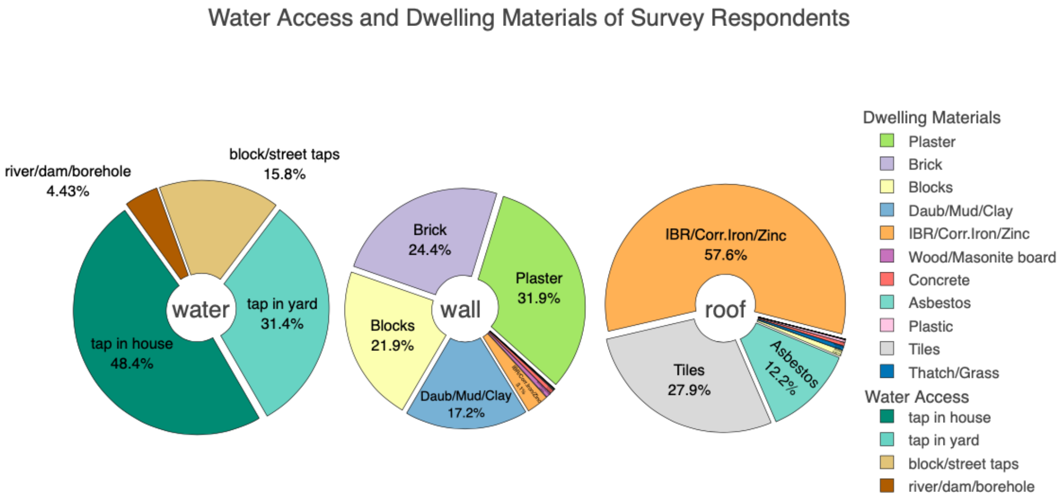


FIGURE 3.5: Proportioned survey responses for water access, wall and roof materials

3.1.2 Data Representation

As done in previous studies, all 5 minute observations are averaged over 60 minute intervals, producing 3 295 848 daily load profiles for 14 945 households. Each interval is labeled by the start time, such that $t = 0$ captures interval 00:00:00 - 00:59:59. We discard daily load profiles if they contain any invalid observations. This aggregated dataset has been published as the Domestic Electrical Load Metering, Hourly Data (DELMH) [71].

The daily load profile h of household j on day d is:

$$\text{load profile} = l(t), \text{ where } t = \{0, 1...23\} \tag{3.1}$$

$$h_d^{(j)} = l(t)_d \tag{3.2}$$

$H^{(j)}$ is the array of all 24-element daily load profile vectors h_d for household j.

$$H^{(j)} = \left[h_d^{(j)}\right], \text{where } d = \{1, 2 ... d \text{ days}\} \tag{3.3}$$

d varies for each household and depends both on the duration for which the household was observed and on the number of valid readings in that period. The mean observation duration $\bar{d}$ for all households is 220 days. 61% of households were observed for more than half a year (ie $d > 183$). The maximum number of households observed on a single day was on 23 August 1999 when the electricity consumption of 1245 households was recorded. The median daily household count is 399.

The input array X of all daily load profiles h is

$$X = \left[H^{(j)}\right], \text{where } j = \{1, 2 ... 14945\} \tag{3.4}$$

X has dimensions 3 295 848 × 24. The distribution of annual mean daily demand of all households is shown in Figure 3.6. Half the households consume on average less than 10kWh/day.

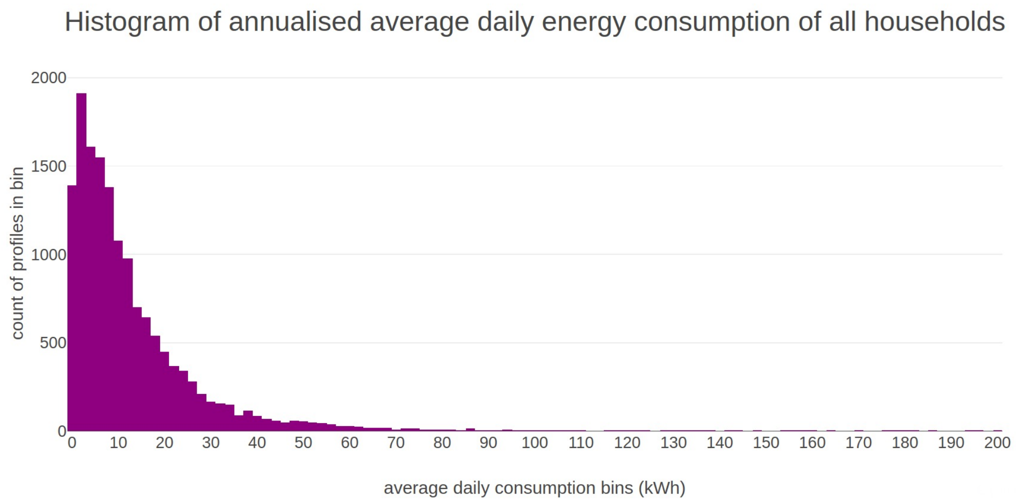


FIGURE 3.6: Histogram of mean daily household power consumption in 10kWh bins

3.2 Domain-led Cluster Evaluation

To produce meaningful clusters, a domain perspective of the objective of the clustering problem is key to inform parameter selection and evaluation. While internal clustering indexes evaluate cluster compactness and distinctness, previous studies have indicated that expert validation remains key to ensure that clusters are fit for purpose for their intended application [45][19][76].

3.2.1 Competency Questions

We performed a domain analysis and consulted with experts to identify criteria and attributes of RDLPs that are important for generating meaningful customer archetypes. From the interviews we established that RDLPs are required to be specific to particular socio-economic contexts, temporal contexts (or loading conditions) and known energy consumption behaviour to be useful for constructing customer archetypes for long term energy planning. Current domain knowledge suggests that energy consumption behaviour is strongly influenced by daily routines, seasonal climatic variability and the consumption category (eg low, medium, high) of a household. Existing customer segments in South Africa are thus categorised accordingly and any new customer archetypes should be differentiable within these contexts to be acceptable to experts.

To formalise these insights we formulated five competency questions that establish the ability of a set of clusters to represent specific contexts and particular energy consumption behaviour for South African households.

1. Can the load shape and demand be deduced from clusters?
2. Do clusters distinguish between low, medium and high consumption consumers?
3. Can clusters represent specific loading conditions for different day types and seasons?
4. Can a zero-consumption profile be represented in the clustering structure?
5. Is the number of households assigned to clusters reasonable, given knowledge of the sample population?

Question 4 was deemed important for considering energy access in low income contexts, as households may go through periods where they cannot afford to buy electricity and thus show no consumption.

The competency questions were then used to frame and develop a set of external evaluation measures that characterise a good clustering structure. These qualitative measures are described in the next section.

3.3 Qualitative Evaluation Measures

A good set of clusters should represent energy consumption behaviour that makes sense in relation to the consumers' contexts and carry sufficient potential for meaning to make it interesting to users. We define this characteristic as the usability of a clustering structure, which relates to competency questions 1, 4 and 5. As an RDLP can be derived for all clusters (see equation 4.5), there is no need to provide a measure for question 1 – it is true for all our experiments. Question 4 requires a manual evaluation based on expert judgement and is evaluated as being either true, or false. Question 5 is calculated as the percentage of clusters whose membership exceeds a threshold value.

Beyond being usable, clusters also need to be expressive. The ability of individual clusters to convey specific meaning is particularly important in contexts where populations are highly variabile. Expressivity requires firstly that the RDLP of a cluster is representative of the energy consumption behaviour of the individual daily load profiles that are members of that cluster. The time and magnitude of daily peak demand and the total daily consumption are significant features in a daily load profile. The representativity is thus calculated as the peak coincidence ratio and the mean consumption error of peak and total demand.

Secondly, an expressive cluster must convey specific information about the demand profiles of a group of consumers (question 2) or a temporal loading condition (question 3). We call the capability of a cluster to represent a specific context homogeneity. A perfectly homogeneous cluster represents a single context, e.g. daily load profiles of low consumption households on Sundays in summer. All available information is embedded in the cluster, and assigning it to a new daily load profile will only yield insights about the profile, not about the cluster. This is desirable, as it entails that the RDLP of the cluster is a good proxy for its member profiles. Cluster entropy can be calculated to establish the information embedded in a cluster and thus homogeneity. The lower the entropy, the more information is embedded in the cluster, the more homogeneous (specific) the cluster, the better the cluster.

The external evaluation measures are described in detail below. We use k_x to denote a single cluster in clustering structure k.

3.3.1 Mean Consumption Error

The total daily demand and peak daily demand for an actual daily load profile $l(t)_d$ and a predicted cluster representative daily load profile $l(t)_x$ are given by the equations below.

$$d^{(j)}_{d_{total}} = \sum_{t=0}^{23} l(t)_d \text{ and } d^{(j)}_{d_{max}} = l(t)^{max}_d \tag{3.5}$$

$$d^{(R)}_{x_{total}} = \sum_{t=0}^{23} l(t)_x \text{ and } d^{(R)}_{x_{max}} = l(t)^{max}_x \tag{3.6}$$

Four error metrics are calculated to characterise the extent of deviation between the total and peak demand of a cluster, and those of its member profiles. Mean absolute percentage error (MAPE) and median absolute percentage error (MdAPE) are well known error metrics, with MdAPE being more robust to outliers than MAPE, which is very sensitive to extreme values. Both metrics penalise overprediction more heavily than underprediction.

The median log accuracy ratio (MdLQ) and median symmetric accuracy (MdSymA) overcome some of these drawbacks [58]. The log-transformation tends to induce symmetry in positively skewed distributions, thus reducing bias. Interpreting MdLQ is not intuitive, a problem overcome by MdSymA which can be interpreted as a percentage error, similar to MAPE. The same formulae are used to calculate peak demand and total consumption errors.

None of the consumption error metrics are defined for zero-profiles where $d_d^{(j)} = 0$. We replace these values with NaN and exclude them from the evaluation.

Absolute Percentage Error

$$mape = 100 \times \frac{1}{N} \sum_{1}^{N} \frac{|d_d^{(j)} - d_x^{(R)}|}{d_d^{(j)}}, \text{ where } N \text{ are all } h_d^{(j)} \text{ assigned to } k_x \tag{3.7}$$

$$mdape = 100 \times median\left(\frac{|d_d^{(j)} - d_x^{(R)}|}{d_d^{(j)}} \right) \text{ for all } h_d^{(j)} \text{ assigned to } k_x \tag{3.8}$$

Median Log Accuracy ratio

$$Q_d^{(j)} = \frac{d_x^{(R)}}{d_d^{(j)}} \tag{3.9}$$

$$mdlq = median\big(log(Q_d^{(j)})\big) \text{ for all } h_d^{(j)} \text{ assigned to } k_x \tag{3.10}$$

Median Symmetric Accuracy

$$mdsyma = 100 \times \left(\exp\left(median\big(|log(Q_d^{(j)})|\big)\right) - 1\right) \text{ for all } h_d^{(j)} \text{ assigned to } k_x \tag{3.11}$$

3.3.2 Mean Peak Coincidence Ratio

The mean peak coincidence ratio for a single cluster is a value between 0 and 1 that represents the ratio of mean peak overlap to the count of peaks in cluster k_x. The closer the ratio is to 1, the higher the peak coincidence and the better the cluster.

The magnitude of the peak is not taken into account in calculating the mean peak coincidence ratio.

For each daily load profile $l(t)_d$ the peaks are identified as all those values that are greater than half the maximum daily load profile value:

$$PeakTimes_d^{(j)}, PeakValues_d^{(j)} = i, l(i)_d \text{ , where } i = \{0, 1, ...23\} \tag{3.12}$$

$$\text{and } l(i)_d > 0.5 \times l(t)_d^{max}$$

The python package peakutils was used to extract the peak values and peak times for all daily load profiles and all representative daily load profiles. The mean peak coincidence was calculated from the intersection of the actual and cluster peak times.

$$MeanPeakCoincidence_x = \frac{1}{\#h_d^{(j)}} \times \#\big(PeakTimes_d^{(j)} \cap PeakTimes_x^{(R)}\big) \tag{3.13}$$

$$\text{for all } h_d^{(j)} \text{ assigned to } k_x$$

3.3.3 Entropy as a Measure of Cluster Homogeneity

We define four entropy measures – weekday, monthly, total daily demand and total peak demand entropy – to quantify how specific a cluster is to these particular contexts. We select weekday and monthly rather than daytype and seasonal contexts, as they do not require weightings. Daytypes and seasons can be easily derived from weekdays and months. Entropy H is used to quantify the homogeneity of clusters and is calculated as follows:

$$H_x^{(f)} = -\sum_{i=1}^{n} p(v_i) \log_2(p(v_i)) \tag{3.14}$$

Here $i = \{1...n\}$ are the values of a feature f and $p(v_i)$ is the probability that cluster k_x is assigned to daily load profiles with value v_i for feature f. For example, $H_x^{(weekday)}$ expresses the homogeneity of a cluster with regards to dau of the week, with $f = weekday$ and $i = \{Mon, Tues, Wed, Thurs, Fri, Sat, Sun\}$, where $p(Sun)$ is the likelihood that cluster k_x is assigned to daily load profiles that are used on a Sunday.

To calculate peak and total daily demand entropy, we created percentile demand bins. Thus the values of feature $f = peak\_demand$ are $i = \{0...99\}$ and $p(59)$ is the likelihood that cluster k_x is assigned to daily load profiles with peak demand corresponding to that of the 60th peak demand percentile.

3.4 Cluster Scoring Matrix

The evaluation measures described above calculate the performance of a single cluster. With the exception of the mean peak coincidence ratio the score S^m of a qualitative measure for clustering structure k is the weighted mean of the scores S^m_x of all clusters k_x where $N > 10490$. For the mean peak coincidence ratio the weighted sum is calculated instead. Clusters with small member size were excluded as they are too specific to be of general use and tend to bias results as they frequently perform better than large clusters. The threshold for N was selected as a value equivalent to 5% of households using a particular cluster for 14 days.

We use the external evaluation measures as relative indexes. To select the best set of clusters, we create a scoring matrix that weights all qualitative evaluation measures according to their importance. For each measure, experiments are ranked, with 1 being assigned to the best performing clustering structure. A weighted score is computed by multiplying the rank with its corresponding weight for each measure. The scores can be interpreted as penalty points allocated to experiments based on their ranked performance by weighted measures. Measures that are deemed more important to overall performance are given larger weights, as these penalise low ranked clustering structures more severely. The total score is the sum of the weighted scores for all measures. The lower the total score, the better the clustering structure.

Category	**Evaluation measure**		**Question**	**Weight**
usable	sensible count per cluster	threshold	5	2
	zero-profile representation		4	1
expressive representative	consumption error	total	1	6
		peak	1	6
	peak coincidence		1	3
expressive homogeneous	temporal entropy	weekday	3	4
		monthly	3	4
	demand entropy	total daily	2	5
		peak daily	2	5

TABLE 3.1: Qualitative Clustering Scoring Matrix

3.5 Summary

In this chapter we present the South African DELS dataset. We formally specified existing expert knowledge into a set of five competency questions that specify attributes of good clusters. Good clusters should be usable and expressive, and expressive clusters are representative and homogeneous. Qualitative evaluation measures have been developed to quantify these high level categories. All qualitative measures have been weighted and combined into a cluster scoring matrix that will be used for evaluating experiments. The next chapter provides an overview of the experiment setup, algorithms, normalisation and pre-binning techniques and the traditional quantitative metrics that are used for evaluating experiment.

Chapter 4

Clustering

The design of clustering experiments is presented in this chapter. After a brief overview, the chapter details the normalisation, pre-binning and clustering techniques under comparison. Performance metrics are established and the development of representative daily load profiles is described.

4.1 Overview of Clustering Process

The clustering process is shown in Figure 4.1.

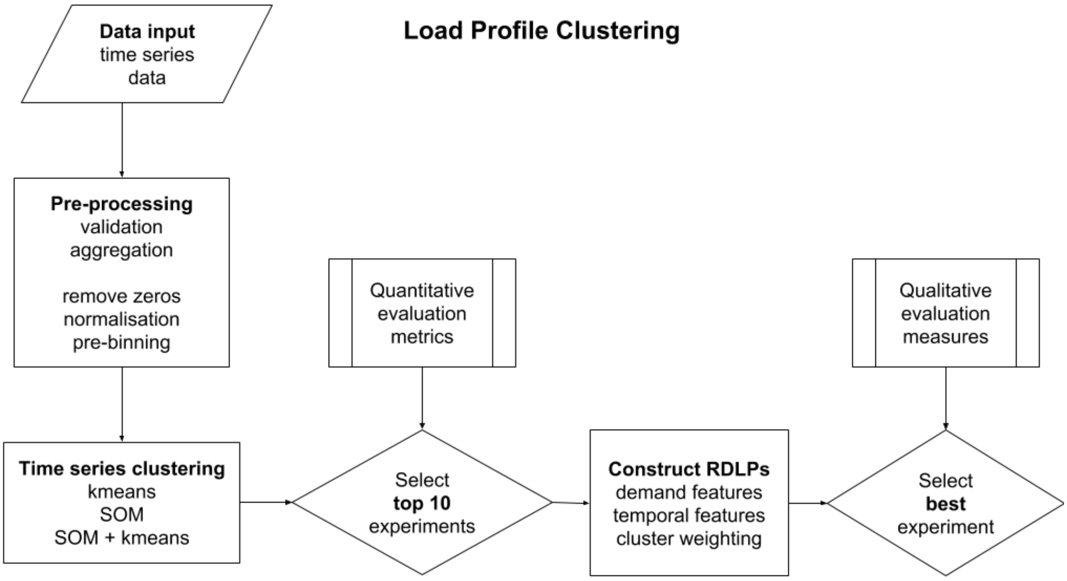


FIGURE 4.1: Overview of time series clustering method

All valid daily load profiles are preprocessed as described in Section 3.1.2. Depending on the experiment, the input data is further processed by removing zeros, applying a normalisation algorithm and pre-binning. Each algorithm is then initialised with the relevant cluster ranges. The experiment results are recorded and the quantitative evaluation metrics are calculated to select the best 10 experiments. For these experiments a RDLP is produced for each cluster and weights are calculated based on the cluster size. Demand and temporal features for individual daily load profiles and RDLPs are then extracted. Finally, qualitative measures are used to rank the top 10 experiments and select the best ranked experiment.

4.2 Normalisation

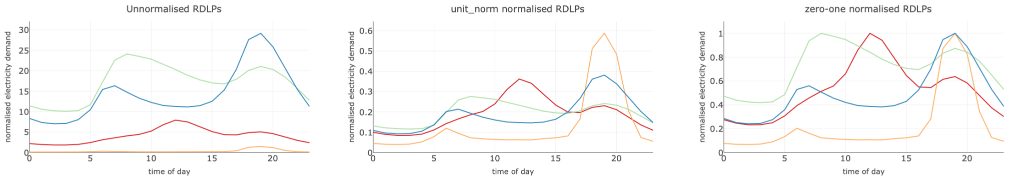
It is known that normalisation has a considerable influence on clustering results [47]. We compare four algorithms that are used in the energy domain (Table 4.1). The normalised daily load profile for household j observed on day d is denoted as $n_d^{(j)}$. Figure 4.2 shows the transformation that different load profiles undergo under the normalisation algorithms.

Normalisation	Equation	Comments
Unit norm (u)	$n_d^{(j)} = \frac{h_d^{(j)}}{\lvert h_d^{(j)} \rvert}$	Scales input vectors individually to unit norm
De-minning (d)	$n_d^{(j)} = \frac{l(t)_d - l(t)_d^{min}}{\lvert l(t)_d - l(t)_d^{min} \rvert}$	Proposed by [45], deminning subtracts the daily minimum demand from each hourly value. All values are then divided by the deminned daily total.
Zero-one (z)	$n_d^{(j)} = \frac{h_d^{(j)}}{l(t)_d^{max}}$	Also known as min-max scaler, zero-one normalisation scales all values to a range [0, 1]. Retains the profile shape, but is sensitive to outliers.
SA norm (sa)	$n_d^{(j)} = \frac{h_d^{(j)}}{\frac{1}{24} \times \sum_{t=0}^{23} l(t)_d}$	Introduced for comparison, as it is frequently used by South African experts to develop customer segmentation models. Normalises all input vectors to a mean of 1. Profile shape is retained, but the approach is sensitive to outliers.

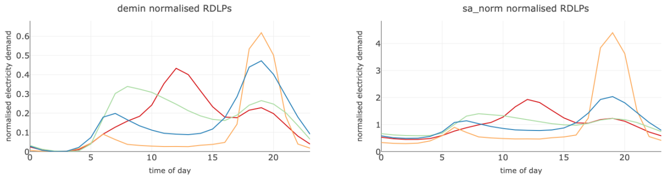
TABLE 4.1: Data normalisation algorithms and descriptions

4.3 Pre-binning

Two different approaches are applied to pre-bin all daily load profiles. For pre-binning by average monthly consumption, we define 8 expert-approved bin ranges based on South African electricity tariff ranges. Pre-binning by integral k-means is a data-driven approach based on the work of [79].



(A) Unnormalised (B) Unit norm (C) Zero-one



(D) Demin (E) SA norm

FIGURE 4.2: Normalisation effects on RDLP representation

4.3.1 Pre-binning by average monthly consumption

The average monthly consumption (AMC) for household j over a one year period is calculated as follows:

$$AMC^{(j)} = \frac{1}{12} \sum_{month=1}^{12} \sum_{d=1}^{month_{end}} \sum_{t=0}^{23} 230 \times l(t)_d \text{ kWh} \tag{4.1}$$

All the daily load profiles, $H^{(j)}$ of household j are assigned to one of 8 consumption bins based on its $AMC^{(j)}$ value. Individual household identifiers are removed from X after pre-binning.

bin	AMC	
1	0 - 1 kWh	no consumption
2	2 - 50 kWh	lifeline tariff - free basic electricity
3	51 - 150 kWh	
4	151 - 400 kWh	
5	401 - 600 kWh	
6	601 - 1200 kWh	
7	1201 - 2500 kWh	
8	2501 - 4000 kWh	

TABLE 4.2: AMC bins based on South African electricity tariffs

4.3.2 Pre-binning by integral k-means

For the simple case where t represents hourly values, pre-binning by integral k-means follows these steps:

1. Construct a new sequence $c(t)$ from the cumulative sum of normalised profile $n_d^{(j)}$

2. Append $l(t)_d^{max}$ to $c(t)$ to ensure that both peak demand and relative demand increase are taken into consideration
3. Gather all features in array X^C and remove individual household identifiers
4. Use the k-means algorithm to cluster X^C into $k = 8$ bins, corresponding to the number of bins created for AMC pre-binning

Early experiments found unit norm to be a promising normalisation technique. Step 1 of the pre-binning thus normalised profiles with unit norm.

4.4 Clustering Algorithms and Experiments

Based on clustering approaches described in previous research, variations of k-means, self-organising maps (SOM) and a combination of the two algorithms are implemented to cluster X. Due to the large size of the dataset, we choose Euclidean distance as the distance measure for the k-means algorithm. Each algorithm is initialised with different sets of parameter values, normalisation and pre-processing steps. The method evolved through iteration, with each experiment building on the results of the previous experiments. Table 4.3 summarises the different experiments that were performed.

Experiment	Algorithm	Cluster ranges	Normalisation	Pre-binning	Drop zeros
1	k-means (test)	$n\{5, 8, 11, ...136\}$	none	none	
2	k-means	$n\{5, 8, 11, ...136\}$	none, u, d, z, sa		
	SOM	$d\{5, 7, 9, ...29\}$	none, u, d, z, sa		
	SOM+k-means	$d\{30, 40, ...90\}, n$	none, u, d, z, sa		
3	k-means	$n\{5, 8, 11, ...136\}$	none, u, d, z, sa		True
	SOM	$d\{5, 7, 9, ...29\}$	none, u, d, z, sa		
	SOM+k-means	$d\{30, 40, ...90\}, n$	none, u, d, z, sa		
4	k-means	$n\{2, 3, ...10\}$	none, u, d, z, sa	AMC	
	SOM	$d\{2, 3, 4, 5\}$	none, u, d, z, sa		
	SOM+k-means	$d\{4, 7, 11, ...20\}, n$	none, u, d, z, sa		
5	k-means	$n\{2, 3, ...19\}$	none, u, d, z, sa	AMC	
	SOM+k-means	$d\{4, 7, 11, ...20\}, n$	none, u, d, z, sa		
6	k-means	$n\{2, 3, ...19\}$	none, u, d, z, sa	AMC	True
7	k-means	$n\{2, 3, ...19\}$	none, u, d, z, sa	integral k-means	
8	k-means	$n\{2, 3, ...19\}$	none, u, d, z, sa	integral k-means	True

TABLE 4.3: Experiment details

Due to South Africa's geographic spread and economic inequality, significant variability in national energy consumption patterns is anticipated. We thus allow for a relatively large number of clusters, while taking the following three factors into consideration:

1. Fewer clusters ease interpretation and are thus preferable to many clusters
2. Increments should be sufficiently large to allow for timeous algorithm run times, while also being sufficiently small to discern clustering performance
3. Maximum number of clusters is 220, based on population diversity and existing expert models which account for 11 socio-demographic groups, 2 seasons, 2 daytypes and 5 climatic zones

The k-means algorithm is initialised with a range of n clusters, producing $k^{(i)} = \{k_1^{(i)}...k_{n_i}^{(i)}\}$ for n_i in n. The SOM algorithm is initialised as a square map with dimensions $d_i \times d_i$ for d_i in range d. The SOM algorithm produces $k^{(i)} = \{k_1^{(i)}...k_{d_i \times d_i^{(i)}}\}$ for d_i in d. The cluster ranges produced by SOM span a greater range and increase the number of clusters k in large increments, which has the advantage of testing edge cases, but has the drawback of making it difficult to discern the best number of clusters $k^{(i)}$. Combining SOM and k-means first creates a $d \times d$ map, which acts as a form of dimensionality reduction on X. For each d, k-means then clusters the map into n clusters. The mapping only makes sense if d^2 is greater than n.

For experiments with pre-binning, clustering is done independently within each bin, thus performing a two-stage clustering process. The maximum acceptable number of clusters per bin is considerably smaller and the range of n is chosen accordingly. The coarse-grained clustering increments of SOM do not make it well suited to the requirement of fewer clusters and pre-binning is only done with k-means.

4.5 Quantitative Clustering Metrics

Cluster compactness and distinctness are two important attributes that characterise a good clustering structure. In a compact cluster, the cluster members lie close to the cluster centroid and to each other, while the clusters in a distinct clustering structure are different from each other. We select three common clustering metrics, the Mean Index Adequacy (MIA), the Davies-Bouldin Index (DBI) and the Silhouette Index, to evaluate the experiments. MIA quantifies cluster compactness, and the DBI and Silhouette Index measure both compactness and distinctness. MIA and DBI have a lower bound of 0, and a lower score is an indication of better clusters. The Silhouette Index has a range from -1 to 1. A negative score is an indication of poor clustering, where a large number of cluster members would potentially be better assigned to a neighbouring cluster. An average score close to 1 indicates a good set of clusters.

Comparing experiments across the three metrics is challenging, as performance across metrics may not always be consistent. In this study cluster evaluation is conducted on a relative rank basis, not by absolute values. Thus, the three metrics can be combined into a single Combined Index (CI) to ease the evaluation process. This CI index is described next.

4.5.1 Combined Index

The CI is calculated from the product of the DBI, MIA and inverse Silhouette Index. It provides an indication of the performance of experiments across all metrics and is defined as follows:

$$CI = log\left(\sum_{bin=1}^{bins}\left(Ix_{bin} \times \frac{N_{bin}}{N_{total}}\right)\right), \text{where } N \text{ is the count of } h_d^{(j)} \tag{4.2}$$

$$Ix = \begin{cases} \text{undefined} & \text{if } DBI, MIA, SilhouetteIndex \leq 0 \\ \dfrac{DBI \times MIA}{SilhouetteIndex} & \text{otherwise} \end{cases} \tag{4.3}$$

The CI is used as a relative index to avoid having to interpret multiple metrics simultaneously. Ix is an interim score that computes the product of the DBI, MIA and inverse Silhouette Index. The CI is the log of the weighted sum of Ix across all experiment bins. A lower CI is desirable and an indication of a better clustering structure. The logarithmic relationship between Ix and the CI means that the CI is negative when Ix is between 0 and 1, 0 when $Ix = 1$ and greater than 0 otherwise.

The log function is only defined for values greater than 0. As the lower bound of the DBI and MIA is 0 and a negative Silhouette Index is an indication of poor clustering, the Ix score is undefined for all scores equal to or below 0, so that the input to Equation 4.2 is valid. The Ix increases linearly with the DBI and MIA. When these scores are low, so is the Ix. However, as both metrics evaluate cluster compactness, we anticipate them to increase simultaneously. Thus, if cluster compactness deteriorates, the Ix should be affected significantly. Neither DBI nor MIA has an upper bound, which is thus also true for the Ix. The Silhouette Index on the other hand is inversely related to Ix. When the Silhouette Index is close to 1, clusters are good and the Silhouette Index has only a marginal influence on Ix. The closer the Silhouette Index is to 0, the greater Ix becomes.

For experiments with pre-binning, the experiment with the lowest Ix score in each bin is selected, as it represents the best clustering structure for that bin. For experiments without pre-binning, $bins = 1$ and $N_{bin} = N_{total}$. Weighting Ix of each bin is important to account for the size of cluster membership in that bin.

The ten experiments with the lowest CI score are selected for the qualitative evaluation. This makes the CI a convenient measure for selecting the best clustering structures from hundreds of experiments.

4.6 From Clusters to Representative Daily Load Profiles

The clustering algorithm predicts a cluster $k_x^{(i)}$ for each normalised daily load profile $n_d^{(j)}$. The representative daily load profile $r_x^{(i)}$ that cluster $k_x^{(i)}$ symbolises, is the mean of all de-normalised daily load profiles $h_d^{(j)}$ assigned to that cluster. RDLPs are calculated for the top 10 clustering structures.

$$r_x^{(i)} = l(t)_x, \text{ where } t = \{0, 1...23\} \tag{4.4}$$

$$r_x^{(i)} = \frac{1}{N}\sum_{1}^{N} h_d^{(j)}, \text{ where } N \text{ are all } n_d^{(j)} \text{ assigned to cluster } k_x^{(i)} \tag{4.5}$$

The set of representative daily load profiles $R^{(i)}$ for all clusters in $k^{(i)}$ is $\{r_1^{(i)}...r_{n_i}^{(i)}\}$. The output of the clustering experiments is the selection of the best clustering structure $k^{(i)}$ that symbolises the best set of representative daily load profiles R for X.

4.7 Summary

This chapter presents an overview of the clustering process, the four normalisation and two pre-binning techniques. Eight experiments and their parameters have been defined. The development of the Combined Index, which combines the DBI, MIA and Silhouette Index for easier evaluation has been described. Finally we show how clusters will be used to generate RDLPs. The next chapter presents the results and analysis of the experiments.

Chapter 5

Results and Analysis

The results of the clustering experiments are presented in this chapter. The first section discusses the performance of experiments based on the Combined Index (CI) score and presents the top 10 experiments. The second section presents the results of the cluster scoring matrix and closely examines the usability and expressivity of three of the top 10 experiments. The code used to run the experiments is available online [1].

5.1 Quantitative Evaluation of Clustering Experiments

We implemented our experiments in python 3.6.5 using k-means algorithms from scikit-learn (0.19.1) and self-organising maps from the SOMOCLU (1.7.5) libraries. In total 2083 individual experiments were conducted. The CI scores for all experiments are plotted as a percentage distribution in Figure 5.1. For experiments with pre-binning, the best clustering structure for each bin was selected based on the lowest Ix score in that bin. The CI was then computed as the weighted sum of Ix across all bins. 65.5% experiments have a score below 4 and over 97.1% of experiments have a score below 6.5.

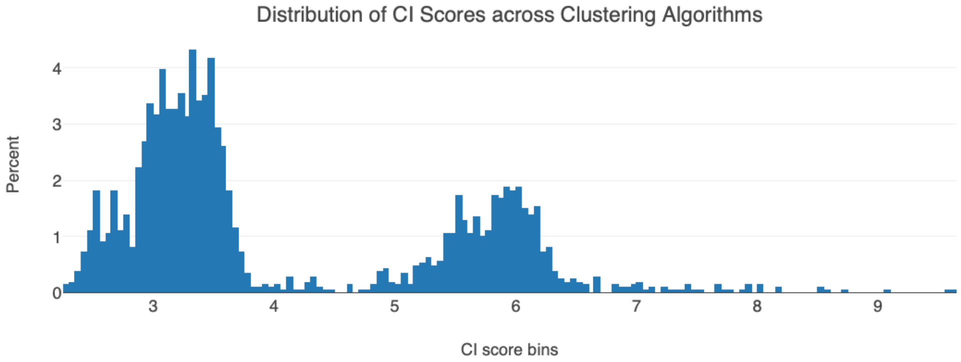


FIGURE 5.1: Distribution of CI scores across experiments

Figure 5.2 visualises the Ix scores for all experiments. Both axes are a log scale. For experiments with pre-binning, Ix is averaged across all bins initialised with the same

---

[1] https://github.com/wiebket/del_clustering

number of clusters. Scores range from a low of 2.282296 to a maximum of 9.626502. The plot shows two distinct bands of experiments. Experiments in the top band have not been normalised, or normalised with SA norm. These experiments have high scores and correspond to the long tail in the distribution of scores in Figure 5.1. Experiments in the bottom band have been normalised with unit norm, deminning or zero-one.

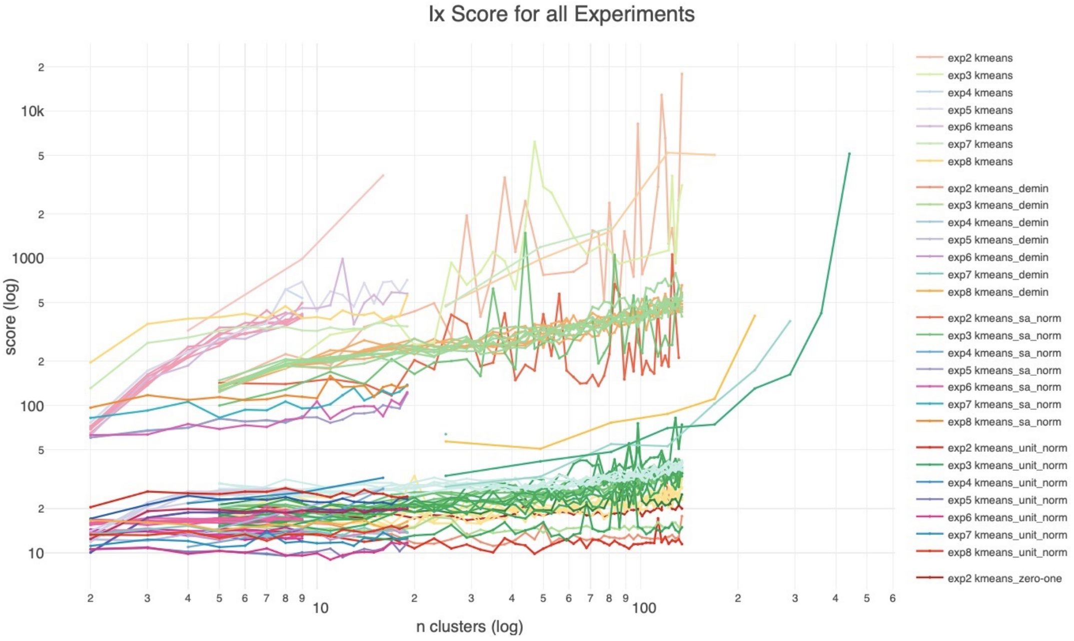


FIGURE 5.2: Ix scores for all experiments

5.1.1 Performance of Normalisation, Pre-binning and Algorithms

The percentage distributions of CI scores across normalisation, pre-binning and algorithm types are shown in Figures 5.3 to 5.5. From the histograms it is clear that normalisation and pre-binning improve clustering results. It is however not immediately evident which normalisation and pre-binning approaches are best.

Figure 5.3 shows the distribution of scores for all experiments across the four normalisation algorithms and experiments without normalisation An individual percentage distribution is calculated for each group of experiments that uses the same normalisation algorithm. Most of the results of the experiments without normalisation have scores above 5. Normalisation clearly improves the CI score. Unit norm has the highest percentage of experiments with the best CI scores, while a couple of zero-one outliers are also top performing. The shape of the distributions makes it difficult to determine whether unit norm or deminning is more likely to produce the best results. It is also not clear whether normalisation or some other experimental parameters are responsible for improved performance. Of the normalisation algorithms SA norm performs worst and shows very limited improvement over unnormalised experiments.

