LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT
School of Engineering Science
Industrial Engineering and Management
Business Analytics

Mike Mustonen

FINDING UNUSUAL ENERGY CONSUMPTION PROFILES FROM LARGE SCALE DATA

Examiner(s): Professor Pasi Luukka
Post doc Christoph Lohrmann

ABSTRACT

LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT
School of Engineering Science
Industrial Engineering and Management
Business Analytics

Mike Mustonen

Finding unusual energy consumption profiles from large scale data

Master's Thesis

2020

70 pages, 35 figures, 7 tables.

Examiners: Professor Pasi Luukka
Post doc Christoph Lohrmann

Keywords: K-means clustering, K-medoids, Load profiling, Davies-Bouldin index, Silhouette index, Calinski-Habarasz index, Anomaly detection.

Our energy consumption is increasing every year. The new sensor-based smart measurement devices will provide an opportunity to measure this consumption accurately. This study focuses on finding anomalies from time-series consumption data gathered from various buildings operating in the field of grocery and retail business. The K-means clustering algorithm was used to profile different customers based on their consumption patterns. Investigating these profiles with their respective clusters and applying other statistical methods, possible anomalous locations were extracted for further examination for specialists with domain knowledge.

ACKNOWLEDGEMENTS

I would like to thank to me, myself and I for finishing this study and also the supervisors for guidance and fact checking. Time is the most valuable asset in life, live like there is no tomorrow. Peace.

01.05.2020
Mike Mustonen

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ARMA	Autoregressive-moving-average
CCA	Curvilinear Component Analysis
CDI	The Clustering Dispersion Index
DBI	The Davies-Bouldin index
DFT	Discrete Fourier Transforms
DWT	Discrete Wavelet Transforms
EU	European Union
IQR	The Interquartile Range
LSTM	Long Short-Term Memory
MDI	The Modified Dunn index
PAA	Piecewise Constant
PCA	Principal Component Analysis
PLA	Piecewise Linear Approximations
RLM	Robust Linear Model
SAX	Symbolic Aggregate Approximations
SI	The Scatter index
SOM	Self-Organizing Map
SSB	The Average Between-Cluster Sum of Squares
SSW	The Average Within-Cluster Sum of Squares

1 INTRODUCTION

In this chapter, the background for energy consumption is described, problem definition is presented, objectives and delimitations for this study are introduced and the structure of the thesis is given.

1.1 Background

In this fragile and ever-changing world of the 21st century, sustainability has become one of the most important topics. Our demand and consumption for energy have already achieved its limits concerning the preservation of planet earth as we now know it. According to the European Union (EU) [1] the building sector represents 40% of the total energy consumption.

One solution to promote energy efficiency and decrease consumption is installing sensor-based smart meters and monitoring consumption data. Smart meters are meters that can record energy consumption remotely. By investigating this data and finding high consumption profiles or abnormalities, valuable insight can potentially be achieved, and necessary measures could be executed to decrease consumption. A report made by Energiavirasto [2] stated that "by the end of 2018, more than 99 percent of consumption places in Finland had already a smart meter."

According to Motiva [3] there were 3923 grocery stores in Finland in 2011. In 2003, it was estimated that the field of grocery store consumed 1150 GWh of electricity, which is 1.3% of total electricity consumption in Finland. Fourty percent of consumption comes from refrigeration equipment, 25% from lightning and the rest 35% from heating or cooling systems and other consumption sources [3].

Investigating consumption data might reveal imbalances and anomalies and by finding and correcting these anomalies huge potential savings can be made. One approach to gain a deeper understanding of consumption patterns is to obtain load profiles by clustering customers. These clusters can then be used for load forecasting or finding unusual consumption behavior [4]. Understanding consumption behavior is a good starting point. A deeper examination of load profiles within a specific domain is desirable for achieving valuable results.

1.2 Problem definition

A large amount of data is collected and saved from smart meters. The raw data itself is not valuable unless it can be utilized to draw conclusions out of it. In an ideal situation, an automated system would compare measured consumption data and possible other measured variables within a given time interval and create an output of IDs of possible anomalous targets for further investigation. In this study, the consumption data is given in the format of time-series.

Time-series data requires a notable amount of analysis and modification to obtain valuable insights due to its high dimensionality. A problem arises, when trying to find abnormal consumption patterns or anomalies manually, since this is extremely time consuming and thus not efficient. It might be possible to build an automated system that will help end-users to manually investigate a subset of instances that are considered anomalous from a large amount of data as an output of this system.

1.3 Objectives and delimitations

The objectives of this thesis are to find anomalies and irregular consumption patterns from energy consumption data. To achieve the goals following steps are taken:

- Cluster the heating consumption data using the k-means algorithm to obtain load profiles. Use statistical methods to extract possible anomalies from obtained clusters.
- Cluster electricity consumption data using the k-means algorithm to obtain load profiles. Find objects with the high baseload during night time and consider them as anomalies. Use statistical methods to extract possible anomalies from obtained clusters.

This study is limited to use only outside temperature as a weather variable, though other variables such as solar radiation amount and wind speed also could be used as has been done by Lundström [5]. Consumption values are aggregated to the main level from sub-meters and their possible locations are not considered. The possible effects of condensing heat affecting heating consumption are not considered due to the lack of measurements and knowledge. The

final product is aimed to be robust and computationally efficient to be easily implemented for use.

1.4 Structure of the thesis

This masters' thesis consists of six main chapters. The introduction serves as an introduction for energy consumption and provides information on the subject and background. It also defines the problem and the objectives and their limitations. Related work investigates previous research done on a similar domain. A short review of the studies that are related to this study is represented in this chapter.

Methods describe the clustering techniques and approach implemented in this study. This chapter provides more of a theoretical approach and gives a summary of the used algorithms and methods. Implementation describes what tools and packages are implemented throughout this study. Experiments and results provide a description of the experiments and their workflow and a summary of implementation. In results, findings are further visualized, discussed and analyzed. The discussion and conclusions summarize steps taken to achieve the results in this study and reflect the future.

2 RELATED WORK

This chapter contains a review of some of the previous studies and theories related to this study. At the end of each subchapter, an implementation for this study is presented.

2.1 Energy load classification

There has been a lot of research made related to consumption or load profiling by using smart meter data [6-9]. According to Elexon [10] load profiling can be described as the consumption pattern of electricity of individuals or groups over a certain period.This period can be anything from 15 minutes to a span of 3 years. Similarly, the frequency of measurements can vary from seconds to hourly readings. Figure 1 presents an artificially created load profile of electricity usage over a span of six months with consumption measured hourly just to visualize characteristics of consumption data.

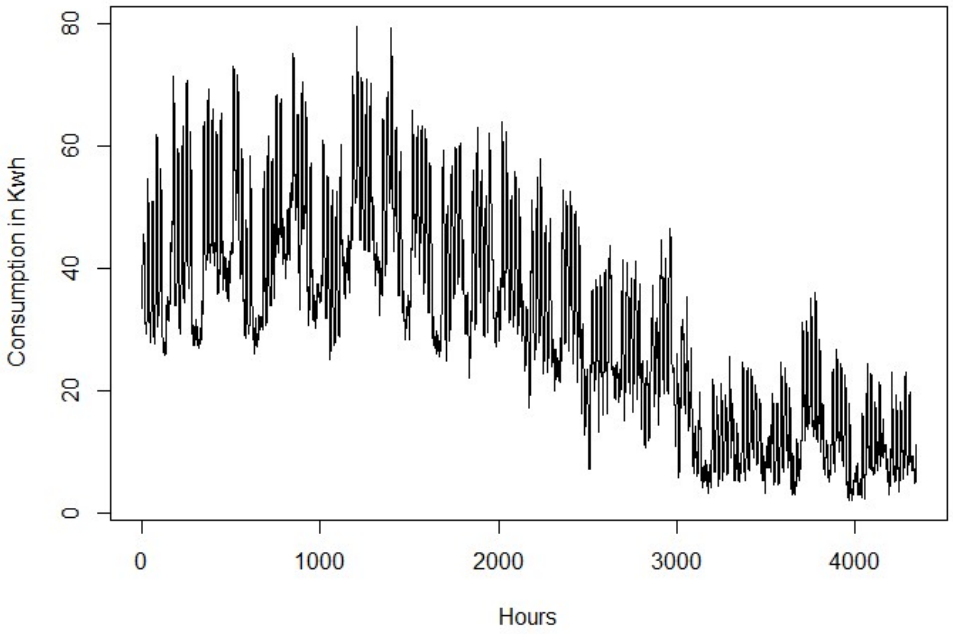

Figure 1. Artificially created load profile of electricity consumption over six months with consumption measured hourly.

According to Yang *et al.* [11] load profiles can be used in load classification where most similar load profiles are portioned into the same groups by the chosen clustering algorithm. These load patterns will then represent the characteristics of a certain cluster [11].

A review done by Yang *et al.* [11] examined load classification methods in smart grids. In this paper, the process for load classification was divided into five stages. These five stages include load data preparation, load data clustering preparation, load data clustering implementation, understanding and evaluation of load classification results, and applications of load classification results. This process model can be seen in Figure 2. A similar approach until the load data clustering implementation phase will be implemented in this study.

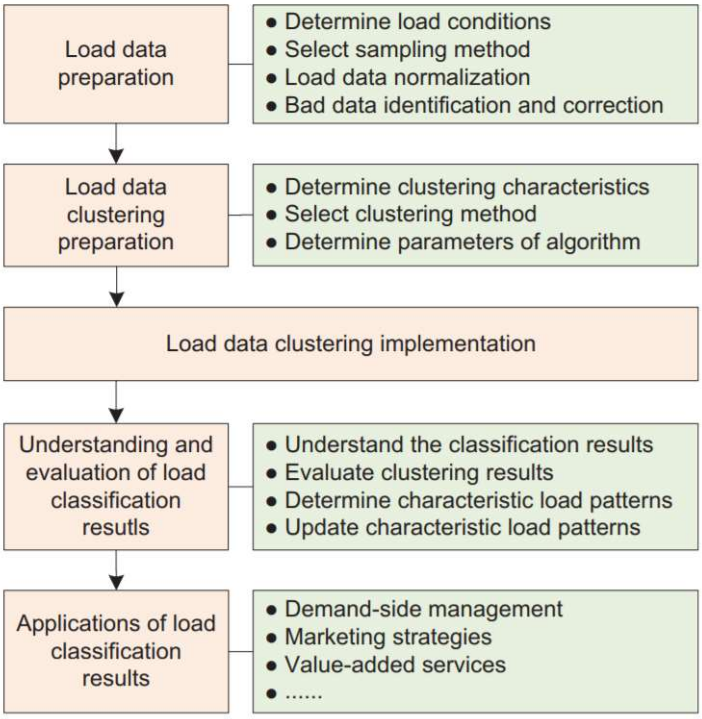

Figure 2. Process model of load classification. Reproduced from Yang *et al.* [11] with the permission of the publisher.

Yang *et al.* [11] also concluded that no clustering algorithm is superior to any other but K-means, fuzzy c-means, hierarchical clustering, and SOM, were referred to as well-known clustering algorithms and they were more commonly used for classification. In addition to these more commonly used methods, new methods such as Support Vector clustering, FaiNet, and iterative refinement clustering have been studied and applied to load classification.

Identification of low-quality data and processing, normalization of data, data sampling and reduction methods were considered to have an important role in the preparation process of load classification. Also, effective and efficient methods of evaluation should be implemented for understanding and analysis of classification results [11].

It can be seen that various different consumption patterns of consumers can be used for load classification with smart grids and they can help to understand different consumption patterns of consumers more efficiently. From the obtained profiles necessary actions such as optimizing the grid can possibly be implemented in the future more accurately.

Another interesting study done by Napoli *et al.* [12] compared the performance of several clustering algorithms such as modified follow-the-leader, hierarchical clustering of different types, K-means, fuzzy K-means, and SOM. They also compared different data reduction techniques such as Sammon maps, principal component analysis (PCA) and curvilinear component analysis (CCA). Their goal on data reduction was the possible savings on consumed data storage space along with the savings on used memory and reduction of computational complexity. Modified follow-the-leader along with the hierarchical clustering emerged as the most effective methods. Based on their results with reduction, Sammon maps indicated good robustness compared to PCA and CCA. Data reduction techniques were validated generally to be effective and acceptable [12].

They also discovered that some indications for the optimal number of clusters can be obtained from the clustering validity indicators. To produce a clear and reasonable amount of information for their end product with easy management, the number of classes was aimed to be feasible and not too high. So although validity indicators such as the clustering dispersion indicator (CDI), Modified Dunn index (MDI), the scatter index (SI) and the Davies–Bouldin index (DBI) suggested a higher number of clusters they obtained best results with 15 clusters. It is important to note that while validity indicators are important and provide valuable information their results should be always observed carefully within the context of the desired outcome [12]. Below, we can see an example of the performance of the clustering validity indicators with different clustering techniques under test for k values interval between 5 and 100.

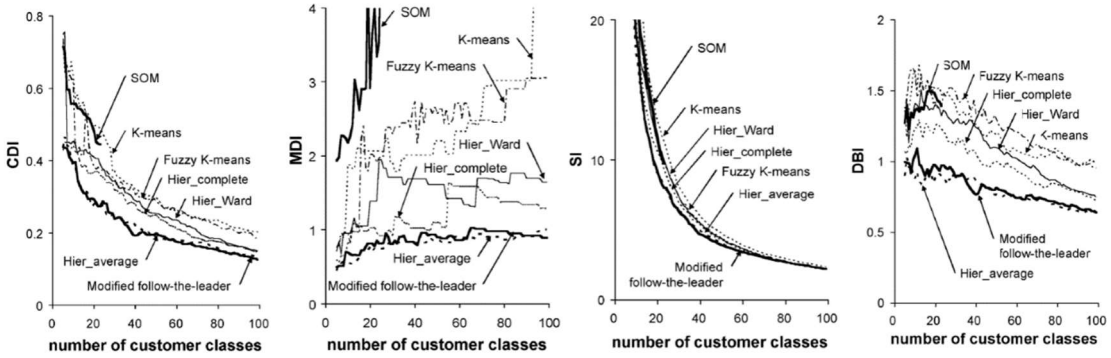

Figure 3. Performance of the clustering validity indicators with different clustering techniques under test for k values interval between 5 and 100. Reproduced from Napoli *et al.* [12] with the permission of the publisher.

Räsänen *et al.* [13] had similar findings with validity indices when performing clustering to obtain load profiles in the context of electricity load management. They used K-means, Self-Organizing Map (SOM), and Hierarchical clustering for a large dataset of 3989 customers located in Finland. They concluded that the balancing of performance and interpretability of deciding the optimal number of clusters is crucial. Deciding the optimal number of clusters is a complex problem and the number of classes should be limited in size but also be representative. They ended up using the Davies-Bouldin index to determine the optimal number of clusters [13].

Laurinec & Luka [14] also compared time-series depictions for clustering. In this study, they used novel representation methods that were based on models such as robust linear regression, generalized additive model, Holt-Winters exponential smoothing, and the median daily profile. They concluded that the best representations were Piecewise linear approximation (PLA) and model-based representations, particularly the robust linear model (RLM) and Median [14].

In another study by Laurinec & Luka [15] used K-means to perform the load classification of customers. The optimal number of k was determined by using the Davies-Bouldin index. Five different forecast methods were used for the clustering-based forecasting method. These were

seasonal naïve method, multiple linear regression, random forest, conditional inference trees, and triple exponential smoothing. The clustering-based forecast method outperformed a completely disaggregated approach. It was important to notice that the load classification phase plays an important role also in forecasting [15].

The inspiration for using k-means for load classification in this work was derived from these previous studies. The K-means algorithm is generally considered to be fast and efficient, and it has the ability to handle large data sets with easy implementation. It is also preferable to study some different representation methods for the data and implement them in order to reduce dimensionality. To obtain the optimal number of clusters some validity indices need to be used. However, it is important to see investigate these results carefully, since increasing the number of clusters may not bring any additional value.

2.2 Anomalies in consumption data

There are many different approaches in order to find anomalous behavior in data depending on how one defines anomaly in the given context. Cui & Wang [16] examined five different models to detect anomalies in the school's electricity consumption. These models were Autoregressive model, Autoregressive-Moving-Average model, Polynomial regression model, Gaussian kernel distribution model, and Gaussian distribution model. Model evaluation criteria contained common criteria such as recall, precision, and false-negative rate. A hybrid model that combined polynomial regression and gaussian distribution was chosen based on validation and experiments. The model obtained desired results and anomalies were found; however, it needed to be trained manually before detecting anomalies [16].

Another study done by Chahla *et al.* [17] used K-means clustering to represent energy consumption behavior and Long Short Term Memory (LSTM) to predict power consumption in the upcoming hour. They used an unsupervised approach due to the nature of the data, which was collected from Pecan street's data port. A hybrid model combining LSTM and K-means algorithm successfully detected anomalies, but domain knowledge was needed to verify the results [17].

Clustering-based approaches to detect anomalies are also widely used. Chandola *et al.* [18] represented a wide review in a survey of different methods. Clustering-based techniques that followed two-step approaches such as K-means and SOM assumed that "Normal data instances lie close to their closest cluster centroid, while anomalies were far away from their closest cluster centroid." Another group clustering-based techniques such as Cluster-Based Local Outlier Factor (CBLOF) relied on assumption that "Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse clusters.". Clustering-based techniques can be used in an unsupervised mode and their implementation was fast. The disadvantage of using these techniques was that many of them were built to perform clustering and not optimized for detecting anomalies. To obtain anomalies by using clustering-based techniques was only effective when possible anomalies were not creating significant clusters by themselves [18].

Anomalies can be discovered with many different approaches as can be seen by these previous studies and there is no ultimate solution that fits for all. In this study, a clustering-based approach to detect anomalies is implemented since it can be used in an unsupervised mode and has been widely used in other studies as well. It is important to notice that clustering-based techniques were not created to detect anomalies and have their limitations within this context.

2.2.1 Defining anomalies

According to Abraham & Chuang [19], anomaly detection is a process of detecting patterns in a given data set that are differing from the behavior considered to be normal. Anomaly detection has been implemented in many applications and research areas such as insurance/credit card fault detection, image processing, health-care monitoring, and network intrusion [18]. A study done by Chandola *et al.* [18] concentrates on finding abnormal consumption profiles and patterns from the energy consumption data of various buildings.

Anomalies in time-series data can be divided into three categories:

- Point anomalies: A point anomaly is a single independent instance of data that does not fit within data considered to be normal [18].

- Contextual: Anomalies are considered to be contextual when specific values change abruptly with respect to their temporally contiguous values [18].

- Collective: Collective anomalies are described to have unusual shapes within the whole or large portion of time-series [20].

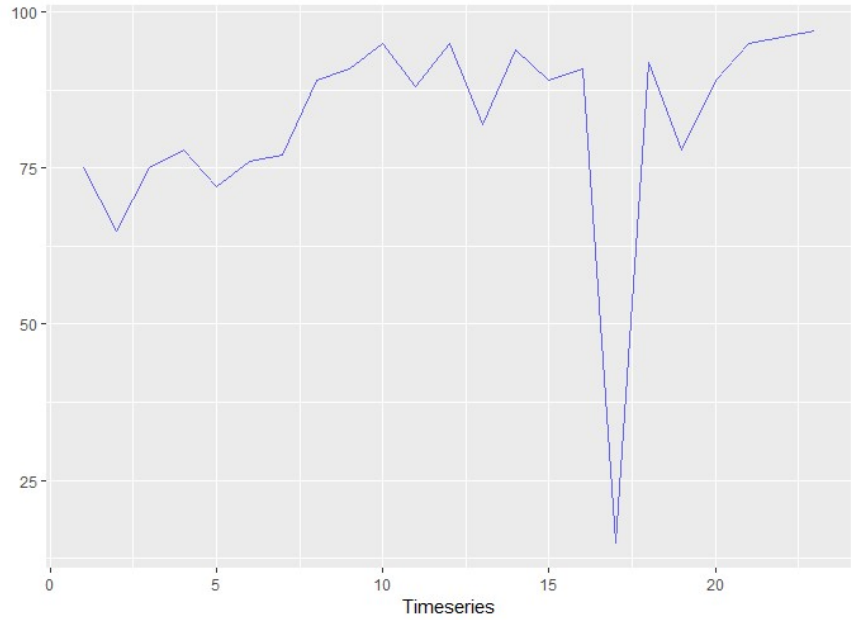

Figure 4. Representation of a point anomaly

This study focuses on collective anomalies that might reveal unusual electricity consumption patterns for different locations. These locations are introduced in more detail later in chapter 5.1 that describes the data being used in this study. By detecting these possible anomalies further measures such as optimizing heating or cooling systems, for example, can be adapted. This is important in terms of energy efficiency.

2.2.2 Detecting anomalies

Chandola *et al*. [18] concluded that labels within the data describe their nature of being normal or anomalous. Labeling is usually done manually by humans which makes detailed and depictive labeled data to be expensive. Getting labeled data for anomalous instances is far more difficult compared with normal ones. According to them, anomaly detection can be divided into

three different categories based on the labels available. Following three categories represent these different approaches:

- Supervised anomaly detection: Supervised detection model relies on the training data set that has labeled instances for both anomalous and normal. One common way of implementing supervised detection would be to build a predictive model for normal vs. anomaly classes. After this, the unseen data will be compared against the labeled instances to decide which class it belongs to.

- Semi-supervised anomaly detection: Semi-supervised detection model assumes that the training data set has labeled instances only for the normal data. Unknown samples are classified as outliers when their behavior is far from that of the known normal samples. This approach is more widely applicable since it does not require labels for the anomaly classes.

- Unsupervised anomaly detection: Unsupervised detection models do not rely on training data and thus are most widely adopted [18]. These models presume that normal instances are way more frequent than anomalous ones. By using a sample of the unlabeled data as training data, some of the semi-supervised techniques can also be implemented to operate in an unsupervised way. [18].

2.2.3 Challenges

As stated by Chandola. *et al* [18] an ideal solution to anomaly detection would be to define a normal region and consider objects outside of this region as anomalies. Several factors make this seemingly simple approach more complex such as:

- Defining an absolute normal that is suitable for every scenario is very difficult.
- In many domains normal behavior is constantly evolving so that the current situation may not represent the normal state in the future.
- There is often a lot of noise within the data that can be similar to the actual anomalies [18].

Because of these factors, detecting true anomalies can be very challenging. Domain knowledge from a specific field must be applied when defining normal and possible anomalies to obtain the best possible results. By inspecting only patterns and numbers, it is possible to find a lot of deviations from most of the data, which is considered to be “normal”. However, without domain knowledge, it is possible that these deviations cannot be described as anomalies before they are investigated more thoroughly according to the specific context.

3 METHODS

This chapter provides a short but detailed description of the methods and techniques used in this study. Clustering time-series and methods introduce different clustering methods and provide a background for the k-means algorithm. Validity indices, representations, distance measures, and normalizations used in this study are also represented with a short background.

3.1 Clustering time-series

Aghabozorgi *et al.* [21] described time-series as a sequence of continuous real-valued elements. It is a type of temporal data ordered chronologically, which is usually large in terms of the number of observations.

They also concluded that clustering is a data mining technique where similar data are distributed into collections of high similarity. Clusters can be formed by grouping objects with high similarity with other objects within the collection, and that are dissimilar with objects in other collections. Interesting patterns that might be valuable can be found by the clustering of time-series. These can be either frequent or rare patterns. By representing time-series cluster structures by visualizing them, it can help users quickly to understand the nature of the data, possible anomalies, and other regularities [21].

3.2 Clustering Methods

Aghabozorgi *et al.* [21] divided clustering methods into five categories based on their clustering criterion. Different categories and their respective methods can be seen in Figure 5 below. Some of the methods are used more often for load classification, but the superiority of one method over another depends on the application [21].

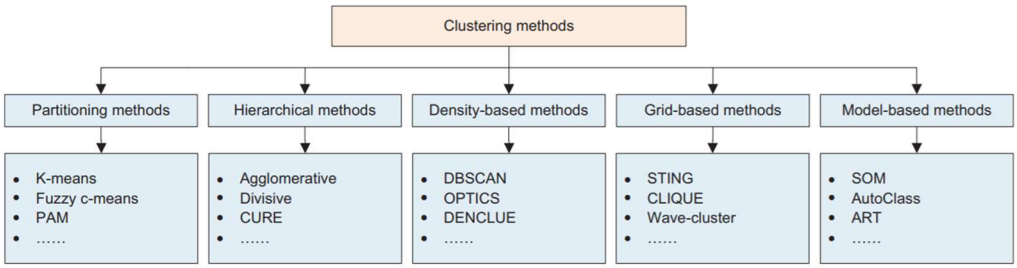

Figure 5. Different clustering methods used for load classification reproduced from Aghabozorgi *et al.* [21] with the permission of the publisher.