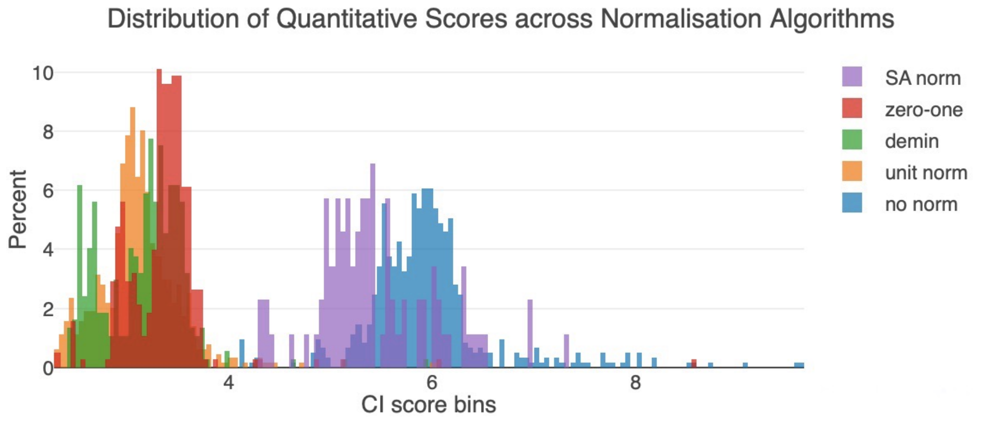


FIGURE 5.3: Distribution of CI scores across normalisation algorithms

Figure 5.4 shows the impact of pre-binning on the CI scores. Pre-binning by average monthly consumption (AMC) produces the most results with the best scores. Integral k-means yields a higher percentage of top results, though none are best performing. It is not possible to determine with certainty which of the pre-binning approaches is better, but it is clear that pre-binning improves clustering scores as a whole.

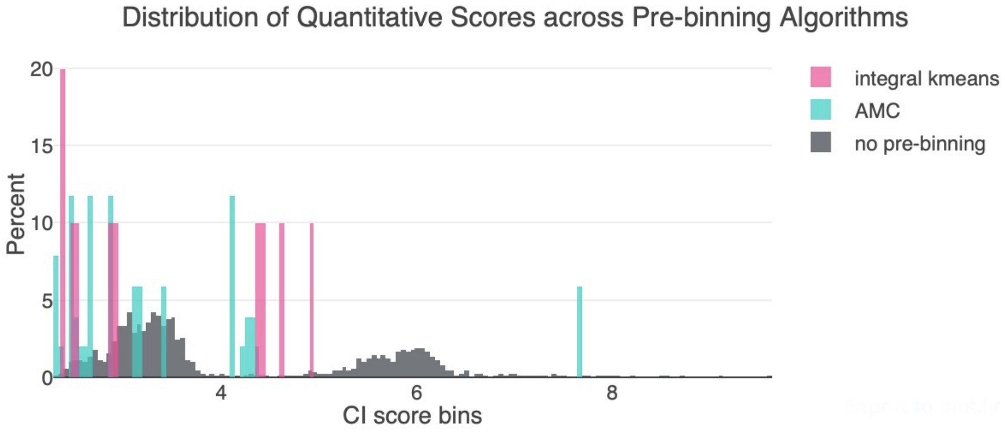


FIGURE 5.4: Distribution of CI scores across pre-binning approaches

Figure 5.5 shows the impact of the choice of clustering algorithm on the CI score. While the figure clearly shows that the k-means algorithm outperforms other algorithms, analysing Figure 5.2 in detail reveals some nuances. Without normalisation, SOM+k-means performs better than k-means on its own, which could be due to the dimensionality-reducing effect of the SOM. With normalisation k-means performs best, followed by SOM+k-means and lastly SOM. SOM frequently had a negative Silhouette Index, which is an indication of incorrect cluster assignment. The CI score was undefined for those experiments.

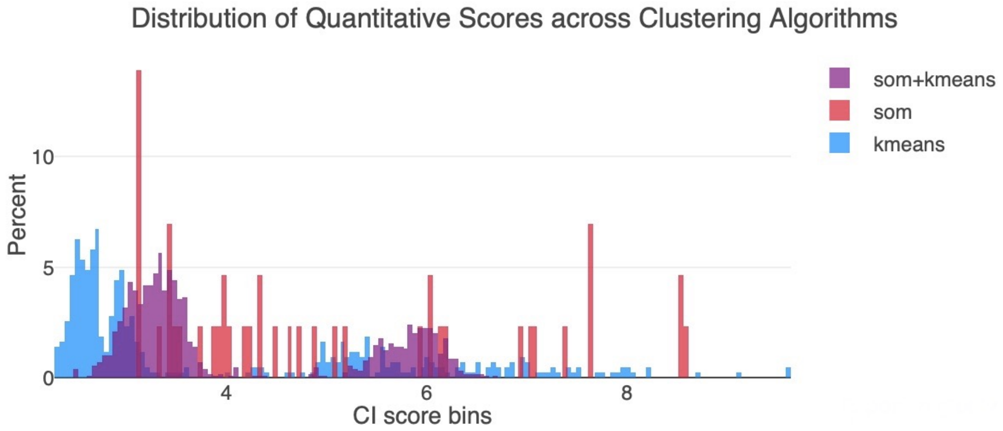
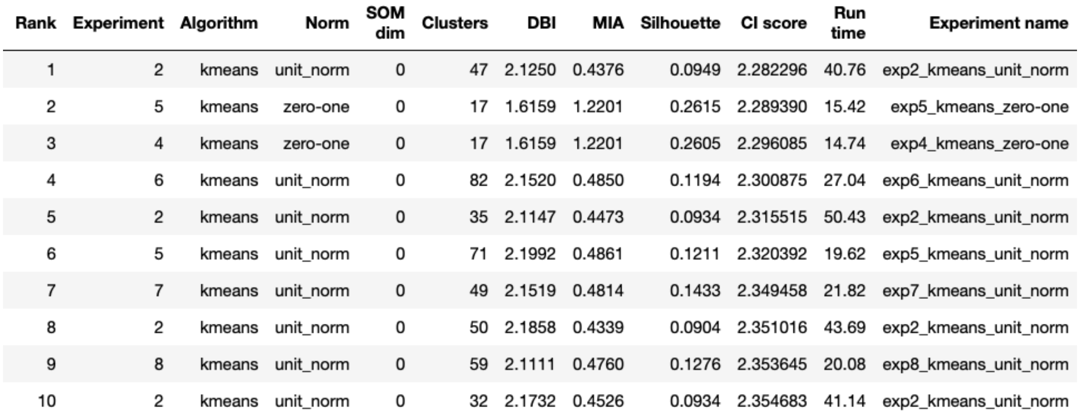


FIGURE 5.5: Distribution of CI scores across clustering algorithms

5.1.2 Top 10 Experiments

The 10 top ranked experiments based on the CI score are shown in Figure 5.6. All of the experiments have been normalised with unit norm, with the exception of two experiments that have been normalised with zero-one. All variations of pre-binning (including no pre-binning) are included in the top results. K-means is the uncontested best clustering algorithm.

The scores are difficult to interpret and are most effectively used as a relative index. Even so, the percentage point difference between the best and tenth best experiment is only 3.2. Selecting the best set of clusters based on these scores alone would be challenging and meaningless.



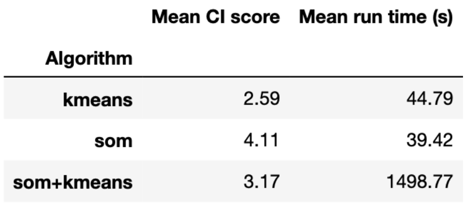
Rank	Experiment	Algorithm	Norm	SOM dim	Clusters	DBI	MIA	Silhouette	CI score	Run time	Experiment name
1	2	kmeans	unit_norm	0	47	2.1250	0.4376	0.0949	2.282296	40.76	exp2_kmeans_unit_norm
2	5	kmeans	zero-one	0	17	1.6159	1.2201	0.2615	2.289390	15.42	exp5_kmeans_zero-one
3	4	kmeans	zero-one	0	17	1.6159	1.2201	0.2605	2.296085	14.74	exp4_kmeans_zero-one
4	6	kmeans	unit_norm	0	82	2.1520	0.4850	0.1194	2.300875	27.04	exp6_kmeans_unit_norm
5	2	kmeans	unit_norm	0	35	2.1147	0.4473	0.0934	2.315515	50.43	exp2_kmeans_unit_norm
6	5	kmeans	unit_norm	0	71	2.1992	0.4861	0.1211	2.320392	19.62	exp5_kmeans_unit_norm
7	7	kmeans	unit_norm	0	49	2.1519	0.4814	0.1433	2.349458	21.82	exp7_kmeans_unit_norm
8	2	kmeans	unit_norm	0	50	2.1858	0.4339	0.0904	2.351016	43.69	exp2_kmeans_unit_norm
9	8	kmeans	unit_norm	0	59	2.1111	0.4760	0.1276	2.353645	20.08	exp8_kmeans_unit_norm
10	2	kmeans	unit_norm	0	32	2.1732	0.4526	0.0934	2.354683	41.14	exp2_kmeans_unit_norm

FIGURE 5.6: Ten best experiments by the CI score

5.1.3 Experiment Run Times

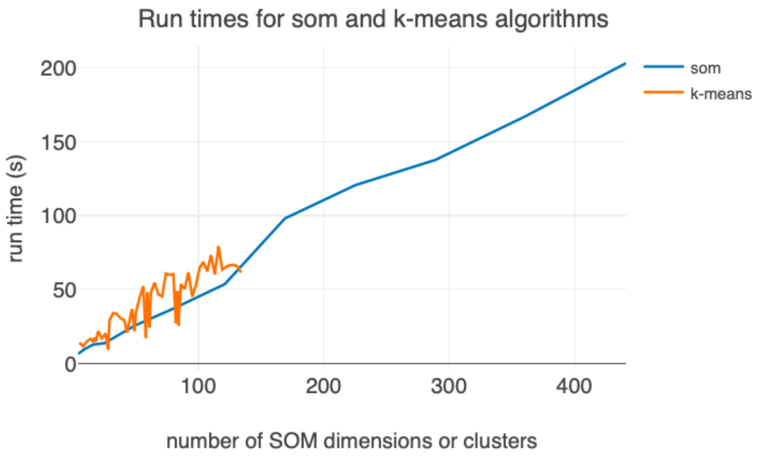
For both the k-means and SOM algorithms the batch fit time increases linearly with dimensionality, as Figure 5.7b demonstrates. When SOM dimensions are low, the run times of the two algorithms are comparable.

For SOM+k-means the SOM is used for dimensionality reduction, not for producing the final clustering structure. The SOM dimensions explored are thus considerably greater, which has a significant impact on increasing experiment run times. This can be seen in Figure 5.7a, which shows the mean run time for each algorithm averaged across all experiments.



	Mean CI score	Mean run time (s)
Algorithm		
kmeans	2.59	44.79
som	4.11	39.42
som+kmeans	3.17	1498.77

(A) Summary of mean experiment run times (seconds)



(B) Cluster run time for k-means and SOM algorithms

FIGURE 5.7: Comparison of experiment run times

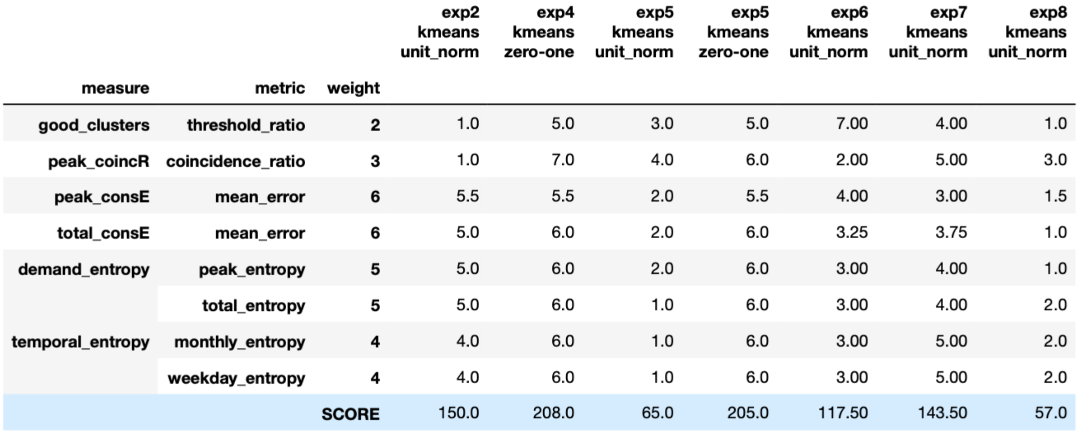
5.2 Qualitative Evaluation of Cluster Sets

Qualitative scores have been calculated for the best of the top 10 experiments generated from a pre-binning, normalisation and algorithm combination. Thus for experiment 2 with k-means and unit norm, only the top clustering structure with $n = 47$ has been evaluated further. Clusters with membership below a threshold of 10490 have been removed and performance is weighted by cluster size to account for the overall effect that a particular cluster has on the experiment.

5.2.1 Cluster Scoring Matrix Results

The results of the qualitative evaluation are captured in Figure 5.8. Experiments with pre-binning (experiments 5, 6, 7, 8) and unit norm normalisation perform better than the remaining experiments when qualitative measures are taken into consideration.

The top two experiments in Table 5.1, Exp 8 (k-means, unit norm) and Exp 5 (k-means, unit norm), lie only 8 points apart. Their positioning between best and second best experiment is largely influenced by the threshold value, the weights that have been assigned to the evaluation measures and the ranking method. These two experiments comfortably outperform the third best experiment, which has double the score. Interestingly, the only difference between experiment 5 and 6 is that the former, better performing experiment, contained zero-valued profiles, while the latter removed them. For experiment 7 and 8 the reverse is true. Experiment 8 removed zero-valued profiles and significantly outperformed experiment 7, which retained them.



measure	metric	weight	exp2 kmeans unit_norm	exp4 kmeans zero-one	exp5 kmeans unit_norm	exp5 kmeans zero-one	exp6 kmeans unit_norm	exp7 kmeans unit_norm	exp8 kmeans unit_norm
good_clusters	threshold_ratio	2	1.0	5.0	3.0	5.0	7.00	4.00	1.0
peak_coincR	coincidence_ratio	3	1.0	7.0	4.0	6.0	2.00	5.00	3.0
peak_consE	mean_error	6	5.5	5.5	2.0	5.5	4.00	3.00	1.5
total_consE	mean_error	6	5.0	6.0	2.0	6.0	3.25	3.75	1.0
demand_entropy	peak_entropy	5	5.0	6.0	2.0	6.0	3.00	4.00	1.0
	total_entropy	5	5.0	6.0	1.0	6.0	3.00	4.00	2.0
temporal_entropy	monthly_entropy	4	4.0	6.0	1.0	6.0	3.00	5.00	2.0
	weekday_entropy	4	4.0	6.0	1.0	6.0	3.00	5.00	2.0
		SCORE	150.0	208.0	65.0	205.0	117.50	143.50	57.0

FIGURE 5.8: Cluster Scoring Matrix

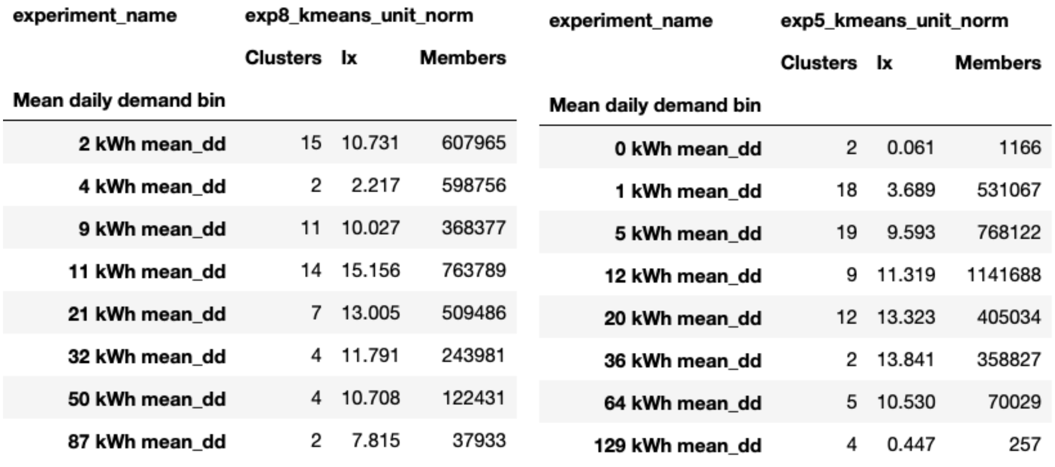
Rank	Score	Exp.	Algorithm	Normalisation	Pre-binning	Drop 0
1	57.0	8	k-means	unit norm	integral k-means	True
2	65.0	5	k-means	unit norm	AMC	
3	117.5	6	k-means	unit norm	AMC	True
4	143.5	7	k-means	unit norm	integral k-means	
5	150.0	2	k-means	unit norm		
6	205.0	5	k-means	zero-one	AMC	
7	208.0	4	k-means	zero-one	AMC	

TABLE 5.1: Experiments ranked by qualitative scores

Contrasting the results of the quantitative and qualitative evaluation, Exp 5 (k-means, zero-one) had the second best run based on the quantitative CI score but was ranked second last during qualitative evaluation. Exp 8 (k-means, unit norm) on the other hand only ranked ninth by quantitative score, yet convincingly claimed the top position based on qualitative measures. We closely examine the qualitative evaluation measures for Exp 5 (k-means, zero-one), Exp 5 (k-means, unit norm) and Exp 8 (k-means, unit norm) in the sections that follow. These three experiments have been selected to validate whether the re-ranking of experiments by the qualitative measures is justified, and to assess the difference in clustering structures produced by the top two experiments.

5.2.2 Analysis of Cluster Usability

Cluster usability assesses the capacity of RDLPs to represent consumption behaviour that makes sense in relation to existing knowledge about consumers, and to have potential for meaning. Usable RDLPs are well-shaped and premised on well-sized clusters that are neither so large that they are devoid of meaning, nor so small that they only represent an insignificant amount of member profiles.

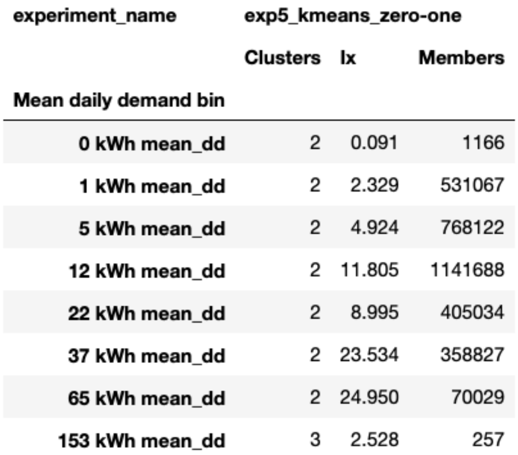


experiment_name	exp8_kmeans_unit_norm		
	Clusters	Ix	Members
Mean daily demand bin			
2 kWh mean_dd	15	10.731	607965
4 kWh mean_dd	2	2.217	598756
9 kWh mean_dd	11	10.027	368377
11 kWh mean_dd	14	15.156	763789
21 kWh mean_dd	7	13.005	509486
32 kWh mean_dd	4	11.791	243981
50 kWh mean_dd	4	10.708	122431
87 kWh mean_dd	2	7.815	37933

(A) exp 8 k-means unit norm

experiment_name	exp5_kmeans_unit_norm		
	Clusters	Ix	Members
Mean daily demand bin			
0 kWh mean_dd	2	0.061	1166
1 kWh mean_dd	18	3.689	531067
5 kWh mean_dd	19	9.593	768122
12 kWh mean_dd	9	11.319	1141688
20 kWh mean_dd	12	13.323	405034
36 kWh mean_dd	2	13.841	358827
64 kWh mean_dd	5	10.530	70029
129 kWh mean_dd	4	0.447	257

(B) exp 5 k-means unit norm



experiment_name	exp5_kmeans_zero-one		
	Clusters	Ix	Members
Mean daily demand bin			
0 kWh mean_dd	2	0.091	1166
1 kWh mean_dd	2	2.329	531067
5 kWh mean_dd	2	4.924	768122
12 kWh mean_dd	2	11.805	1141688
22 kWh mean_dd	2	8.995	405034
37 kWh mean_dd	2	23.534	358827
65 kWh mean_dd	2	24.950	70029
153 kWh mean_dd	3	2.528	257

(C) exp 5 k-means zero-one

FIGURE 5.9: Bin details and cluster scores of three experiments

The ranked threshold ratio in Figure 5.8 provides a relative measure for how well-sized the clusters of different experiments are. Exp 8 (k-means, unit norm) was ranked best, Exp 5 (k-means, unit norm) in the middle and Exp 5 (k-means, zero-one) second worst based, on threshold ratio. Figure 5.9 shows that the bin sizes are more evenly spread for Exp 8 (k-means, unit norm), which was pre-binned with integral k-means, than for Exp 5 (k-means, zero-one) and Exp 5 (k-means, unit norm), which were pre-binned based on average monthly consumption. The bin values are the mean total daily consumption (kWh) of all the daily load profiles that are members of clusters in that bin. Intuitively, the threshold ratio ranking makes sense in relation to the number of clusters and cluster membership size of the three experiments.

Regardless of the experiment, half the bins have a mean daily consumption of 12kWh or less, and over half the daily load profiles are assigned to these bins. This corresponds to the mean daily consumption distribution of the dataset (see Figure 3.6). From the individual bin scores it is interesting to observe that bins with small numbers of clusters or few members tend to have a very low quantitative score.

Description of Representative Daily Load Profiles and Member Profiles

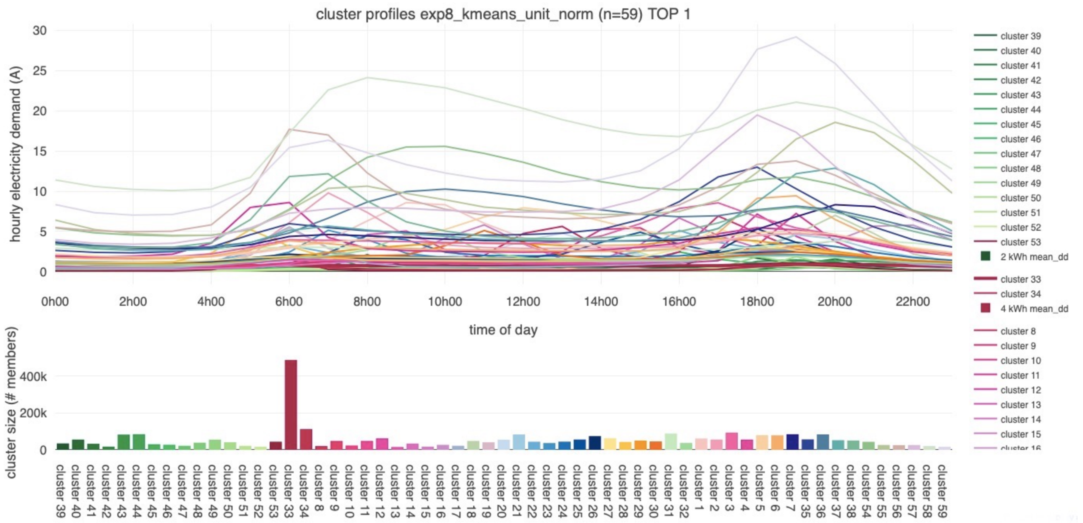
All RDLPs for Exp 5 (k-means, zero-one), Exp 5 (k-means, unit norm) and Exp 8 (k-means, unit norm) are captured in Figure 5.10. In each subfigure the top plot captures the cluster RDLPs and the bottom plot shows the member count per cluster.

Exp 8 (k-means, unit norm) has 59 clusters, varying between 2 and 15 clusters per bin (Figure 5.10a). With the exception of *Cluster 33* which accounts for roughly 15% of all daily load profiles, cluster membership for the remaining clusters varies in a range from 15 000 to 100 000 members. *Cluster 33* is one of only two clusters in the 4kWh mean_dd bin, which has the second largest bin membership for the experiment. Exp 5 (k-means, unit norm) has more clusters and a smaller range of cluster sizes than Exp 8 (k-means, unit norm). However, it contains some very small clusters that are not usable (*Cluster 1, 2, Cluster 68 - 71*). The demand range of RDLPs of the two cluster sets is similar and several clusters appear to be comparable between them.

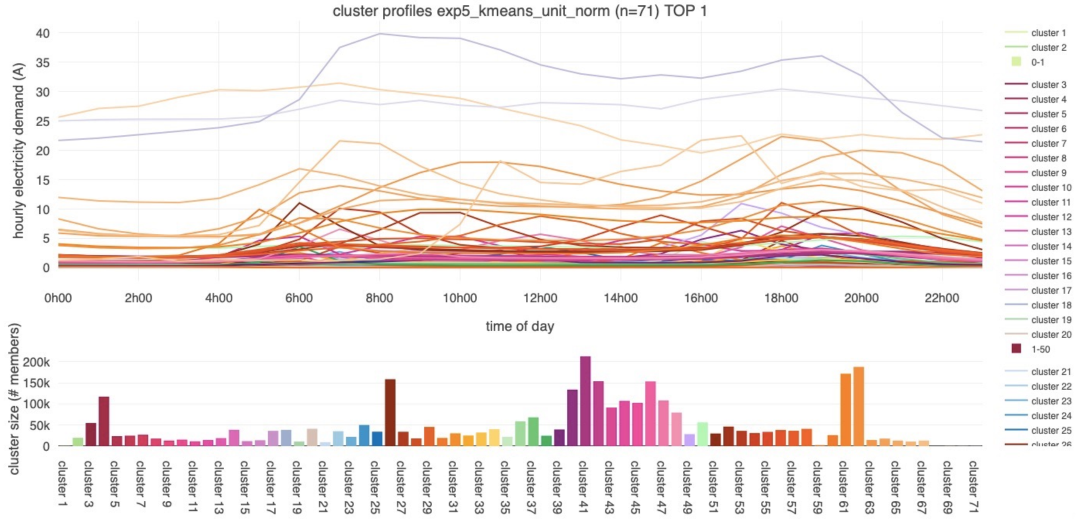
By contrast, Exp 5 (k-means, zero-one) has only 18 clusters and on average 2.125 clusters per bin (Figure 5.10c). Five of the 18 clusters have very few members and appear hidden in the bar plot. The next three larger clusters account for less than 3% of all load profiles. Over half of all load profiles belong to only three clusters, *Cluster 5*, *Cluster 6* and *Cluster 9*. The individual RDLPs lack distinguishing features and are not usable.

Exp 8 (k-means, unit norm) contains the most distinct shapes, while Exp 5 (k-means, zero-one) has very flat shapes that lack characteristic features. Exp 5 (k-means, zero-one) and Exp 5 (k-means, unit norm) do not capture the peak consumption ranges of member profiles as well as Exp 8 (k-means, unit norm). *Cluster 26, 36* and *Cluster 41* of Exp 5 (k-means, unit norm) contain member profiles whose peaks exceed those of the cluster RDLP by 3 to 4 times. *Cluster 6, 8* and *Cluster 9* of Exp 5 (k-means, zero-one) are even less compact, with RDLPs representing neither the member shapes nor the consumption ranges.

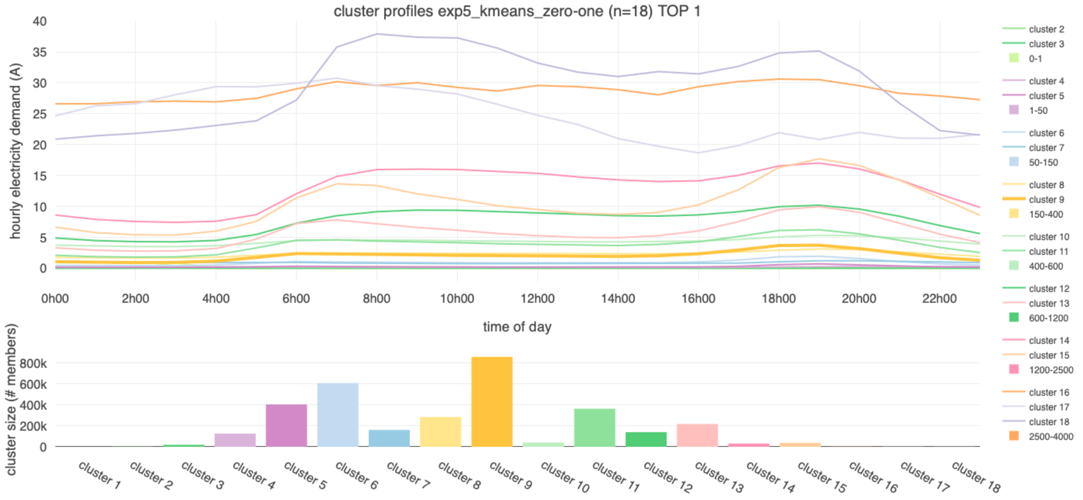
Exp 8 (k-means, unit norm) does not contain a zero profile. *Cluster 3* is the zero-profile of Exp 5 (k-means, zero-one). Exp 5 (k-means, unit norm) has a zero-profile (*Cluster 2*) and two bins with profiles that appear to be zero proxies - *Cluster 15* and *Cluster 59*.



(A) exp 8 (k-means, unit norm, integral k-means pre-binning)



(B) exp 5 (k-means, unit norm, AMC pre-binning)



(C) exp 5 (k-means, zero-one norm, AMC pre-binning)

FIGURE 5.10: Representative Daily Load Profiles of three experiments

The RDLPs of the 15 largest clusters of Exp 5 (k-means, zero-one), Exp 5 (k-means, unit norm) and Exp 8 (k-means, unit norm) and 250 randomly sampled member profiles of those clusters are shown in Figures 5.11 to 5.13. These plots provide an intuition for how well-shaped the different RDLPs are, and provide a visual representation of cluster distinctness and compactness measured by the internal indexes.

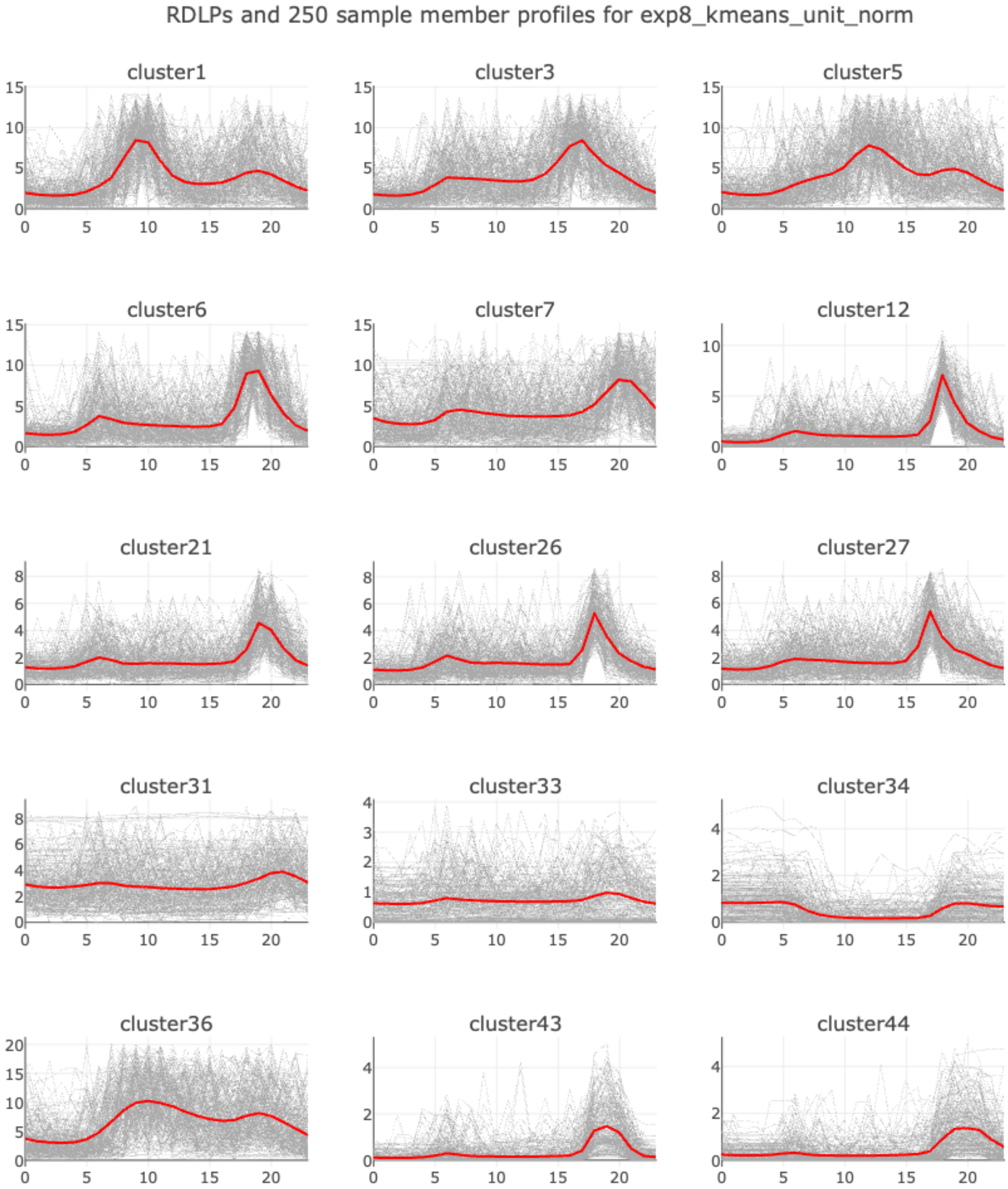


FIGURE 5.11: 15 largest clusters for exp 8 (k-means, unit norm, integral k-means pre-binning)

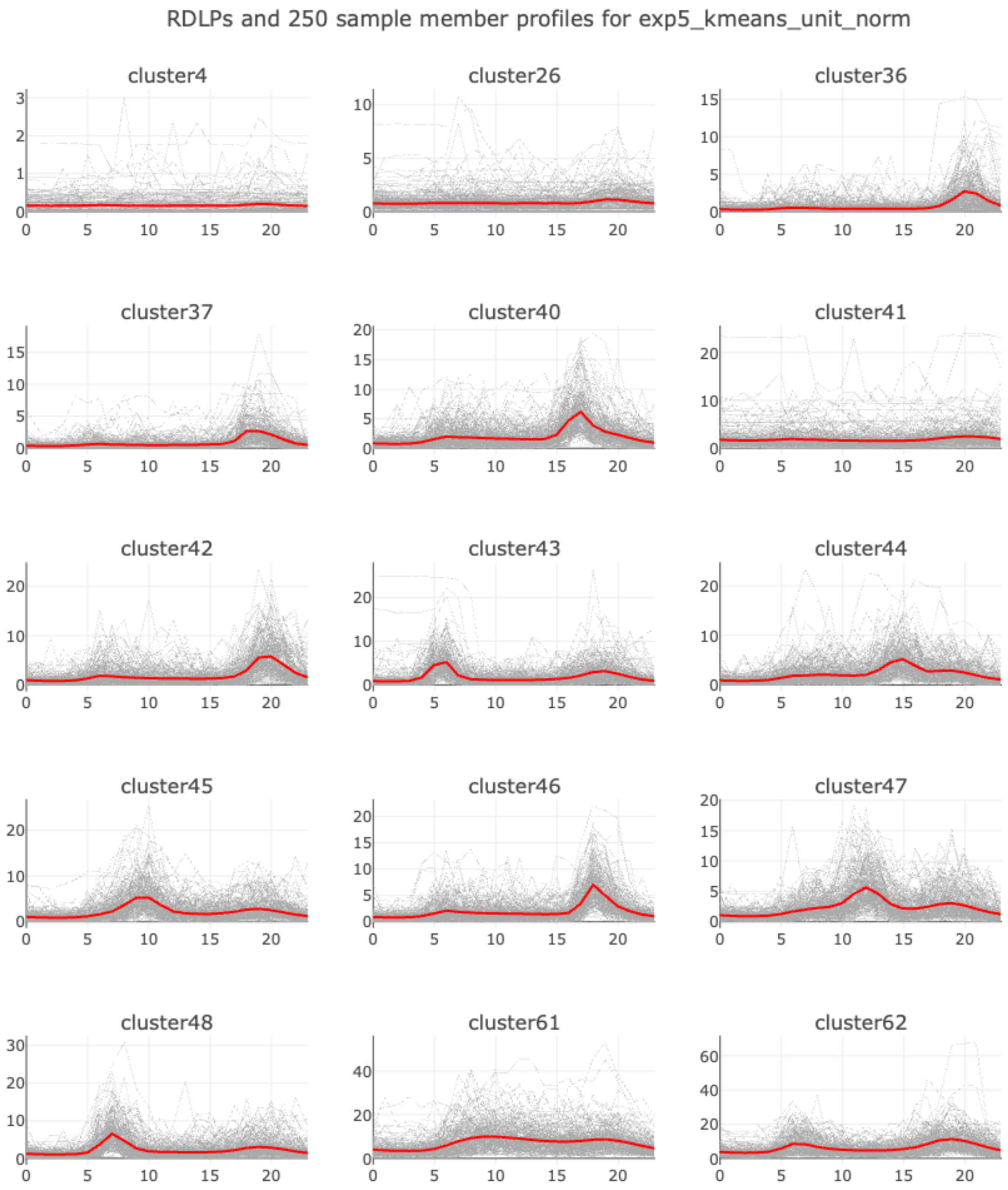


FIGURE 5.12: 15 largest clusters for exp 5 (k-means, unit norm, AMC pre-binning)

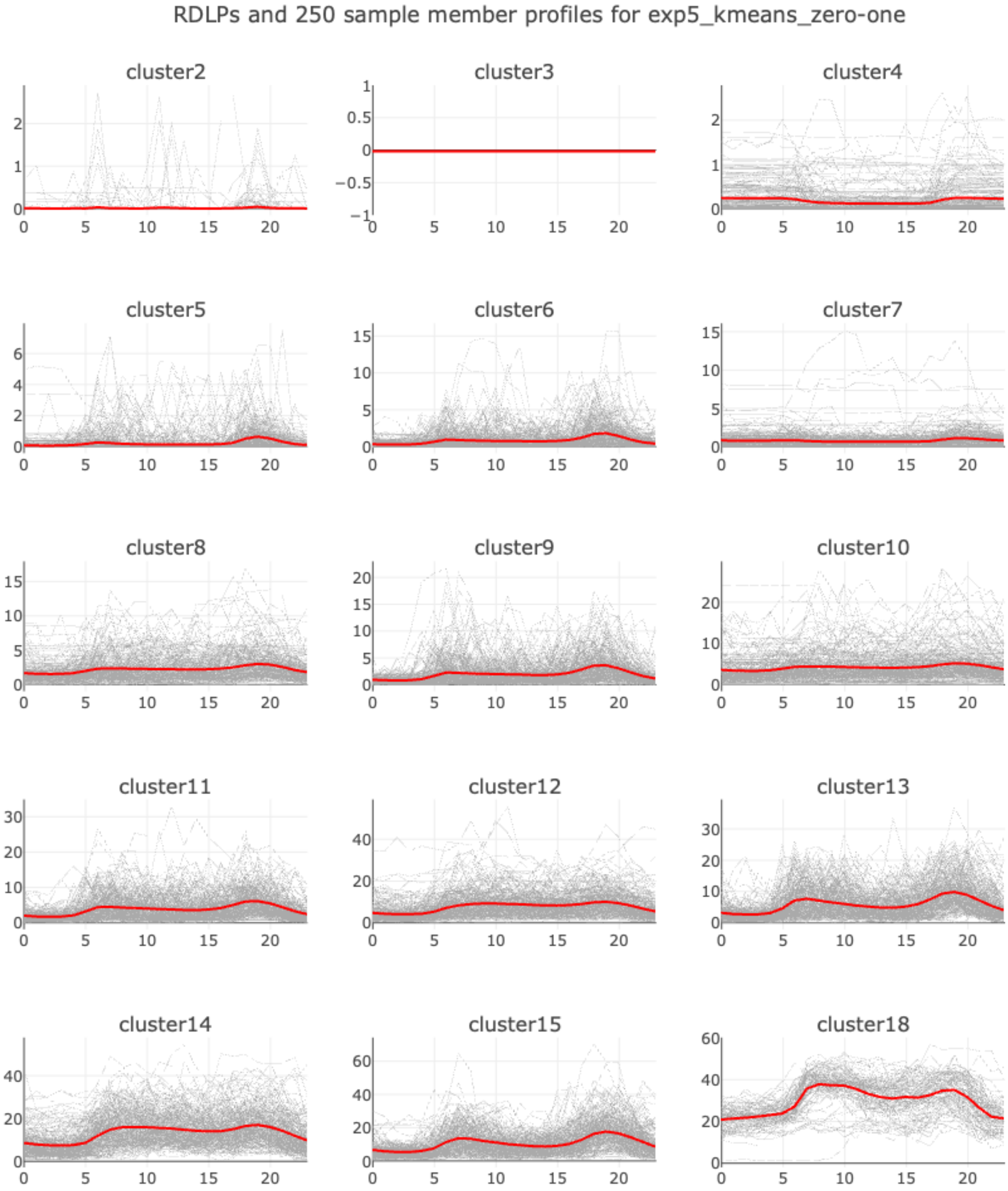


FIGURE 5.13: 15 largest clusters for exp 5 (k-means, zero-one norm, AMC pre-binning)

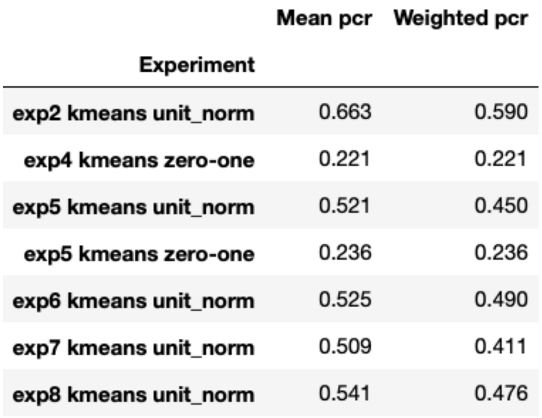
5.2.3 Analysis of Cluster Expressivity

We defined cluster expressivity as the ability of a cluster to convey specific information about the consumption behaviour and loading conditions of member profiles. Clusters with good expressivity are representative and homogeneous. The metrics used to determine representativity and homogeneity are described in Section 3.3).