In this study, load classification is done by using the K-means algorithm. K-means clustering algorithm is one of the oldest and most widely used partitional clustering algorithms. It was developed by numerous researchers but most notably by J.B MacQueen in 1967 [22].

The K-means algorithm is an unsupervised machine learning algorithm that finds k non-overlapping clusters. Clusters are then represented by their centroids, which typically is the mean of the points within the cluster. This is achieved by first selecting centroids randomly based on a user-defined number of k and assigning data points to the closest centroid. The centroid of the cluster is updated based on the points within the cluster. This is repeated iteratively until convergence or the maximum number of iterations is achieved. In this context, convergence means that there is no change in the location of the cluster centroid or their membership values.

Let x be a sample to be clustered into k clusters in set of clusters C where $C = \{1,2, \dots, k\}$. The number of data points is denoted as n. The objective function J of the K-means algorithm is to minimize total within-cluster variance, which is defined as follows:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_{ij} - \mu_j||^2 \tag{1}$$

$||x_{ij} - \mu_j||^2$ is the distance metric used, between the data point x_{ij} and the cluster center μ_j. Euclidean distance will be used in this study and the formula is given in Equation (2). Euclidean distance is a distance measure that can be used with different clustering algorithms.

It is the shortest distance between two points in an N-dimensional Euclidean space. It is used to measure the distance between two data objects. Euclidean distance is only appropriate for continuous numerical variables.

$$d(x_l, \mu_l) = \sqrt{\sum_{l=1}^{N} (x_l - \mu_l)^2} \tag{2}$$

The execution of the k-means algorithm is described below:

1. A user defines the desired number of clusters as k.
2. Cluster centroids are randomly generated.
3. The distances between each data point and the cluster centroids are calculated.
4. Data points are then assigned to the closest centroid (for the chosen distance measure).
5. Calculate new cluster centroids μ_j corresponding to the new sets C_j where $|C_j|$ represents the cardinality of j^{th} cluster by calculating the means with the formula:

$$\mu_j = \frac{1}{|C_j|} \sum_{i=1}^{C_j} x_i \tag{3}$$

6. Repeat from step 5 until the cluster assignments stop changing and convergence is achieved.

3.3 The optimal number of clusters

Different approaches to obtain the optimal number of clusters can be adopted by using validity indices. However, as stated by Dent *et al.* [23] there is no single best validation tool. One way to evaluate validation indices is to compare them simultaneously. The next chapters represent three different approaches for deciding the optimal number of clusters.

3.3.1 Elbow method

The elbow method is an approach to determine the optimal number of clusters. The basic idea of K-means, described in chapter 3.2.1, is to minimize clusters total within-cluster sum of square (WSS) for each given k. Elbow method utilizes the total WSS as a function of the number of the clusters. The general idea is to choose the optimal number of clusters so that adding another cluster does not improve the total WSS substantially. The execution of the elbow method is as follows:

1. Apply the k-means algorithm with a predefined interval of k's.
2. For each k, calculate the WSS.
3. Plot the values of WSS against a number of k's.
4. The location of a bend where the curvature is at the maximum is considered the optimal number of clusters.

Below in Figure 6 we can see an image of plotted WSS values concerning predefined k values. The optimal number of clusters would be 2 in this case.

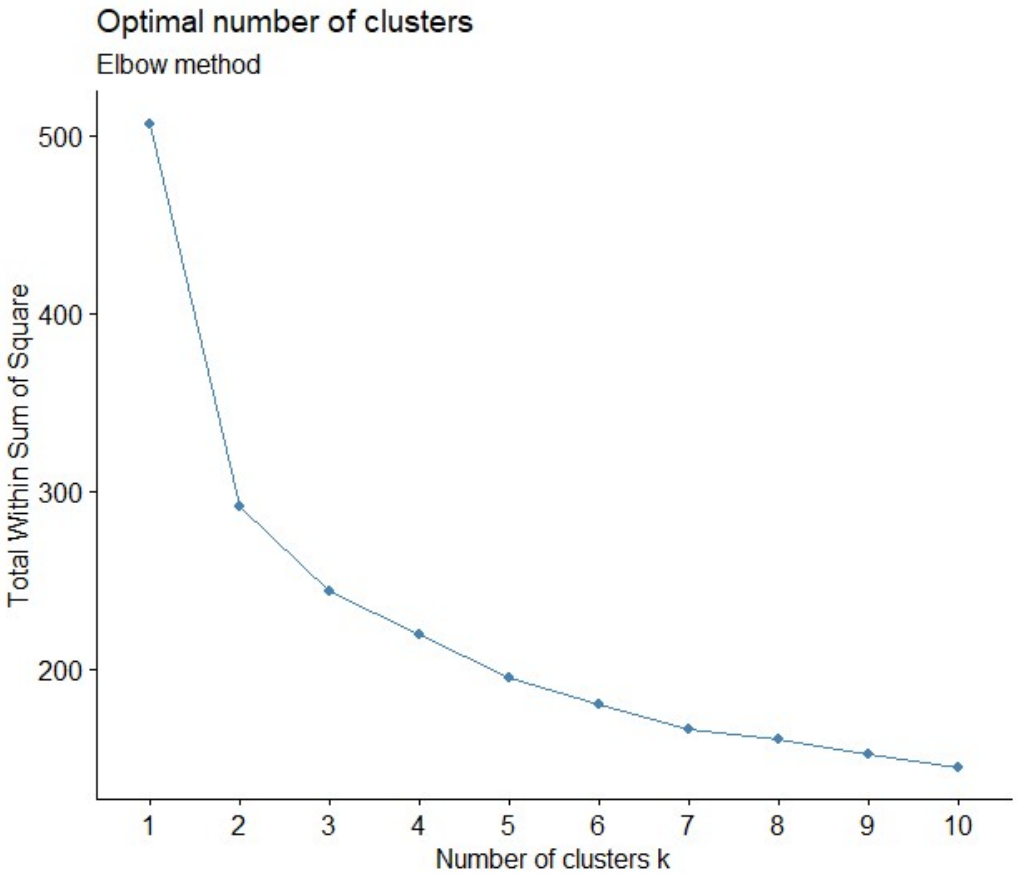

Figure 6. The optimal number of clusters using the elbow method.

3.3.2 Silhouette index

Silhouette index is used to study the separation distance between clusters. The silhouette value is a similarity measure between objects within its cluster compared to the other clusters. One way of interpreting silhouette values can be done by plotting values. The plot will visualize how close each point in one cluster is to other neighboring clusters. Silhouette values lie within [-1,1].

The definition of the silhouette index was adopted from Starczewski & Krzyżak [24].
Let $C = c_{1,\ldots,}c_k$ be set of clusters with c_k indicating the k^{th} cluster in a given data set. Silhouette index is based on the *silhouette width,* which is expressed as follows:

$$S(x) = \frac{b(\mathrm{x}) - a(x)}{\max\big(a(x), b(x)\big)} \tag{4}$$

Where $a(x)$ indicates the within-cluster mean distance defined as the average distance between an observation x which belongs to C_k and rest of the patterns x_k belonging to the same cluster, which is

$$a(x) = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} d(x, x_k) \tag{5}$$

n_k indicates the number of objects in C_k. The smallest mean distances x to the objects x_i belonging to the other cluster C_e, is denoted by $b(\mathrm{x})$, where $i = 1, \dots, k$ and $i \neq k$. Smallest distance is expressed as follows:

$$b(x) = \min_k \big(mean(x, x_i)\big) \tag{6}$$

The mean distance for C_e can be written as:

$$mean(x, x_i) = \frac{1}{n_k} \sum_{i=1}^{n_k} d(x, x_i) \tag{7}$$

and n_k indicates a number of objects in C_k. So, the average *silhouette width* for the given cluster C_k is expressed as:

$$S = \frac{1}{n}\sum_{i=1}^{n} S(x_i) \tag{8}$$

Below in Figure 7, the visualization of silhouette widths can be seen. The thickness of different colored areas indicates the size of the cluster. It can be seen, that some of the objects could be clustered into the wrong cluster. This is indicated as a "leak" below zero in the respective cluster.

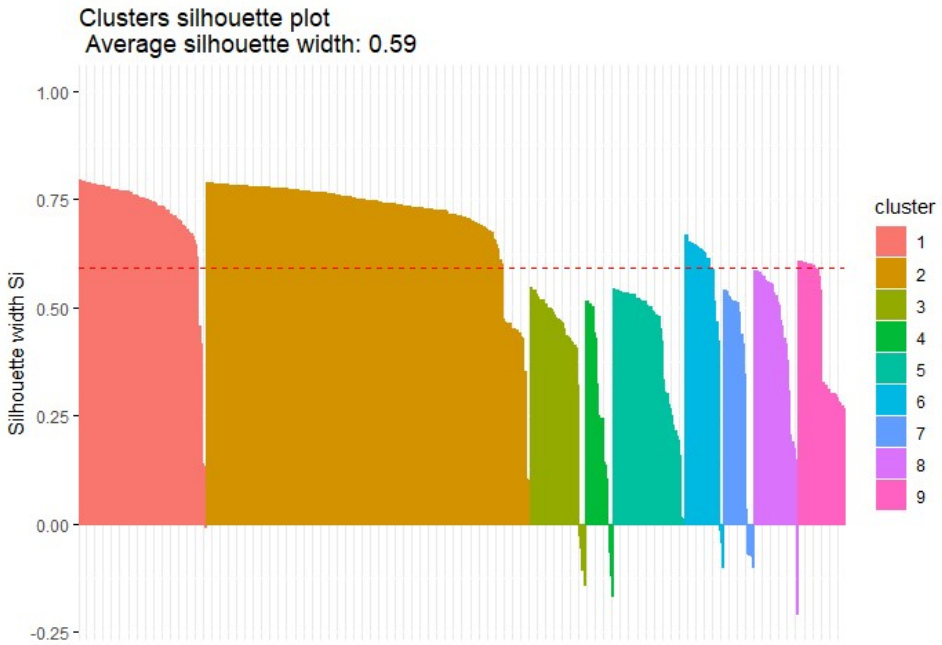

Figure 7. Silhouette widths with an average silhouette width as a red dotted line.

3.3.3 Calinski-Harabasz index

The Calinski-Harabasz index was introduced in 1974 by Calinski and Harabasz [25]. According to Ünlü & Xanthopoulos "It is a validation index based on the average sum of squares between and within a cluster." The following formal definition below is adapted from Ünlü & Xanthopoulos [26] and expressed as:

$$CH = \frac{SSB}{SSW} \times \frac{(n-k)}{(k-1)} \tag{9}$$

where SSB indicates the between-cluster sum of squares and SSW is the within-cluster sum of squares. k stands for the number of clusters and whereas n is the number of observations.

SSB is computed with the following equation:

$$SSB = \sum_{j=1}^{k} n_j \, ||\mu_j - \mu||^2 \tag{10}$$

where μ_j represents the centroid of cluster j. $\| \mu_j - \mu \|^2$is the Euclidean distance between the centroid and the mean of all data points μ.

The formula for obtaining SSW is expressed as follows:

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_{ij} - \mu_j||^2 \tag{11}$$

where$\| x_{ij} - \mu_j \|^2$ represents the Euclidean distance between the sample and the centroid of a cluster. High Calinski-Harabasz index value indicates a good clustering result, though high SSB and low SSW values produce a well-partitioned cluster [26].

3.4 Representations of the time-series data

Time-series are often extremely large and could be high dimensional as described in chapter 3.1. This produces several problems and limitations with computational power making certain approaches inefficient. According to Krawczak & Szkatuła [27] the main goal of reducing dimensionality is to preserve sufficient information for new reduced representation. There are several approaches for reducing dimensionality.

Some of the possible representation methods presented in four groups:

- Nondata adaptive: Piecewise Constant (PAA) or Discrete Wavelet Transforms (DWT).
- Data adaptive: Symbolic Aggregate Approximations (SAX) or Piecewise Linear Approximation (PLA).
- Model-based: Autoregressive-moving-average (ARMA) or Average values [27].
- Data dictated: Clipping [28].

A review done by Fu [29] concluded that sampling is often considered one of the simplest methods for the reduction of time-series. By using sampling with a ratio of m/n, where m denotes the length of a time-series, and n indicates the dimension after the reduction of dimensionality. A disadvantage of this method is that reduced series is often too distorted compared to the original series. Using Figure 8 as an example, sampling would result in a loss of all the peaks that can be seen in the image on left. Another simple approach would be using the average value for each segment to represent the underlying data points. In this case, a segment represents whatever time interval chosen for example an hour or a day [28]. More advanced methods such as Discrete Fourier Transforms (DFT), PCA, PAA, DWT, and SAX also exist. Selected representation often depends on the similarity objective, which depends on the problem [29].