Evaluation of Cluster Representativity

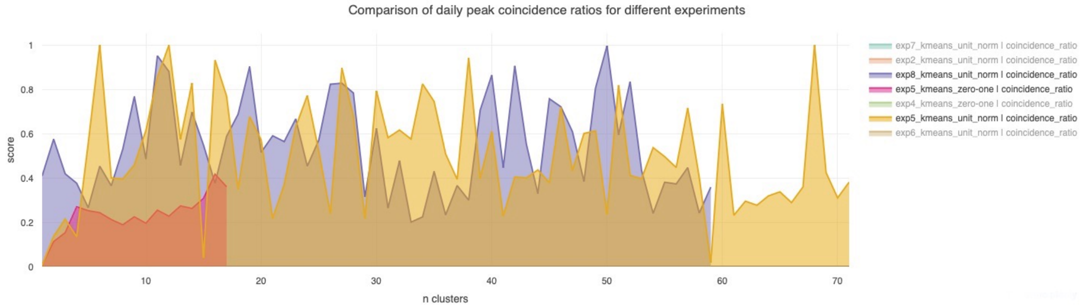
Figure 5.15 visualises the scores of the four peak consumption error metrics, the peak coincidence ratio, total demand entropy and weekday entropy to illustrate how we evaluated whether clusters represent their member profiles. The scores in the figure are unweighted and the threshold constraint has been removed to demonstrate the performance of individual clusters. Three experiments are shown. Exp 5 (k-means, unit norm) and Exp 8 (k-means, unit norm) show comparable performance across most measures. Exp 5 (k-means, zero-one) performs considerably worse.

The mean peak coincidence ratio for most experiments is poor. Figure 5.14 tabulates both the unweighted and the weighted mean ratio for the experiments under evaluation. Clusters with membership size below the threshold have been removed. The metric can be interpreted directly as the average number that peak times between daily load profiles and the RDLP of the cluster they are assigned to overlap. Exp 2 (k-means, unit norm) has the largest overlap, which is still low at 0.59. For all other experiments, less than half the peaks overlap. The experiments normalised with zero-one perform extremely poorly, with less than a quarter of peaks coinciding.

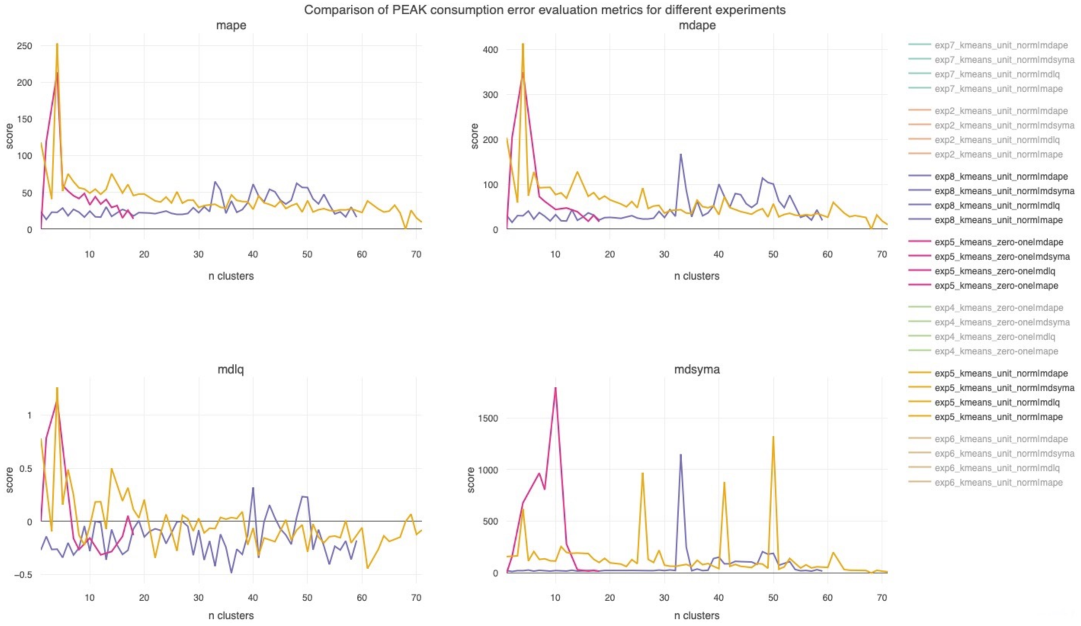


Experiment	Mean pcr	Weighted pcr
exp2 kmeans unit_norm	0.663	0.590
exp4 kmeans zero-one	0.221	0.221
exp5 kmeans unit_norm	0.521	0.450
exp5 kmeans zero-one	0.236	0.236
exp6 kmeans unit_norm	0.525	0.490
exp7 kmeans unit_norm	0.509	0.411
exp8 kmeans unit_norm	0.541	0.476

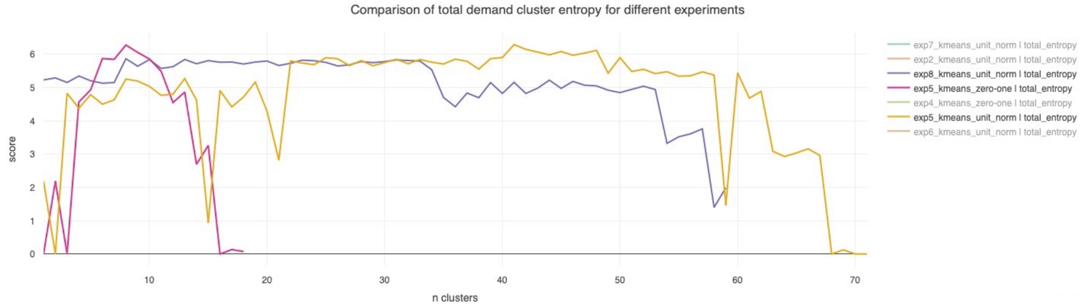
FIGURE 5.14: Unweighted and weighted mean peak coincidence ratio



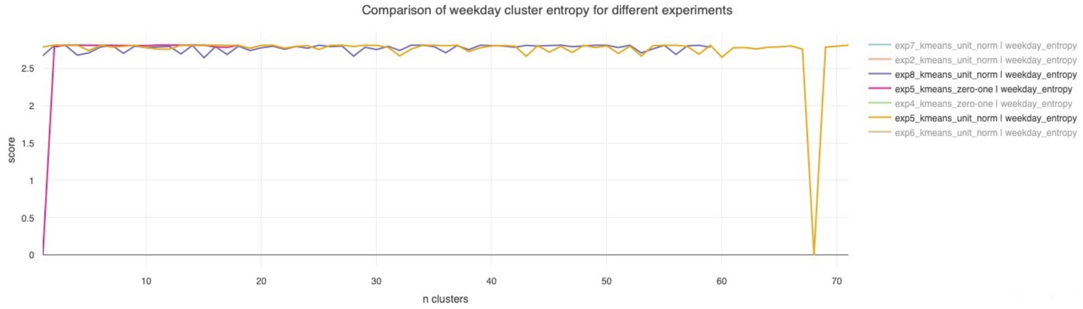
(A) Peak coincidence ratios



(B) Peak consumption error metrics



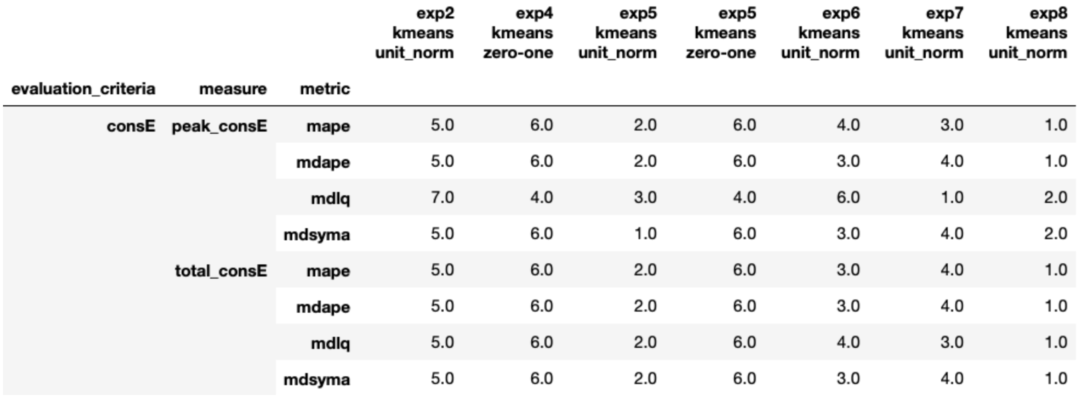
(C) Total demand entropy



(D) Weekday entropy

FIGURE 5.15: Visual representation of qualitative measures of three experiments

The experiment ranking across the consumption error metrics is depicted in Figure 5.16. For the total consumption error measure the ranking is stable across metrics. Peak consumption error is relatively consistent, with the exception of the median log accuracy ratio (mdlq) which shifts the ranking by two positions for most of the experiments.



evaluation_criteria	measure	metric	exp2 kmeans unit_norm	exp4 kmeans zero-one	exp5 kmeans unit_norm	exp5 kmeans zero-one	exp6 kmeans unit_norm	exp7 kmeans unit_norm	exp8 kmeans unit_norm
consE	**peak_consE**	**mape**	5.0	6.0	2.0	6.0	4.0	3.0	1.0
		mdape	5.0	6.0	2.0	6.0	3.0	4.0	1.0
		mdlq	7.0	4.0	3.0	4.0	6.0	1.0	2.0
		mdsyma	5.0	6.0	1.0	6.0	3.0	4.0	2.0
	total_consE	**mape**	5.0	6.0	2.0	6.0	3.0	4.0	1.0
		mdape	5.0	6.0	2.0	6.0	3.0	4.0	1.0
		mdlq	5.0	6.0	2.0	6.0	4.0	3.0	1.0
		mdsyma	5.0	6.0	2.0	6.0	3.0	4.0	1.0

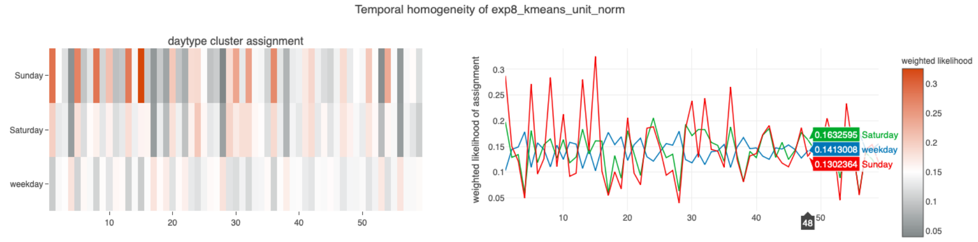
FIGURE 5.16: Experiment ranking across consumption error metrics

Analysing the peak consumption error metrics in Figure 5.15b reveals that the median symmetric accuracy (mdsyma) provides the most interpretable peak consumption error metric. It is dominated by some large clusters - *Cluster 33* for Exp 8 (k-means, unit norm) and *Cluster 4, 26, 41* for Exp 5 (k-means, unit norm). For the latter *Cluster 50* which is average sized has a surprisingly large peak consumption error. The scores for Exp 5 (k-means, zero-one) do not correspond to cluster size, as the consumption error is undefined for zero-valued load profiles contained in several of the clusters.

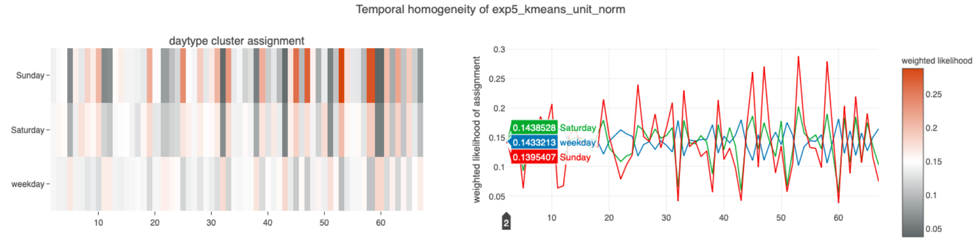
Clusters in Exp 8 (k-means, unit norm) at the lower and upper end of the consumption spectrum (*Cluster 35 - 59*) have a lower total demand entropy and higher weekday entropy. These clusters are thus more specific with regards to the demand percentile in which they are used, and less specific about the day on which they are used. A similar trend can be observed for Exp 5 (k-means, unit norm). *Cluster 1 - 21* the very low consumption clusters, and *Cluster 61 - 71* the very high consumption clusters, have a lower demand entropy and higher weekday entropy (see Subfigures 5.15c and 5.15d).

Visualising Cluster Homogeneity - Temporal Entropy

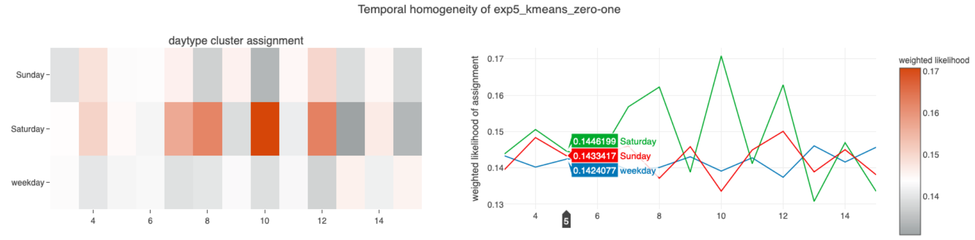
We investigate cluster homogeneity more closely by examining the temporal entropy of the three experiments in Figure 5.17. The figure shows the entropy for a daytype feature with values workday, Saturday and Sunday. The daytype entropy is the weighted weekday entropy which has been used in the cluster scoring matrix. The reduced feature values make daytype entropy simpler to understand and explain visually than weekday entropy. The representations are comparable.



(A) Temporal entropy of exp 8 (k-means, unit norm, integral k-means pre-binning)



(B) Temporal entropy of exp 5 (k-means, unit norm, AMC pre-binning)



(C) Temporal entropy of exp 5 (k-means, zero-one norm, AMC pre-binning)

FIGURE 5.17: Temporal homogeneity of three experiments

The heatmap on the left hand side of each Subfigure shows the weighted likelihood that member profiles of a cluster are used on a particular daytype. The darker the red, the greater the probability that a cluster is used on that daytype. The darker the grey, the less likely that a cluster is used on a particular day. The likelihoods are weighted against the random likelihood of a cluster being used on any day of the week. The light white colour marks the midpoint between a likelihood that is greater and less than random assignment.

In the linegraph on the right hand side each daytype is represented by a trace. The x-axis contains the cluster names in increasing order. The higher the peak of the trace, the more likely that to profiles used on that daytype are assigned to that cluster. The lower the peak, the less likely that this is the case. *Cluster 15* of Exp 8 (k-means, unit norm) is a good example of a cluster that has a very high likelihood of being used on a Sunday, an almost random likelihood of being used on a Saturday and a less than random likelihood of being used on a Sunday. This cluster thus has high temporal homogeneity.

Both Exp 5 (k-means, unit norm) and Exp 8 (k-means, unit norm) show good peakiness for different daytypes, giving an intuition that some of the clusters are specific to workdays, others to weekends, Saturdays or Sundays. Only few clusters like *Cluster*

14, 39, 46, 57 for Exp 8 (k-means, unit norm) or *Cluster 15, 16, 34, 35* for Exp 5 (k-means, unit norm) do not have a greater likelihood of being used on one daytype over another. Moreover, they show specific preferences for all daytypes. Exp 5 (zero-norm) on the other hand indicates random likelihood of assignment for *Cluster 5, 6, 9, 11*, which happen to be its largest clusters. Six of the remaining clusters (one third of all clusters) are most likely to be used on a Saturday. Only two clusters show a very marginal higher likelhood of being used on a weekday. The temporal homogeneity of Exp 5 (k-means, zero-one) is poor.

Exp 5 (k-means, unit norm) and Exp 8 (k-means, unit norm) both present strong qualitative results. Their individual RDLPs are expressive, featured and distinct, which promises that they will be useful for constructing customer archetypes. The RDLPs of Exp 5 (k-means, zero-one) on the other hand are too few and too generic to carry sufficient information to represent our dataset.

5.3 Summary

This chapter presents and analyses the results of the experiments, closely examining both the CI scores and the ranking of the cluster scoring matrix. The sets of RDLPs from three of the top 10 experiments are visualised and discussed. In the next chapter we demonstrate how the RDLPs of Exp 8 (k-means, unit norm) can be used in a real world application. Exp 5 (k-means, unit norm) has several high consumption clusters that do not meet the threshold value and is thus less useable to cover a range of customer archetypes than Exp 8.

Chapter 6

Application

In this chapter the representative daily load profiles developed in the previous chapter are tested by reconstructing customer archetypes currently used for long term energy forecasting in South Africa. The chapter starts by outlining the methodology and features that characterise typical residential energy consumers. Limitations of the new archetypes are discussed and a benchmark is specified. Finally, some of the archetypes are presented.

6.1 Constructing Customer Archetypes

To test the RDLPs of Exp 8 (k-means, unit norm) in a real world application, customer archetypes are created and compared against a benchmark of existing expert archetypes. Figure 6.1 presents an overview of the method followed for creating customer archetypes.

Household survey data is joined with cluster and temporal data of the best experiment to characterise the RDLPs by socio-demographic attributes. We combine clusters, temporal features, spatial features and socio-demographic features into a new dataset F for classification. Multinomial logistic regression is used to build a model that quantifies the likelihood that a particular feature value characterises a cluster. The attributes of customer archetypes are based on expert-defined criteria. Clusters characterised by the set of values that corresponds to the most distinguishing attributes of an archetype are then gathered as the set of RDLPs for that customer archetype. Finally the new customer archetypes are compared against a benchmark of comparable, typical RDLPs extracted from archetypes previously constructed by experts.

6.1.1 Socio-demographic Feature Extraction

The sparse socio-demographic data was condensed by creating a single feature vector per survey question, with all categorical responses represented as feature values.

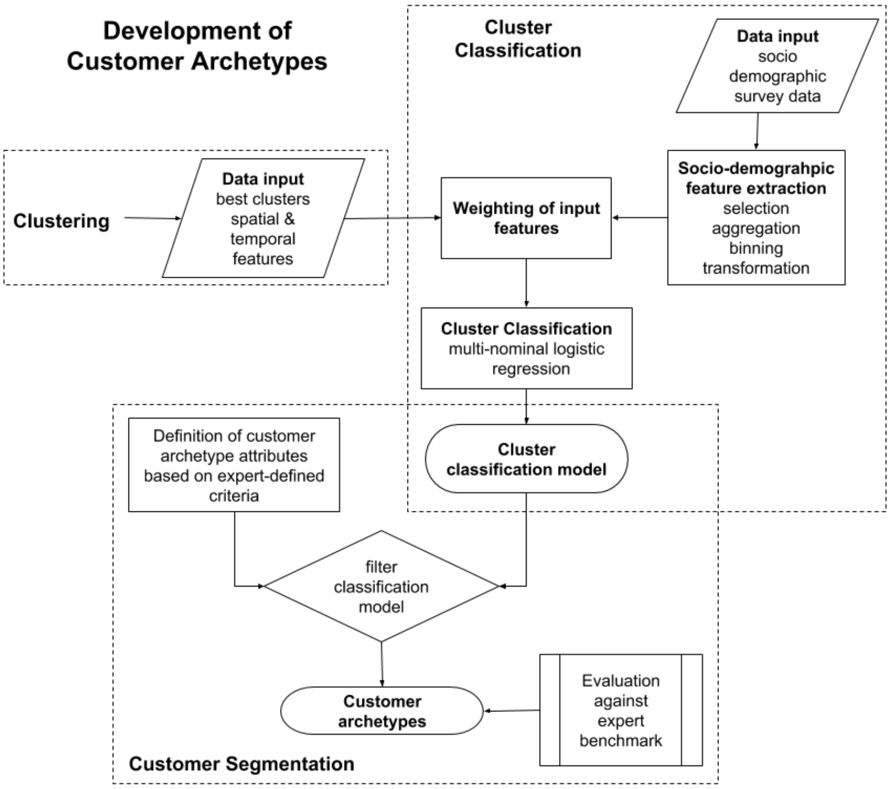

Development of Customer Archetypes
Cluster Classification
Data input socio demographic survey data
Clustering
Data input best clusters spatial & temporal features
Weighting of input features
Socio-demograhpic feature extraction selection aggregation binning transformation
Cluster Classification multi-nominal logistic regression
Definition of customer archetype attributes based on expert-defined criteria
Cluster classification model
filter classification model
Customer archetypes
Evaluation against expert benchmark
Customer Segmentation

FIGURE 6.1: Method for Constructing Customer Archetype

Feature selection from the questions of the annual household survey is informed by socio-demographic characteristics that experts consider to be significant contributing factors to electricity consumption. These characteristics can be categorised as economic, dwelling, connection, appliance and occupant features. All feature values are represented as nominal or category values. They are presented below.

Economic Features

The monthly household income is the combined income of salaries and external financial sources, such as pensions, remittances, the sale of agricultural produce and small business revenues. Early surveys from 1994 - 1999 only asked about salary information. All recorded income values are adjusted for inflation using the CPI index supplied by Statistics South Africa, referenced to December 2016. The following equation is used to adjust the monthly income:

$$MonthlyIncome_{adjusted} = MonthlyIncome_{year} / CPI_{year}$$

Household income ranges used by domain experts for existing customer archetypes are used as a guideline to create monthly income bins. Some high income households chose not to report their income and it reflects as R0[1]. In the dataset these households can still be identified as high income households based on the large floor area of the dwelling.

Dwelling Features

Dwelling floor area ranges are derived from ranges currently used by experts. These roughly correspond with small, medium, large and estate-sized homes in South Africa.

Connection Features

First-time electrified households accumulate appliances over time. Domain experts assume that household electricity demand stabilises 15 years after electrification. Taking this into consideration we set the nominal values for the *years_electrified* feature to ranges of 0-5 years, 5-10 years, 10-15 years and over 15 years to distinguish between newly electrified, medium-term electrified, mature electrified and long-term electrified consumers.

The circuit breaker size of the main switch limits the instantaneous maximum electrical demand of a household. South Africa's electrification programme supplied electrical connections up to 20A free of charge to households, while the standard connection is 60A. Connections between 20A and 60A, and connections over 60A are not common in the dataset. The nominal values for *cb_size* are thus selected as <20A, 21-60A and <61A to reflect typical connection types.

Appliance Features

Geysers are a major contributor towards household electricity consumption and the count of geysers in the household is thus represented as a feature in the appliance category. A geyser value of -1 indicates that the geyser is broken.

Occupant Features

Features in the occupant category contain a count of the number of people living in a household with that attribute (eg *part_time_employed* = 3-4 means that 3 to 4 adults living in a household were part time employed at the time the survey was conducted).

---

[1] Domain experts that managed the data collection process highlighted that survey respondents were not obliged to respond to all questions. Some respondents withheld information on household income. This was usually only the case for households in wealthy areas, which by inference were assumed to have a high income

Adults are people over the age of 16. The employment status of children is not taken into consideration.

6.1.2 Feature Weighting and Data Input

To construct the input dataset F for classification, we replace each daily load profile $h_d^{(j)}$ in X by its most representative cluster k_x from Exp 8 (k-means, unit norm), and assign it to $f_d^{(j)}$, a feature vector representing household j on day d. Individual date stamps are transformed to season and day type features, as daily routines and seasonal climatic variability have a known impact on household electricity consumption. The dataset can be reduced by creating one weighted instance per cluster, season and day type for each household. The instance weight is computed as the count of days d assigned to cluster k_x for a season and day type.

$$f_d^{(j)} = k_x + f_{season} + f_{daytype} \tag{6.1}$$

$$f_{xsd}^{(j)} = k_x + f_{season} + f_{daytype} + weight_{k_x,season,daytype}^{(j)} \tag{6.2}$$

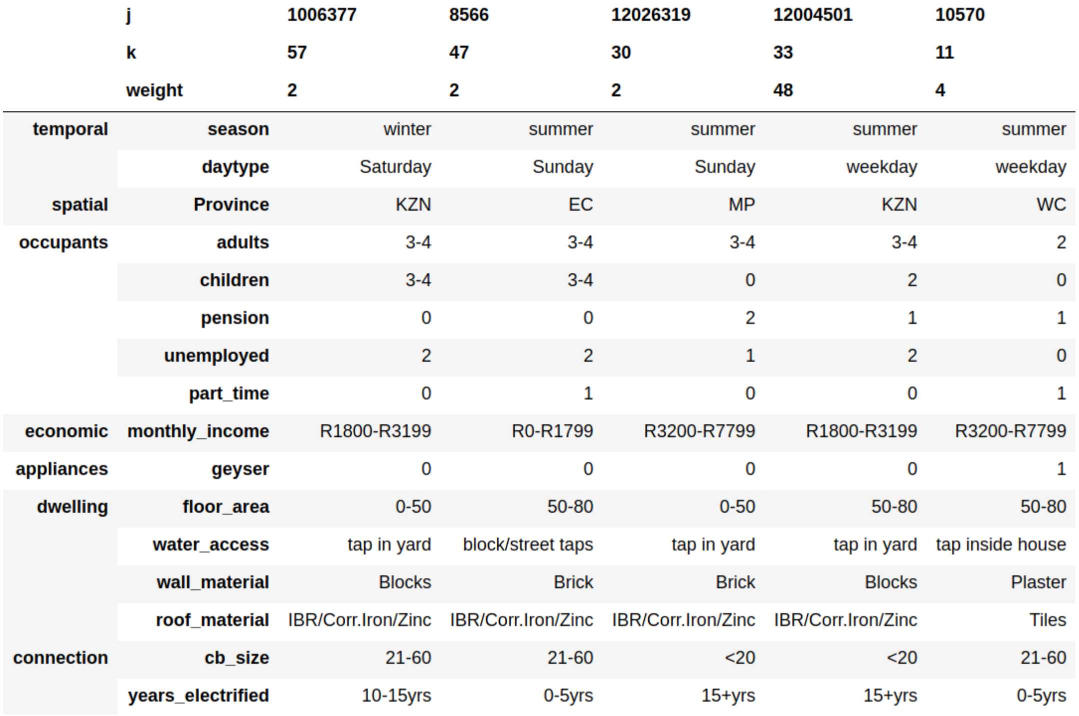
$$\begin{aligned} weight_{k_x,season,daytype}^{(j)} =& \# f_d^{(j)}, \text{ where} \\ & season \in \{winter, summer\}, \\ & daytype \in \{weekday, Friday, Saturday, Sunday\}, \\ & k_x \in k \text{ and has been assigned to } f_d^{(j)} \end{aligned}$$

We join $f_{xsd}^{(j)}$ with all socio-demographic features into a 16 element feature vector $f_{xsd}^{(j)'}$ and gather all feature vectors of all households into F, which is used as input to the cluster classification. If $weight_{k_x,season,daytype}^{(j)} = 0$, $f_{xsd}^{(j)}$ is not included in F.

$$f_{xsd}^{(j)'} = f_{xsd}^{(j)} + f_{spatial}^{(j)} + f_{occupants}^{(j)} + f_{economic}^{(j)} + f_{appliances}^{(j)} + f_{dwelling}^{(j)} + f_{connection}^{(j)} \tag{6.3}$$

$$F = [f'_{xsd}] \tag{6.4}$$

A random sample of the feature input is shown in Figure 6.2. Appendix A contains a table of all features, the feature values and the count of daily load profiles per value prior to weighting.



	j	1006377	8566	12026319	12004501	10570
	k	57	47	30	33	11
	weight	2	2	2	48	4
temporal	season	winter	summer	summer	summer	summer
	daytype	Saturday	Sunday	Sunday	weekday	weekday
spatial	Province	KZN	EC	MP	KZN	WC
occupants	adults	3-4	3-4	3-4	3-4	2
	children	3-4	3-4	0	2	0
	pension	0	0	2	1	1
	unemployed	2	2	1	2	0
	part_time	0	1	0	0	1
economic	monthly_income	R1800-R3199	R0-R1799	R3200-R7799	R1800-R3199	R3200-R7799
appliances	geyser	0	0	0	0	1
dwelling	floor_area	0-50	50-80	0-50	50-80	50-80
	water_access	tap in yard	block/street taps	tap in yard	tap in yard	tap inside house
	wall_material	Blocks	Brick	Brick	Blocks	Plaster
	roof_material	IBR/Corr.Iron/Zinc	IBR/Corr.Iron/Zinc	IBR/Corr.Iron/Zinc	IBR/Corr.Iron/Zinc	Tiles
connection	cb_size	21-60	21-60	<20	<20	21-60
	years_electrified	10-15yrs	0-5yrs	15+yrs	15+yrs	0-5yrs

FIGURE 6.2: Random sample of 5 $f_{xsd}^{(j)'}$ from classification data input F

6.1.3 Expert Defined Attributes of Customer Archetypes

Experts use observations of physical household attributes related to dwelling type and infrastructure, socio-demographics, and load consumption to form a dominant impression about the household's customer class. For example, take a household that is a 'traditional hut' built with 'tradtional construction methods' such as clay and thatch. The income of the household is below R1500 and derived primarily from pensions and subsistence farming. There may be water reticulation to communal stand pipes, but frequently water is collected from a nearby river or dam. These observations lead to the determination of a customer class, in this case a rural customer. The existing process identifies energy consumption behaviour of households by the class name, which implies the associated class load profile. This process is shown in Figure 6.3.

After first electrification, appliance ownership of a household increases as consumers grow accustomed to energy services and fully exploit their benefits. After 15 years of electrification households are assumed to have acquired all of the appliances that they require and energy consumption reaches a steady state with annualised mean load curves flattening out [41]. Important considerations in developing countries are thus the demand growth of newly electrified consumers and the more stable energy consumption behaviour of long term electrified customers. Energy consumption behaviour also shows regional variation that corresponds with climatic conditions and daylight hours.

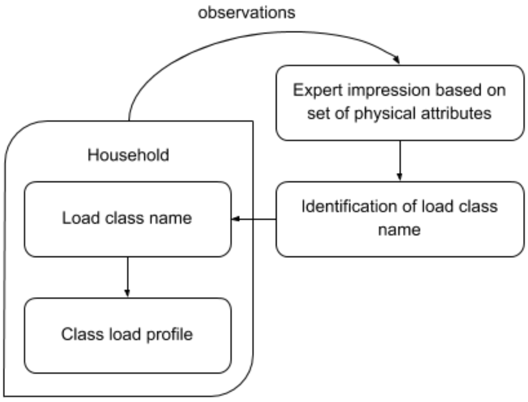

observations
Expert impression based on set of physical attributes
Household
Load class name
Identification of load class name
Class load profile

FIGURE 6.3: Current expert schema for creating customer classes [35]

We use the customer class attributes previously defined by experts to characterise our customer archetypes, so that we can benchmark our results against existing customer classes. Table 6.1 lists the attributes that we use to construct comparable customer archetypes.

Archetype	Water	Wall material	Floor area	Income
rural	river/dam	daub/mud/clay	0-50	R0-R1.8k
informal	street taps, tap in yard	corr.iron/zinc	0-50	R1.8-R3.2k
township	tap in house	asbestos, blocks, brick	50-80	R3.2k-R7.8k
lower middle	tap in house	asbestos, blocks, brick	80-150	R7.8k-R11.6k
upper middle	tap in house	brick	150-250	R19k-R24.5k

TABLE 6.1: Attributes of customer archetypes

6.1.4 Classification Implementation and Archetype Construction

Multinomial logistic regression (MNLR) is a classifier that can be used to predict categorical variables when more than two categories are present. We use WEKA's experimenter to build a classification model using MNLR to quantify the likelihood that a particular feature value characterises a cluster. We initiate the MNLR algorithm with conjugate gradient descent to run for up to 10 iterations. One of the outputs of the model is a table of odds ratios, which we save to Excel for further analysis.

The odds ratio represents the odds that cluster k_x will be assigned to a household with a particular attribute, compared with the odds that k_x will be assigned to a household that does not have that attribute. An odds ratio less than 1 indicates a negative association between k_x and the attribute, while a ratio greater than 1 indicates a positive association between the two. For example, *Cluster 39* has an odds ratio of 1.32 for feature *floor_area* with value *0-50*, which indicates that this cluster is more likely to be used for households with a floor area of 0 - 50m^2 than for those that have a different floor area.

To create customer archetypes, we set a threshold value of 1.05 for the odds ratio to restrict our selection to clusters and attributes with strong positive associations. Attributes are obtained from Table 6.1 and clusters are selected if those feature values that correspond to the attributes have an odds ratio greater than the threshold. The RDLPs of these clusters are gathered to construct the RDLPs of the customer archetype. The filtered odds ratio tables for several customer archetypes are included in Appendix B for illustration purposes. Green fields indicate an odds ratio greater than 1.05, while grey colours indicate an odds ratio below that. The darker the colour, the higher (or lower) the odds ratio.

6.2 Expert Benchmark

The benchmark is a set of RDLPs for customer classes (archetypes) developed by experts. It is derived from the models contained in the Distribution Pre-Electrification Tool (DPET). The customer classes in the DPET represent the average consumption behaviour of a group of similar households in the same location [41], under the same loading condition. A customer class is characterised by exactly six RDLPs that specify mean hourly energy demand over a 24 hour period for a workday, Saturday and Sunday day types in summer and winter seasons. Thus, if a group of residential energy consumers contains some households with a 7am morning peak and others with a 6pm evening peak on winter weekdays, both these profiles will be combined in a single winter weekday RDLP that has a 7am morning and a 6pm evening peak.

Data for the benchmark is retrieved for a specific location for every customer class. Hourly RDLP values are obtained with a lookup using the DPET Software released in 2013 with 2016 income values and all default settings. Table 6.2 lists the DPET parameters that were provided as input to the software to obtain load data for each archetype used in the benchmark. These attributes have been adopted from Table 6 in the Geo-based Load Forecast (GLF) Standard [35].

Profile selection in the DPET is done based solely on the mean income of a community of households, not by class. Classes must thus be derived from income information. This limits the information that can be retrieved. For example, customer classes within the same income bracket are indistinguishable from each other. The version of the software that we obtained does not provide data for high income households and estates, with a monthly income above R20 000.

Archetype	Municipality (Province)	Mean income	Electrified
rural	Ehlanzeni District (MP)	R1000	0 - 5 yrs
informal settlement	Capricorn District (LIM)	R2000	5 - 10 yrs
informal settlement	Ehlanzeni District (MP)	R2000	0 - 5 yrs
informal settlement	Nelson Mandela Metro (EC)	R2000	0 - 5 yrs
township	Johannesburg Metro (GP)	R5500	12 - 15 yrs
lower middle	eThekwini Metro (KZN)	R10000	6 - 7 yrs
upper middle	eThekwini Metro (KZN)	R15500	12 - 15 yrs
upper middle	City of Cape Town Metro (WC)	R15500	10 - 15 yrs

TABLE 6.2: Parameters used for selecting load data from the DPET software to construct archetypes

6.3 Evaluation of New Customer Archetypes

The new customer archetypes are evaluated in an intrinsic and an external manner. An archetype requires a RDLP for each day type in each season, so that it can be used to represent a household's energy demand throughout the year. The intrinsic evaluation examines the temporal coverage, seasonal and day type exclusivity of the archetype. The external evaluation compares the new archetype's RDLP shapes, peak times and energy demand for specific loading conditions against those of the benchmark archetype.

Five scenarios have been defined to qualify the extent of temporal coverage of an archetype, as this limits the representative strength of individual archetypes. The scenarios and sample customer archetypes to which they apply are described and listed in Table 6.4. Their strengths and limitations are discussed in the following sections. Some additional archetypes are included in Appendix C.

EC (Eastern Cape)	KZN (KwaZulu Natal)	NC (Northern Cape)
FS (Free State)	LIM (Limpopo)	NW (North West)
GP (Gauteng)	MP (Mpumalanga)	WC (Western Cape)

TABLE 6.3: Provincial abbrevations

6.3.1 Archetype with Full Temporal Coverage

Description

Figure 6.4a depicts the RDLPs of an archetype for lower middle class customers that have been electrified for over 15 years in the KwaZulu Natal province in South Africa. This archetype has piped water access to the house. The dwellings have a floor area between $80m^2$ and $150m^2$ with walls constructed from asbestos, blocks or bricks. Households earn between R7 800 and R11 600 per month. Seven clusters showed a strong correlation with this archetype. Table 6.5 characterises the clusters

Scenario	**Description**	**Archetype**
Full coverage	at least one RDLP for each loading condition	lower middle class, long-term electrified, KZN
Partial coverage	one weekend RDLP missing	informal settlement, newly electrified, MP
Scattered coverage	one weekday RDLP or more than two weekend RDLPs missing	informal settlement, medium-term electrified, LIM
Abundant coverage	more than 3 RDLPs for more than 2 loading conditions	rural, newly electrified, MP
No coverage	no RDLPs	

TABLE 6.4: Evaluation scenarios for new customer archetypes

by day type and season. The odds ratios are shown in Appendix B.1. As a whole, the RDLPs of this archetype were found to be reasonable in relation to expected customer behaviour.

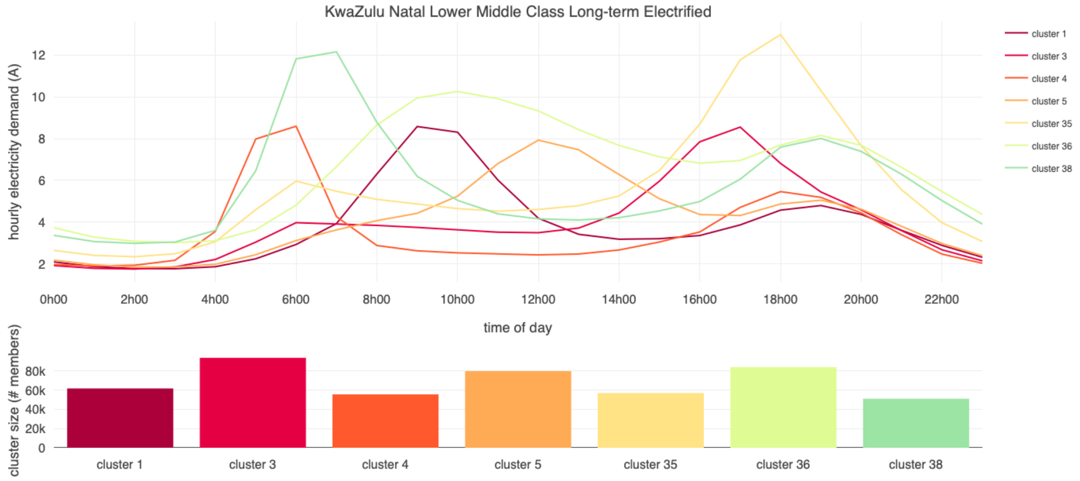
Winter		**Summer**	
Cluster	**Daytype**	**Cluster**	**Daytype**
3	weekday	**1**	Saturday, Sunday
35	weekday	**4**	weekday, Friday
36	Saturday, Sunday	**5**	Saturday, Sunday
38	weekday, Friday		

TABLE 6.5: Temporal attributes of clusters for archetype in Fig 6.4a

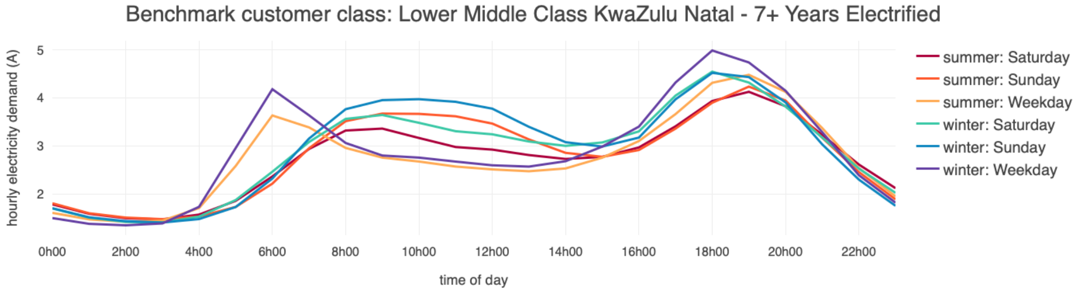
Interpretation and Discussion

Each day type in each season is represented by at least one cluster. Full temporal coverage like this is desirable. For this archetype both work day and weekend clusters, and winter and summer clusters, are mutually exclusive. There are 3 winter work day clusters (*Cluster 3, 35, 38*), one summer work day cluster (*Cluster 4*), 1 winter weekend cluster (*Cluster 36*) and two summer weekend clusters (*Cluster 1, 5*).

All work day clusters resemble a typical 'out of home' shape, with either a high morning or evening peak and lower consumption throughout the day. This is expected for a lower middle class household, where adults are typically blue collar workers that have a fixed work routine. *Clusters 1, 5* and *36* show a strong correlation with weekends. *Cluster 1* and *36* are indicative of a slow starting day when there is no job to rush to. *Cluster 5* with its peak at 12pm is typical for families that have a strong tradition of a shared family lunch on weekends. The shapes of the benchmark RDLPs in Figure 6.4b have both a morning and an evening peak, rather than a single distinct peak like the new archetype. As the benchmark represents the aggregate consumption of a group of households, the shapes of the new archetype would more



(A) RDLPs for Customer Archetype



(B) Benchmark

FIGURE 6.4: Medium-term electrified lower middle class household in KwaZulu Natal

closely resemble the benchmark if we were to aggregate the RDLPs for common loading conditions, such as the three winter work day RDLPs.

KwaZulu Natal lies in the East of South Africa, and subsequently has an earlier sunrise and sunset than most other parts of the country. Work day morning peaks are between 5am and 7am, and evening peaks between 5pm and 7pm. The summer work day *Cluster 4* has an earlier morning peak than the winter weekday clusters. Winter weekday *Cluster 3* and *35*, as well as summer work day *Cluster 4* with its early morning peak, show an earlier evening peak. The peak times of the benchmark RDLPs correspond approximately to those of the new customer archetype.

Except for *Cluster 3* which has a similar maximum demand to the summer clusters, the winter clusters show a higher energy demand throughout the day. The peak demand values of the new archetype are twice the value of the benchmark for most RDLPs. As with the cluster shape, this effect is likely to smoothen out when aggregating across a group of households, rather than considering individual RDLPs.

Limitations

The benchmark model only extends to 7 years after electrification for this archetype. We thus selected a medium-term electrified customer from the benchmark, even though the clusters are unlikely to be used for households electrified for this length of time (see Section 6.3.5). The benchmark is thus not an entirely accurate comparison for our new archetype, which may account for some of the difference between the RDLPs.

6.3.2 Archetype with Partial Temporal Coverage

Description

Figure 6.5b depicts the RDLPs of an archetype for newly electrified informal settlement customers in Mpumalanga. This archetype gets water from street taps or a tap in the yard. The floor area of dwellings is less than $50m^2$. Walls are constructed from corrugated iron or zinc and the household income is between R1 800 and R3 200 per month. Six clusters showed a strong correlation with this archetype. Table 6.6 characterises the clusters by day type and season. The odds ratios for the archetype are shown in Appendix B.2. While this archetype lacks some detail, as a whole the RDLPs are reasonable in relation to expected customer behaviour.

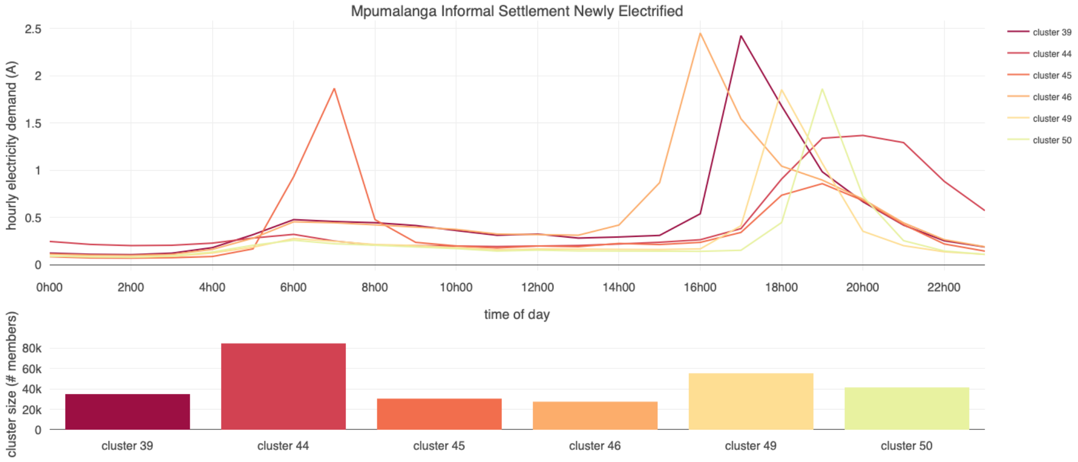
Winter		Summer	
Cluster	**Daytype**	**Cluster**	**Daytype**
39	weekday	**44**	Friday, Saturday
45	weekday, Friday	**50**	any
46	weekday		
49	Saturday, Sunday		

TABLE 6.6: Temporal attributes of clusters for archetype in Fig 6.5a

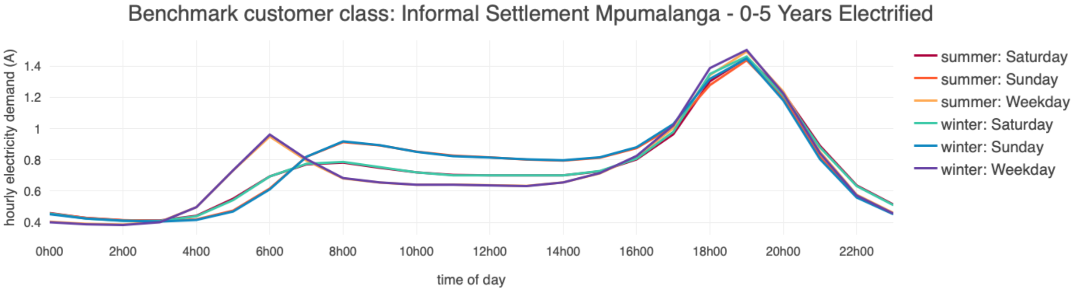
Interpretation and Discussion

With the exception of summer Sundays, all loading conditions are represented by at least one cluster. This gives the archetype partial temporal coverage. Winter and summer clusters are mutually exclusive. Distinct RDLPs cover work days and weekends in winter, but not in summer. All day type odds ratios of *Cluster 50* lie below the threshold value of 1.05, and the likelihood of use is similar for all day types in summer.

Considering the informal context and the low energy requirements in the hot summer period (this type of customer is highly unlikely to own an air cooler), it is possible that this archetype has a similar consumption pattern on all summer days. All cluster shapes indicate limited day time energy consumption. *Cluster 44* which is used on



(A) RDLPs for Customer Archetype



(B) Benchmark

FIGURE 6.5: Newly electrified informal settlement household in Mpumalanga

Fridays and Saturdays in summer has a late, long peak that may be indicative of social activities and entertainment.

The shapes of the benchmark RDLPs in Figure 6.6b are very similar for summer and winter seasons. The most distinguishing features of the benchmark RDLPs are the peak times. The morning peak for weekday profiles is at 6am, while the evening peak is consistently at 7pm. This is different to the peak times of the new archetype, which are dispersed between 3pm and 8pm.

Mpumalanga is a province with mild winter climate. Nonetheless there is a difference in day time and peak demand between the summer clusters *Cluster 44* and *50* and the winter clusters. The peak demand values of the new archetype are almost twice the value of the benchmark for most RDLPs, while the day time base load is only half. As with the previous archetype, these effects are likely to smoothen out when aggregating across a group of households.

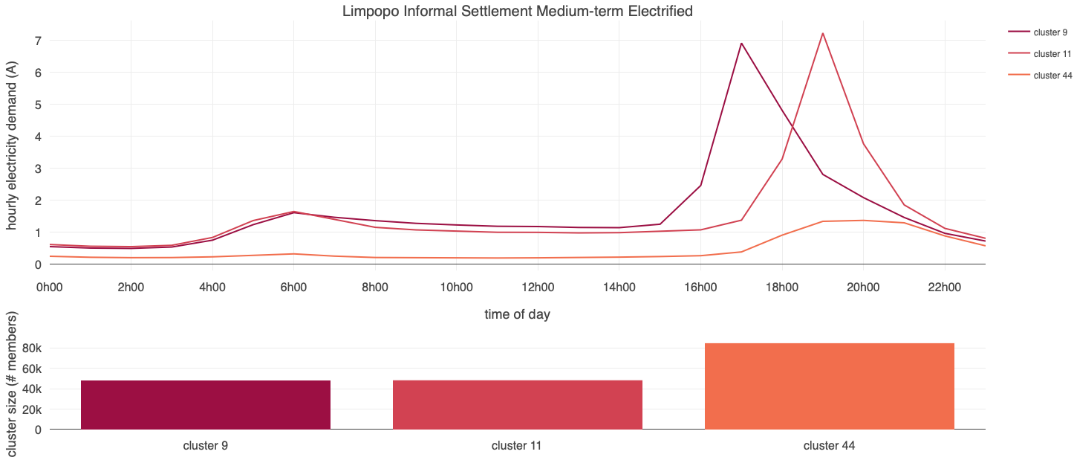
6.3.3 Archetype with Scattered Temporal Coverage

Description

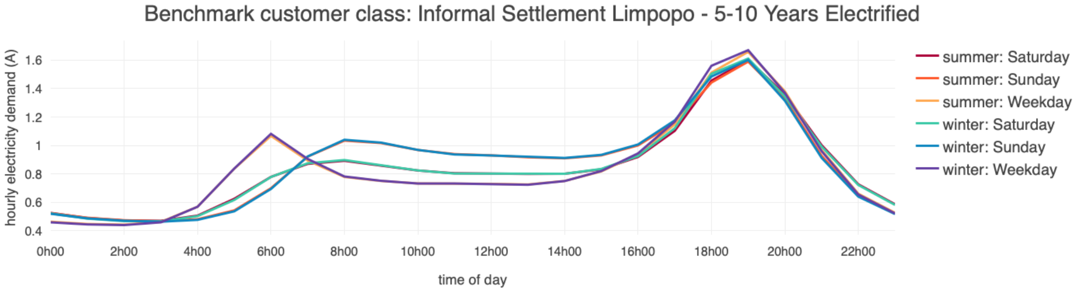
Figure 6.6a depicts the RDLPs of an archetype for medium-term electrified informal settlement customers in Limpopo. This archetype gets water from street taps or a tap in the yard. The floor area of dwellings is less than $50m^2$. Walls are constructed from corrugated iron or zinc and the household income is between R1 800 and R3 200 per month. Three clusters showed a strong correlation with this archetype. Table 6.7 characterises the clusters by day type and season. The odds ratios for the archetype are shown in Appendix B.2.

Winter		**Summer**	
Cluster	**Daytype**	**Cluster**	**Daytype**
9	weekday	**11**	weekday, Friday
		44	Friday, Saturday

TABLE 6.7: Temporal attributes of clusters for archetype in Fig 6.6a



(A) RDLPs for Customer Archetype



(B) Benchmark

FIGURE 6.6: Medium-term electrified informal settlement household in Limpopo

Interpretation and Discussion

Only three loading conditions are captured in this archetype. Considering that the summer and winter benchmark profiles are almost the same for each day type, this could be a reasonable set of RDLPs. However, the clusters do not indicate shared use across seasons and the peak demand of *Cluster 9* and *11* is more than four times that of the benchmark, which is surprising. The temporal coverage of this archetype is too scattered to make it unusable for practical applications.

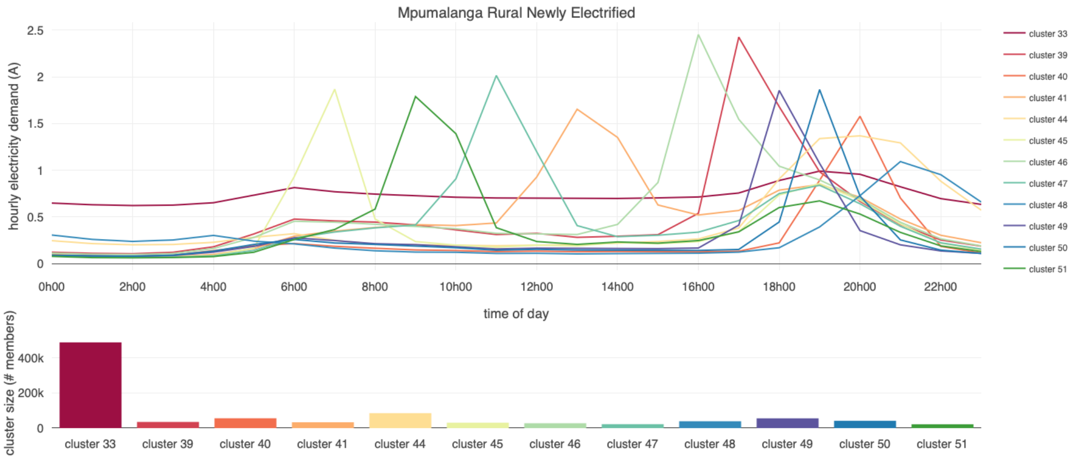
6.3.4 Archetype with Abundant Temporal Coverage

Description

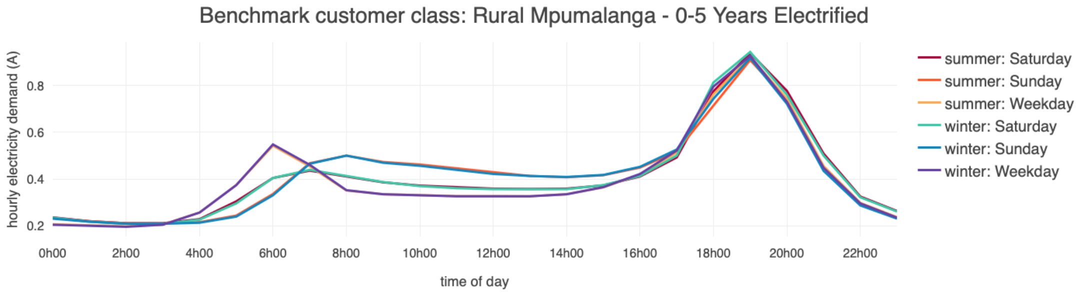
Figure 6.7b depicts the RDLPs of an archetype for newly electrified rural customers in Mpumalanga. This archetype does not have water access to the home and is highly likely to obtain water from a nearby river or dam. The floor area of dwellings is less than 50m^2. Traditionally walls are constructed from daub, mud or clay. Household may live a subsistence lifestyle, or have an income up to R1 800.

Winter		**Summer**	
Cluster	**Daytype**	**Cluster**	**Daytype**
39	weekday	**33**	Friday, Saturday, Sunday
45	weekday, Friday	**40**	weekday, Friday, Sunday
46	weekday	**41**	Saturday, Sunday
49	Saturday, Sunday	**44**	Friday, Saturday
		47	Saturday, Sunday
		48	Friday, Saturday
		50	any
		51	Saturday, Sunday

TABLE 6.8: Temporal attributes of clusters for archetype in Fig 6.7a



(A) RDLPs for Customer Archetype



(B) Benchmark

FIGURE 6.7: Newly electrified rural household in Mpumalanga

Twelve clusters showed a strong correlation with this archetype. Table 6.8 characterises the clusters by day type and season. The odds ratios for the archetype are shown in Appendix B.4. While it may be challenging to apply this archetype due to the large number of RDLPs, it nonetheless represents customer behaviour in line with expectations.

Interpretation and Discussion

The clusters cover all seasons and day types, with an abundance of Friday, Saturday and Sunday clusters. Out of the 8 summer clusters, 7 are used on Saturdays and 6 on Sundays. While the winter clusters are mutually exclusive across work days and weekends, this is not true for summer clusters.

The shapes of the benchmark RDLPs create the impression that the behaviour of this archetype follows a structured and routine pattern. The new archetype, on the other hand, provides useful insights that the behaviour of this customer class may not be that predictable. Rural customers are not bound by urban constraints such as traffic and office hours, allowing for a greater distribution of peak times. Despite the variety of shapes, it is important to consider that *Cluster 33* with its large membership may account for most summer Friday, Saturday and Sunday profiles.

For very low consuming households single appliance usage, e.g. boiling water or ironing, can occur sporadically and account for the entire day's peak demand. *Cluster 45* is the only weekday clusters with a morning peak. All other weekday clusters peak between 4pm and 8pm. On weekends peaks occur much more sporadically.

The peak demand of most RDLPs for the new archetype is similar and lies between 1.5A and 2A. The demand range of *Cluster 33* more closely resembles that of the benchmark.

6.3.5 Limitations of New Customer Archetypes

The NRS Load Research data collection process was designed to target specific customer archetypes that have been electrified for a particular length of time in regions of interest. The data collected thus does not cover all archetypes. Archetypes for which no data was collected cannot be constructed, as the odds ratio of clusters lies below the threshold value for their attributes. Table 6.9 lists the provincial coverage and duration of electrification of archetypes that have at least 3 clusters with an odds ratio greater than the threshold. Customer archetypes outside of this provincial coverage and electrification duration cannot be constructed and are missing.

Archetype	Provincial coverage	Electrified for
rural	EC, FS, LIM, MP, NC, NW	0 - 10 years
informal	EC, FS, LIM, MP, NC, NW, WC	0 - 10 years
township	EC, FS, GP, KZN, LIM, NC, NW, WC	over 10 years
lower middle	GP, KZN, NW, WC	over 10 years
upper middle	GP, KZN, NW, WC	over 10 years

TABLE 6.9: Coverage limitations of new customer archetypes

6.4 Summary

This chapter details the construction of customer archetypes from RDLPs. It starts with an overview of the process and the features included in the classification input data. Next, the attributes of customer archetypes as defined by experts, are described. The chapter further details the implementation of a multinomial logistic regression model for classifying the clusters and continues to present the current expert archetypes that are used as benchmark. Finally new customer archetypes are presented and discussed for five different temporal coverage scenarios.

Chapter 7

Discussion

The overarching goal of this study was to select the optimal set of clusters from South Africa's Domestic Electrical Load Study (DELS) dataset in order to create residential energy customer archetypes. This chapter summarises and reflects on the successes and limitations of the approach taken and the results produced. The discussion focuses on the dataset, the effect of normalisation, pre-binning and clustering techniques, the qualitative evaluation framework that was developed and the application of the clusters to generating representative daily load profiles and customer archetypes.

7.1 Dataset

This section discusses characteristics and quality concerns of the energy consumption and survey data that may have influenced the performance of clustering techniques, cluster classification and the archetypes that were created.

7.1.1 Overview

The DELS dataset presents the most comprehensive collection of household energy consumption data in South Africa. The dataset consists of metered voltage, power and current readings observed at 5 minute intervals and survey data collected annually. Over the 20 year period between 1994 and 2014, 14 945 households were metered for approximately one year. This study used only the energy demand (current) observations. The dataset has been described in detail in Section 3.1.

Daily load profiles that contained any invalid readings were removed. Thus, if a daily load profile contained one invalid 5 minute reading, the entire profile was discarded. As the analysis was primarily focused on clustering daily load profiles, their individual integrity was deemed important, but complete coverage of single households was not. Households were observed for different lengths of time, with 61% of households having valid readings for more than half a year. Some households may thus have contributed a single valid profile to the input data, while others contributed 365 valid profiles.

The data input to the clustering techniques contained 3 295 848 daily load profiles. As the dataset had been used previously by experts to construct customer archetypes, it was expected that the profiles have a high tendency to form clusters. Seasonal patterns, daytype related patterns and diurnal patterns are typical for residential energy consumers. Due to the large geographic area that South Africa covers, a phase shift of approximately 1 hour between households located in the East and the West of the country was also expected.

7.1.2 Energy Consumption Data Quality

In contrast to most other studies in the domain, daily load profiles were not preselected based on loading condition. Consequently, the input data had a high degree of variability across temporal and geographic dimensions. Moreover, the population that was observed ranged from rural, low income households to very affluent urban households. The highest consuming households used 20 to 40 times as much energy per day on average as the lowest consuming households. A part of the research objective was to evaluate the effect of pre-binning on datasets with high variance. All apparent outliers in the dataset were thus retained in an attempt to gain insights into the consumption behaviour of very low consuming households, without discarding the less frequent high consuming households.

For lower income households and households with prepaid electricity meters it is possible to have days with no electricity consumption. Zero-profiles were thus not considered a data quality issue. As they tend to be removed in other studies, experiments were conducted both with and without zero-valued profiles to observe their effect on clustering results.

7.1.3 Survey Data Quality

Due to the rigorous historic data validation process no causal checks were applied to validate that the household survey data is correct. All customers in the dataset were assumed to be residential households. However, while extracting features for classification, it was observed that this was not always the case. Some households with more than 10 residents and more than 5 geysers emerged. It is likely that this type of customer represents a hostel or guesthouse rather than a residential household. Furthermore, the occupant attribute contained a ’0 adults’ value. While these could be child-headed households, it is more probable that both these examples present potential data quality problems. In the future it would be useful to remove households from the dataset that do not have the attributes of pre-defined customer archetypes before clustering.

An anticipated data quality challenged noted by experts was that high income households sometimes indicated their income as R0. The attributes of high and low income customers were however sufficiently distinct that this did not affect our ability to assign clusters to customer archetypes based on other attributes, such as water access and dwelling materials.

The household surveys were collected once a year, generally between June and August over the winter period. Metering on the other hand was conducted over approximately a 12 month period that frequently spanned over two calendar years. The survey year and the timestamps of metered observations were thus out of sync in some instances. No survey data was collected in 2007 and 2013.

7.2 Normalisation, Pre-binning and Clustering Analysis

To select a clustering structure that produces good RDLPs, this research compared the results of clustering and normalisation algorithms and evaluated the effect that pre-binning daily load profiles has on clustering results. This section summarises the most important findings related to the performance of clustering techniques, normalisation and pre-binning.

7.2.1 Overview

The clustering techniques implemented in previous research served as the starting point for the cluster analysis. However, the particular characteristics of the DELS dataset necessitated that clustering algorithms yield meaningful results even when feature ranges are large. The results show that normalisation and pre-binning were key to producing good clusters. The k-means algorithm undoubtedly performed better than the SOM and the multi-step technique that combines SOM with k-means. The full results have been described in Section 5.1.

7.2.2 Evaluating Cluster Compactness and Distinctness

A Combined Index (CI) was developed to effectively compare the results of 2083 experiments across several metrics. The CI was used as a relative index to avoid having to interpret individual scores and was useful for selecting the top 10 clustering structures.

The CI was calculated as the log of the weighted sum of Ix, the product of the DBI, MIA and inverse Silhouette Index, across all experiment bins. It is described in detail in Section 4.5.1. Choosing a logarithmic function was appropriate, as selecting the best experiments required the ability to discern between low Ix scores, while high scores were condensed to make the range of scores more manageable. The distribution of CI scores across all experiments has been visualised in Figure 5.1. The DBI, MIA and Silhouette Index, which were calculated from within cluster and between cluster distances, tended to favour clusters and bins with small membership. The same was also true for Ix, as demonstrated in Figure 5.9. For CI scores to take cluster membership into consideration, weighting Ix by cluster size was thus critical.

Good cluster compactness and distinctness are critical attributes of a good clustering structure. The challenge with cluster evaluation was not finding clusters with good (i.e. low) scores, but rather that using the CI scores alone was insufficient for selecting the best clustering structure with confidence. This confirms the conclusions drawn by previous studies.

7.2.3 Performance of Clustering Techniques

To compare the impact of different normalisation, pre-binning and algorithms, the distributions of CI scores have been visualised across techniques for all experiments in Figures 5.3 to 5.5. As expected, normalisation significantly impacted clustering results. There is a distinct difference in performance between experiments normalised with algorithms that transform daily load profiles to values between 0 and 1 (unit norm, de-minning and zero-one normalisation) and those that do not (SA norm and unnormalised experiments). Unit norm was the best normalisation for most experiments. SA norm performed the worst. This was no surprise, as the Euclidean distance measure and the error metrics would be severely impacted by the larger values that this normalisation permits.

While normalisation preserved only the profile shape, pre-binning retained some information about the magnitude of consumption. This proved to be important, considering the high variance in energy consumption in the dataset. As a whole, a greater number of experiments with pre-binning had good scores than experiments without pre-binning. The top performing experiment based on the CI score however was experiment 2 (k-means, unit norm), which had no pre-binning. This experiment was not only the top performer, but four out of the 10 top experiments in Figure 5.6 were variations of this experiment with different numbers of clusters.

AMC pre-binning performed better than pre-binning by integral k-means, with the former having four and the latter two experiments in the top 10 scores. The scores of the experiments pre-binned by AMC were also lower. Pre-binning by integral k-means produced 8 clusters to match the number of bins used with AMC. In future, it would be useful to compare a larger range of clusters.

Comparing the clustering algorithms, k-means outperformed the SOM and SOM+k-means techniques for almost all experiments. This is best observed in the distribution of CI scores by algorithm in Figure 5.5. As the dataset was large and high dimensional, with fixed time series length and regular sampling intervals, this result corresponds with the suggestions made in the cluster analysis literature [26] and with the results of previous studies as shown in Table 2.2. Some previous authors found the SOM a promising approach [19] [55]. In this research the square map initialised with the SOM may have resulted in a clustering structure too coarse to capture the variability in the dataset. SOM+k-means had the drawback of slow run times when the SOM dimension was high, as Figure 5.7a shows. Due to the poor results and slow run times of SOM and SOM+k-means they were not implemented for most of the experiments with pre-binning. The Euclidean distance measure was used in all algorithms. It would be interesting to see if algorithms that use DTW produce significantly different clusters.

While the CI scores provide an indication of the performance of clustering techniques, the top 10 scores lie very close together. From Figure 5.6 it is clear that unit norm and k-means produced the best clustering structures. The difference between the best and the tenth best scores was however only 3.1 percentage points. As a quantitative interpretation of the scores was not possible, their use was limited to serving as a relative index indicating which experiments warrant further evaluation.

7.2.4 Dataset Effects

Piecewise Aggregate Approximation was used to reduce the dataset from 5 minute intervals to hourly intervals. This reduced the dataset to a manageable size, removed noise and was in line with sampling resolution used in many previous studies. It would be interesting to explore whether a higher resolution of 15 minute intervals yields similar results.

Experiments pre-binned with average monthly consumption (AMC) did not take the number of profiles into consideration when calculating the average. Consequently, households may have been allocated to bins based on partially observed years, which can have the effect of under or over-estimating mean consumption. For example, the average monthly consumption of a household observed for 8 summer months is likely to be lower than if the same household was observed for an entire year. The higher consumption of the winter months would raise the average and the profiles of the household could have been allocated to a different bin.

7.3 Qualitative Evaluation Framework

The first objective of this research was to develop a qualitative evaluation framework to facilitate the selection of the clustering structure that is most suitable for a specific real world application in the long term energy forecasting domain. This section summarises and critically reflects on the qualitative evaluation metrics that were selected, defined and applied.

7.3.1 Overview

The qualitative evaluation framework was based on a set of competency questions developed from insights gained in expert interviews. The questions specify attributes of good clusters, which were then used to define qualitative evaluation measures. Two high level evaluation categories were identified, usability and expressivity. For each of these categories several measures were defined. All measures were weighted based on their perceived importance, and combined into a cluster scoring matrix. For each measure the cluster scoring matrix ranked experiments according to their scores. A total score was then calculated for each experiment based on the sum of weighted ranks for all measures. The qualitative evaluation measures and cluster scoring matrix are described in detail in Sections 3.3 and 3.4 respectively.

While the results still required user validation, the cluster scoring matrix greatly assisted in selecting the best experiment by reducing the uncertainty inherent in identifying the best clustering structure from the CI scores alone. The results in Figure 5.8 and Table 5.1 were quite different to the ranking of experiments by CI scores in Figure 5.6. A summary of findings is presented below and in greater detail in Section 5.2.

7.3.2 Cluster Selection based on Qualitative Measures

The zero-one normalisation algorithm scored favourably during the quantitative evaluation for some top experiments, but the resultant number of clusters it produced were too few and too uniform to be sufficiently expressive. Consequently these experiments were penalised heavily during the qualitative evaluation process. The two experiments normalised with zero-one scored above 200 points (lower score is better). This was 50 points higher than the next worst experiment and almost four times the score of the best experiment. This is an interesting observation, as most existing studies have relied on zero-one normalisation without exploring alternatives.

Experiments with pre-binning and unit norm had much better overall scores than those without. There appears to be only a slight advantage for pre-binning by integral k-means versus pre-binning by AMC, which was unexpected. Integral k-means seemed like the more sophisticated pre-binning approach and it was anticipated that it would produce better results.