In Figure 8 below, we can see the original time-series of 336 data points on the left side and reduced one by using DWT on the right side. It can be seen that the consumption pattern is preserved after dimensionality reduction.

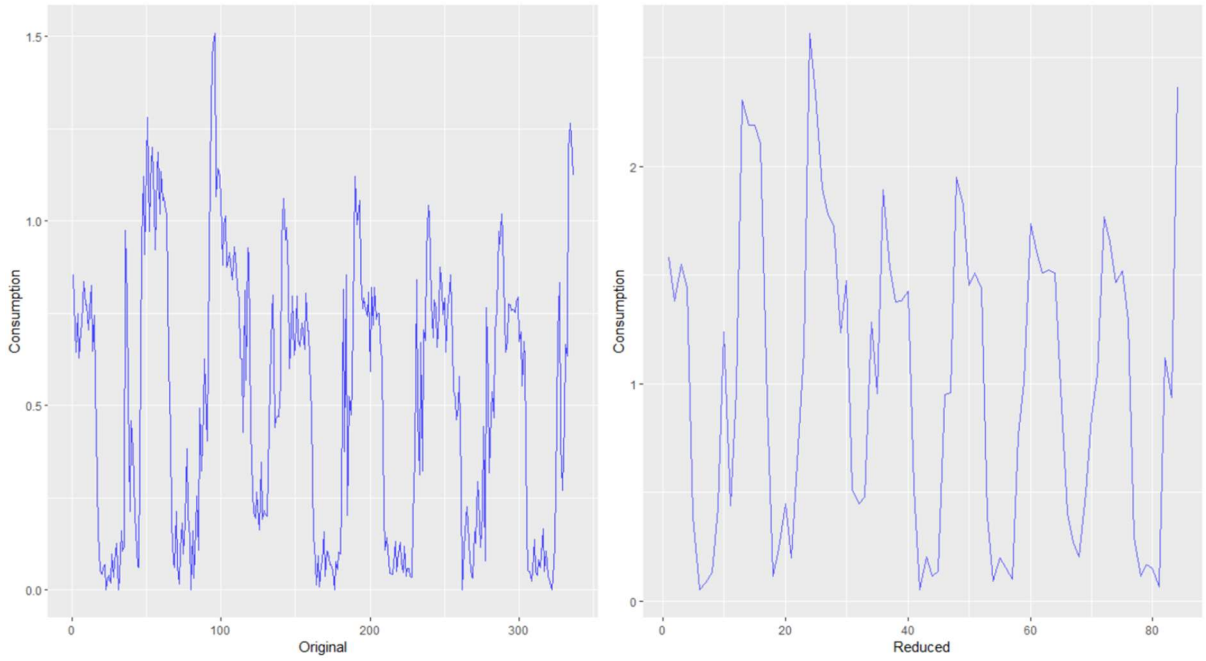

Figure 8. Dimensionality reduction of time-series by using DWT. On the left side, the original pattern with 336 data points. On the right side, the reduced one with 84 data points.

According to Jain *et al.* [30] normalization is one of the most common pre-processing techniques when building a regression or classification model. One goal of normalization is to transform values to have a common scale when values have different ranges. There are multiple approaches for the normalization of the data that include min-max normalization, z-score normalization, decimal scaling, etc. [30].

As claimed by Investopedia [31], z-score describes the number of standard deviations a data point is from its mean. Z-score is also widely known as a standard score because it allows a comparison between different kinds of variables by standardizing the distribution to have zero mean and standard deviation of one. Z-score is calculated by the first calculating the sample mean μ and standard deviation σ. The final calculation is given in the formula below, where x represents a data point in a sample set

$$z = \frac{(x - \mu)}{\sigma} \tag{12}$$

After performing z-score normalization the given data set has zero mean and standard deviation of 1 [31]. This study applied z-score normalization for the data before clustering with the K-means algorithm.

4 IMPLEMENTATION WITH R

The following chapter provides a short but detailed description of the tools and packages used in this study.

According to the R-project [32] R is a programming language and open-source environment for statistical computing and graphics. It is similar to the S language and environment developed at Bell laboratories by John chambers. Much code written for S runs unaltered in R but there are some differences. R is an integrated suite of software facilities for data manipulation, calculation, and graphical display. R is available as Free Software and supported operating systems are a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and macOS [32]. In this study RStudio was being used to run R. Rstudio provides a graphical user interface to be used as an editor for R code and console to execute it. Below in Figure 9, we can see a user face image of RStudio and in Table 1 a list of all R packages and their meta-information used in this study.

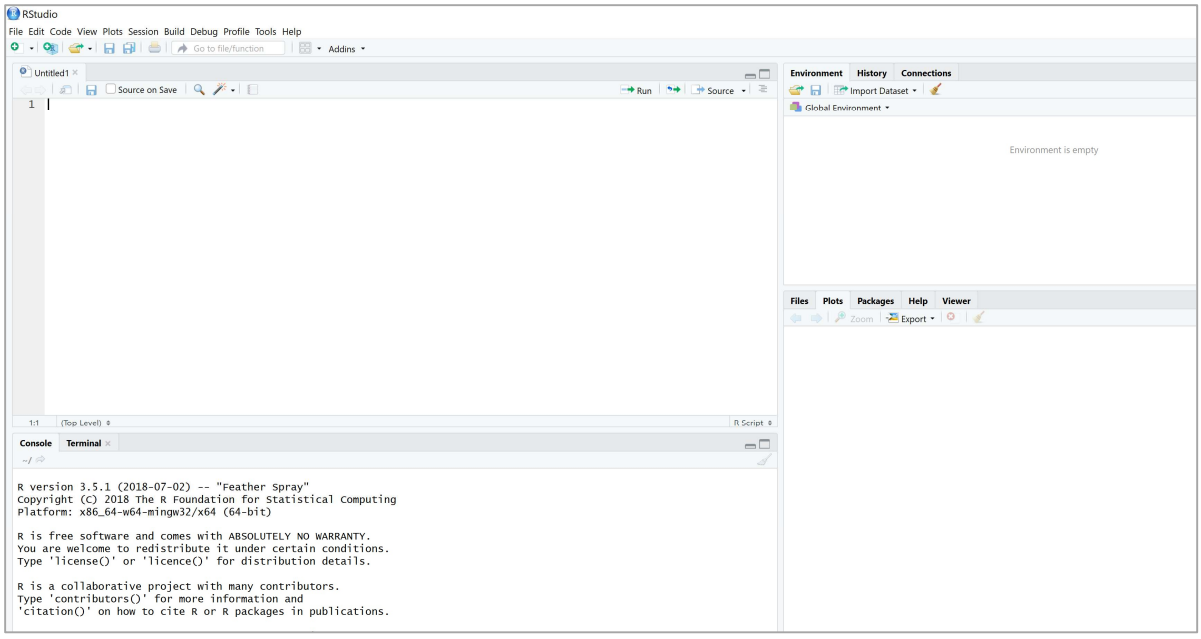

Figure 9. The user interface of RStudio.

Table 1. Table of all used R packages

Name	Version	By	Source
ggplot2	V3.2.1	Hadley Wickham	https://www.rdocumentation.org/packages/ggplot2
tidyr	V0.8.3	Hadley Wickham	https://www.rdocumentation.org/packages/tidyr
readr	V1.3.1	James Hester	https://www.rdocumentation.org/packages/readr
dplyr	V0.7.8	Hadley Wickham	https://www.rdocumentation.org/packages/dplyr
data.table	V1.12.4	Matt Dowle	https://www.rdocumentation.org/packages/data.table
NbClust	V3.0	Malika Charrad	https://www.rdocumentation.org/packages/NbClust
TSrepr	V1.0.3	Peter Laurinec	https://www.rdocumentation.org/packages/TSrepr
tseries	V0.10-47	Kurt hornik	https://www.rdocumentation.org/packages/tseries
corrplot	V0.84	Taiyun Wei	https://www.rdocumentation.org/packages/corrplot
caret	V6.0-85	Max Kuhn	https://www.rdocumentation.org/packages/caret
factoextra	V1.0.6	Alboukadel Kassambra	https://www.rdocumentation.org/packages/factoextra
TSclust	V1.2.4	Pablo Montero	https://www.rdocumentation.org/packages/TSclust
reshape2	V1.3.4	Hadley Wickham	https://www.rdocumentation.org/packages/reshape2

5 EXPERIMENTS AND RESULTS

The experiment section was divided into three different parts and it follows the original objectives presented in chapter 1.3. The chapter starts with a short description of data and pre-processing before continuing into actual experiments.
In the first and second parts, electricity and heating consumption data are clustered by using the K-means algorithm. The final stage investigates results achieved by clustering and isolates possible anomalous objects as final results.

The following experiments were executed on R studio Version 1.1.456. The operating system was Windows 10. The Processor was Intel i7-6820HQ CPU @ 2.70GHz, 4 Cores, 8 Logical Processors. The RAM amount was 16GB.

5.1 Data description and pre-processing

The company is a market-leading ISO 50001-certified energy data management system provider in the North. It has over 100 000 metering points and over 1000 customers. The company provided selected data for 4152 different locations in Finland.

Data contained a variety of measurements for different fields of business. However, the majority of the data was from the field of grocery and retail with a building type of 11. For this study, a subset of data with a building type of 11, was used, since the majority and these buildings generally operate in the same field. This resulted in 2167 unique locations. The data was gathered specifically for this work from a relational database through Azure API and it describes actual customers.

Data can be divided into two different categories which are:

1. Hourly measured one-dimensional electricity consumption data for 1 year from 1st of January 2018 to 1st of January 2019.

2. Multi-dimensional additional metadata information about locations:
    i. Enegia Id - String (Categorical)
    ii. Profile name - String (Categorical)
    iii. Name of location - String (Categorical)
    iv. Year of build – Integer (Numerical)
        1. Range from 0 to 2019
    v. Heating type – String (Categorical)
        1. District Heating
        2. Electric Heating
        3. Electric Geothermal Heating
        4. Gas Heating
        5. Geothermal Heating
        6. Oil Heating
        7. Other Heating
        8. Unknown
    vi. Area/m2 – Double (Numerical)
        1. Range from 0 to 69309
    vii. Volume/m3 – Double (Numerical)
        1. Range from 0 to 355847
    viii. Postal code – Integer (Numerical)
    ix. Building type - String (Categorical)

The problem with the additional metadata was that there were a lot of missing values or values of constant zero. Year of the build had 1567 zero values, Volume had 1820 zero values, Area had only 37 zero values and Heating type had 441 missing values. In order to describe the data, these missing values or constant zeros needed some modifications or they needed to be removed. Since there were a lot of missing values or constant zeros compared to the overall values, these missing and constant values were removed. Except for the Heating type the missing value was replaced with text “Unknown”. Postal codes were also replaced with a range from 1 to 21 describing “Suuralue” in Finland. This resulted in a set of 478 locations with all the information necessary. Visualizations of the distribution of the reduced set of 478 locations can be seen in Figures 10, 11, 12, and 13.