The two pre-binning approaches responded very differently to the inclusion of zero-profiles. For AMC pre-binning a special bin was included for very low demand between 0 and 1A. When zero profiles were included, this bin contained two clusters, one zero cluster and one very small cluster with 226 members that had a demand just above 0A. When zero profiles were removed, this bin contained 16 clusters, of which 15 had less than 10 members. The second structure was penalised heavily during evaluation. Pre-binning by integral k-means, on the other hand, scored a lot higher when zero profiles were removed, indicating that the clustering structure produced with integral k-means may have been severely affected by including zero profiles in the dataset.

The two best experiments were k-means pre-binned with integral k-means, normalised with unit norm, with zero profiles removed and k-means pre-binned with AMC, normalised with unit norm, including zero profiles. While pre-binning by integral k-means performed a bit better, the ranking of the experiments was very sensitive to the weighting of evaluation measures and the threshold value that excludes clusters with small membership. For threshold values between 9488 and 10005 the scores of the two experiments were equal. If the threshold was dropped below 9488, AMC pre-binning scored better. Ultimately the experiment pre-binned with integral k-means was selected, as the other clustering structure had several high consumption clusters that did not meet the threshold value. It was thus considered less useable to cover a range of customer archetypes.

7.3.3 Evaluating the Cluster Usability Measures

The cluster usability category consisted of two measures. A numeric value represented the fraction of clusters with membership size above a threshold, and a binary value indicated whether zero-profiles can be represented. The threshold value was set to 10490, which specifies a minimum of 5% of households using an individual cluster for a minimum of 14 days. The threshold impacted results on two levels. Firstly, the fraction of clusters with membership size above the threshold had importance as a stand-alone measure in the matrix. Secondly, the threshold also adjusted all other measures. Small clusters tended to have better scores due to the smaller sample size. Removing them was an additional important function of the threshold. The threshold that was set provided a reasonable approximation for minimum cluster membership. It should be retained as a measure in the cluster scoring matrix to satisfy competency question 5.

No score was assigned to the zero-profile representation, as its weight of 1 and maximum value of 1 made its impact on the total score negligible. Given the discussion in the previous section about the effect of zero-profiles on the different experiments, future applications should remove all zero-profiles from the input data to conform to standard data cleaning techniques. A special zero-cluster can then be assigned to daily load profiles with 0 value to reintroduce them into the clustering structure after the clustering. This will guarantee that competency question 4 is satisfied and the measure can be removed from the usability category.

7.3.4 Evaluating the Cluster Expressivity Measures

The cluster expressivity category had two groups of measures. Peak coincidence, peak and total consumption errors measure representativity, while temporal and demand entropy measure cluster homogeneity. The measures have been visualised for some experiments in Figure 5.15. None of the consumption error metrics are defined for zero profiles, which were excluded from the evaluation. This may have affected results. The consumption error score was calculated as the average cluster ranking by MAPE, MdAPE, the median log accuracy ratio and median symmetric accuracy. Figure 5.16 shows that most of the experiments ranked consistently across the metrics. We found the median symmetric accuracy to be the most interpretable of these metrics and would limit the consumption error to this metric in future.

Peaks were defined as all those values greater than half the maximum daily load profile value. Experiment 2 (k-means, unit norm) had considerably better peak overlap than the remaining experiments. This is an interesting indication that experiment 2 (k-means, unit norm), in which data was normalised but not pre-binned, may have been stronger at clustering by shape than by demand. The technique applied to calculate the peak coincidence ratio was a rough approximation. More nuanced techniques have been suggested in other studies [79] and should be implemented in the future. Nonetheless, the mean peak coincidence ratio provided a useful indication for the degree of overlap of maximum demand and should be retained as a measure.

Entropy proved to be a useful measure to evaluate both demand-related and temporal homogeneity. The clusters had the tendency to rank similarly across entropy measures. While this may be a specific characteristic of the dataset, consistent homogeneity improved our confidence in these clusters.

Seasonality was anticipated to influence clusters. In line with expert definitions, the four month period from May to August was categorised as winter. In practice, this is a gross approximation. South Africa has experienced considerable shifts in weather patterns over the past three decades. Moreover, the different climatic zones experience the onset of winter at different times and the cold months last for different durations. Cold winter weather is also largely influenced by cold fronts. Rather than considering monthly entropy, it would be beneficial to consider temperature effects and daylight hours in the future.

7.4 Customer Archetypes

The final objective of this research was to use the clusters and their associated RDLPs to construct customer archetypes that can be used for applications in long term energy planning. This section provides an overview of the resultant archetypes and reviews their attributes. The features that were extracted are discussed and the strengths and limitations of the classification technique that was used are described.

7.4.1 Overview

The DELS dataset contained socio-demographic survey data covering 180 features for 58% of metered households. This survey data was used to classify the clusters of Experiment 8 (k-means, unit norm) based on socio-demographic characteristics that experts consider to be significant contributing factors to electricity consumption. The RDLPs of the classified clusters were then used to construct customer archetypes with similar attributes to existing benchmark archetypes developed by experts. The method for constructing customer archetypes from the clusters is described in Section 6.1.

The RDLPs of the customer archetypes were developed, analysed and compared against expert benchmarks. The evaluation examined the temporal coverage, seasonal and daytype exclusivity, RDLP shapes and energy demand of the new archetypes. Based on the extent of temporal coverage, five scenarios were defined for full, partial, scattered, abundant and no temporal coverage (see Table 6.4). For each scenario a sample archetype is presented in Section 6.3. Some general observations are discussed below.

The archetypes with full and partial temporal coverage also had good seasonal and daytype exclusivity. The shapes, peak times and energy demand of the individual RDLPs aligned with expectations. The archetype with abundant temporal coverage was more challenging to interpret. Even though clusters were exclusive to seasons, only some were exclusive to work day or weekend daytypes. Nonetheless, this may be indicative of a behavioural attribute of rural customers, which are likely to be less constrained by routines typical to urban environments. In the scenario of scattered temporal coverage, interpreting the archetype was challenging as too many RDLPs were missing. This may have been the results of certain feature design choices, which are discussed in 7.4.3. For all archetypes discrepancies existed between them and the benchmark. Possible explanations for the differences are discussed in Section 7.4.4.

Some archetypes that experts have constructed in the past could not be recreated, as the sample design of the study constrained the data in the dataset. The data-imposed limitations around geographic coverage and time since electrification are captured in Table 6.9. Archetypes outside these bounds had no temporal coverage and could not be represented.

As a whole, clusters were mostly specific to particular loading conditions, socio-demographic features, dwelling features and geographic contexts. The RDLPs of the archetypes exhibited East-West effects, daytype effects, seasonal effects, and discerned between high and low energy consumers, thus satisfying the first three competency questions posed in Section 3.2.1. While several previous studies indicated the possibility of creating detailed customer archetypes by characterising representative daily load profiles with socio-demographic customer features, only few have implemented this. To our knowledge this work presents the first comprehensive attempt to generate and evaluate RDLPs that can be used to create customer archetypes on a national scale in a developing country.

7.4.2 Attributes of Customer Archetypes

The attributes of customer archetypes were selected based on discriminating characteristics defined by experts (see Table 6.1). The four attributes used in this research were water access, the dwelling's wall material, the floor area of the dwelling and the household income. These attributes uniquely described the customer archetypes. They are however static and have not been revised in several years to reflect changing household attributes. For example, it is known that many rural areas develop when young people move to the cities, find work, earn money, and return home to renovate the houses of their parents. Increasingly rural dwellings thus no longer rely on traditional building materials like mud and clay. The attributes of the rural customer archetype however have not been updated to reflect this change.

It is likely that the two decades over which the dataset spans contain such socio-demographic shifts. The attributes may thus be more relevant for early years in the dataset than for the most recent years. Classifying the clusters year-on-year rather than characterising all 20 years simultaneously could provide interesting insights into how household consumption and customer attributes have changed over time.

7.4.3 Feature Extraction

The high level feature categories comprised temporal, spatial, occupant, economic, appliance, dwelling and connection related features. In total 16 socio-demographic features, a cluster and a weight were included in the classification input data. The features and range values were selected specifically to correspond with attribute ranges used by experts, so that the characterised clusters could be used to create comparable archetypes based on current expert definitions. All features have been listed in the Appendix in Figure A.1.

Feature extraction posed several challenges, some of which have been discussed as data quality challenges in Section 7.1.2. Using province as a proxy for the climatic region provided a more granular geographic feature, but may be one of the reasons why some customer archetypes had insufficient RDLPs assigned to them. Based on early discussions with experts, Friday was treated as a daytype of its own, separate from other work days. This granularity was not available in current expert archetypes. While it may provide interesting insights for experts, it hindered the comparison against the benchmark. Public holidays were included in the analysis. No obvious negative effects were observed, however, this may explain why many Saturday clusters were also frequently used on Fridays. It is advisable to take this into account and remove public holidays in future studies.

7.4.4 Expert Benchmark

The quality of the customer archetypes was evaluated against an equivalent expert benchmark. The benchmark consists of location specific archetypes developed by experts. A benchmark archetype contained a total of six RDLPs, one RDLP for each loading condition. Data for the benchmark was retrieved with a lookup from the DPET software, as described in Section 6.2. Only minimal attributes could be selected in the benchmark model. The benchmark attributes thus did not provide an exact match for the attributes against which the new customer archetypes were developed. Table 6.2 lists the assumptions made in this study.

The expert benchmark RDLPs were developed to represent a community of similar households, rather than individual households. This posed several constraints on their ability to present an equivalent comparative RDLP to the new archetypes. As mentioned above, benchmark archetypes had exactly six RDLPs per archetype. The shapes of the RDLPs were considerably more uniform and smoother than the RDLPs generated through cluster analysis. Frequently a benchmark RDLP would have both a morning and an evening peak, whereas the new customer archetype would represent this consumption behaviour with two separate RDLPs. The effects of aggregation not only influenced the RDLP shapes, but also the energy demand. Typically, the peak demand in the benchmark RDLPs was half that of the new RDLPs under the same loading conditions.

Due to the maximum monthly household income being limited to R20 000 in the DPET software, no benchmark was extracted for certain archetypes, like very high earning households in low-density estates. The RDLPs that were generated have been used to construct these archetypes, but they could not be evaluated.

7.4.5 Classification Technique

Multinomial Logistic Regression (MNLR) was used to classify clusters based on the socio-demographic attributes of households that use them. The technique was applied in previous studies and was selected as the resultant odds ratios could provide a probabilistic view of which socio-demographic features are most strongly associated with individual clusters.

The odds ratio tables have been included in Appendix B. While they were useful for constructing customer archetypes, their interpretation was challenging and ambiguous. Clusters were assigned to customer archetypes if the odds ratio of features that are attributes of that archetype exceeded a threshold of 1.05. The selection of the threshold was motivated by a 5% increase in the likelihood of assignment being considered significant. Frequently clusters had odds ratios greater than 1.05 for a number of features of neighbouring customer archetypes. This introduced two complexities. On the one hand it made the archetype construction process uncertain, especially when too many or too few clusters were associated with that archetype. Secondly, it made ′rare′ attributes that were only used very infrequently appear to have a high likelihood of occurring when in reality the likelihood may have been skewed by their small sample size. One approach to overcoming this challenge in future would be to weight the odds ratios by their frequency count. This could provide both an indication of the likelihood of association between a cluster and an attribute, and of the likelihood that households have that attribute.

Overall, the archetypes were promising and mostly displayed energy consumption behaviour in line with expectations. Their creation and evaluation were knowledge and time intensive. To aid evaluation, it would be useful if figures like Figure 6.4a visualised the frequency that clusters are used within an archetype, rather than the entire dataset. Alternative classification techniques should be considered in the future. Bayesian networks may be ideally suited for this task, offering a graphical representation of relations between attributes and customer archetypes, and an intuitive interpretation of uncertainty.

Chapter 8

Conclusion

This work compares and evaluates clustering techniques for generating representative daily load profiles (RDLPs) that are characteristic of residential energy consumers in South Africa. Cluster evaluation and selection was guided by a qualitative cluster scoring matrix, which ranked clustering structures with good compactness and distinctness by their usability and expressivity. This approach greatly aided the typically challenging cluster selection process and ensured that the RDLPs are suitable for application in a real-world, long-term energy planning scenario.

Different algorithms, normalisation and pre-binning techniques were evaluated to determine the best clustering structure. Unsurprisingly, k-means performed better on the large dataset than self-organising maps (SOM) and a multi-step algorithm combining SOM and k-means. The traditional clustering metrics indicated that pre-binning and normalisation to a range between zero and one generally produce better clustering structures. However, the scores of top experiments were so close together, that selecting an experiment based on the relative position alone was not justifiable. Previous studies observed similar challenges and typically relied on expert validation to select the best set of clusters.

The qualitative framework that was developed presented a promising alternative to successfully evaluate clustering structures against competency questions developed with experts. The qualitative cluster scoring matrix clearly indicated that unit norm and pre-binning produces the most expressive and usable clusters. While the best experiment was pre-binned with integral k-means, both pre-binning approaches produced comparable scores that are primarily influenced by the weights assigned to different evaluation measures and the threshold determining the minimum cluster membership.

The RDLPs that were generated from the best clusters were used to construct customer archetypes that represent a wide variety of households spanning rural, informal and wealthy urban areas across five climatic zones and two timezones. While there are discrepancies between benchmark archetypes developed by experts and the archetypes generated from the RDLPs, these are in part due to the limitations and constraints of the dataset and the benchmark. The next section suggests methodological improvements to create better archetypes in the future.

To our knowledge this is the first work that applies state of the art cluster analysis techniques to the residential energy domain in a developing country context. While the analysis is limited to the electricity sector, similar approaches may be promising in other residential utility domains, such as the water sector.

8.1 Future Work

In the course of this research several improvements, alternative approaches and opportunities for further investigation were identified. While time did not permit their exploration, they are captured below to indicate future areas of research.

Data cleaning and quality control Minimal data cleaning and quality control was done in this research. For future research concerned with constructing customer archetypes from the DELS dataset it is suggested that the daily load profiles of households that do not have attributes of pre-defined customer archetypes are removed from the dataset. It is further suggested that public holidays are removed.

Data representation The data in this research was represented as standard daily hourly load profiles. Investigating the effect of clustering different time resolutions, in particular at *15 minute and 30 minute intervals*, would be a useful adaption in a subsequent study.

Clustering techniques More rigorous analysis of pre-binning techniques, including integral k-means with different numbers of clusters, is warranted. While the type of dataset is well suited to clustering with k-means, alternative partitional clustering algorithms such as *k-medoids* should be explored, as well as algorithms that use *Dynamic Time Warping*.

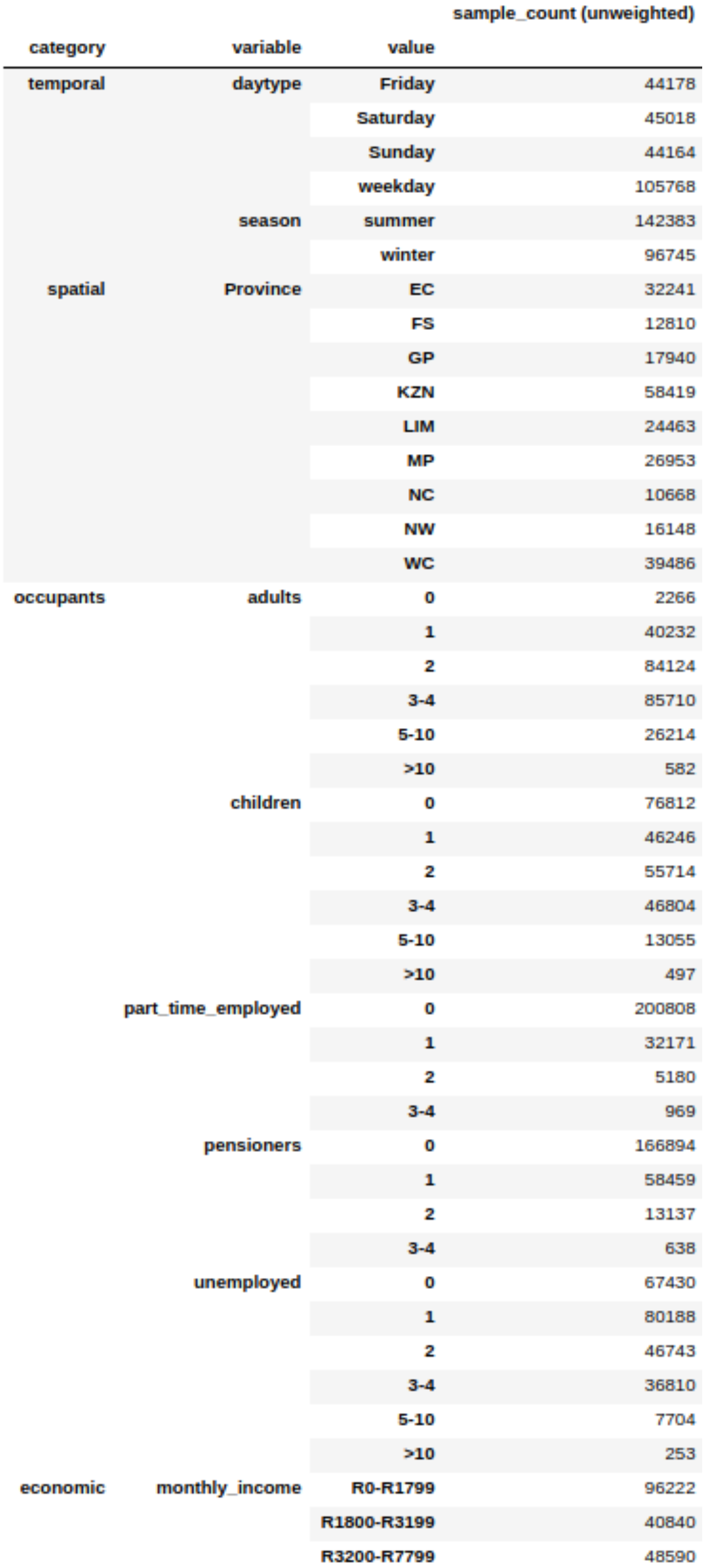
Qualitative evaluation The qualitative evaluation framework and cluster scoring matrix are a promising approach to improve cluster selection and would benefit from further development. More nuanced approaches for calculating the peak coincidence ratio have been developed in other studies and should be incorporated. Homogeneity measures should be extended to include temperature and daylight effects in addition to seasonal effects.

Customer archetypes Further research is required to improve both the process of creating customer archetypes and the archetypes. Future work should compare different classification algorithms for characterising the cluster dictionary, and formalise its representation with semantics. Bayesian networks may be well suited for the task of creating customer archetypes from the clusters. Year-on-year should be considered to capture changing archetype attributes. For multinomial logistic regression, odds-ratios could be weighted by frequency count to indicate the overall likelihood of cluster use.

Alternative applications of RDLPs Finally, the RDLPs present an opportunity for deeper analysis of long term changes and short term volatility of customer behaviour. [51] presents an approach to how RDLPs could be used to analyse customer variability.

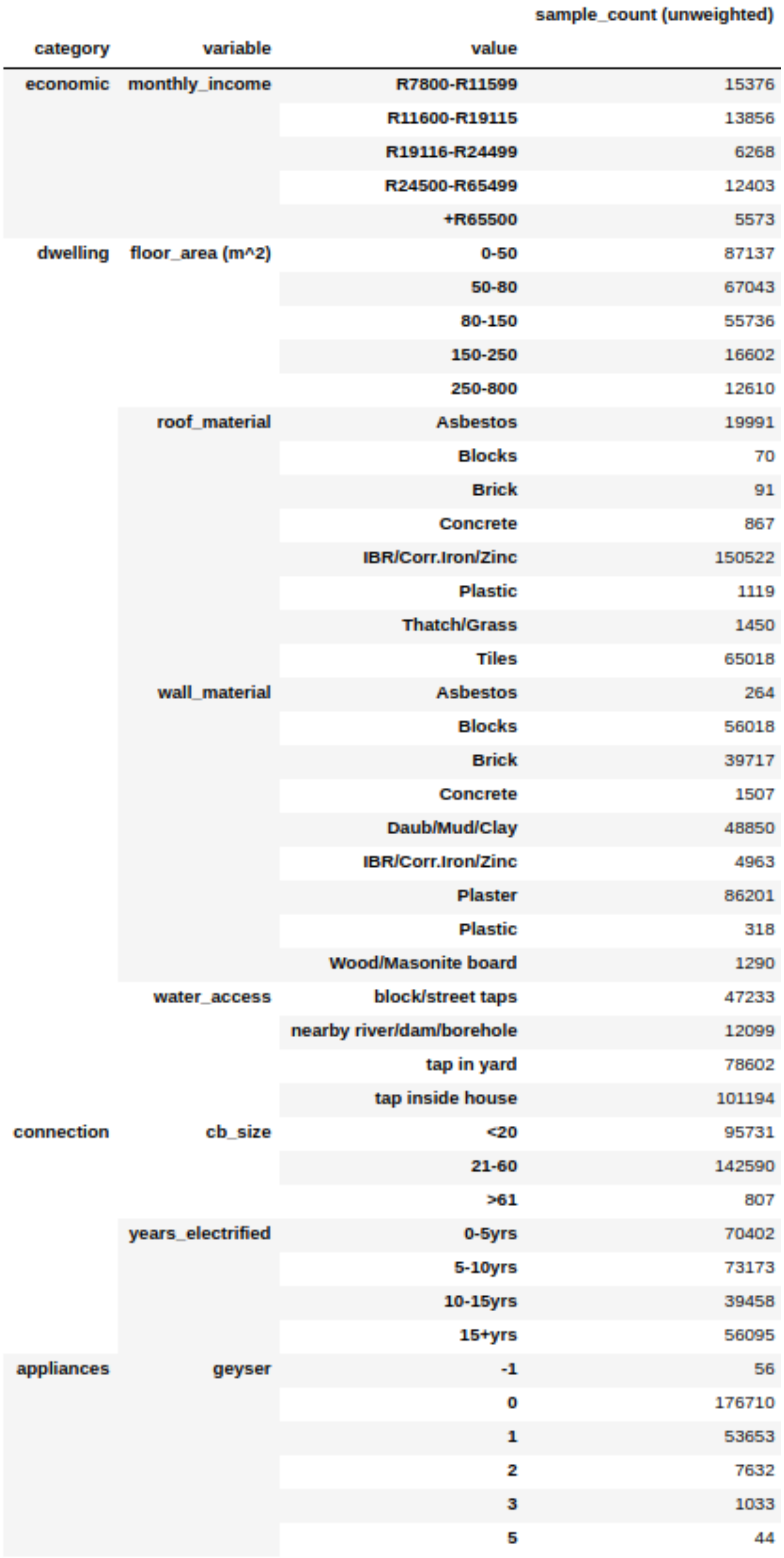
Appendix A

Features for Cluster Classification



category	variable	value	sample_count (unweighted)
temporal	daytype	Friday	44178
		Saturday	45018
		Sunday	44164
		weekday	105768
	season	summer	142383
		winter	96745
spatial	Province	EC	32241
		FS	12810
		GP	17940
		KZN	58419
		LIM	24463
		MP	26953
		NC	10668
		NW	16148
		WC	39486
occupants	adults	0	2266
		1	40232
		2	84124
		3-4	85710
		5-10	26214
		>10	582
	children	0	76812
		1	46246
		2	55714
		3-4	46804
		5-10	13055
		>10	497
	part_time_employed	0	200808
		1	32171
		2	5180
		3-4	969
	pensioners	0	166894
		1	58459
		2	13137
		3-4	638
	unemployed	0	67430
		1	80188
		2	46743
		3-4	36810
		5-10	7704
		>10	253
economic	monthly_income	R0-R1799	96222
		R1800-R3199	40840
		R3200-R7799	48590

FIGURE A.1: Table of features, their values and the count of daily load profiles per value prior to weighting instances



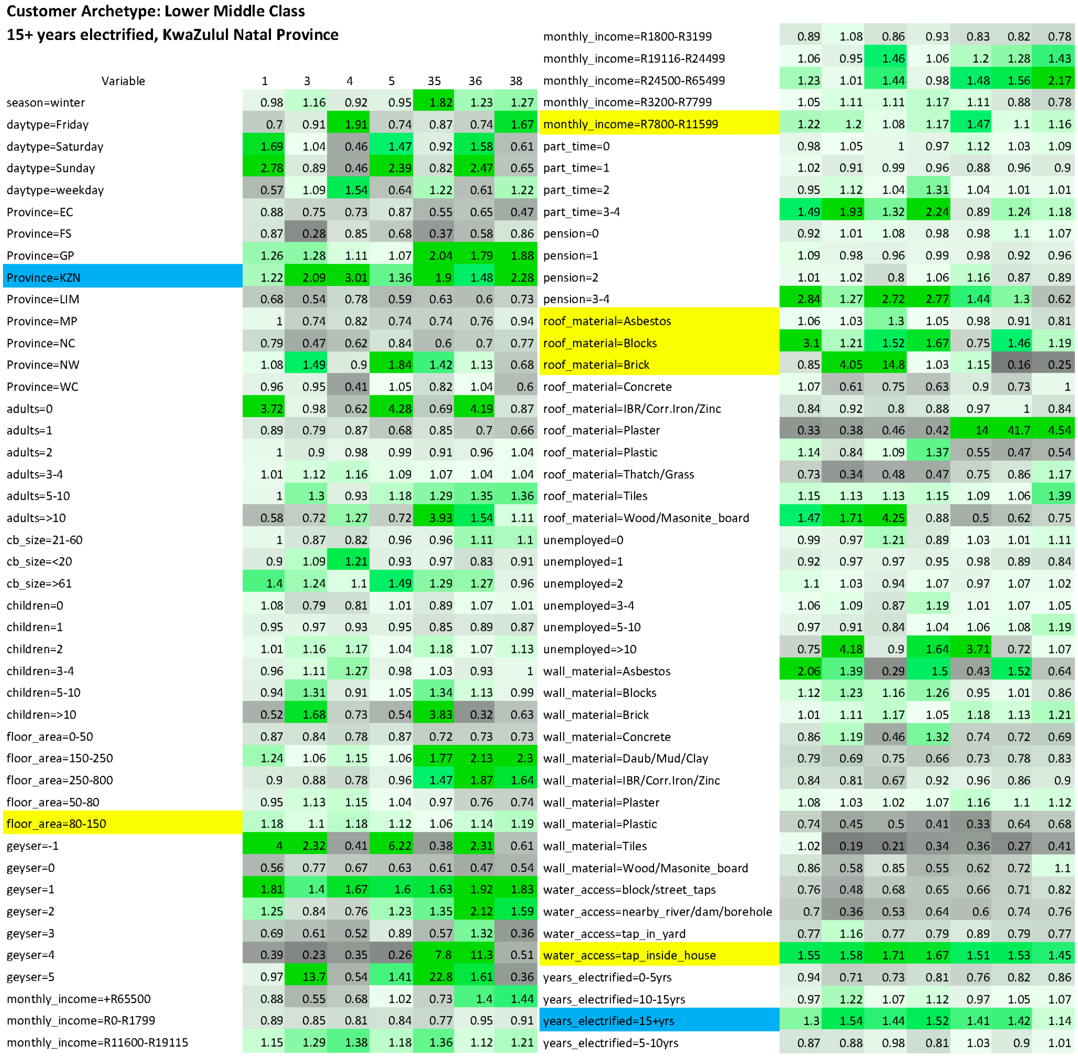
category	variable	value	sample_count (unweighted)
economic	monthly_income	R7800-R11599	15376
		R11600-R19115	13856
		R19116-R24499	6268
		R24500-R65499	12403
		+R65500	5573
dwelling	floor_area (m^2)	0-50	87137
		50-80	67043
		80-150	55736
		150-250	16602
		250-800	12610
	roof_material	Asbestos	19991
		Blocks	70
		Brick	91
		Concrete	867
		IBR/Corr.Iron/Zinc	150522
		Plastic	1119
		Thatch/Grass	1450
		Tiles	65018
	wall_material	Asbestos	264
		Blocks	56018
		Brick	39717
		Concrete	1507
		Daub/Mud/Clay	48850
		IBR/Corr.Iron/Zinc	4963
		Plaster	86201
		Plastic	318
		Wood/Masonite board	1290
	water_access	block/street taps	47233
		nearby river/dam/borehole	12099
		tap in yard	78602
		tap inside house	101194
connection	cb_size	<20	95731
		21-60	142590
		>61	807
	years_electrified	0-5yrs	70402
		5-10yrs	73173
		10-15yrs	39458
		15+yrs	56095
appliances	geyser	-1	56
		0	176710
		1	53653
		2	7632
		3	1033
		5	44

FIGURE A.1: Table of features continued

Appendix B

Odds Ratios for Clusters

B.1 Lower middle class, 15+ years electrified, KwaZulu Natal

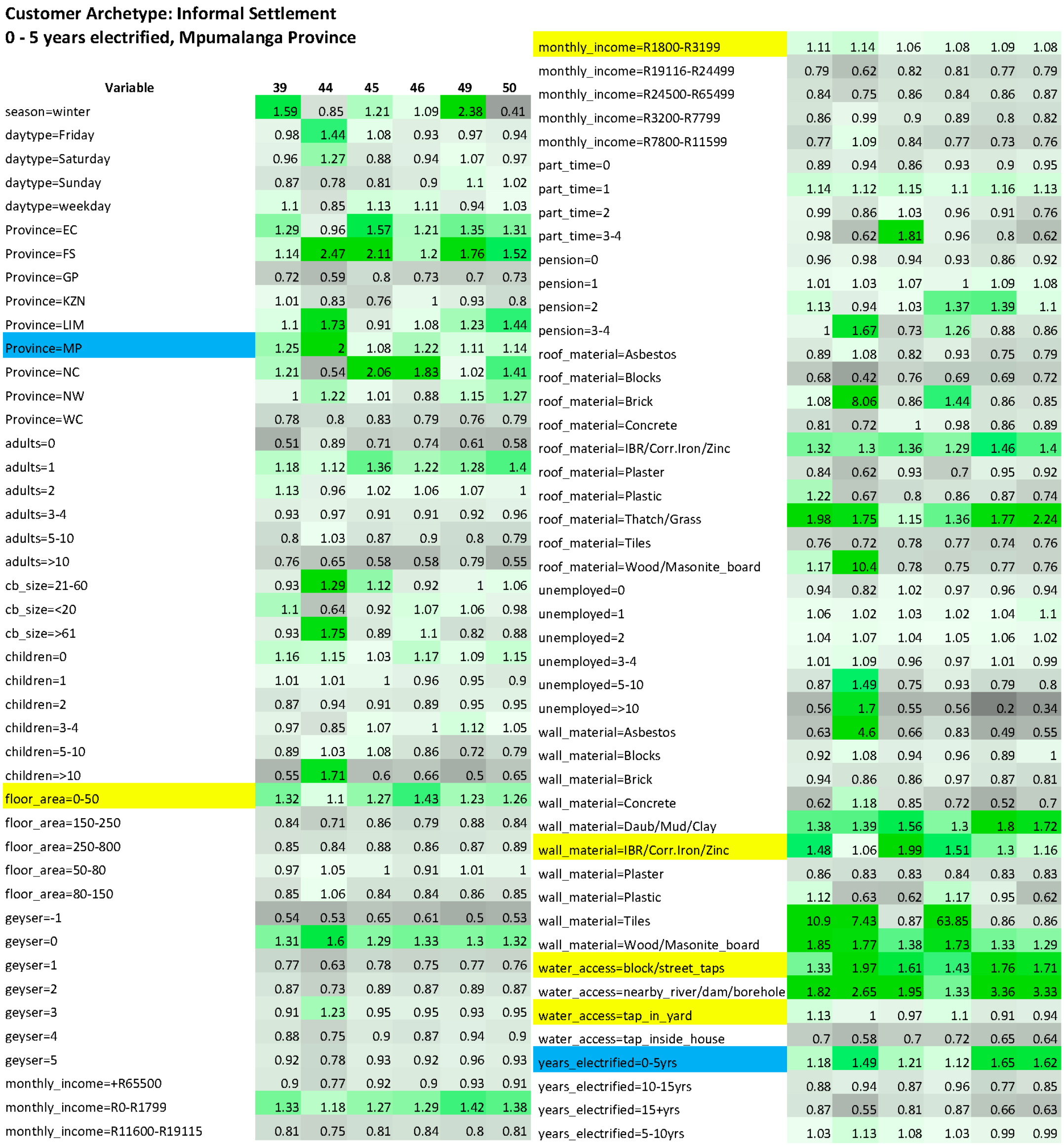


Customer Archetype: Lower Middle Class
15+ years electrified, KwaZulul Natal Province

Variable	1	3	4	5	35	36	38
season=winter	0.98	1.16	0.92	0.95	1.82	1.23	1.27
daytype=Friday	0.7	0.91	1.91	0.74	0.87	0.74	1.67
daytype=Saturday	1.69	1.04	0.46	1.47	0.92	1.58	0.61
daytype=Sunday	2.78	0.89	0.46	2.39	0.82	2.47	0.65
daytype=weekday	0.57	1.09	1.54	0.64	1.22	0.61	1.22
Province=EC	0.88	0.75	0.73	0.87	0.55	0.65	0.47
Province=FS	0.87	0.28	0.85	0.68	0.37	0.58	0.86
Province=GP	1.26	1.28	1.11	1.07	2.04	1.79	1.88
Province=KZN	1.22	2.09	3.01	1.36	1.9	1.48	2.28
Province=LIM	0.68	0.54	0.78	0.59	0.63	0.6	0.73
Province=MP	1	0.74	0.82	0.74	0.74	0.76	0.94
Province=NC	0.79	0.47	0.62	0.84	0.6	0.7	0.77
Province=NW	1.08	1.49	0.9	1.84	1.42	1.13	0.68
Province=WC	0.96	0.95	0.41	1.05	0.82	1.04	0.6
adults=0	3.72	0.98	0.62	4.28	0.69	4.19	0.87
adults=1	0.89	0.79	0.87	0.68	0.85	0.7	0.66
adults=2	1	0.9	0.98	0.99	0.91	0.96	1.04
adults=3-4	1.01	1.12	1.16	1.09	1.07	1.04	1.04
adults=5-10	1	1.3	0.93	1.18	1.29	1.35	1.36
adults=>10	0.58	0.72	1.27	0.72	3.93	1.54	1.11
cb_size=21-60	1	0.87	0.82	0.96	0.96	1.11	1.1
cb_size=<20	0.9	1.09	1.21	0.93	0.97	0.83	0.91
cb_size=>61	1.4	1.24	1.1	1.49	1.29	1.27	0.96
children=0	1.08	0.79	0.81	1.01	0.89	1.07	1.01
children=1	0.95	0.97	0.93	0.95	0.85	0.89	0.87
children=2	1.01	1.16	1.17	1.04	1.18	1.07	1.13
children=3-4	0.96	1.11	1.27	0.98	1.03	0.93	1
children=5-10	0.94	1.31	0.91	1.05	1.34	1.13	0.99
children=>10	0.52	1.68	0.73	0.54	3.83	0.32	0.63
floor_area=0-50	0.87	0.84	0.78	0.87	0.72	0.73	0.73
floor_area=150-250	1.24	1.06	1.15	1.06	1.77	2.13	2.3
floor_area=250-800	0.9	0.88	0.78	0.96	1.47	1.87	1.64
floor_area=50-80	0.95	1.13	1.15	1.04	0.97	0.76	0.74
floor_area=80-150	1.18	1.1	1.18	1.12	1.06	1.14	1.19
geyser=-1	4	2.32	0.41	6.22	0.38	2.31	0.61
geyser=0	0.56	0.77	0.67	0.63	0.61	0.47	0.54
geyser=1	1.81	1.4	1.67	1.6	1.63	1.92	1.83
geyser=2	1.25	0.84	0.76	1.23	1.35	2.12	1.59
geyser=3	0.69	0.61	0.52	0.89	0.57	1.32	0.36
geyser=4	0.39	0.23	0.35	0.26	7.8	11.3	0.51
geyser=5	0.97	13.7	0.54	1.41	22.8	1.61	0.36
monthly_income=+R65500	0.88	0.55	0.68	1.02	0.73	1.4	1.44
monthly_income=R0-R1799	0.89	0.85	0.81	0.84	0.77	0.95	0.91
monthly_income=R11600-R19115	1.15	1.29	1.38	1.18	1.36	1.12	1.21
monthly_income=R1800-R3199	0.89	1.08	0.86	0.93	0.83	0.82	0.78
monthly_income=R19116-R24499	1.06	0.95	1.46	1.06	1.2	1.28	1.43
monthly_income=R24500-R65499	1.23	1.01	1.44	0.98	1.48	1.56	2.17
monthly_income=R3200-R7799	1.05	1.11	1.11	1.17	1.11	0.88	0.78
monthly_income=R7800-R11599	1.22	1.2	1.08	1.17	1.47	1.1	1.16
part_time=0	0.98	1.05	1	0.97	1.12	1.03	1.09
part_time=1	1.02	0.91	0.99	0.96	0.88	0.96	0.9
part_time=2	0.95	1.12	1.04	1.31	1.04	1.01	1.01
part_time=3-4	1.49	1.93	1.32	2.24	0.89	1.24	1.18
pension=0	0.92	1.01	1.08	0.98	0.98	1.1	1.07
pension=1	1.09	0.98	0.96	0.99	0.98	0.92	0.96
pension=2	1.01	1.02	0.8	1.06	1.16	0.87	0.89
pension=3-4	2.84	1.27	2.72	2.77	1.44	1.3	0.62
roof_material=Asbestos	1.06	1.03	1.3	1.05	0.98	0.91	0.81
roof_material=Blocks	3.1	1.21	1.52	1.67	0.75	1.46	1.19
roof_material=Brick	0.85	4.05	14.8	1.03	1.15	0.16	0.25
roof_material=Concrete	1.07	0.61	0.75	0.63	0.9	0.73	1
roof_material=IBR/Corr.Iron/Zinc	0.84	0.92	0.8	0.88	0.97	1	0.84
roof_material=Plaster	0.33	0.38	0.46	0.42	14	41.7	4.54
roof_material=Plastic	1.14	0.84	1.09	1.37	0.55	0.47	0.54
roof_material=Thatch/Grass	0.73	0.34	0.48	0.47	0.75	0.86	1.17
roof_material=Tiles	1.15	1.13	1.13	1.15	1.09	1.06	1.39
roof_material=Wood/Masonite_board	1.47	1.71	4.25	0.88	0.5	0.62	0.75
unemployed=0	0.99	0.97	1.21	0.89	1.03	1.01	1.11
unemployed=1	0.92	0.97	0.97	0.95	0.98	0.89	0.84
unemployed=2	1.1	1.03	0.94	1.07	0.97	1.07	1.02
unemployed=3-4	1.06	1.09	0.87	1.19	1.01	1.07	1.05
unemployed=5-10	0.97	0.91	0.84	1.04	1.06	1.08	1.19
unemployed=>10	0.75	4.18	0.9	1.64	3.71	0.72	1.07
wall_material=Asbestos	2.06	1.39	0.29	1.5	0.43	1.52	0.64
wall_material=Blocks	1.12	1.23	1.16	1.26	0.95	1.01	0.86
wall_material=Brick	1.01	1.11	1.17	1.05	1.18	1.13	1.21
wall_material=Concrete	0.86	1.19	0.46	1.32	0.74	0.72	0.69
wall_material=Daub/Mud/Clay	0.79	0.69	0.75	0.66	0.73	0.78	0.83
wall_material=IBR/Corr.Iron/Zinc	0.84	0.81	0.67	0.92	0.96	0.86	0.9
wall_material=Plaster	1.08	1.03	1.02	1.07	1.16	1.1	1.12
wall_material=Plastic	0.74	0.45	0.5	0.41	0.33	0.64	0.68
wall_material=Tiles	1.02	0.19	0.21	0.34	0.36	0.27	0.41
wall_material=Wood/Masonite_board	0.86	0.58	0.85	0.55	0.62	0.72	1.1
water_access=block/street_taps	0.76	0.48	0.68	0.65	0.66	0.71	0.82
water_access=nearby_river/dam/borehole	0.7	0.36	0.53	0.64	0.6	0.74	0.76
water_access=tap_in_yard	0.77	1.16	0.77	0.79	0.89	0.79	0.77
water_access=tap_inside_house	1.55	1.58	1.71	1.67	1.51	1.53	1.45
years_electrified=0-5yrs	0.94	0.71	0.73	0.81	0.76	0.82	0.86
years_electrified=10-15yrs	0.97	1.22	1.07	1.12	0.97	1.05	1.07
years_electrified=15+yrs	1.3	1.54	1.44	1.52	1.41	1.42	1.14
years_electrified=5-10yrs	0.87	0.88	0.98	0.81	1.03	0.9	1.01

FIGURE B.1: Odds Ratios filtered for selection of lower middle class, 15+ years electrified, KwaZulu Natal Customer Archetype

B.2 Informal settlement, 0 - 5 years electrified, Mpumalanga

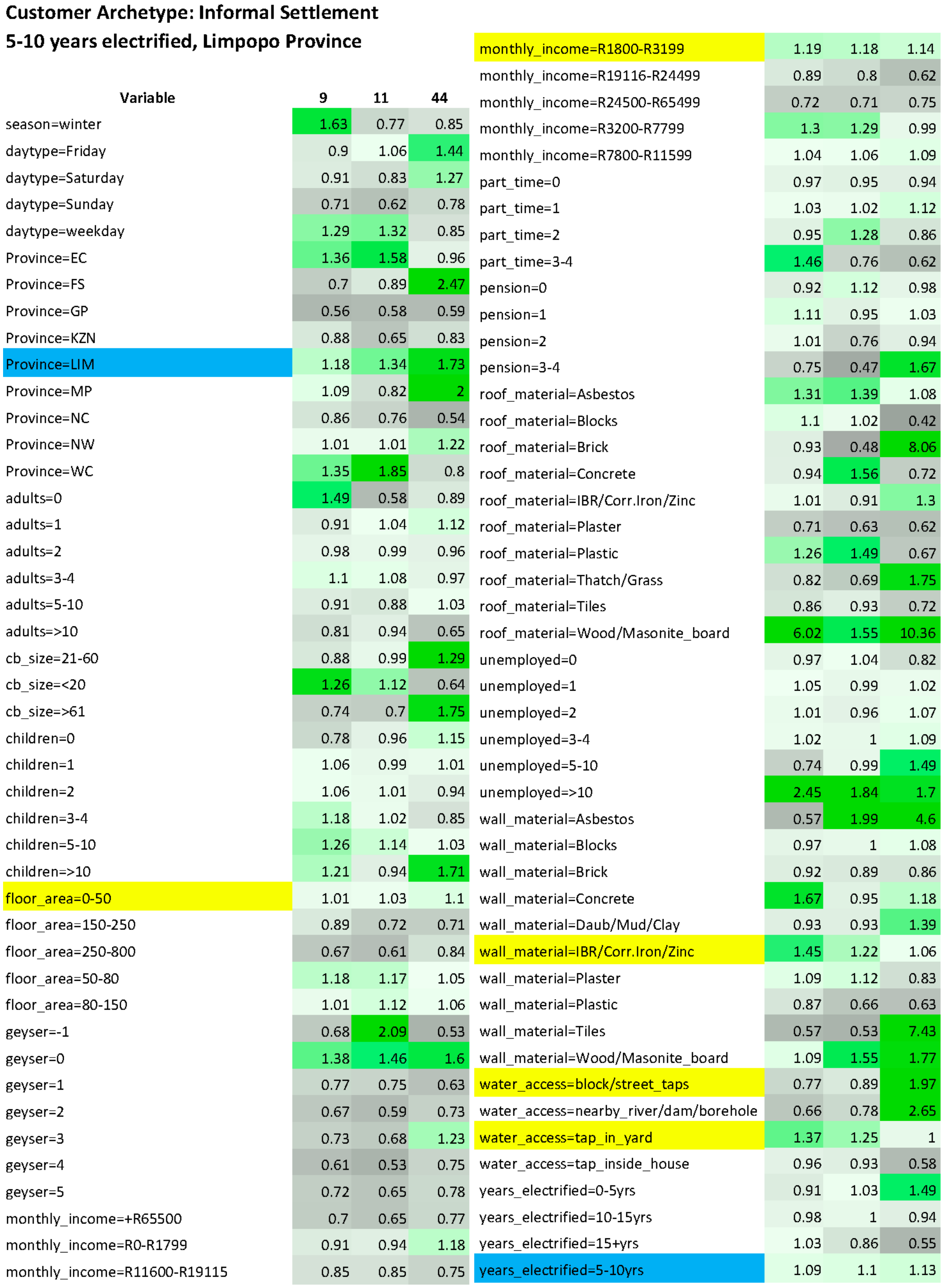


Customer Archetype: Informal Settlement
0 - 5 years electrified, Mpumalanga Province

Variable	39	44	45	46	49	50
season=winter	1.59	0.85	1.21	1.09	2.38	0.41
daytype=Friday	0.98	1.44	1.08	0.93	0.97	0.94
daytype=Saturday	0.96	1.27	0.88	0.94	1.07	0.97
daytype=Sunday	0.87	0.78	0.81	0.9	1.1	1.02
daytype=weekday	1.1	0.85	1.13	1.11	0.94	1.03
Province=EC	1.29	0.96	1.57	1.21	1.35	1.31
Province=FS	1.14	2.47	2.11	1.2	1.76	1.52
Province=GP	0.72	0.59	0.8	0.73	0.7	0.73
Province=KZN	1.01	0.83	0.76	1	0.93	0.8
Province=LIM	1.1	1.73	0.91	1.08	1.23	1.44
Province=MP	1.25	2	1.08	1.22	1.11	1.14
Province=NC	1.21	0.54	2.06	1.83	1.02	1.41
Province=NW	1	1.22	1.01	0.88	1.15	1.27
Province=WC	0.78	0.8	0.83	0.79	0.76	0.79
adults=0	0.51	0.89	0.71	0.74	0.61	0.58
adults=1	1.18	1.12	1.36	1.22	1.28	1.4
adults=2	1.13	0.96	1.02	1.06	1.07	1
adults=3-4	0.93	0.97	0.91	0.91	0.92	0.96
adults=5-10	0.8	1.03	0.87	0.9	0.8	0.79
adults=>10	0.76	0.65	0.58	0.58	0.79	0.55
cb_size=21-60	0.93	1.29	1.12	0.92	1	1.06
cb_size=<20	1.1	0.64	0.92	1.07	1.06	0.98
cb_size=>61	0.93	1.75	0.89	1.1	0.82	0.88
children=0	1.16	1.15	1.03	1.17	1.09	1.15
children=1	1.01	1.01	1	0.96	0.95	0.9
children=2	0.87	0.94	0.91	0.89	0.95	0.95
children=3-4	0.97	0.85	1.07	1	1.12	1.05
children=5-10	0.89	1.03	1.08	0.86	0.72	0.79
children=>10	0.55	1.71	0.6	0.66	0.5	0.65
floor_area=0-50	1.32	1.1	1.27	1.43	1.23	1.26
floor_area=150-250	0.84	0.71	0.86	0.79	0.88	0.84
floor_area=250-800	0.85	0.84	0.88	0.86	0.87	0.89
floor_area=50-80	0.97	1.05	1	0.91	1.01	1
floor_area=80-150	0.85	1.06	0.84	0.84	0.86	0.85
geyser=-1	0.54	0.53	0.65	0.61	0.5	0.53
geyser=0	1.31	1.6	1.29	1.33	1.3	1.32
geyser=1	0.77	0.63	0.78	0.75	0.77	0.76
geyser=2	0.87	0.73	0.89	0.87	0.89	0.87
geyser=3	0.91	1.23	0.95	0.95	0.93	0.95
geyser=4	0.88	0.75	0.9	0.87	0.94	0.9
geyser=5	0.92	0.78	0.93	0.92	0.96	0.93
monthly_income=+R65500	0.9	0.77	0.92	0.9	0.93	0.91
monthly_income=R0-R1799	1.33	1.18	1.27	1.29	1.42	1.38
monthly_income=R11600-R19115	0.81	0.75	0.81	0.84	0.8	0.81
monthly_income=R1800-R3199	1.11	1.14	1.06	1.08	1.09	1.08
monthly_income=R19116-R24499	0.79	0.62	0.82	0.81	0.77	0.79
monthly_income=R24500-R65499	0.84	0.75	0.86	0.84	0.86	0.87
monthly_income=R3200-R7799	0.86	0.99	0.9	0.89	0.8	0.82
monthly_income=R7800-R11599	0.77	1.09	0.84	0.77	0.73	0.76
part_time=0	0.89	0.94	0.86	0.93	0.9	0.95
part_time=1	1.14	1.12	1.15	1.1	1.16	1.13
part_time=2	0.99	0.86	1.03	0.96	0.91	0.76
part_time=3-4	0.98	0.62	1.81	0.96	0.8	0.62
pension=0	0.96	0.98	0.94	0.93	0.86	0.92
pension=1	1.01	1.03	1.07	1	1.09	1.08
pension=2	1.13	0.94	1.03	1.37	1.39	1.1
pension=3-4	1	1.67	0.73	1.26	0.88	0.86
roof_material=Asbestos	0.89	1.08	0.82	0.93	0.75	0.79
roof_material=Blocks	0.68	0.42	0.76	0.69	0.69	0.72
roof_material=Brick	1.08	8.06	0.86	1.44	0.86	0.85
roof_material=Concrete	0.81	0.72	1	0.98	0.86	0.89
roof_material=IBR/Corr.Iron/Zinc	1.32	1.3	1.36	1.29	1.46	1.4
roof_material=Plaster	0.84	0.62	0.93	0.7	0.95	0.92
roof_material=Plastic	1.22	0.67	0.8	0.86	0.87	0.74
roof_material=Thatch/Grass	1.98	1.75	1.15	1.36	1.77	2.24
roof_material=Tiles	0.76	0.72	0.78	0.77	0.74	0.76
roof_material=Wood/Masonite_board	1.17	10.4	0.78	0.75	0.77	0.76
unemployed=0	0.94	0.82	1.02	0.97	0.96	0.94
unemployed=1	1.06	1.02	1.03	1.02	1.04	1.1
unemployed=2	1.04	1.07	1.04	1.05	1.06	1.02
unemployed=3-4	1.01	1.09	0.96	0.97	1.01	0.99
unemployed=5-10	0.87	1.49	0.75	0.93	0.79	0.8
unemployed=>10	0.56	1.7	0.55	0.56	0.2	0.34
wall_material=Asbestos	0.63	4.6	0.66	0.83	0.49	0.55
wall_material=Blocks	0.92	1.08	0.94	0.96	0.89	1
wall_material=Brick	0.94	0.86	0.86	0.97	0.87	0.81
wall_material=Concrete	0.62	1.18	0.85	0.72	0.52	0.7
wall_material=Daub/Mud/Clay	1.38	1.39	1.56	1.3	1.8	1.72
wall_material=IBR/Corr.Iron/Zinc	1.48	1.06	1.99	1.51	1.3	1.16
wall_material=Plaster	0.86	0.83	0.83	0.84	0.83	0.83
wall_material=Plastic	1.12	0.63	0.62	1.17	0.95	0.62
wall_material=Tiles	10.9	7.43	0.87	63.85	0.86	0.86
wall_material=Wood/Masonite_board	1.85	1.77	1.38	1.73	1.33	1.29
water_access=block/street_taps	1.33	1.97	1.61	1.43	1.76	1.71
water_access=nearby_river/dam/borehole	1.82	2.65	1.95	1.33	3.36	3.33
water_access=tap_in_yard	1.13	1	0.97	1.1	0.91	0.94
water_access=tap_inside_house	0.7	0.58	0.7	0.72	0.65	0.64
years_electrified=0-5yrs	1.18	1.49	1.21	1.12	1.65	1.62
years_electrified=10-15yrs	0.88	0.94	0.87	0.96	0.77	0.85
years_electrified=15+yrs	0.87	0.55	0.81	0.87	0.66	0.63
years_electrified=5-10yrs	1.03	1.13	1.08	1.03	0.99	0.99

FIGURE B.2: Odds Ratios filtered for selection of informal settlement, 0 - 5 years electrified, Mpumalanga Customer Archetype

B.3 Informal settlement, 5 - 10 years electrified, Limpopo

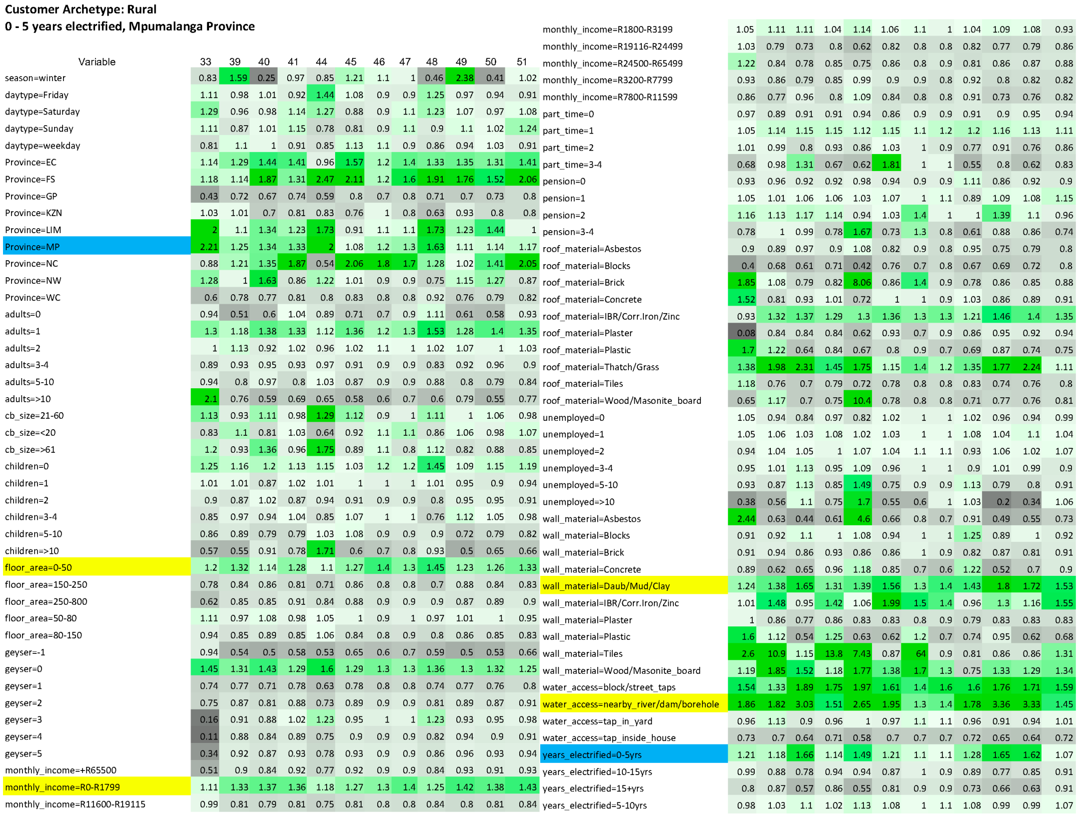


Customer Archetype: Informal Settlement
5-10 years electrified, Limpopo Province

Variable	9	11	44
season=winter	1.63	0.77	0.85
daytype=Friday	0.9	1.06	1.44
daytype=Saturday	0.91	0.83	1.27
daytype=Sunday	0.71	0.62	0.78
daytype=weekday	1.29	1.32	0.85
Province=EC	1.36	1.58	0.96
Province=FS	0.7	0.89	2.47
Province=GP	0.56	0.58	0.59
Province=KZN	0.88	0.65	0.83
Province=LIM	1.18	1.34	1.73
Province=MP	1.09	0.82	2
Province=NC	0.86	0.76	0.54
Province=NW	1.01	1.01	1.22
Province=WC	1.35	1.85	0.8
adults=0	1.49	0.58	0.89
adults=1	0.91	1.04	1.12
adults=2	0.98	0.99	0.96
adults=3-4	1.1	1.08	0.97
adults=5-10	0.91	0.88	1.03
adults=>10	0.81	0.94	0.65
cb_size=21-60	0.88	0.99	1.29
cb_size=<20	1.26	1.12	0.64
cb_size=>61	0.74	0.7	1.75
children=0	0.78	0.96	1.15
children=1	1.06	0.99	1.01
children=2	1.06	1.01	0.94
children=3-4	1.18	1.02	0.85
children=5-10	1.26	1.14	1.03
children=>10	1.21	0.94	1.71
floor_area=0-50	1.01	1.03	1.1
floor_area=150-250	0.89	0.72	0.71
floor_area=250-800	0.67	0.61	0.84
floor_area=50-80	1.18	1.17	1.05
floor_area=80-150	1.01	1.12	1.06
geyser=-1	0.68	2.09	0.53
geyser=0	1.38	1.46	1.6
geyser=1	0.77	0.75	0.63
geyser=2	0.67	0.59	0.73
geyser=3	0.73	0.68	1.23
geyser=4	0.61	0.53	0.75
geyser=5	0.72	0.65	0.78
monthly_income=+R65500	0.7	0.65	0.77
monthly_income=R0-R1799	0.91	0.94	1.18
monthly_income=R11600-R19115	0.85	0.85	0.75
monthly_income=R1800-R3199	1.19	1.18	1.14
monthly_income=R19116-R24499	0.89	0.8	0.62
monthly_income=R24500-R65499	0.72	0.71	0.75
monthly_income=R3200-R7799	1.3	1.29	0.99
monthly_income=R7800-R11599	1.04	1.06	1.09
part_time=0	0.97	0.95	0.94
part_time=1	1.03	1.02	1.12
part_time=2	0.95	1.28	0.86
part_time=3-4	1.46	0.76	0.62
pension=0	0.92	1.12	0.98
pension=1	1.11	0.95	1.03
pension=2	1.01	0.76	0.94
pension=3-4	0.75	0.47	1.67
roof_material=Asbestos	1.31	1.39	1.08
roof_material=Blocks	1.1	1.02	0.42
roof_material=Brick	0.93	0.48	8.06
roof_material=Concrete	0.94	1.56	0.72
roof_material=IBR/Corr.Iron/Zinc	1.01	0.91	1.3
roof_material=Plaster	0.71	0.63	0.62
roof_material=Plastic	1.26	1.49	0.67
roof_material=Thatch/Grass	0.82	0.69	1.75
roof_material=Tiles	0.86	0.93	0.72
roof_material=Wood/Masonite_board	6.02	1.55	10.36
unemployed=0	0.97	1.04	0.82
unemployed=1	1.05	0.99	1.02
unemployed=2	1.01	0.96	1.07
unemployed=3-4	1.02	1	1.09
unemployed=5-10	0.74	0.99	1.49
unemployed=>10	2.45	1.84	1.7
wall_material=Asbestos	0.57	1.99	4.6
wall_material=Blocks	0.97	1	1.08
wall_material=Brick	0.92	0.89	0.86
wall_material=Concrete	1.67	0.95	1.18
wall_material=Daub/Mud/Clay	0.93	0.93	1.39
wall_material=IBR/Corr.Iron/Zinc	1.45	1.22	1.06
wall_material=Plaster	1.09	1.12	0.83
wall_material=Plastic	0.87	0.66	0.63
wall_material=Tiles	0.57	0.53	7.43
wall_material=Wood/Masonite_board	1.09	1.55	1.77
water_access=block/street_taps	0.77	0.89	1.97
water_access=nearby_river/dam/borehole	0.66	0.78	2.65
water_access=tap_in_yard	1.37	1.25	1
water_access=tap_inside_house	0.96	0.93	0.58
years_electrified=0-5yrs	0.91	1.03	1.49
years_electrified=10-15yrs	0.98	1	0.94
years_electrified=15+yrs	1.03	0.86	0.55
years_electrified=5-10yrs	1.09	1.1	1.13

FIGURE B.3: Odds Ratios filtered for selection of informal settlement, 5 - 10 years electrified, Limpopo Customer Archetype

B.4 Rural, 0 - 5 years electrified, Mpumalanga

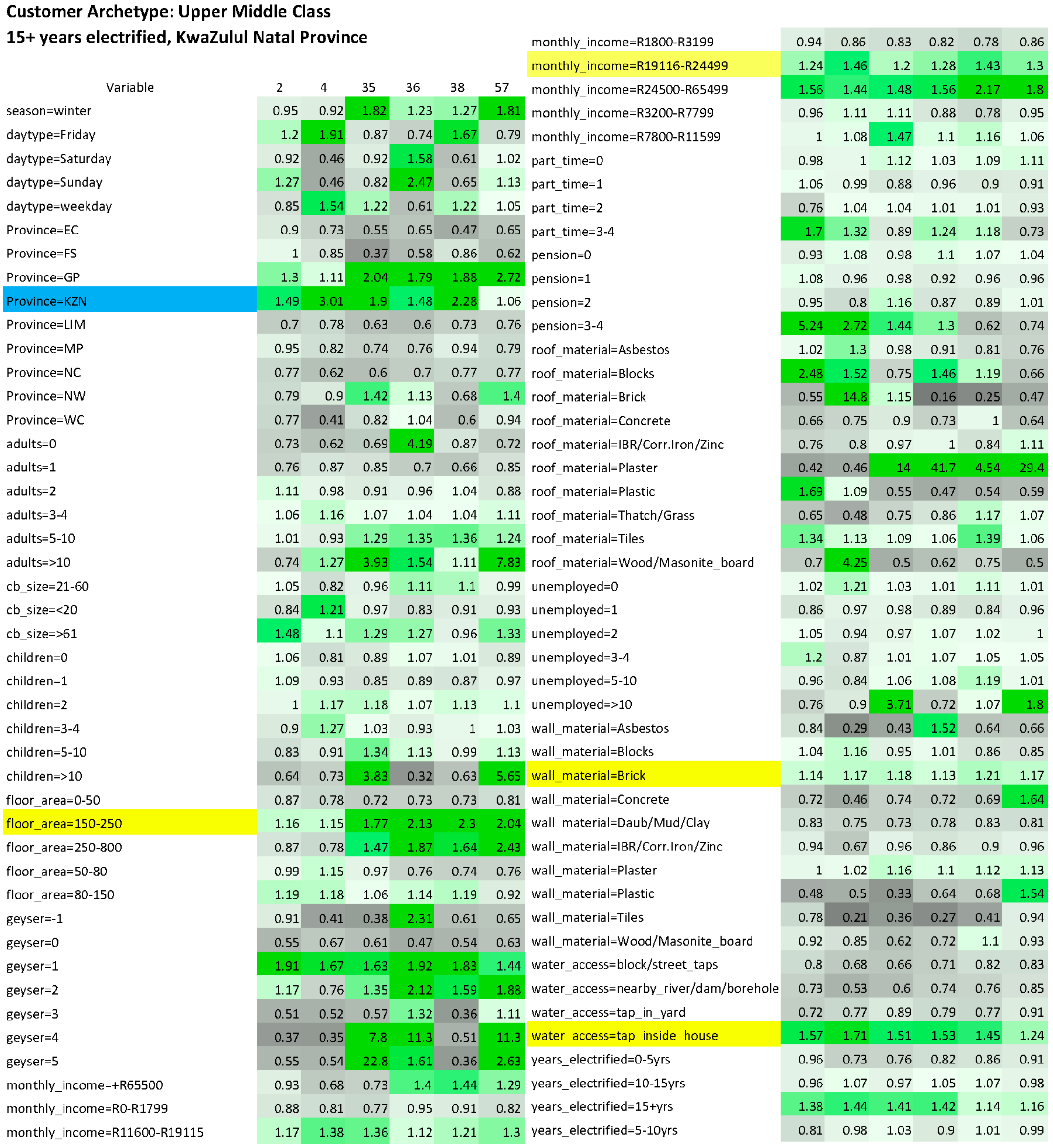


Customer Archetype: Rural
0 - 5 years electrified, Mpumalanga Province

Variable	33	39	40	41	44	45	46	47	48	49	50	51
season=winter	0.83	1.59	0.25	0.97	0.85	1.21	1.1	1	0.46	2.38	0.41	1.02
daytype=Friday	1.11	0.98	1.01	0.92	1.44	1.08	0.9	0.9	1.25	0.97	0.94	0.91
daytype=Saturday	1.29	0.96	0.98	1.14	1.27	0.88	0.9	1.1	1.23	1.07	0.97	1.08
daytype=Sunday	1.11	0.87	1.01	1.15	0.78	0.81	0.9	1.1	0.9	1.1	1.02	1.24
daytype=weekday	0.81	1.1	1	0.91	0.85	1.13	1.1	0.9	0.86	0.94	1.03	0.91
Province=EC	1.14	1.29	1.44	1.41	0.96	1.57	1.2	1.4	1.33	1.35	1.31	1.41
Province=FS	1.18	1.14	1.87	1.31	2.47	2.11	1.2	1.6	1.91	1.76	1.52	2.06
Province=GP	0.43	0.72	0.67	0.74	0.59	0.8	0.7	0.8	0.71	0.7	0.73	0.8
Province=KZN	1.03	1.01	0.7	0.81	0.83	0.76	1	0.8	0.63	0.93	0.8	0.8
Province=LIM	2	1.1	1.34	1.23	1.73	0.91	1.1	1.1	1.73	1.23	1.44	1
Province=MP	2.21	1.25	1.34	1.33	2	1.08	1.2	1.3	1.63	1.11	1.14	1.17
Province=NC	0.88	1.21	1.35	1.87	0.54	2.06	1.8	1.7	1.28	1.02	1.41	2.05
Province=NW	1.28	1	1.63	0.86	1.22	1.01	0.9	0.9	0.75	1.15	1.27	0.87
Province=WC	0.6	0.78	0.77	0.81	0.8	0.83	0.8	0.8	0.92	0.76	0.79	0.82
adults=0	0.94	0.51	0.6	1.04	0.89	0.71	0.7	0.9	1.11	0.61	0.58	0.93
adults=1	1.3	1.18	1.38	1.33	1.12	1.36	1.2	1.3	1.53	1.28	1.4	1.35
adults=2	1	1.13	0.92	1.02	0.96	1.02	1.1	1	1.02	1.07	1	1.03
adults=3-4	0.89	0.93	0.95	0.93	0.97	0.91	0.9	0.9	0.83	0.92	0.96	0.9
adults=5-10	0.94	0.8	0.97	0.8	1.03	0.87	0.9	0.9	0.88	0.8	0.79	0.84
adults=>10	2.1	0.76	0.59	0.69	0.65	0.58	0.6	0.7	0.6	0.79	0.55	0.77
cb_size=21-60	1.13	0.93	1.11	0.98	1.29	1.12	0.9	1	1.11	1	1.06	0.98
cb_size=<20	0.83	1.1	0.81	1.03	0.64	0.92	1.1	1.1	0.86	1.06	0.98	1.07
cb_size=>61	1.2	0.93	1.36	0.96	1.75	0.89	1.1	0.8	1.12	0.82	0.88	0.85
children=0	1.25	1.16	1.2	1.13	1.15	1.03	1.2	1.2	1.45	1.09	1.15	1.19
children=1	1.01	1.01	0.87	1.02	1.01	1	1	1	1.01	0.95	0.9	0.94
children=2	0.9	0.87	1.02	0.87	0.94	0.91	0.9	0.9	0.8	0.95	0.95	0.91
children=3-4	0.85	0.97	0.94	1.04	0.85	1.07	1	1	0.76	1.12	1.05	0.98
children=5-10	0.86	0.89	0.79	0.79	1.03	1.08	0.9	0.9	0.9	0.72	0.79	0.82
children=>10	0.57	0.55	0.91	0.78	1.71	0.6	0.7	0.8	0.93	0.5	0.65	0.66
floor_area=0-50	1.2	1.32	1.14	1.28	1.1	1.27	1.4	1.3	1.45	1.23	1.26	1.33
floor_area=150-250	0.78	0.84	0.86	0.81	0.71	0.86	0.8	0.8	0.7	0.88	0.84	0.83
floor_area=250-800	0.62	0.85	0.85	0.91	0.84	0.88	0.9	0.9	0.9	0.87	0.89	0.9
floor_area=50-80	1.11	0.97	1.08	0.98	1.05	1	0.9	1	0.97	1.01	1	0.95
floor_area=80-150	0.94	0.85	0.89	0.85	1.06	0.84	0.8	0.9	0.8	0.86	0.85	0.83
geyser=-1	0.94	0.54	0.5	0.58	0.53	0.65	0.6	0.7	0.59	0.5	0.53	0.66
geyser=0	1.45	1.31	1.43	1.29	1.6	1.29	1.3	1.3	1.36	1.3	1.32	1.25
geyser=1	0.74	0.77	0.71	0.78	0.63	0.78	0.8	0.8	0.74	0.77	0.76	0.8
geyser=2	0.75	0.87	0.81	0.88	0.73	0.89	0.9	0.9	0.81	0.89	0.87	0.91
geyser=3	0.16	0.91	0.88	1.02	1.23	0.95	1	1	1.23	0.93	0.95	0.98
geyser=4	0.11	0.88	0.84	0.89	0.75	0.9	0.9	0.9	0.82	0.94	0.9	0.91
geyser=5	0.34	0.92	0.87	0.93	0.78	0.93	0.9	0.9	0.86	0.96	0.93	0.94
monthly_income=+R65500	0.51	0.9	0.84	0.92	0.77	0.92	0.9	0.9	0.84	0.93	0.91	0.93
monthly_income=R0-R1799	1.11	1.33	1.37	1.36	1.18	1.27	1.3	1.4	1.25	1.42	1.38	1.43
monthly_income=R11600-R19115	0.99	0.81	0.79	0.81	0.75	0.81	0.8	0.8	0.84	0.8	0.81	0.84
monthly_income=R1800-R3199	1.05	1.11	1.11	1.04	1.14	1.06	1.1	1	1.04	1.09	1.08	0.93
monthly_income=R19116-R24499	1.03	0.79	0.73	0.8	0.62	0.82	0.8	0.8	0.82	0.77	0.79	0.86
monthly_income=R24500-R65499	1.22	0.84	0.78	0.85	0.75	0.86	0.8	0.9	0.81	0.86	0.87	0.88
monthly_income=R3200-R7799	0.93	0.86	0.79	0.85	0.99	0.9	0.9	0.8	0.92	0.8	0.82	0.82
monthly_income=R7800-R11599	0.86	0.77	0.96	0.8	1.09	0.84	0.8	0.8	0.91	0.73	0.76	0.82
part_time=0	0.97	0.89	0.91	0.91	0.94	0.86	0.9	0.9	0.91	0.9	0.95	0.94
part_time=1	1.05	1.14	1.15	1.15	1.12	1.15	1.1	1.2	1.2	1.16	1.13	1.11
part_time=2	1.01	0.99	0.8	0.93	0.86	1.03	1	0.9	0.77	0.91	0.76	0.86
part_time=3-4	0.68	0.98	1.31	0.67	0.62	1.81	1	1	0.55	0.8	0.62	0.83
pension=0	0.93	0.96	0.92	0.92	0.98	0.94	0.9	0.9	1.11	0.86	0.92	0.9
pension=1	1.05	1.01	1.06	1.06	1.03	1.07	1	1.1	0.89	1.09	1.08	1.15
pension=2	1.16	1.13	1.17	1.14	0.94	1.03	1.4	1	1	1.39	1.1	0.96
pension=3-4	0.78	1	0.99	0.78	1.67	0.73	1.3	0.8	0.61	0.88	0.86	0.74
roof_material=Asbestos	0.9	0.89	0.97	0.9	1.08	0.82	0.9	0.8	0.95	0.75	0.79	0.8
roof_material=Blocks	0.4	0.68	0.61	0.71	0.42	0.76	0.7	0.8	0.67	0.69	0.72	0.8
roof_material=Brick	1.85	1.08	0.79	0.82	8.06	0.86	1.4	0.9	0.78	0.86	0.85	0.88
roof_material=Concrete	1.52	0.81	0.93	1.01	0.72	1	1	0.9	1.03	0.86	0.89	0.91
roof_material=IBR/Corr.Iron/Zinc	0.93	1.32	1.37	1.29	1.3	1.36	1.3	1.3	1.21	1.46	1.4	1.35
roof_material=Plaster	0.08	0.84	0.84	0.84	0.62	0.93	0.7	0.9	0.86	0.95	0.92	0.94
roof_material=Plastic	1.7	1.22	0.64	0.84	0.67	0.8	0.9	0.7	0.69	0.87	0.74	0.75
roof_material=Thatch/Grass	1.38	1.98	2.31	1.45	1.75	1.15	1.4	1.2	1.35	1.77	2.24	1.11
roof_material=Tiles	1.18	0.76	0.7	0.79	0.72	0.78	0.8	0.8	0.83	0.74	0.76	0.8
roof_material=Wood/Masonite_board	0.65	1.17	0.7	0.75	10.4	0.78	0.8	0.8	0.71	0.77	0.76	0.81
unemployed=0	1.05	0.94	0.84	0.97	0.82	1.02	1	1	1.02	0.96	0.94	0.99
unemployed=1	1.05	1.06	1.03	1.08	1.02	1.03	1	1	1.08	1.04	1.1	1.04
unemployed=2	0.94	1.04	1.05	1	1.07	1.04	1.1	1.1	0.93	1.06	1.02	1.07
unemployed=3-4	0.95	1.01	1.13	0.95	1.09	0.96	1	1	0.9	1.01	0.99	0.9
unemployed=5-10	0.93	0.87	1.13	0.85	1.49	0.75	0.9	0.9	1.13	0.79	0.8	0.91
unemployed=>10	0.38	0.56	1.1	0.75	1.7	0.55	0.6	1	1.03	0.2	0.34	1.06
wall_material=Asbestos	2.44	0.63	0.44	0.61	4.6	0.66	0.8	0.7	0.91	0.49	0.55	0.73
wall_material=Blocks	0.91	0.92	1.1	1	1.08	0.94	1	1	1.25	0.89	1	0.92
wall_material=Brick	0.91	0.94	0.86	0.93	0.86	0.86	1	0.9	0.82	0.87	0.81	0.91
wall_material=Concrete	0.89	0.62	0.65	0.96	1.18	0.85	0.7	0.6	1.22	0.52	0.7	0.9
wall_material=Daub/Mud/Clay	1.24	1.38	1.65	1.31	1.39	1.56	1.3	1.4	1.43	1.8	1.72	1.53
wall_material=IBR/Corr.Iron/Zinc	1.01	1.48	0.95	1.42	1.06	1.99	1.5	1.4	0.96	1.3	1.16	1.55
wall_material=Plaster	1	0.86	0.77	0.86	0.83	0.83	0.8	0.9	0.79	0.83	0.83	0.83
wall_material=Plastic	1.6	1.12	0.54	1.25	0.63	0.62	1.2	0.7	0.74	0.95	0.62	0.68
wall_material=Tiles	2.6	10.9	1.15	13.8	7.43	0.87	64	0.9	0.81	0.86	0.86	1.31
wall_material=Wood/Masonite_board	1.19	1.85	1.52	1.18	1.77	1.38	1.7	1.3	0.75	1.33	1.29	1.34
water_access=block/street_taps	1.54	1.33	1.89	1.75	1.97	1.61	1.4	1.6	1.6	1.76	1.71	1.59
water_access=nearby_river/dam/borehole	1.86	1.82	3.03	1.51	2.65	1.95	1.3	1.4	1.78	3.36	3.33	1.45
water_access=tap_in_yard	0.96	1.13	0.9	0.96	1	0.97	1.1	1.1	0.96	0.91	0.94	1.01
water_access=tap_inside_house	0.73	0.7	0.64	0.71	0.58	0.7	0.7	0.7	0.72	0.65	0.64	0.72
years_electrified=0-5yrs	1.21	1.18	1.66	1.14	1.49	1.21	1.1	1.1	1.28	1.65	1.62	1.07
years_electrified=10-15yrs	0.99	0.88	0.78	0.94	0.94	0.87	1	0.9	0.89	0.77	0.85	0.91
years_electrified=15+yrs	0.8	0.87	0.57	0.86	0.55	0.81	0.9	0.9	0.73	0.66	0.63	0.91
years_electrified=5-10yrs	0.98	1.03	1.1	1.02	1.13	1.08	1	1.1	1.08	0.99	0.99	1.07