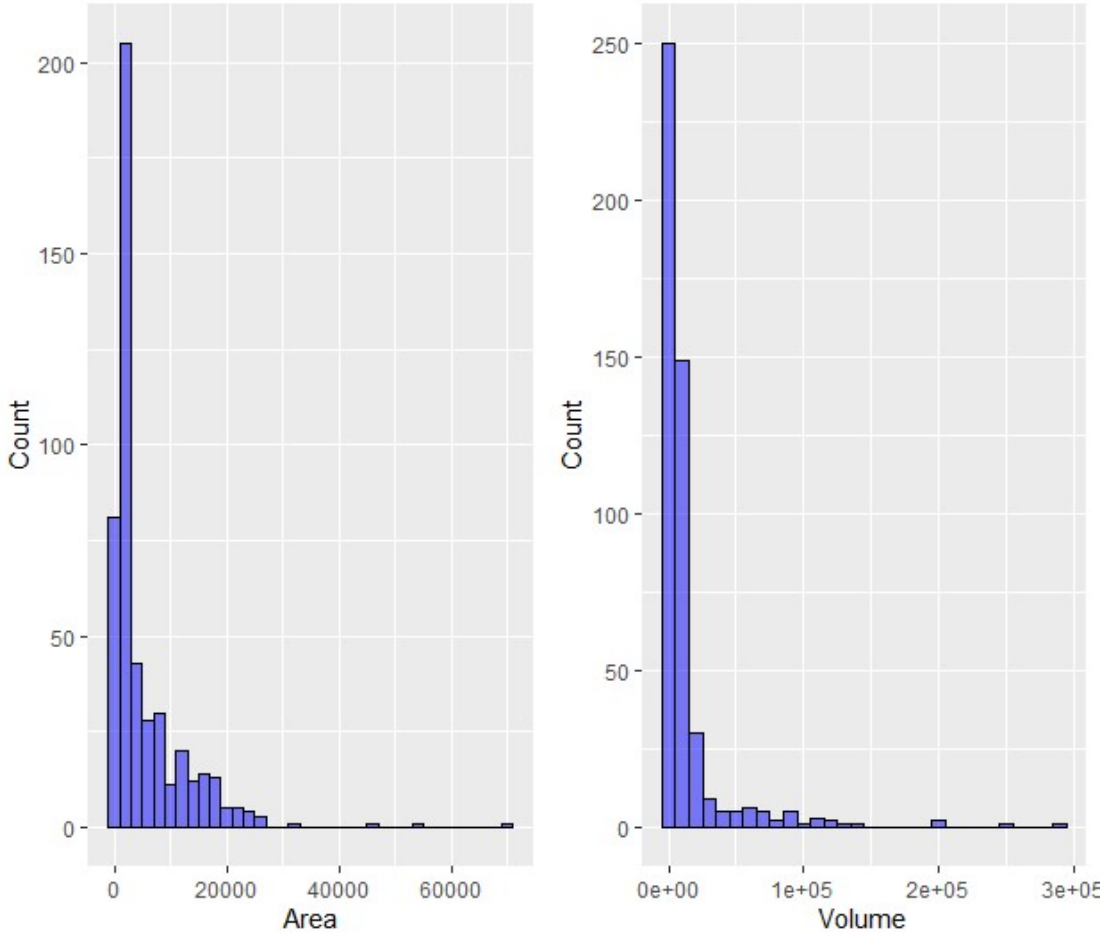

Figure 10. On the left side histogram of Area and on the right side histogram of Volume.

From the histograms of Area and Volume in Figure 10 above can be seen that the majority of the data has Area less than 30000 and Volume less than 100000.

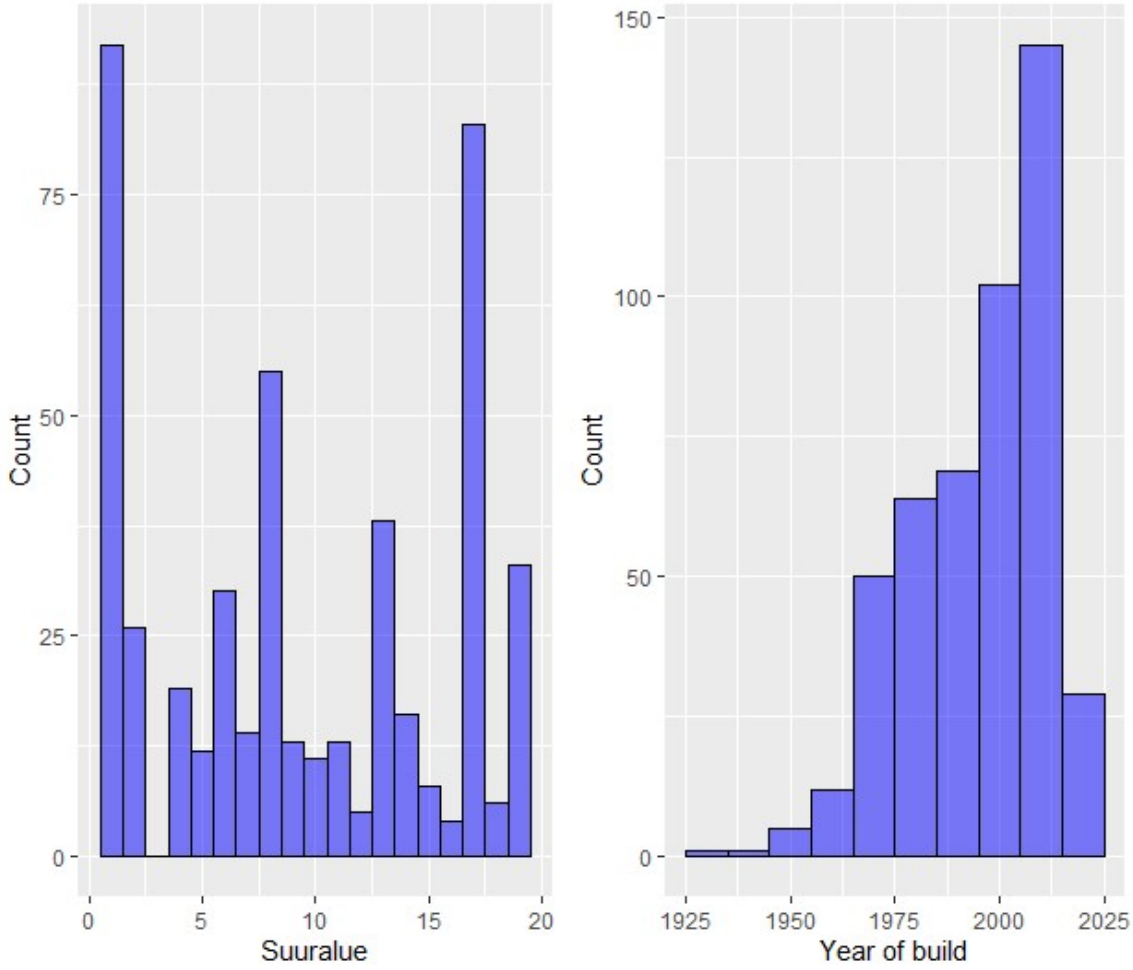

Figure 11. On the left side histogram of Suuralue and on the right side histogram of Year of build.

From the histograms in Figure 11 above can be seen that almost all regions of Suuralue were presented in the data. However, there were some more dominant Suuralue such as 1, 8, 13, 17 and 19. It can also be seen that since the removal of zero values the earliest year of build was 1925. The majority of the data has a Year of build between 1970 and 2020.

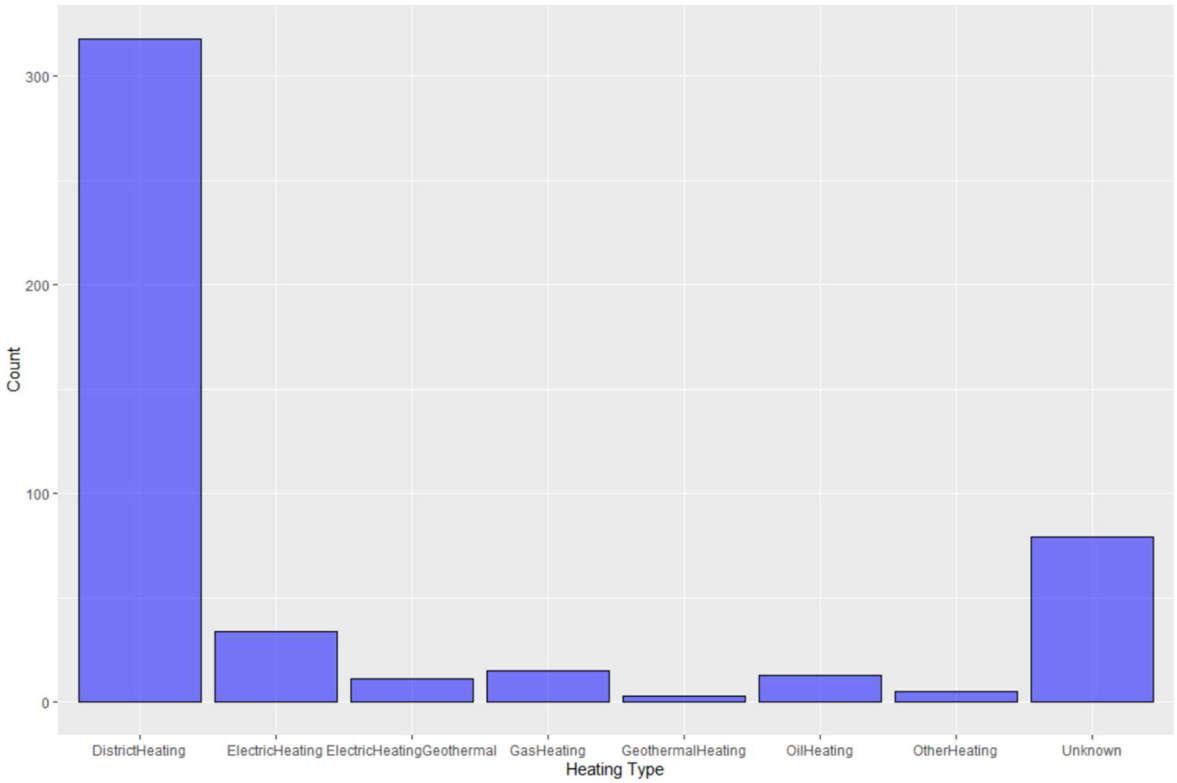

Figure 12. Histogram of Heating type.

From the histogram 12 above can be seen that Heating type is quite unevenly distributed. The majority of the locations have District Heating. The reduced set of 478 locations contained 79 Unknown values for Heating type.

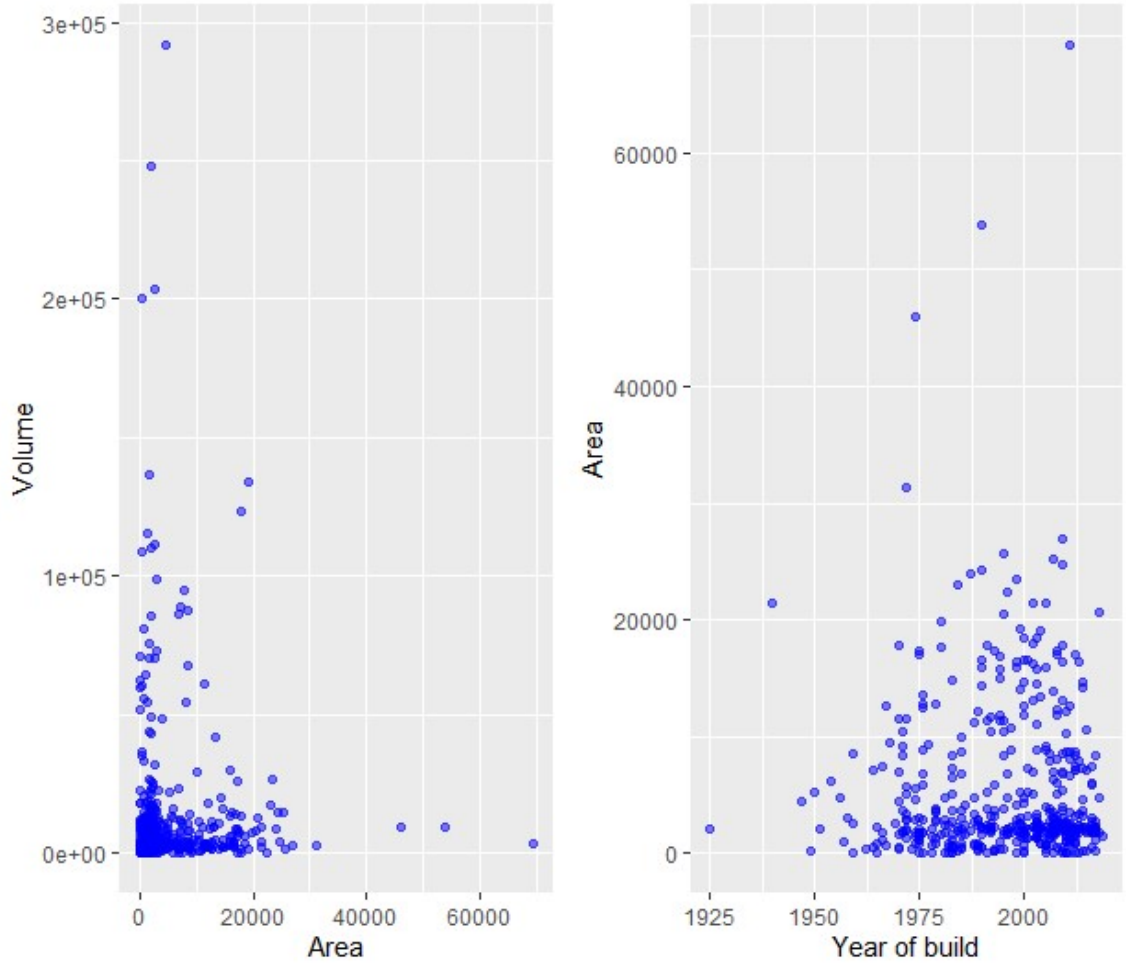

Figure 13. On the left side Area vs. Volume scatter plot and on the right side Year of build vs. Area scatter plot.

Scatter plots in Figure 13 above indicated that there were some locations with high Volume and low Area and vice versa. This could indicate that there are some errors in the measurements. Locations with high Volume and big Area can be considered to be shopping centers or similar locations. A minor trend of building smaller locations can be seen at approx. from 1960 until 2020. It is important to notice that these findings were based on a reduced snapshot of 478 locations and thus did not describe the true characteristics of the additional metadata completely. However, the data of the reduced set was considered to be quite versatilely distributed and enough to get a basic idea of the data.

The extracted subset of additional metadata of 2167 locations was used to filter consumption data to include only these specific enegiaid's with the building type of 11.

Consumption data was provided as a time-series with an interval of one year and a frequency of one hour. Consumption data also contained outside temperature measurement in Celsius from the nearest weather station with a frequency of one hour and a status column indicating whether the store was open or closed. Data was divided into two different sets, which are:

- Electricity consumption in kWh. This data set will be later referred to as a data set Electricity.
- Heating consumption in kWh. This data set will be later referred to as a data set Heating.

Some of the locations contained "manual" measurements that had only 1 to 5 consecutive values throughout the whole time interval. These locations were removed by using a threshold of 100 consecutive values in a row.

There were also locations, where consumption data was constant. This was discovered by calculating variance for the consumption measurements and extracting the ones with zero variance. After extracting these locations they were removed. Figure 14 below represents this type of measurement for one randomly selected location.

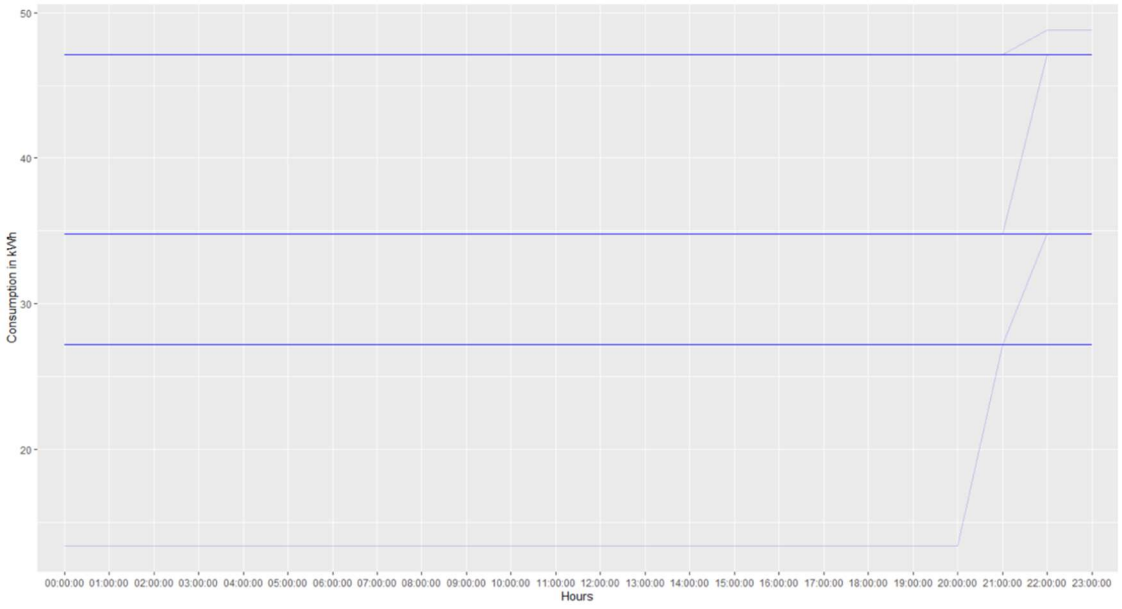

Figure 14. Representation of "manual" consecutive measurements for one location by the hour.

The consumption data measured hourly is divided into sets containing 24 measurements because 24 describes a period of one day. The average for this data is then computed across all

the data for each hour and the average values replace the original ones. The dimension of the data is then reduced from n to 24 dimensions depending on a defined time interval.

Due to the limitations of computational power due to the massive size of 9.2Gb for the whole data set the time interval was set to be three months starting from 1.10.2018 until 31.12.2018. The final data set Heating contained 531 unique locations and the final data set Electricity contained 2037 unique locations. The chosen interval had 2209 hourly measurements. The dimensions were reduced by calculating the hourly average as described above. The dimension of the data set Heating was then reduced from 531x2209 to 531x24. Similarly, the dimension of the data set Electricity was then reduced from 2037x2209 to 2037x24.

Both data sets Electricity and Heating were standardized by using Z-score normalization before performing clustering by using the K-means algorithm. The representation of this reduced and standardized time-series for one randomly selected location can be seen in Figure 11 below.

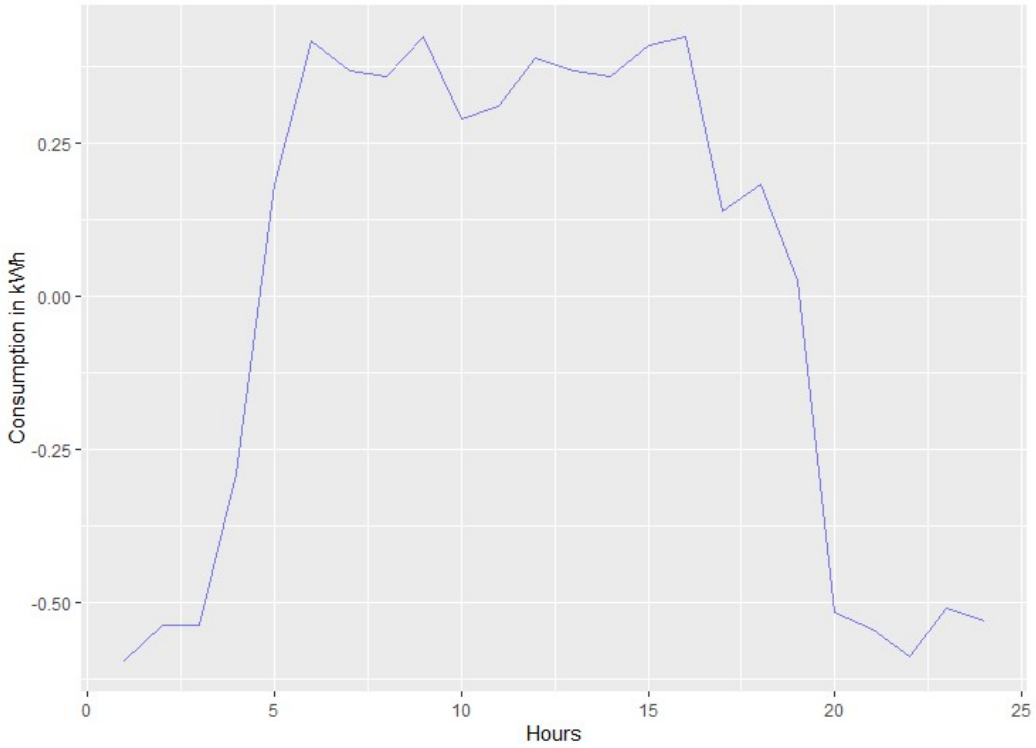

Figure 15. Representation of a reduced and standardized time-series for one location by the hour.

The location in Figure 15 shows a common consumption pattern. It has consumption peaks both in the morning and in the evening with only a little variance during the peak hours between 06:00 to approx. 17:00. Despite the reduction, the main characteristics were preserved.

5.2 Implementation

5.2.1 K-means clustering for heating data set Heating

This experiment was divided into two different main stages according to the objective set of the thesis. In the first one, the heating consumption data set Heating was clustered by using the K-means algorithm. The separation between weekdays and weekends was not implemented since most of the locations were operating 7 days a week. This has a possible effect when calculating the average for each hour within the chosen time interval for the locations that are not operating on weekends. However, it is considered to have a very minor effect within this context.

K-means algorithm was implemented by using 'kmeans' function from R's core package of stats. By default, it uses Euclidean distance and Hartigan and Wong algorithm since it generally does a better job than MacQueen, Lloyd or Forgy [33]. The function outputs k number of clusters, where k is user-specified. To obtain the optimal number of clusters three different methods were being used. These were the Elbow method, Silhouette index, and Calinski-Harabasz index. All methods were run for an interval of k values from 2 to 10. The maximum iteration was set to 1000 and initialization was set to 25. After this, the K-means algorithm was implemented with chosen k value to obtain the final model. Distances to cluster centers were also calculated for each location by using the Equation 2.

Next, the correlation between outside temperature and heating consumption was calculated. This step was done because of the prior investigations done by the company had revealed that some of the locations with high negative correlation had problems with heating and the ventilation of heating. Finally, to extract possible anomalous locations a subset was chosen by interpreting clustering results, distances to cluster centers and correlation coefficient values. Locations with a high negative correlation of less than -0.85 with outside temperature were

considered to need further investigation. A Threshold of -0.85 for correlation was selected to obtain only highly correlated locations and thus avoid extracting unnecessary many of them. Any other threshold could also have been used. Locations far away from cluster centers for each cluster were also extracted for deeper investigation.

The final workflow for this stage is described below:

1. Load and preprocess the data.
2. Find the optimal number of clusters by using the Elbow method, silhouette index, and Calinski-Harabasz index.
3. Execute K-means with the optimal number of k.
4. Calculate distance for each location from cluster centers.
5. Calculate the correlation between outside temperature and heating consumption for each object.
6. Interpret the results and choose a subset for further investigation by using the conditions described above.

5.2.2 K-means clustering for electricity data set Electricity

In the second main stage, the electricity consumption data set Electricity was clustered by using the K-means algorithm. Details and conclusions about the possible minor effect described in chapter 5.2.1 regarding the weekday and weekend separation were the same with this stage. Implementation of 'kmeans' function as well as its parameters was identical to the ones described in the first stage. The optimal number of clusters was also decided similarly by using the Elbow method, Silhouette index and Calinski-Harabasz index with an interval of k values from 2 to 10.

Kmeans algorithm was implemented with chosen k value to receive the final model. Distances to cluster centers were also calculated using Euclidean Distance. Next, the status column indicating whether the store is open or closed during the measured hour was used to create a subset of data to represent opening hours. Locations, where status was set to open at 02:00 and 03:00 were considered to be open 24 hours a day. From this subset of data, only the Enegiaid's of locations were extracted as a result for later use. It was important to consider the fact that

some stores were operating 24 hours a day. To obtain locations with unusually high night-time electricity consumption, that were not operating 24 hours a day, the ones that operate 24 hours a day needed to be marked and known. Otherwise, these locations might show up as possible anomalies although they could be considered as ‘normal’.

Finally, to obtain possible anomalous locations a subset was chosen by interpreting the clustering results and distances to cluster centers. Locations belonging to clusters showing only a little variance between day and night consumption were considered to need deeper examination. Locations far away from cluster centers for each cluster were also extracted for further examination.

The final workflow for this stage is described below:

1. Load and preprocess the data.
2. Find the optimal number of clusters by using the Elbow method, silhouette index and Calinski-Harabasz index.
3. Execute K-means with the optimal number of k.
4. Calculate distance for each location from cluster centers.
5. Create a subset of data based on operating hours with the status of open to extract the ones that are operating 24 hours a day.
6. Interpret the results and choose a subset for further investigation by using the conditions described above.

5.3 Results

5.3.1 The optimal number of clusters for data set Heating

Three evaluation methods were used to find the optimal number of clusters. These were the elbow method, silhouette index, and Calinski-Harabasz index. Values for the number of clusters were predefined between 2 to 10. The results can be seen in Table 2 and Figure 16 below. Bold indicates the optimal value in the table.

Table 2. Results of evaluation methods for data set Heating.