FIGURE B.4: Odds Ratios filtered for selection of rural, 0 - 5 years electrified, Mpumalanga Customer Archetype

B.5 Upper middle class, 15+ years electrified, KwaZulu Natal

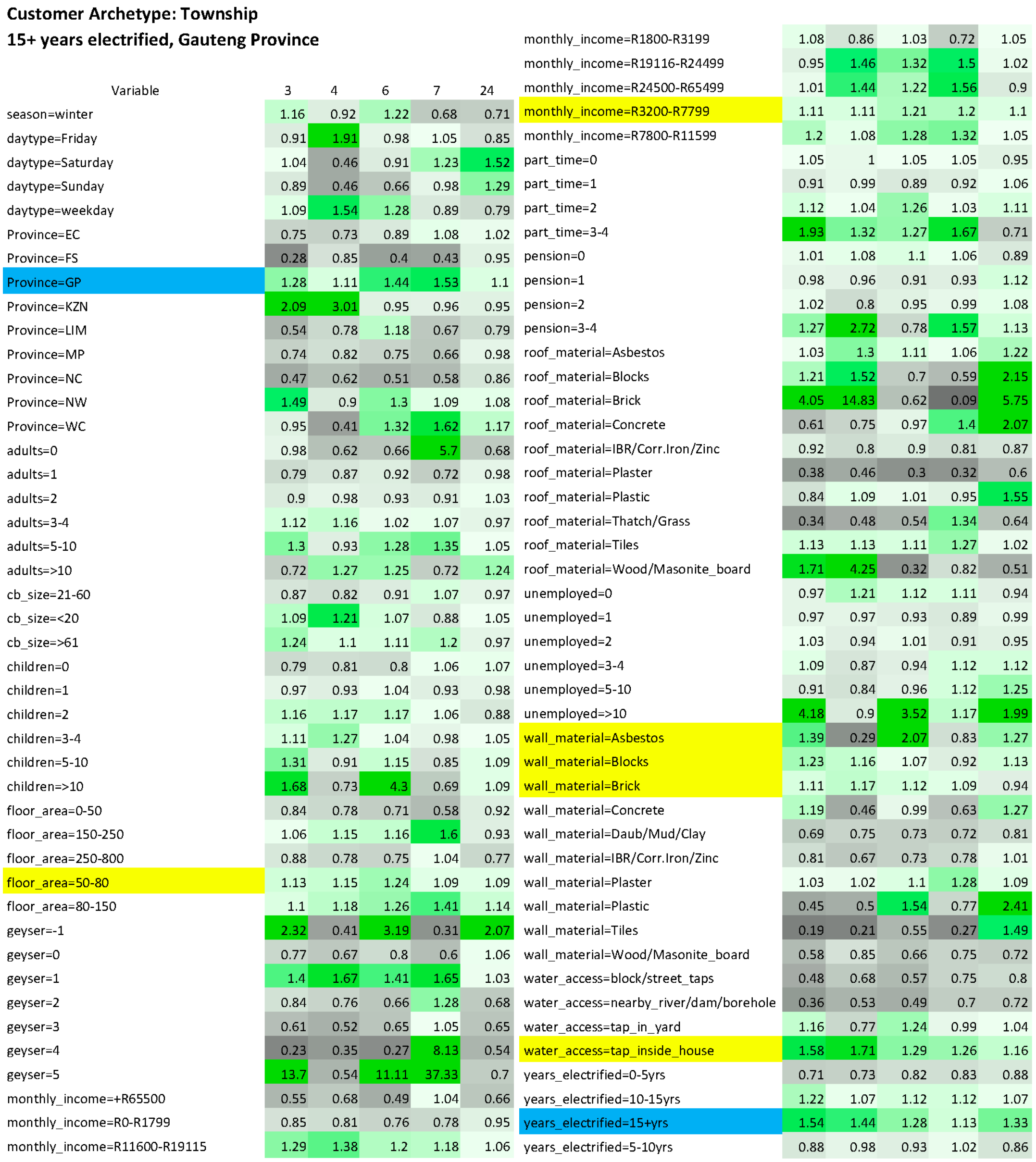


Customer Archetype: Upper Middle Class
15+ years electrified, KwaZulul Natal Province

Variable	2	4	35	36	38	57
season=winter	0.95	0.92	1.82	1.23	1.27	1.81
daytype=Friday	1.2	1.91	0.87	0.74	1.67	0.79
daytype=Saturday	0.92	0.46	0.92	1.58	0.61	1.02
daytype=Sunday	1.27	0.46	0.82	2.47	0.65	1.13
daytype=weekday	0.85	1.54	1.22	0.61	1.22	1.05
Province=EC	0.9	0.73	0.55	0.65	0.47	0.65
Province=FS	1	0.85	0.37	0.58	0.86	0.62
Province=GP	1.3	1.11	2.04	1.79	1.88	2.72
Province=KZN	1.49	3.01	1.9	1.48	2.28	1.06
Province=LIM	0.7	0.78	0.63	0.6	0.73	0.76
Province=MP	0.95	0.82	0.74	0.76	0.94	0.79
Province=NC	0.77	0.62	0.6	0.7	0.77	0.77
Province=NW	0.79	0.9	1.42	1.13	0.68	1.4
Province=WC	0.77	0.41	0.82	1.04	0.6	0.94
adults=0	0.73	0.62	0.69	4.19	0.87	0.72
adults=1	0.76	0.87	0.85	0.7	0.66	0.85
adults=2	1.11	0.98	0.91	0.96	1.04	0.88
adults=3-4	1.06	1.16	1.07	1.04	1.04	1.11
adults=5-10	1.01	0.93	1.29	1.35	1.36	1.24
adults=>10	0.74	1.27	3.93	1.54	1.11	7.83
cb_size=21-60	1.05	0.82	0.96	1.11	1.1	0.99
cb_size=<20	0.84	1.21	0.97	0.83	0.91	0.93
cb_size=>61	1.48	1.1	1.29	1.27	0.96	1.33
children=0	1.06	0.81	0.89	1.07	1.01	0.89
children=1	1.09	0.93	0.85	0.89	0.87	0.97
children=2	1	1.17	1.18	1.07	1.13	1.1
children=3-4	0.9	1.27	1.03	0.93	1	1.03
children=5-10	0.83	0.91	1.34	1.13	0.99	1.13
children=>10	0.64	0.73	3.83	0.32	0.63	5.65
floor_area=0-50	0.87	0.78	0.72	0.73	0.73	0.81
floor_area=150-250	1.16	1.15	1.77	2.13	2.3	2.04
floor_area=250-800	0.87	0.78	1.47	1.87	1.64	2.43
floor_area=50-80	0.99	1.15	0.97	0.76	0.74	0.76
floor_area=80-150	1.19	1.18	1.06	1.14	1.19	0.92
geyser=-1	0.91	0.41	0.38	2.31	0.61	0.65
geyser=0	0.55	0.67	0.61	0.47	0.54	0.63
geyser=1	1.91	1.67	1.63	1.92	1.83	1.44
geyser=2	1.17	0.76	1.35	2.12	1.59	1.88
geyser=3	0.51	0.52	0.57	1.32	0.36	1.11
geyser=4	0.37	0.35	7.8	11.3	0.51	11.3
geyser=5	0.55	0.54	22.8	1.61	0.36	2.63
monthly_income=+R65500	0.93	0.68	0.73	1.4	1.44	1.29
monthly_income=R0-R1799	0.88	0.81	0.77	0.95	0.91	0.82
monthly_income=R11600-R19115	1.17	1.38	1.36	1.12	1.21	1.3
monthly_income=R1800-R3199	0.94	0.86	0.83	0.82	0.78	0.86
monthly_income=R19116-R24499	1.24	1.46	1.2	1.28	1.43	1.3
monthly_income=R24500-R65499	1.56	1.44	1.48	1.56	2.17	1.8
monthly_income=R3200-R7799	0.96	1.11	1.11	0.88	0.78	0.95
monthly_income=R7800-R11599	1	1.08	1.47	1.1	1.16	1.06
part_time=0	0.98	1	1.12	1.03	1.09	1.11
part_time=1	1.06	0.99	0.88	0.96	0.9	0.91
part_time=2	0.76	1.04	1.04	1.01	1.01	0.93
part_time=3-4	1.7	1.32	0.89	1.24	1.18	0.73
pension=0	0.93	1.08	0.98	1.1	1.07	1.04
pension=1	1.08	0.96	0.98	0.92	0.96	0.96
pension=2	0.95	0.8	1.16	0.87	0.89	1.01
pension=3-4	5.24	2.72	1.44	1.3	0.62	0.74
roof_material=Asbestos	1.02	1.3	0.98	0.91	0.81	0.76
roof_material=Blocks	2.48	1.52	0.75	1.46	1.19	0.66
roof_material=Brick	0.55	14.8	1.15	0.16	0.25	0.47
roof_material=Concrete	0.66	0.75	0.9	0.73	1	0.64
roof_material=IBR/Corr.Iron/Zinc	0.76	0.8	0.97	1	0.84	1.11
roof_material=Plaster	0.42	0.46	14	41.7	4.54	29.4
roof_material=Plastic	1.69	1.09	0.55	0.47	0.54	0.59
roof_material=Thatch/Grass	0.65	0.48	0.75	0.86	1.17	1.07
roof_material=Tiles	1.34	1.13	1.09	1.06	1.39	1.06
roof_material=Wood/Masonite_board	0.7	4.25	0.5	0.62	0.75	0.5
unemployed=0	1.02	1.21	1.03	1.01	1.11	1.01
unemployed=1	0.86	0.97	0.98	0.89	0.84	0.96
unemployed=2	1.05	0.94	0.97	1.07	1.02	1
unemployed=3-4	1.2	0.87	1.01	1.07	1.05	1.05
unemployed=5-10	0.96	0.84	1.06	1.08	1.19	1.01
unemployed=>10	0.76	0.9	3.71	0.72	1.07	1.8
wall_material=Asbestos	0.84	0.29	0.43	1.52	0.64	0.66
wall_material=Blocks	1.04	1.16	0.95	1.01	0.86	0.85
wall_material=Brick	1.14	1.17	1.18	1.13	1.21	1.17
wall_material=Concrete	0.72	0.46	0.74	0.72	0.69	1.64
wall_material=Daub/Mud/Clay	0.83	0.75	0.73	0.78	0.83	0.81
wall_material=IBR/Corr.Iron/Zinc	0.94	0.67	0.96	0.86	0.9	0.96
wall_material=Plaster	1	1.02	1.16	1.1	1.12	1.13
wall_material=Plastic	0.48	0.5	0.33	0.64	0.68	1.54
wall_material=Tiles	0.78	0.21	0.36	0.27	0.41	0.94
wall_material=Wood/Masonite_board	0.92	0.85	0.62	0.72	1.1	0.93
water_access=block/street_taps	0.8	0.68	0.66	0.71	0.82	0.83
water_access=nearby_river/dam/borehole	0.73	0.53	0.6	0.74	0.76	0.85
water_access=tap_in_yard	0.72	0.77	0.89	0.79	0.77	0.91
water_access=tap_inside_house	1.57	1.71	1.51	1.53	1.45	1.24
years_electrified=0-5yrs	0.96	0.73	0.76	0.82	0.86	0.91
years_electrified=10-15yrs	0.96	1.07	0.97	1.05	1.07	0.98
years_electrified=15+yrs	1.38	1.44	1.41	1.42	1.14	1.16
years_electrified=5-10yrs	0.81	0.98	1.03	0.9	1.01	0.99

FIGURE B.5: Odds Ratios filtered for selection of upper middle class, 15+ years electrified, KwaZulu Natal Customer Archetype

B.6 Township, 15+ years electrified, Gauteng

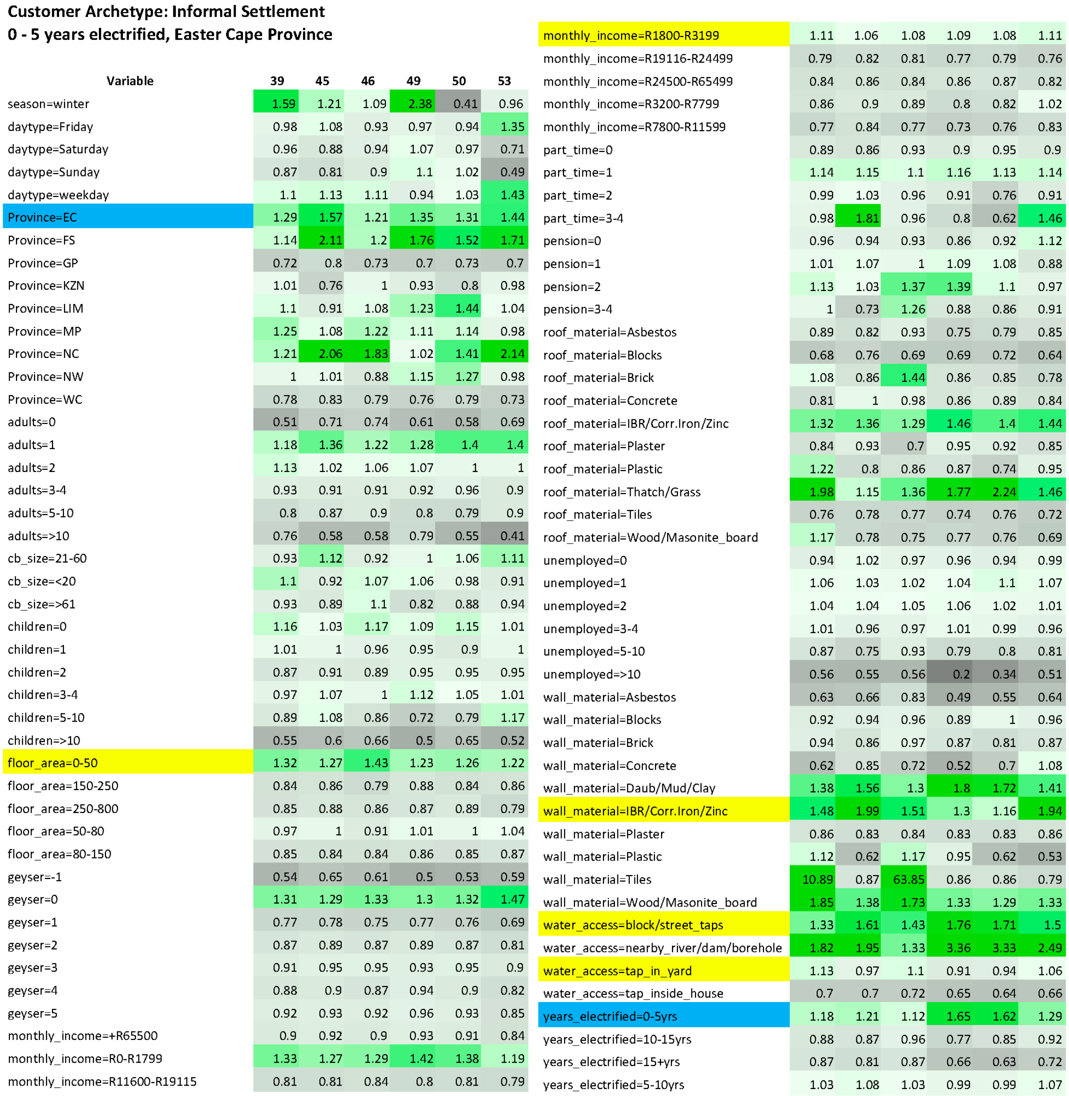


Customer Archetype: Township
15+ years electrified, Gauteng Province

Variable	3	4	6	7	24
season=winter	1.16	0.92	1.22	0.68	0.71
daytype=Friday	0.91	1.91	0.98	1.05	0.85
daytype=Saturday	1.04	0.46	0.91	1.23	1.52
daytype=Sunday	0.89	0.46	0.66	0.98	1.29
daytype=weekday	1.09	1.54	1.28	0.89	0.79
Province=EC	0.75	0.73	0.89	1.08	1.02
Province=FS	0.28	0.85	0.4	0.43	0.95
Province=GP	1.28	1.11	1.44	1.53	1.1
Province=KZN	2.09	3.01	0.95	0.96	0.95
Province=LIM	0.54	0.78	1.18	0.67	0.79
Province=MP	0.74	0.82	0.75	0.66	0.98
Province=NC	0.47	0.62	0.51	0.58	0.86
Province=NW	1.49	0.9	1.3	1.09	1.08
Province=WC	0.95	0.41	1.32	1.62	1.17
adults=0	0.98	0.62	0.66	5.7	0.68
adults=1	0.79	0.87	0.92	0.72	0.98
adults=2	0.9	0.98	0.93	0.91	1.03
adults=3-4	1.12	1.16	1.02	1.07	0.97
adults=5-10	1.3	0.93	1.28	1.35	1.05
adults=>10	0.72	1.27	1.25	0.72	1.24
cb_size=21-60	0.87	0.82	0.91	1.07	0.97
cb_size=<20	1.09	1.21	1.07	0.88	1.05
cb_size=>61	1.24	1.1	1.11	1.2	0.97
children=0	0.79	0.81	0.8	1.06	1.07
children=1	0.97	0.93	1.04	0.93	0.98
children=2	1.16	1.17	1.17	1.06	0.88
children=3-4	1.11	1.27	1.04	0.98	1.05
children=5-10	1.31	0.91	1.15	0.85	1.09
children=>10	1.68	0.73	4.3	0.69	1.09
floor_area=0-50	0.84	0.78	0.71	0.58	0.92
floor_area=150-250	1.06	1.15	1.16	1.6	0.93
floor_area=250-800	0.88	0.78	0.75	1.04	0.77
floor_area=50-80	1.13	1.15	1.24	1.09	1.09
floor_area=80-150	1.1	1.18	1.26	1.41	1.14
geyser=-1	2.32	0.41	3.19	0.31	2.07
geyser=0	0.77	0.67	0.8	0.6	1.06
geyser=1	1.4	1.67	1.41	1.65	1.03
geyser=2	0.84	0.76	0.66	1.28	0.68
geyser=3	0.61	0.52	0.65	1.05	0.65
geyser=4	0.23	0.35	0.27	8.13	0.54
geyser=5	13.7	0.54	11.11	37.33	0.7
monthly_income=+R65500	0.55	0.68	0.49	1.04	0.66
monthly_income=R0-R1799	0.85	0.81	0.76	0.78	0.95
monthly_income=R11600-R19115	1.29	1.38	1.2	1.18	1.06
monthly_income=R1800-R3199	1.08	0.86	1.03	0.72	1.05
monthly_income=R19116-R24499	0.95	1.46	1.32	1.5	1.02
monthly_income=R24500-R65499	1.01	1.44	1.22	1.56	0.9
monthly_income=R3200-R7799	1.11	1.11	1.21	1.2	1.1
monthly_income=R7800-R11599	1.2	1.08	1.28	1.32	1.05
part_time=0	1.05	1	1.05	1.05	0.95
part_time=1	0.91	0.99	0.89	0.92	1.06
part_time=2	1.12	1.04	1.26	1.03	1.11
part_time=3-4	1.93	1.32	1.27	1.67	0.71
pension=0	1.01	1.08	1.1	1.06	0.89
pension=1	0.98	0.96	0.91	0.93	1.12
pension=2	1.02	0.8	0.95	0.99	1.08
pension=3-4	1.27	2.72	0.78	1.57	1.13
roof_material=Asbestos	1.03	1.3	1.11	1.06	1.22
roof_material=Blocks	1.21	1.52	0.7	0.59	2.15
roof_material=Brick	4.05	14.83	0.62	0.09	5.75
roof_material=Concrete	0.61	0.75	0.97	1.4	2.07
roof_material=IBR/Corr.Iron/Zinc	0.92	0.8	0.9	0.81	0.87
roof_material=Plaster	0.38	0.46	0.3	0.32	0.6
roof_material=Plastic	0.84	1.09	1.01	0.95	1.55
roof_material=Thatch/Grass	0.34	0.48	0.54	1.34	0.64
roof_material=Tiles	1.13	1.13	1.11	1.27	1.02
roof_material=Wood/Masonite_board	1.71	4.25	0.32	0.82	0.51
unemployed=0	0.97	1.21	1.12	1.11	0.94
unemployed=1	0.97	0.97	0.93	0.89	0.99
unemployed=2	1.03	0.94	1.01	0.91	0.95
unemployed=3-4	1.09	0.87	0.94	1.12	1.12
unemployed=5-10	0.91	0.84	0.96	1.12	1.25
unemployed=>10	4.18	0.9	3.52	1.17	1.99
wall_material=Asbestos	1.39	0.29	2.07	0.83	1.27
wall_material=Blocks	1.23	1.16	1.07	0.92	1.13
wall_material=Brick	1.11	1.17	1.12	1.09	0.94
wall_material=Concrete	1.19	0.46	0.99	0.63	1.27
wall_material=Daub/Mud/Clay	0.69	0.75	0.73	0.72	0.81
wall_material=IBR/Corr.Iron/Zinc	0.81	0.67	0.73	0.78	1.01
wall_material=Plaster	1.03	1.02	1.1	1.28	1.09
wall_material=Plastic	0.45	0.5	1.54	0.77	2.41
wall_material=Tiles	0.19	0.21	0.55	0.27	1.49
wall_material=Wood/Masonite_board	0.58	0.85	0.66	0.75	0.72
water_access=block/street_taps	0.48	0.68	0.57	0.75	0.8
water_access=nearby_river/dam/borehole	0.36	0.53	0.49	0.7	0.72
water_access=tap_in_yard	1.16	0.77	1.24	0.99	1.04
water_access=tap_inside_house	1.58	1.71	1.29	1.26	1.16
years_electrified=0-5yrs	0.71	0.73	0.82	0.83	0.88
years_electrified=10-15yrs	1.22	1.07	1.12	1.12	1.07
years_electrified=15+yrs	1.54	1.44	1.28	1.13	1.33
years_electrified=5-10yrs	0.88	0.98	0.93	1.02	0.86

FIGURE B.6: Odds Ratios filtered for selection of township, 15+ years electrified, Gauteng Customer Archetype

B.7 Informal settlement, 0 - 5 years electrified, Easter Cape

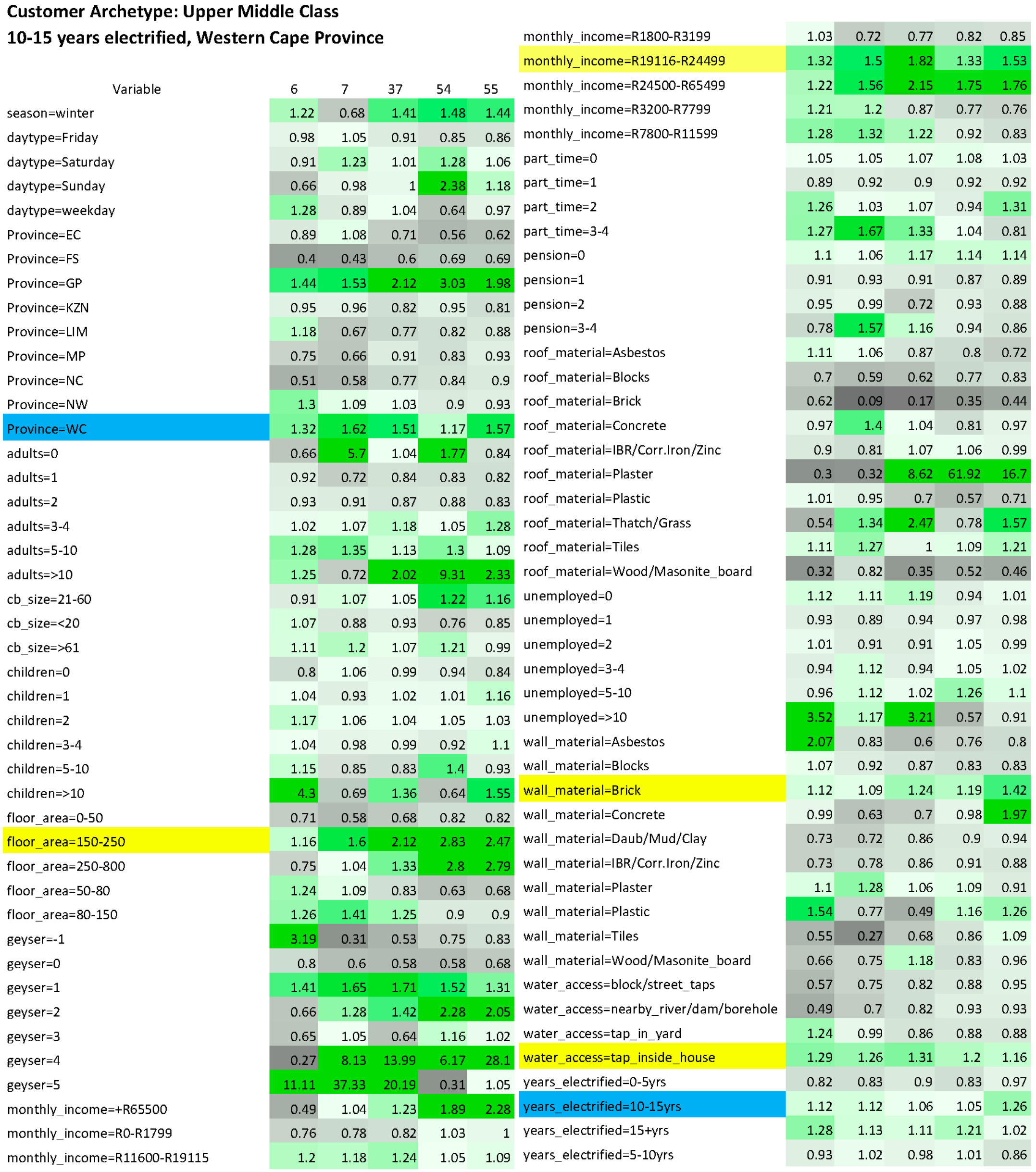


Customer Archetype: Informal Settlement
0 - 5 years electrified, Easter Cape Province

Variable	39	45	46	49	50	53
season=winter	1.59	1.21	1.09	2.38	0.41	0.96
daytype=Friday	0.98	1.08	0.93	0.97	0.94	1.35
daytype=Saturday	0.96	0.88	0.94	1.07	0.97	0.71
daytype=Sunday	0.87	0.81	0.9	1.1	1.02	0.49
daytype=weekday	1.1	1.13	1.11	0.94	1.03	1.43
Province=EC	1.29	1.57	1.21	1.35	1.31	1.44
Province=FS	1.14	2.11	1.2	1.76	1.52	1.71
Province=GP	0.72	0.8	0.73	0.7	0.73	0.7
Province=KZN	1.01	0.76	1	0.93	0.8	0.98
Province=LIM	1.1	0.91	1.08	1.23	1.44	1.04
Province=MP	1.25	1.08	1.22	1.11	1.14	0.98
Province=NC	1.21	2.06	1.83	1.02	1.41	2.14
Province=NW	1	1.01	0.88	1.15	1.27	0.98
Province=WC	0.78	0.83	0.79	0.76	0.79	0.73
adults=0	0.51	0.71	0.74	0.61	0.58	0.69
adults=1	1.18	1.36	1.22	1.28	1.4	1.4
adults=2	1.13	1.02	1.06	1.07	1	1
adults=3-4	0.93	0.91	0.91	0.92	0.96	0.9
adults=5-10	0.8	0.87	0.9	0.8	0.79	0.9
adults=>10	0.76	0.58	0.58	0.79	0.55	0.41
cb_size=21-60	0.93	1.12	0.92	1	1.06	1.11
cb_size=<20	1.1	0.92	1.07	1.06	0.98	0.91
cb_size=>61	0.93	0.89	1.1	0.82	0.88	0.94
children=0	1.16	1.03	1.17	1.09	1.15	1.01
children=1	1.01	1	0.96	0.95	0.9	1
children=2	0.87	0.91	0.89	0.95	0.95	0.95
children=3-4	0.97	1.07	1	1.12	1.05	1.01
children=5-10	0.89	1.08	0.86	0.72	0.79	1.17
children=>10	0.55	0.6	0.66	0.5	0.65	0.52
floor_area=0-50	1.32	1.27	1.43	1.23	1.26	1.22
floor_area=150-250	0.84	0.86	0.79	0.88	0.84	0.86
floor_area=250-800	0.85	0.88	0.86	0.87	0.89	0.79
floor_area=50-80	0.97	1	0.91	1.01	1	1.04
floor_area=80-150	0.85	0.84	0.84	0.86	0.85	0.87
geyser=-1	0.54	0.65	0.61	0.5	0.53	0.59
geyser=0	1.31	1.29	1.33	1.3	1.32	1.47
geyser=1	0.77	0.78	0.75	0.77	0.76	0.69
geyser=2	0.87	0.89	0.87	0.89	0.87	0.81
geyser=3	0.91	0.95	0.95	0.93	0.95	0.9
geyser=4	0.88	0.9	0.87	0.94	0.9	0.82
geyser=5	0.92	0.93	0.92	0.96	0.93	0.85
monthly_income=+R65500	0.9	0.92	0.9	0.93	0.91	0.84
monthly_income=R0-R1799	1.33	1.27	1.29	1.42	1.38	1.19
monthly_income=R11600-R19115	0.81	0.81	0.84	0.8	0.81	0.79
monthly_income=R1800-R3199	1.11	1.06	1.08	1.09	1.08	1.11
monthly_income=R19116-R24499	0.79	0.82	0.81	0.77	0.79	0.76
monthly_income=R24500-R65499	0.84	0.86	0.84	0.86	0.87	0.82
monthly_income=R3200-R7799	0.86	0.9	0.89	0.8	0.82	1.02
monthly_income=R7800-R11599	0.77	0.84	0.77	0.73	0.76	0.83
part_time=0	0.89	0.86	0.93	0.9	0.95	0.9
part_time=1	1.14	1.15	1.1	1.16	1.13	1.14
part_time=2	0.99	1.03	0.96	0.91	0.76	0.91
part_time=3-4	0.98	1.81	0.96	0.8	0.62	1.46
pension=0	0.96	0.94	0.93	0.86	0.92	1.12
pension=1	1.01	1.07	1	1.09	1.08	0.88
pension=2	1.13	1.03	1.37	1.39	1.1	0.97
pension=3-4	1	0.73	1.26	0.88	0.86	0.91
roof_material=Asbestos	0.89	0.82	0.93	0.75	0.79	0.85
roof_material=Blocks	0.68	0.76	0.69	0.69	0.72	0.64
roof_material=Brick	1.08	0.86	1.44	0.86	0.85	0.78
roof_material=Concrete	0.81	1	0.98	0.86	0.89	0.84
roof_material=IBR/Corr.Iron/Zinc	1.32	1.36	1.29	1.46	1.4	1.44
roof_material=Plaster	0.84	0.93	0.7	0.95	0.92	0.85
roof_material=Plastic	1.22	0.8	0.86	0.87	0.74	0.95
roof_material=Thatch/Grass	1.98	1.15	1.36	1.77	2.24	1.46
roof_material=Tiles	0.76	0.78	0.77	0.74	0.76	0.72
roof_material=Wood/Masonite_board	1.17	0.78	0.75	0.77	0.76	0.69
unemployed=0	0.94	1.02	0.97	0.96	0.94	0.99
unemployed=1	1.06	1.03	1.02	1.04	1.1	1.07
unemployed=2	1.04	1.04	1.05	1.06	1.02	1.01
unemployed=3-4	1.01	0.96	0.97	1.01	0.99	0.96
unemployed=5-10	0.87	0.75	0.93	0.79	0.8	0.81
unemployed=>10	0.56	0.55	0.56	0.2	0.34	0.51
wall_material=Asbestos	0.63	0.66	0.83	0.49	0.55	0.64
wall_material=Blocks	0.92	0.94	0.96	0.89	1	0.96
wall_material=Brick	0.94	0.86	0.97	0.87	0.81	0.87
wall_material=Concrete	0.62	0.85	0.72	0.52	0.7	1.08
wall_material=Daub/Mud/Clay	1.38	1.56	1.3	1.8	1.72	1.41
wall_material=IBR/Corr.Iron/Zinc	1.48	1.99	1.51	1.3	1.16	1.94
wall_material=Plaster	0.86	0.83	0.84	0.83	0.83	0.86
wall_material=Plastic	1.12	0.62	1.17	0.95	0.62	0.53
wall_material=Tiles	10.89	0.87	63.85	0.86	0.86	0.79
wall_material=Wood/Masonite_board	1.85	1.38	1.73	1.33	1.29	1.33
water_access=block/street_taps	1.33	1.61	1.43	1.76	1.71	1.5
water_access=nearby_river/dam/borehole	1.82	1.95	1.33	3.36	3.33	2.49
water_access=tap_in_yard	1.13	0.97	1.1	0.91	0.94	1.06
water_access=tap_inside_house	0.7	0.7	0.72	0.65	0.64	0.66
years_electrified=0-5yrs	1.18	1.21	1.12	1.65	1.62	1.29
years_electrified=10-15yrs	0.88	0.87	0.96	0.77	0.85	0.92
years_electrified=15+yrs	0.87	0.81	0.87	0.66	0.63	0.72
years_electrified=5-10yrs	1.03	1.08	1.03	0.99	0.99	1.07

FIGURE B.7: Odds Ratios filtered for selection of informal settlement, 0 - 5 years electrified, Easter Cape Customer Archetype

B.8 Upper middle class, 10 - 15 years electrified, Western Cape



Customer Archetype: Upper Middle Class
10-15 years electrified, Western Cape Province