No. of clusters	Elbow index	Silhouette index	Calinski-Harabasz index
2	710.4506	**0.3297037**	**348.6206**
3	598.3950	0.2208532	255.9972
4	531.6164	0.2084539	218.2377
5	476.1858	0.1845867	193.9876
6	449.4879	0.1661069	170.4239
7	428.3373	0.1628811	154.5281
8	405.6530	0.1731461	144.6182
9	389.5961	0.1579432	134.9537
10	375.5965	0.1532775	126.8546

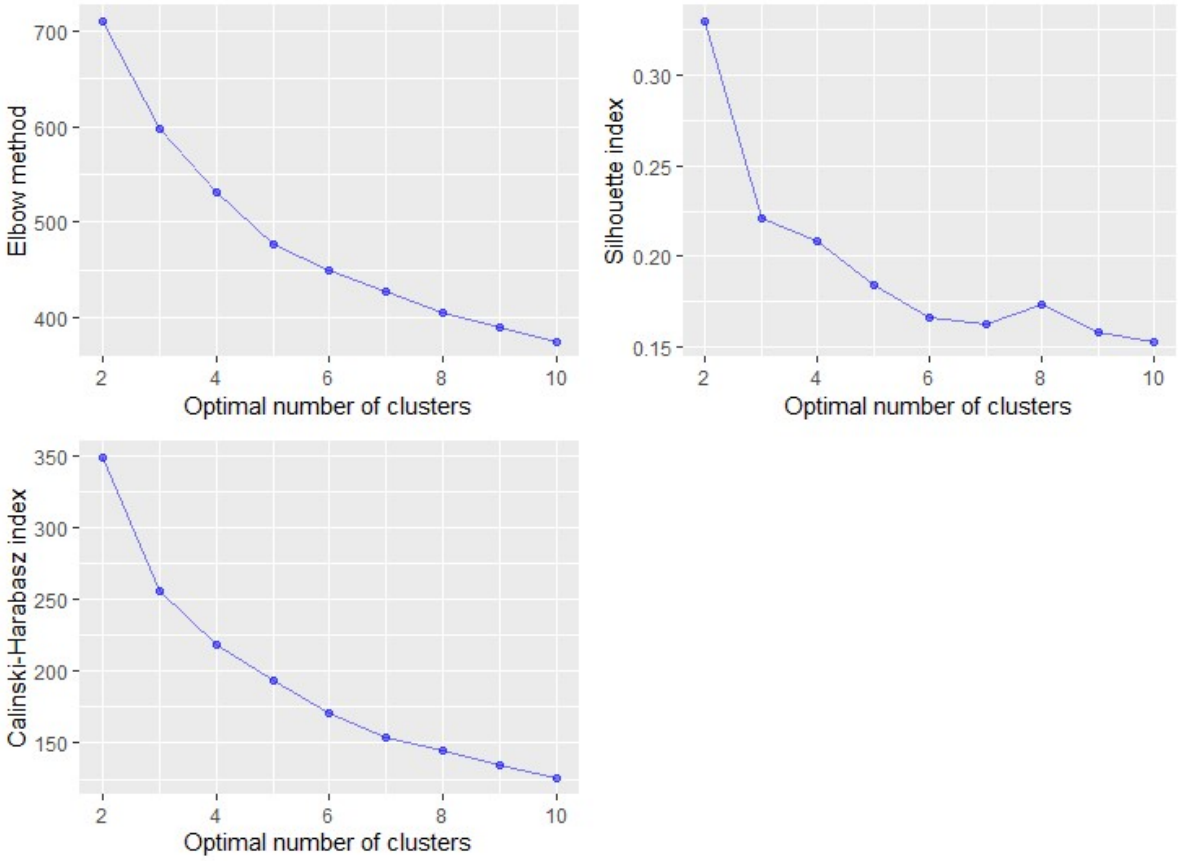

Figure 16. Evaluation metrics for data set Heating.

5.3.2 Clustering results for data set Heating

Both the Silhouette and Calinski-Harabasz index suggested that the optimal number of clusters is 2. Just by plotting cluster results with their respective centers can be seen that 2 is not enough to extract all profiles. The results of the elbow method were inconclusive and optimal value can't be stated confidently. Clustering results for k values of 2, 4, 5 and 8 are represented in Figures 17, 18, 19 and 20 below.

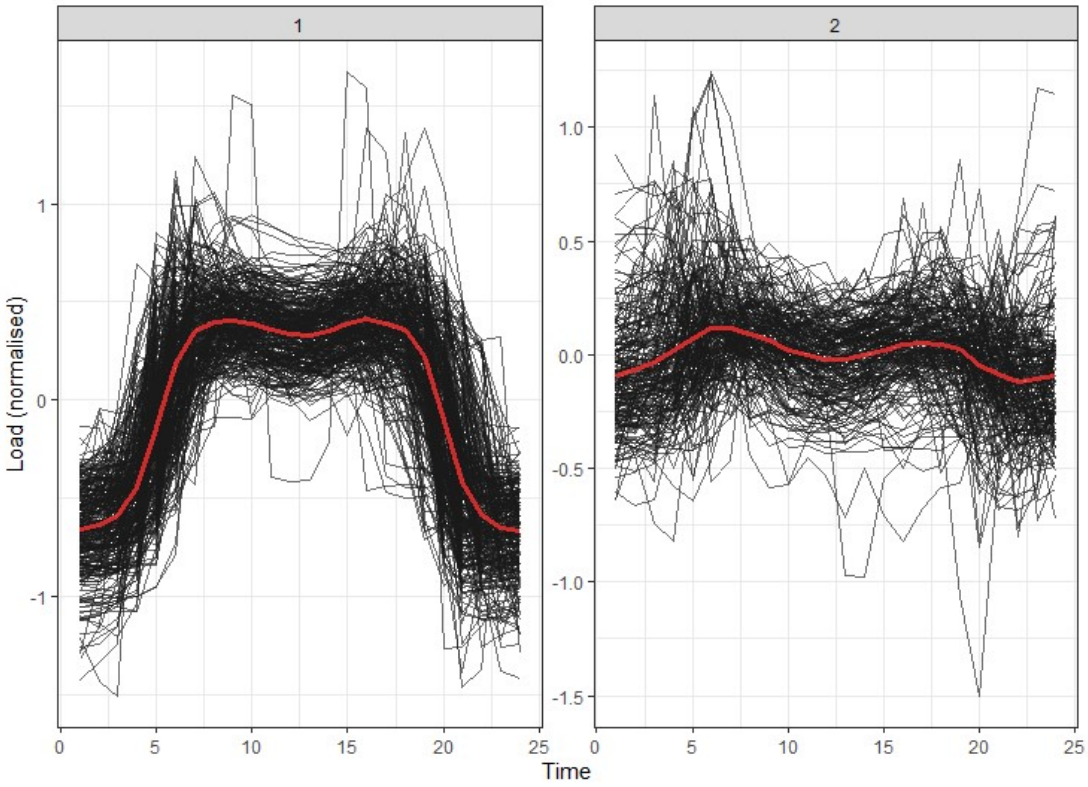

Figure 17. Clustering results with respective cluster centers with k = 2.

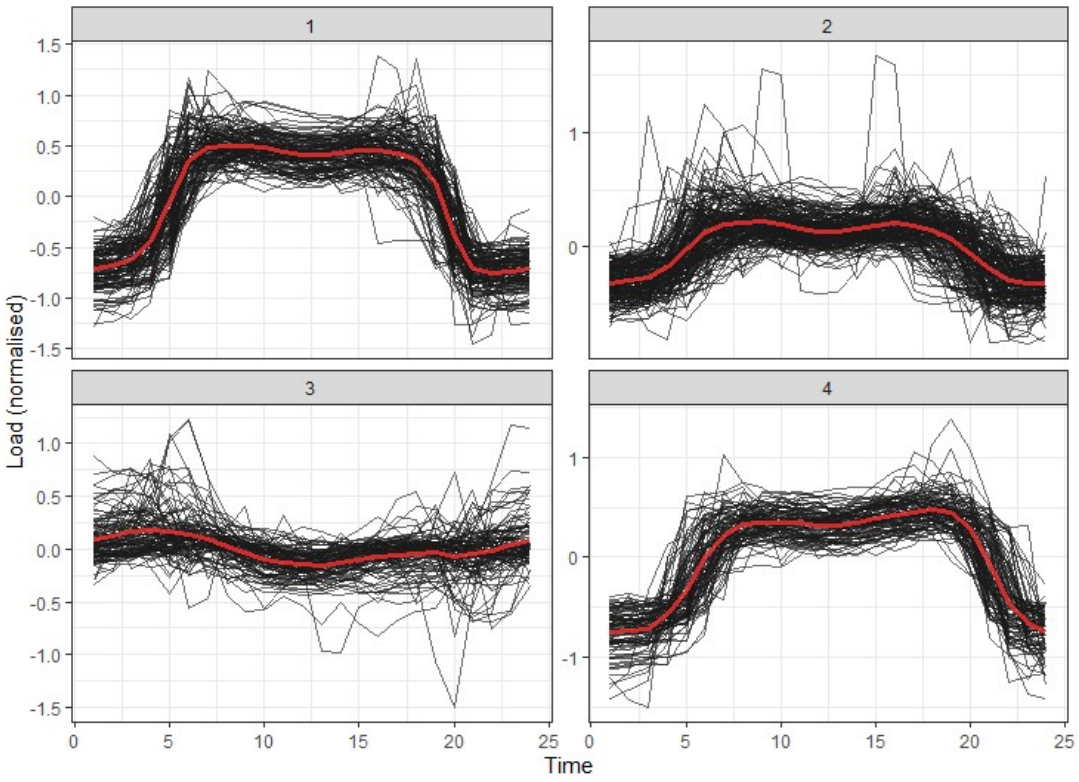

Figure 18. Clustering results with respective cluster centers with k = 4.

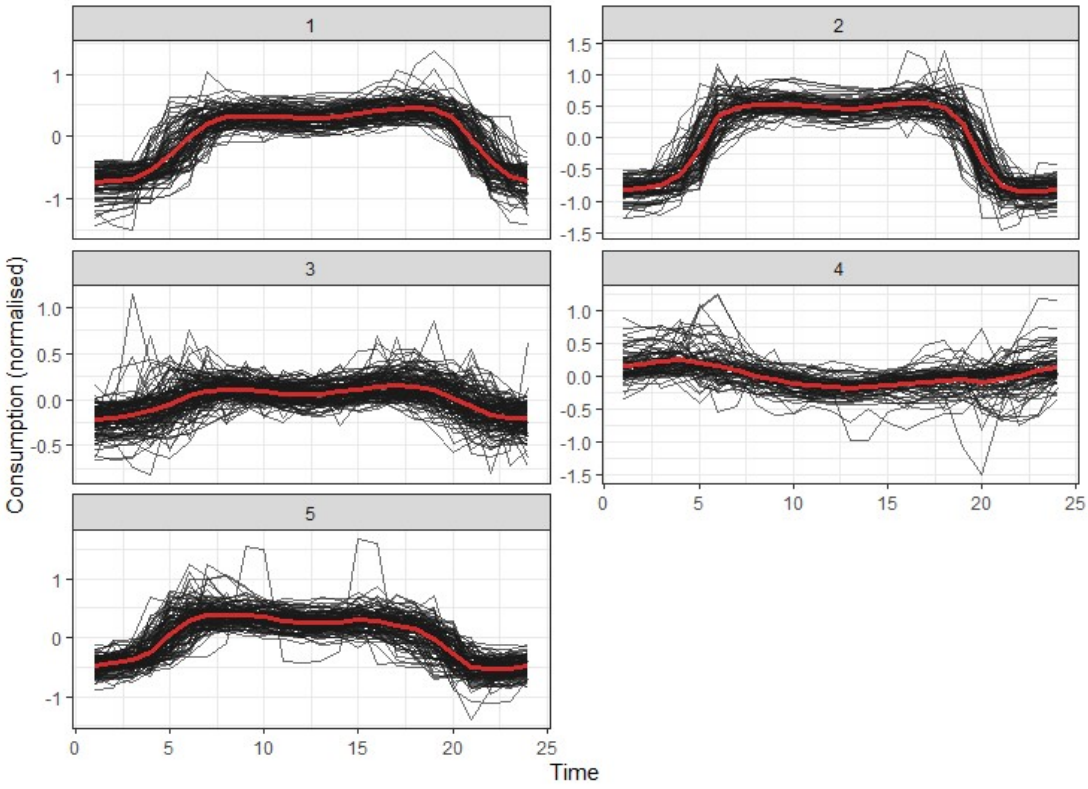

Figure 19. Clustering results with respective cluster centers with k = 5.