Variable	6	7	37	54	55
season=winter	1.22	0.68	1.41	1.48	1.44
daytype=Friday	0.98	1.05	0.91	0.85	0.86
daytype=Saturday	0.91	1.23	1.01	1.28	1.06
daytype=Sunday	0.66	0.98	1	2.38	1.18
daytype=weekday	1.28	0.89	1.04	0.64	0.97
Province=EC	0.89	1.08	0.71	0.56	0.62
Province=FS	0.4	0.43	0.6	0.69	0.69
Province=GP	1.44	1.53	2.12	3.03	1.98
Province=KZN	0.95	0.96	0.82	0.95	0.81
Province=LIM	1.18	0.67	0.77	0.82	0.88
Province=MP	0.75	0.66	0.91	0.83	0.93
Province=NC	0.51	0.58	0.77	0.84	0.9
Province=NW	1.3	1.09	1.03	0.9	0.93
Province=WC	1.32	1.62	1.51	1.17	1.57
adults=0	0.66	5.7	1.04	1.77	0.84
adults=1	0.92	0.72	0.84	0.83	0.82
adults=2	0.93	0.91	0.87	0.88	0.83
adults=3-4	1.02	1.07	1.18	1.05	1.28
adults=5-10	1.28	1.35	1.13	1.3	1.09
adults=>10	1.25	0.72	2.02	9.31	2.33
cb_size=21-60	0.91	1.07	1.05	1.22	1.16
cb_size=<20	1.07	0.88	0.93	0.76	0.85
cb_size=>61	1.11	1.2	1.07	1.21	0.99
children=0	0.8	1.06	0.99	0.94	0.84
children=1	1.04	0.93	1.02	1.01	1.16
children=2	1.17	1.06	1.04	1.05	1.03
children=3-4	1.04	0.98	0.99	0.92	1.1
children=5-10	1.15	0.85	0.83	1.4	0.93
children=>10	4.3	0.69	1.36	0.64	1.55
floor_area=0-50	0.71	0.58	0.68	0.82	0.82
floor_area=150-250	1.16	1.6	2.12	2.83	2.47
floor_area=250-800	0.75	1.04	1.33	2.8	2.79
floor_area=50-80	1.24	1.09	0.83	0.63	0.68
floor_area=80-150	1.26	1.41	1.25	0.9	0.9
geyser=-1	3.19	0.31	0.53	0.75	0.83
geyser=0	0.8	0.6	0.58	0.58	0.68
geyser=1	1.41	1.65	1.71	1.52	1.31
geyser=2	0.66	1.28	1.42	2.28	2.05
geyser=3	0.65	1.05	0.64	1.16	1.02
geyser=4	0.27	8.13	13.99	6.17	28.1
geyser=5	11.11	37.33	20.19	0.31	1.05
monthly_income=+R65500	0.49	1.04	1.23	1.89	2.28
monthly_income=R0-R1799	0.76	0.78	0.82	1.03	1
monthly_income=R11600-R19115	1.2	1.18	1.24	1.05	1.09
monthly_income=R1800-R3199	1.03	0.72	0.77	0.82	0.85
monthly_income=R19116-R24499	1.32	1.5	1.82	1.33	1.53
monthly_income=R24500-R65499	1.22	1.56	2.15	1.75	1.76
monthly_income=R3200-R7799	1.21	1.2	0.87	0.77	0.76
monthly_income=R7800-R11599	1.28	1.32	1.22	0.92	0.83
part_time=0	1.05	1.05	1.07	1.08	1.03
part_time=1	0.89	0.92	0.9	0.92	0.92
part_time=2	1.26	1.03	1.07	0.94	1.31
part_time=3-4	1.27	1.67	1.33	1.04	0.81
pension=0	1.1	1.06	1.17	1.14	1.14
pension=1	0.91	0.93	0.91	0.87	0.89
pension=2	0.95	0.99	0.72	0.93	0.88
pension=3-4	0.78	1.57	1.16	0.94	0.86
roof_material=Asbestos	1.11	1.06	0.87	0.8	0.72
roof_material=Blocks	0.7	0.59	0.62	0.77	0.83
roof_material=Brick	0.62	0.09	0.17	0.35	0.44
roof_material=Concrete	0.97	1.4	1.04	0.81	0.97
roof_material=IBR/Corr.Iron/Zinc	0.9	0.81	1.07	1.06	0.99
roof_material=Plaster	0.3	0.32	8.62	61.92	16.7
roof_material=Plastic	1.01	0.95	0.7	0.57	0.71
roof_material=Thatch/Grass	0.54	1.34	2.47	0.78	1.57
roof_material=Tiles	1.11	1.27	1	1.09	1.21
roof_material=Wood/Masonite_board	0.32	0.82	0.35	0.52	0.46
unemployed=0	1.12	1.11	1.19	0.94	1.01
unemployed=1	0.93	0.89	0.94	0.97	0.98
unemployed=2	1.01	0.91	0.91	1.05	0.99
unemployed=3-4	0.94	1.12	0.94	1.05	1.02
unemployed=5-10	0.96	1.12	1.02	1.26	1.1
unemployed=>10	3.52	1.17	3.21	0.57	0.91
wall_material=Asbestos	2.07	0.83	0.6	0.76	0.8
wall_material=Blocks	1.07	0.92	0.87	0.83	0.83
wall_material=Brick	1.12	1.09	1.24	1.19	1.42
wall_material=Concrete	0.99	0.63	0.7	0.98	1.97
wall_material=Daub/Mud/Clay	0.73	0.72	0.86	0.9	0.94
wall_material=IBR/Corr.Iron/Zinc	0.73	0.78	0.86	0.91	0.88
wall_material=Plaster	1.1	1.28	1.06	1.09	0.91
wall_material=Plastic	1.54	0.77	0.49	1.16	1.26
wall_material=Tiles	0.55	0.27	0.68	0.86	1.09
wall_material=Wood/Masonite_board	0.66	0.75	1.18	0.83	0.96
water_access=block/street_taps	0.57	0.75	0.82	0.88	0.95
water_access=nearby_river/dam/borehole	0.49	0.7	0.82	0.93	0.93
water_access=tap_in_yard	1.24	0.99	0.86	0.88	0.88
water_access=tap_inside_house	1.29	1.26	1.31	1.2	1.16
years_electrified=0-5yrs	0.82	0.83	0.9	0.83	0.97
years_electrified=10-15yrs	1.12	1.12	1.06	1.05	1.26
years_electrified=15+yrs	1.28	1.13	1.11	1.21	1.02
years_electrified=5-10yrs	0.93	1.02	0.98	1.01	0.86

FIGURE B.8: Odds Ratios filtered for selection of upper middle class, 10 - 15 years electrified, Western Cape Customer Archetype

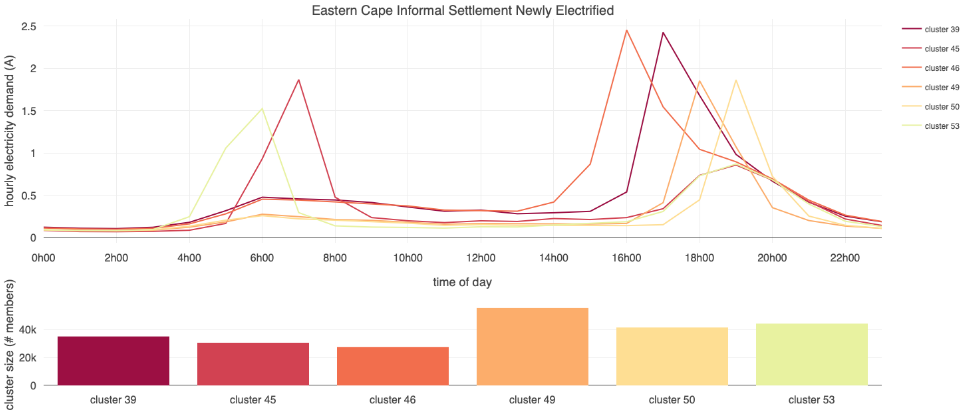
Appendix C

Additional Customer Archetypes

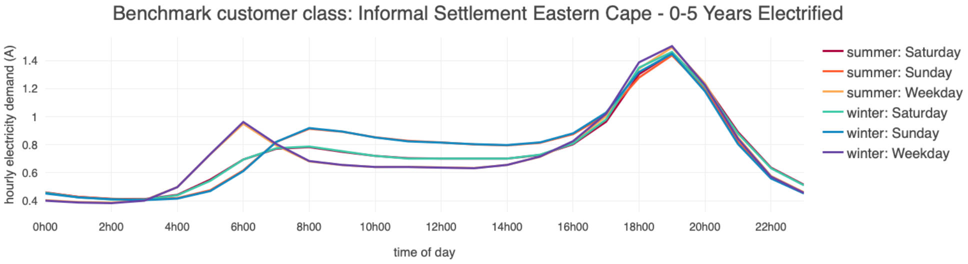
C.1 Newly Electrified, Informal Settlement, Easter Cape

Winter		**Summer**	
Cluster	**Daytype**	**Cluster**	**Daytype**
39	weekday	**50**	any
45	weekday, Friday	**53**	weekday, Friday
46	weekday		
49	Saturday, Sunday		

TABLE C.1: Temporal attributes of clusters for archetype in Fig C.1a



(A) RDLPs for customer archetype



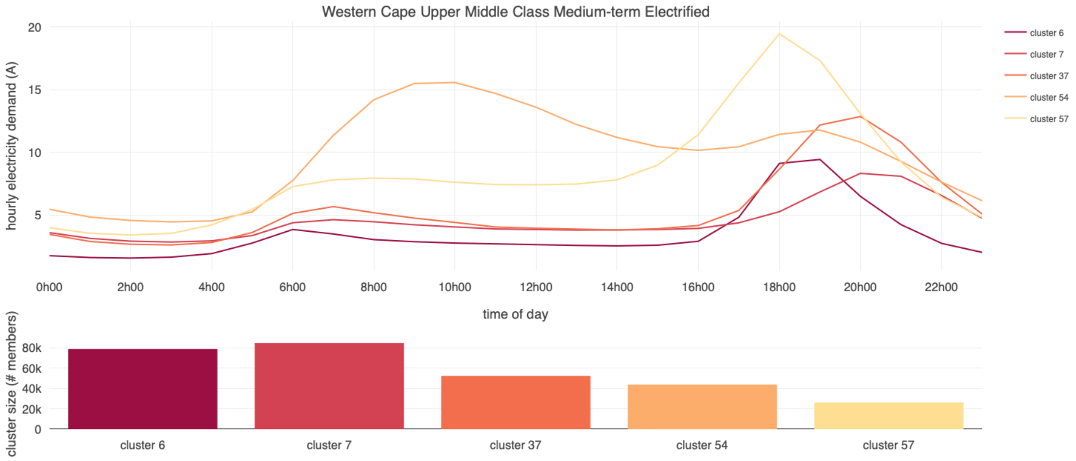
(B) Benchmark

FIGURE C.1: Newly electrified informal settlement household in the Eastern Cape

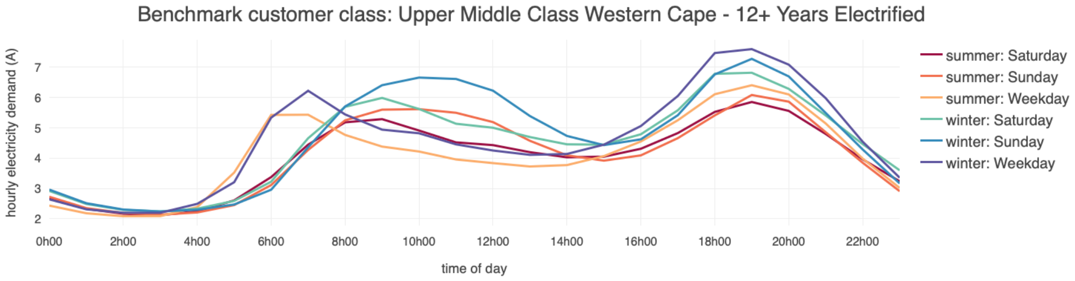
C.2 Long-term Electrified, Upper Middle Class, Western Cape

Winter		Summer	
Cluster	**Daytype**	**Cluster**	**Daytype**
6	weekday	**7**	Saturday
37	any		
54	Saturday, Sunday		
57	Saturday, Sunday		

TABLE C.2: Temporal attributes of clusters for archetype in Fig C.2a



(A) RDLPs for customer archetype



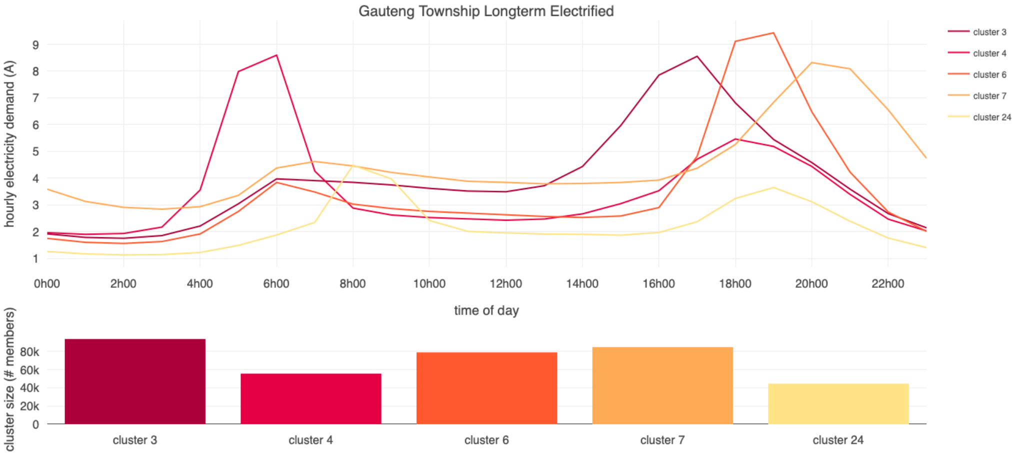
(B) Benchmark

FIGURE C.2: Long-term electrified upper middle class household in the Western Cape

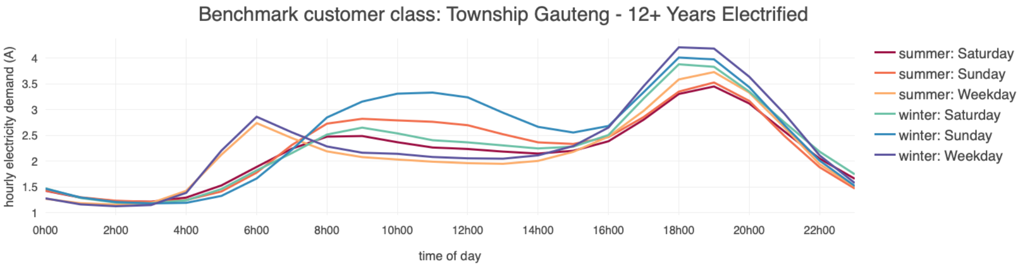
C.3 Long-term Electrified Township Household, Gauteng

Winter		**Summer**	
Cluster	**Daytype**	**Cluster**	**Daytype**
3	weekday	**4**	weekday, Friday
6	weekday	**7**	Saturday
		24	Saturday, Sunday

TABLE C.3: Temporal attributes of clusters for archetype in Fig C.3a



(A) RDLPs for Customer Archetype



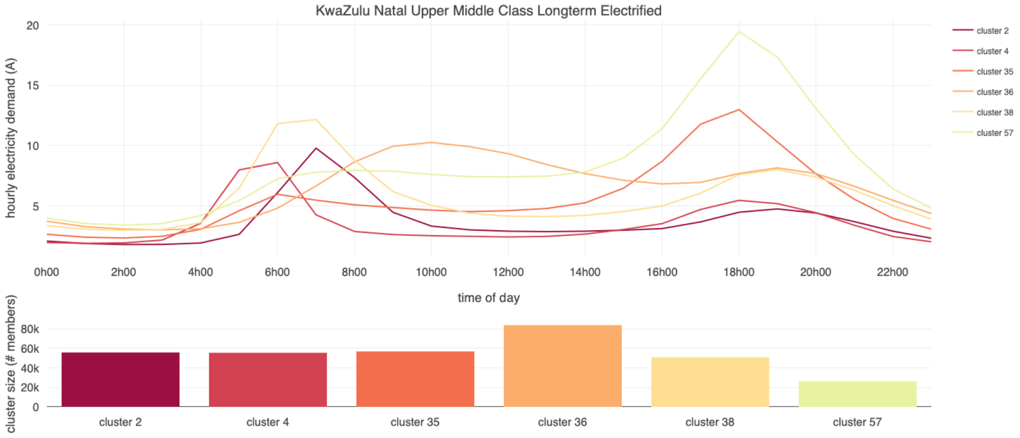
(B) Benchmark

FIGURE C.3: Long-term electrified township household in Gauteng

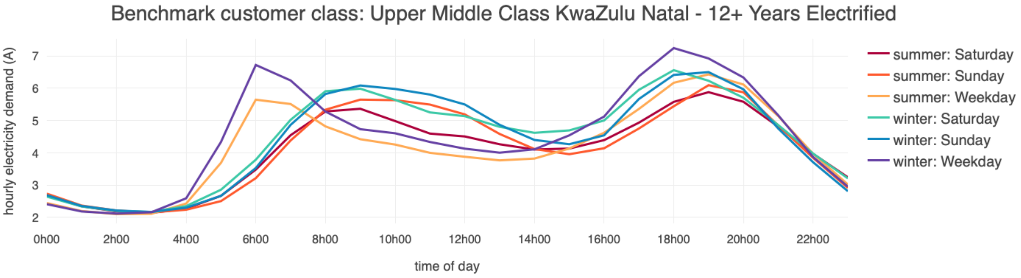
C.4 Long-term Electrified Upper Middle Class Household, KwaZulu Natal

Winter		**Summer**	
Cluster	**Daytype**	**Cluster**	**Daytype**
35	weekday	**2**	Friday, Sunday
36	Saturday, Sunday	**4**	weekday, Friday
38	weekday, Friday		
57	Sunday		

TABLE C.4: Temporal attributes of clusters for archetype in Fig C.4a



(A) RDLPs for Customer Archetype



(B) Benchmark

FIGURE C.4: Long-term electrified upper middle class household in KwaZulu Natal

Bibliography

[1] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Ying Wah. "Time-series clustering – A decade review". In: *Inf. Syst.* 53 (2015), pp. 16–38. DOI: `10.1016/j.is.2015.04.007`. URL: `http://dx.doi.org/10.1016/j.is.2015.04.007`.

[2] Adrian Albert and Mehdi Maasoumy. "Predictive Segmentation of Energy Consumers". In: (2016). URL: `https://web.stanford.edu/{~}adalbert/papers/segmentation.pdf`.

[3] Kadir Amasyali and Nora M. El-Gohary. "A review of data-driven building energy consumption prediction studies". In: *Renew. Sustain. Energy Rev.* 81.March 2017 (2018), pp. 1192–1205. ISSN: 18790690. DOI: `10.1016/j.rser.2017.04.095`. URL: `http://dx.doi.org/10.1016/j.rser.2017.04.095`.

[4] F. M. Andersen, H. V. Larsen, and T. K. Boomsma. "Long-term forecasting of hourly electricity load: Identification of consumption profiles and segmentation of customers". In: *Energy Convers. Manag.* 68 (2013), pp. 244–252. ISSN: 01968904. DOI: `10.1016/j.enconman.2013.01.018`.

[5] Gustavo E.A.P.A. Batista et al. "CID: an efficient complexity-invariant distance for time series". In: *Data Min. Knowl. Discov. Springer* (2013). ISSN: 1011601X. DOI: `10.1007/s10618-013-0312-3`.

[6] Florentin Batrinu et al. "Efficient iterative refinement clustering for electricity customer classification". In: *2005 IEEE Russ. Power Tech, PowerTech* (2005), pp. 1–7. DOI: `10.1109/PTC.2005.4524366`.

[7] Hui Ben and Koen Steemers. "Household archetypes and behavioural patterns in UK domestic energy use". In: *Energy Effic.* 11.3 (2018), pp. 761–771. ISSN: 15706478. DOI: `10.1007/s12053-017-9609-1`.

[8] James C Bezdek and Nikhil R Pal. "Some New Indexes of Cluster Validity". In: 28.3 (1998), pp. 301–315.

[9] S. M. Bidoki et al. "Evaluating different clustering techniques for electricity customer classification". In: *2010 IEEE PES Transm. Distrib. Conf. Expo. Smart Solut. a Chang. World* (2010), pp. 1–5. DOI: `10.1109/TDC.2010.5484234`.

[10] Jacques Booysen, Marcus Dekenah, and Schalk Heunis. *Research Support for GLF*. Tech. rep. Johannesburg: Eskom Holdings Limited, 2013.

[11] Hong An Cao, Christian Beckel, and Thorsten Staake. "Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns". In: *IECON Proc. (Industrial Electron. Conf.* (2013), pp. 4733–4738. ISSN: 1553-572X. DOI: `10.1109/IECON.2013.6699900`.

[12] Charalampos Chelmis. "Big Data Analytics for Demand Response : Clustering Over Space and Time". In: *2015 IEEE Int. Conf. Big Data (Big Data)* (2015), pp. 2223–2232. DOI: `10.1109/BigData.2015.7364011`.

[13] G. Chicco et al. "Customer Characterization Options for Improving the Tariff Offer". In: *IEEE Power Eng. Rev.* 22.11 (2002), p. 60. ISSN: 02721724. DOI: `10.1109/MPER.2002.4311841`.

[14] Gianfranco Chicco. "Overview and performance assessment of the clustering methods for electrical load pattern grouping". In: *Energy* 42.1 (2012), pp. 68–80. ISSN: 03605442. DOI: `10.1016/j.energy.2011.12.031`. URL: `http://dx.doi.org/10.1016/j.energy.2011.12.031`.

[15] Gianfranco Chicco, Roberto Napoli, and Federico Piglione. "A Review of Concepts and Techniques for Emergent Customer Categorisation". In: (2002), pp. 51–58. URL: `http://porto.polito.it/1836917/`.

[16] Gianfranco Chicco, Roberto Napoli, and Federico Piglione. "Application of clustering algorithms and Self Organising Maps to classify electricity customers". In: *2003 IEEE Bol. PowerTech - Conf. Proc.* 1 (2003), pp. 373–379. ISSN: 00448486. DOI: `10.1109/PTC.2003.1304160`.

[17] Gianfranco Chicco, Roberto Napoli, and Federico Piglione. "Comparison Among Clustering Techniques for Electricity Customer Classification". In: *IEEE Trans. POWER Syst.* 21.2 (2006), pp. 1–7. DOI: `10.1109/TPWRS.2006.873122`.

[18] N. Cross and C. T. Gaunt. "Application of rural residential hourly load curves in energy modelling". In: *2003 IEEE Bol. PowerTech - Conf. Proc.* 3 (2003), pp. 853–857. DOI: `10.1109/PTC.2003.1304492`.

[19] The-Hien Dang-Ha, Roland Olsson, and Hao Wang. "Clustering Methods for Electricity Consumers: An Empirical Study in Hvaler-Norway". In: *NIK-2017* (2017). arXiv: `1703.02502`. URL: `http://arxiv.org/abs/1703.02502`.

[20] David L. Davies and Donald W. Bouldin. "A Cluster Separation Measure". In: *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-1.2 (1979), pp. 224–227. ISSN: 01628828. DOI: `10.1109/TPAMI.1979.4766909`.

[21] M Dekenah. "a Load Profile Prediction Model for Residential". In: *energize* May (2010), pp. 46 –49.

[22] Marcus Dekenah. *Guide to Regionalised Domestic Subclasses*. Tech. rep. Johannesburg: M Dekenah Consulting, 2014, pp. 1–45.

[23] Ian Dent and B Sc Hons. "Deriving knowledge of household behaviour from domestic electricity usage metering". In: July (2015). URL: `http://ima.ac.uk/wp-content/uploads/2014/12/thesis{\_}master.pdf`.

[24] Ian Dent et al. "An Approach for Assessing Clustering of Households by Electricity Usage". In: (2014). ISSN: 1556-5068. DOI: `10.2139/ssrn.2828465`. arXiv: `1409.0718`. URL: `http://arxiv.org/abs/1409.0718`.

[25] Ian Dent et al. "Variability of behaviour in electricity load profile clustering; Who does things at the same time each day?" In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 8557 LNAI

(2014), pp. 70–84. ISSN: 16113349. DOI: 10.1007/978-3-319-08976-8_6. arXiv: arXiv:1409.1043v1.

[26] Hui Ding et al. "Querying and Mining of Time Series Data : Experimental Comparison of Representations and Distance Measures". In: *PVLDB* (2008). URL: https://www.researchgate.net/profile/Peter{_}Scheuermann/publication/220538811{_}Querying{_}and{_}mining{_}of{_}time{_}series{_}data{_}Experimental{_}comparison{_}of{_}representations{_}and{_}distance{_}measures/links/53d7c8dc0cf2631430bfc363.pdf.

[27] C Doyle et al. "USING LOAD RESEARCH DATA TO ASSESS DEMAND SIDE MANAGEMENT INTERVENTIONS". In: (2005).

[28] J Du Toit et al. "Customer Segmentation Using Unsupervised Learning on Daily Energy Load Profiles". In: *J. Adv. Inf. Technol.* 7.2 (2016), pp. 69–75. DOI: 10.12720/jait.7.2.69-75. URL: http://www.jait.us/uploadfile/2016/0505/20160505105403530.pdf.

[29] Elexon. "Load Profiles and their use in Electricity Settlement". In: *Profiling* November 2013 (2013), pp. 1–31. URL: http://www.elexon.co.uk/wp-content/uploads/2013/11/load{_}profiles{_}v2.0{_}cgi.pdf.

[30] University of Cape Town Eskom Stellenbosch University. *Domestic Electrical Load Metering-Secure Data 1994-2014. version 1*. 2019. DOI: 10.25828/p3k7-r965. URL: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/760.

[31] University of Cape Town Eskom Stellenbosch University. *Domestic Electrical Load Survey 1994-2014. version 1*. 2019. DOI: 10.25828/kzer-gd88. URL: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/755.

[32] Eugene A Feinberg and Dora Genethliou. "Load Forecasting". In: *Appl. Math. Restructured Electr. Power Syst.* (2006), pp. 269–285. ISSN: 0018-9286. DOI: 10.1007/0-387-23471-3_12.

[33] Vera Figueiredo et al. "An electric energy consumer characterization framework based on data mining techniques". In: *IEEE Trans. Power Syst.* 20.2 (2005), pp. 596–602. ISSN: 08858950. DOI: 10.1109/TPWRS.2005.846234.

[34] Tak Chung Fu. "A review on time series data mining". In: *Eng. Appl. Artif. Intell.* 24.1 (2011), pp. 164–181. ISSN: 09521976. DOI: 10.1016/j.engappai.2010.09.007. arXiv: 1011.1669. URL: http://dx.doi.org/10.1016/j.engappai.2010.09.007.

[35] *Geo-based Load Forecast Standard*. Tech. rep. June 2012. Johannesburg: Eskom, 2012.

[36] A. Grandjean, J. Adnot, and G. Binet. "A review and an analysis of the residential electric load curve models". In: *Renew. Sustain. Energy Rev.* 16.9 (2012), pp. 6539–6565. ISSN: 13640321. DOI: 10.1016/j.rser.2012.08.013. URL: http://dx.doi.org/10.1016/j.rser.2012.08.013.

[37] Ramon Granell, Colin J Axon, and David C H Wallom. “Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles”. In: *IEEE Trans. Power Syst.* 30.6 (2015), pp. 3217–3224. DOI: `10.1109/TPWRS.2014.2377213`.

[38] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Vol. 27. 2. 2001. ISBN: 978-0-387-84857-0. DOI: `10.1198/jasa.2004.s339`. arXiv: `1010.3003`. URL: `http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf{\%}255Cnhttp://www-stat.stanford.edu/{~}tibs/book/preface.ps`.

[39] Marian Hayn, Valentin Bertsch, and Wolf Fichtner. “Electricity load profiles in Europe: The importance of household segmentation”. In: *Energy Res. Soc. Sci.* 3.C (2014), pp. 30–45. ISSN: 22146296. DOI: `10.1016/j.erss.2014.07.002`. URL: `http://dx.doi.org/10.1016/j.erss.2014.07.002`.

[40] Schalk Heunis and Marcus Dekenah. “A load profile prediction model for residential consumers in South Africa”. Johannesburg, 2014.

[41] Schalk Heunis and Marcus Dekenah. *Manual for Eskom Distribution Pre- Electrification Tool (DPET)*. Johannesburg, 2014.

[42] Schalk Heunis and Ron Herman. “A Probabilistic Model for Residential Consumer Loads”. In: *IEEE Trans. Power Syst.* 17.3 (2002), pp. 621–625.

[43] Félix Iglesias and Wolfgang Kastner. “Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns”. In: *Energies* 6.2 (2013), pp. 579–597. ISSN: 19961073. DOI: `10.3390/en6020579`.

[44] Jiawei Han, Micheline Kamber, and Jain Pei. *Data Mining – Concepts & Techniques*. Third. Morgan Kaufmann Publishers, 2012, pp. 1–744. ISBN: 9780123814791. DOI: `10.1016/B978-0-12-381479-1.00001-0`. arXiv: `arXiv:1011.1669v3`.

[45] Ling Jin et al. “Comparison of Clustering Techniques for Residential Energy Behavior Using Smart Meter Data”. In: *AAAI Work. Artif. Intell. Smart Grids Smart Build.* (2017), pp. 260–266.

[46] Ling Jin et al. “Load Shape Clustering Using Residential Smart Meter Data : a Technical Memorandum”. In: September (2016), pp. 1–15.

[47] Eamonn Keogh and Shruti Kasetty. “On the need for time series data mining benchmarks”. In: *Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '02* (2002), p. 102. ISSN: 13845810. DOI: `10.1145/775047.775062`. URL: `http://portal.acm.org/citation.cfm?doid=775047.775062`.

[48] Eamonn Keogh and Michael Pazzani. “Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases”. In: *KAIS* (2000), pp. 1–19.

[49] Caroline Kleist. “Time Series Data Mining Methods : A Review”. In: 533039 (2015).

[50] Teuvo Kohonen. “The Self-organizing Map”. In: *Proc. IEEE* 78.9 (1990), pp. 1464–1480.

[51] Jungsuk Kwac, June Flora, and Ram Rajagopal. "Household energy consumption segmentation using hourly data". In: *IEEE Trans. Smart Grid* 5.1 (2014), pp. 420–430. ISSN: 19493053. DOI: `10.1109/TSG.2013.2278477`.

[52] Jungsuk Kwac et al. "Utility customer segmentation based on smart meter data: Empirical study". In: *2013 IEEE Int. Conf. Smart Grid Commun.* October (2013), pp. 720–725. ISSN: 978-1-4799-1526-2. DOI: `10.1109/SmartGridComm.2013.6688044`. URL: `http://ieeexplore.ieee.org/document/6688044/`.

[53] Peter Laurinec et al. "Adaptive Time Series Forecasting of Energy Consumption using Optimized Cluster Analysis". In: *Icdm* (2016). DOI: `10.1109/ICDMW.2016.159`.

[54] T. Warrent Liao. "Clustering of time series data—a survey". In: *Pattern Recognit.* 38 (2005), pp. 1857 –1874. URL: `https://ac.els-cdn.com/S0031320305001305/1-s2.0-S0031320305001305-main.pdf?{\_}tid=b8457f30-2282-4560-891d-2f9d32fb56bb{\&}acdnat=1541928424{\_}87460d61482c6d900d7a7a3969e99140`.

[55] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. "A clustering approach to domestic electricity load profile characterisation using smart metering data". In: *Appl. Energy* 141 (2015), pp. 190–199. ISSN: 03062619. DOI: `10.1016/j.apenergy.2014.12.039`. URL: `http://dx.doi.org/10.1016/j.apenergy.2014.12.039`.

[56] Vincent Micali and Schalk Heunis. "Statistical methods for time-of-use classifications". In: (2010).

[57] Clayton Miller, Zoltán Nagy, and Arno Schlueter. "A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings". In: *Renew. Sustain. Energy Rev.* 81.December 2018 (2018), pp. 1365–1377. ISSN: 1364-0321. DOI: `10.1016/j.rser.2017.05.124`. URL: `http://dx.doi.org/10.1016/j.rser.2017.05.124`.

[58] Steven Karl Morley. "Alternatives to accuracy and bias metrics based on percentage errors for radiation belt modeling applications". In: 01 (2016). DOI: `10.2172/1260362`. URL: `http://www.osti.gov/servlets/purl/1260362/`.

[59] Jukka V. Paatero and Peter D. Lund. "A model for generating household electricity load profiles". In: *Int. J. Energy Res.* 30.5 (2006), pp. 273–290. ISSN: 0363907X. DOI: `10.1002/er.1136`.

[60] Ioannis P. Panapakidis and Georgios C. Christoforidis. "Optimal Selection of Clustering Algorithm via Multi-Criteria Decision Analysis (MCDA) for Load Profiling Applications". In: *Appl. Sci.* 8.2 (2018), p. 237. ISSN: 2076-3417. DOI: `10.3390/app8020237`. URL: `http://www.mdpi.com/2076-3417/8/2/237`.

[61] S Ramos et al. "Typical Load Profiles in the Smart Grid Context – A Clustering Methods Comparison". In: *2012 IEEE Power Energy Soc. Gen. Meet.* (2012), pp. 1–8. DOI: `10.1109/PESGM.2012.6345565`.

[62] Teemu Räsänen et al. "Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity

use data". In: *Appl. Energy* 87.11 (2010), pp. 3538–3545. ISSN: 03062619. DOI: `10.1016/j.apenergy.2010.05.015`.

[63] Joshua D. Rhodes et al. "Clustering analysis of residential electricity demand profiles". In: *Appl. Energy* 135 (2014), pp. 461–471. ISSN: 03062619. DOI: `10.1016/j.apenergy.2014.08.111`. URL: `http://dx.doi.org/10.1016/j.apenergy.2014.08.111`.

[64] Andrew Rosenberg and Julia Hirschberg. "V-Measure : A conditional entropy-based external cluster evaluation measure". In: June (2007), pp. 410–420.

[65] Warren S. Sarle, Anil K. Jain, and Richard C. Dubes. *Algorithms for Clustering Data*. 1990. DOI: `10.2307/1268876`. arXiv: `tesxx`. URL: `http://www.jstor.org/stable/1268876?origin=crossref`.

[66] Joan Serrà and Josep Ll Arcos. "An Empirical Evaluation of Similarity Measures for Time Series Classification". In: (2014). URL: `https://arxiv.org/pdf/1401.3973.pdf`.

[67] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning : From Theory to Algorithms*. Cambridge University Press, 2014. ISBN: 9781107057135. URL: `http://www.cs.huji.ac.il/{~}shais/UnderstandingMachineLearning`.

[68] Janet Stephenson et al. "Energy cultures: A framework for understanding energy behaviours". In: *Energy Policy* 38.10 (2010), pp. 6120–6129. ISSN: 03014215. DOI: `10.1016/j.enpol.2010.05.069`. arXiv: `/dx.doi.org/10.1016/j.enpol.2010.05.069 [http:]`. URL: `http://dx.doi.org/10.1016/j.enpol.2010.05.069`.

[69] Lukas G. Swan and V. Ismet Ugursal. "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques". In: *Renew. Sustain. Energy Rev.* 13.8 (2009), pp. 1819–1835. ISSN: 13640321. DOI: `10.1016/j.rser.2008.09.033`.

[70] Thanchanok Teeraratkul, Daniel O'Neill, and Sanjay Lall. "Shape-Based Approach to Household Electric Load Curve Clustering and Prediction". In: *IEEE Trans. Smart Grid* 9.5 (2018). ISSN: 19493053. DOI: `10.1109/TSG.2017.2683461`. arXiv: `1702.01414`.

[71] Wiebke Toussaint. *Domestic Electrical Load Metering, Hourly Data 1994-2014. version 1*. 2019. DOI: `10.25828/56nh-fw77`. URL: `https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/759`.

[72] Wiebke Toussaint. *Domestic Electrical Load Survey - Key Variables 1994-2014. version 1*. 2019. DOI: `10.25828/mf8s-hh79`. URL: `https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/758`.

[73] George J. Tsekouras, Nikos D. Hatziargyriou, and Evangelos N. Dialynas. "Two-stage pattern recognition of load curves for classification of electricity customers". In: *IEEE Trans. Power Syst.* 22.3 (2007), pp. 1120–1128. ISSN: 08858950. DOI: `10.1109/TPWRS.2007.901287`.

[74] Geoffrey K.F. Tso and Kelvin K.W. Yau. "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks".

In: *Energy* 32.9 (2007), pp. 1761–1768. ISSN: 03605442. DOI: `10.1016/j.energy.2006.11.010`.

[75] Juha Vesanto, Esa Alhoniemi, and Student Member. "Clustering of the Self-Organizing Map". In: 11.3 (2000), pp. 586–600.

[76] Joaquim L. Viegas et al. "Classification of new electricity customers based on surveys and smart metering data". In: *Energy* 107 (2016), pp. 804–817. ISSN: 03605442. DOI: `10.1016/j.energy.2016.04.065`.

[77] Joaquim L. Viegas et al. "Electricity demand profile prediction based on household characteristics". In: *Int. Conf. Eur. Energy Mark. EEM* 2015-Augus (2015), pp. 0–4. ISSN: 21654093. DOI: `10.1109/EEM.2015.7216746`.

[78] Seunghye J Wilson. "Data representation for time series data mining : time domain approaches". In: 9.February (2017), pp. 1–6. DOI: `10.1002/wics.1392`.

[79] Sharon Xu, Edward Barbour, and Marta C González. "Household Segmentation by Load Shape and Daily Consumption". In: *Proc. of. ACM SigKDD 2017 Conf.* (2017), pp. 1–9. DOI: `10.475/123`. URL: `http://humnetlab.mit.edu/wordpress/wp-content/uploads/2016/03/household-segmentation-load-shape-consumption.pdf`.

[80] Ying Zhao and George Karypis. *Criterion Functions for Document Clustering: Experiments and Analysis*. Tech. rep. Minneapolis: University of Minnesota, 2001.