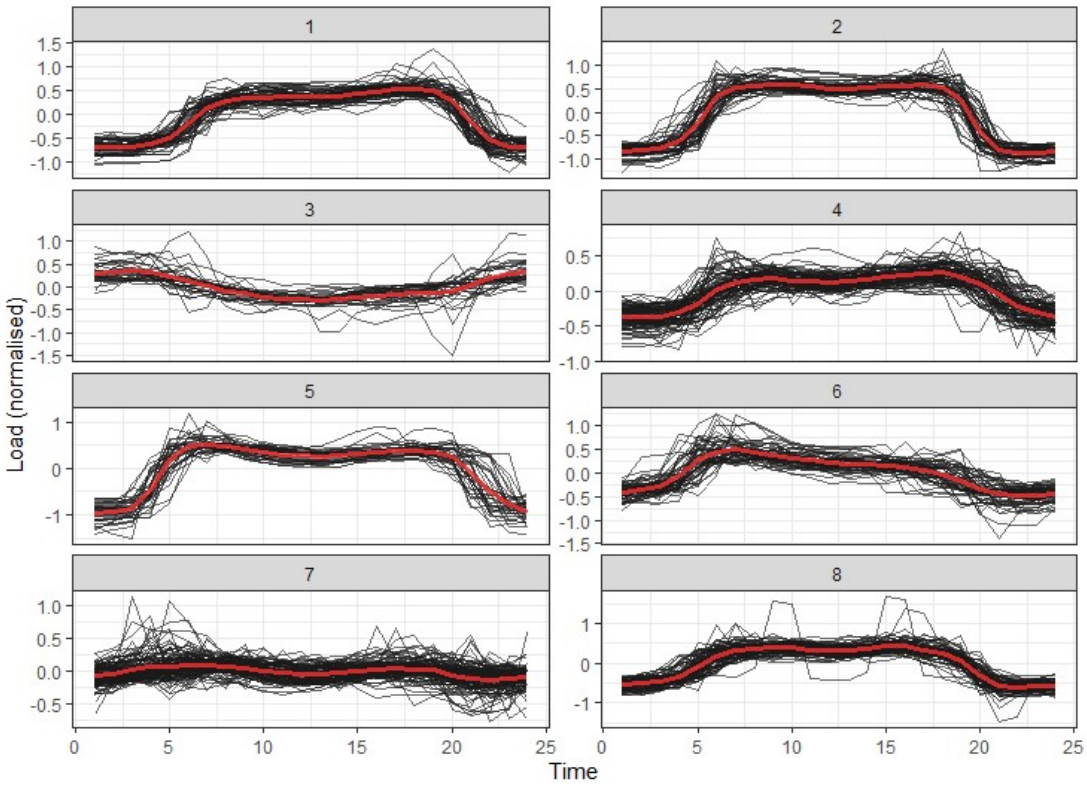

Figure 20. Clustering results with respective cluster centers with k = 8.

K values of 4, 5 and 8 provide more detailed consumption profiles compared to the k-value being 2. To obtain informative profiles, this study continued with 5 clusters. With 5 clusters the members were divided somewhat evenly as can be seen in Table 3 below. Profiles with 2 clusters in Figure 17 or 4 clusters in Figure 18 are considered to be too general and lack information to find possible anomalous locations whereas with 8 clusters in Figure 20 there are unnecessarily many of them. Profiles with quite similar consumption patterns were separately clustered as can be seen for profiles 1 and 2 in Figure 20 for example.

From Figure 19 with 5 clusters, it can be seen that there are locations that show only a little variance in consumption between night and day. All these locations were in cluster 4. Cluster 4 was also interesting because it showed locations where night time consumption was higher than day time. Clusters 1, 2 and 5 represented locations with similar consumption patterns where consumption increased significantly in the morning and then declined in the evening hours. These peaks were a clear indication of opening and closing time for these locations. Throughout the day the consumption remained constant to some extent, although some very

minor decrease in the middle of the day can be seen. Locations in cluster 5 showed on average a bit smaller increase and decrease during the peaks.

Cluster 3 obtained locations where the slope for peaks was substantially lower compared to clusters 1, 2 and 5. Locations in cluster 3 also seemed to have more variation in profiles which can be seen by more dispersed distances from the cluster center.
Locations with unusual consumption patterns for the cluster center existed clearly in every cluster.

Table 3. Cluster size distribution with data set Heating.

Cluster	Members with 2 clusters	Members with 4 clusters	Members with 5 clusters	Members with 8 clusters
1	306	135	101	60
2	225	188	85	57
3		107	134	33
4		101	80	92
5			131	36
6				55
7				115
8				83

5.3.3 Correlation with outside temperature for data set Heating

After clustering, the correlation for heating consumption with outside temperature was calculated for each location. The distribution of correlation and correlation against variance in consumption are represented in Figure 21. The majority of the correlation was negative as expected because when outside temperature decreases heating consumption increases.

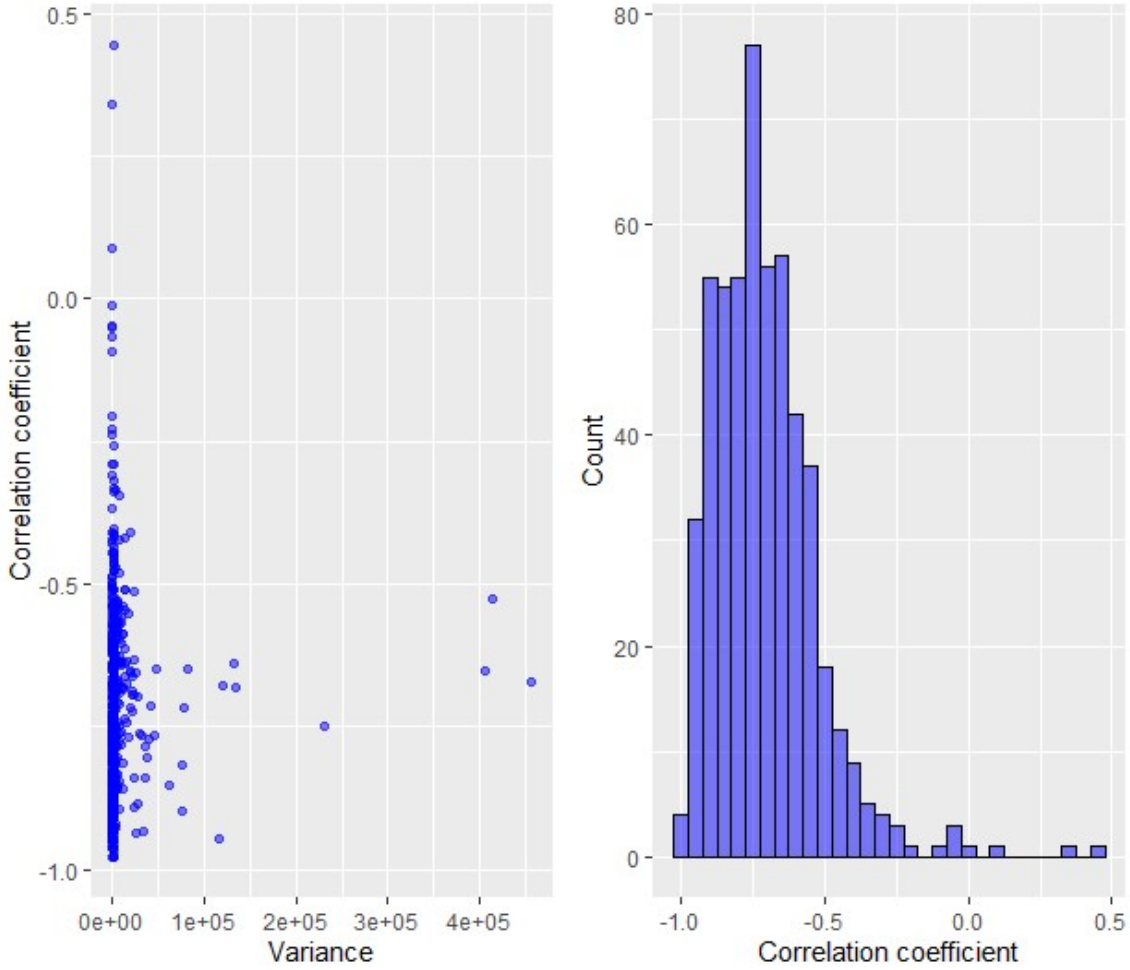

Figure 21. Scatter plot of correlation against variance on the left and distribution of correlation as a histogram on the right.

Only 3 locations have a positive correlation and 8 have variance over 100000. Locations with high negative correlation tend to have low variance but it's not so evident.

5.3.4 Possible anomalous locations for data set Heating

The final stage was to extract possible anomalous locations by conditions and thresholds given at the end of chapter 5.3.1. This part was divided into two different approaches for clarification. In the first approach, locations with a negative correlation lower than or equal to -0.85 in cluster 4 yielded 45 unique locations.

In the second approach, anomalous locations by using distances from cluster centers were extracted by using the Interquartile Range (IQR). IQR can be considered to be a measure of

variability, based on dividing the data set into different quartiles. The values splitting each part are called the first, second, and third quartiles. These are denoted later as Q1, Q2, and Q3. Q1 is the “middle” value in the first half of the data set. Q2 is the median value of the set. Q3 is the “middle” values in the second half of the data set. The maximum value used is 1.5 times IQR above the Q3 and the minimum value used is 1.5 times IQR below the Q1 [34].

This resulted in 20 unique locations. Results for IQR are presented in Table 4 and Figure 22 below in the form of a boxplot. Anomalous locations are shown as black dots. The threshold for anomalous locations was calculated to be greater than 1.74 according to Maximum in Table 4.

Table 4. IQR results for data set Heating.

Minimum	**Q1**	**Q2**	**Q3**	**Maximum**
0.169	0.665	0.860	1.106	1.743

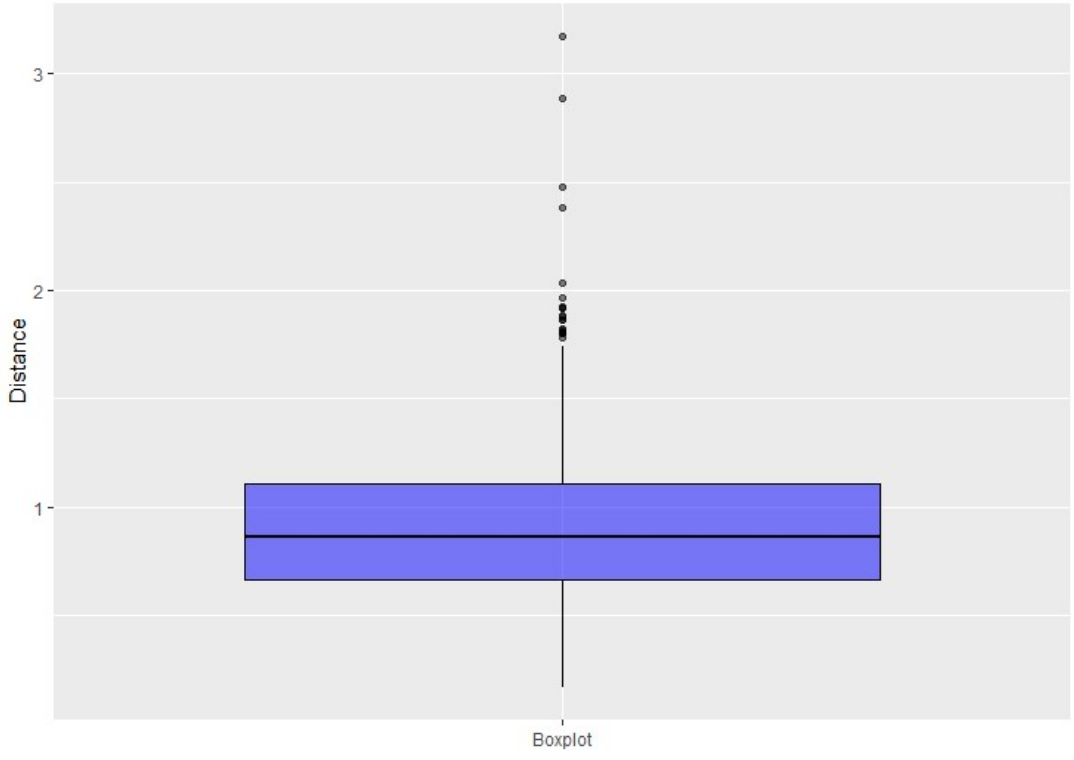

Figure 22. Boxplot representing IQR results for data set Heating.

Cross-checking locations of both results indicated one duplicate value which was then removed. The first approach resulted in 45 locations and the second one 20 with one duplicate value. Thus, the final data set contained 64 unique locations. Visualizations of two randomly selected possible anomalous locations are represented in Figures 23 and 24. For both figures, the individual blue lines represent one particular day within the given time interval of 3 months.

Figure 23 describes the heating consumption of one randomly selected location and temperature variation from the nearest weather station extracted by using the first approach. The mean value is shown as a black line. This selected location has slightly higher consumption during night time. Correlation with temperature is -0.97. In Figure 24 one distance-based anomalous location by using the second approach is represented.

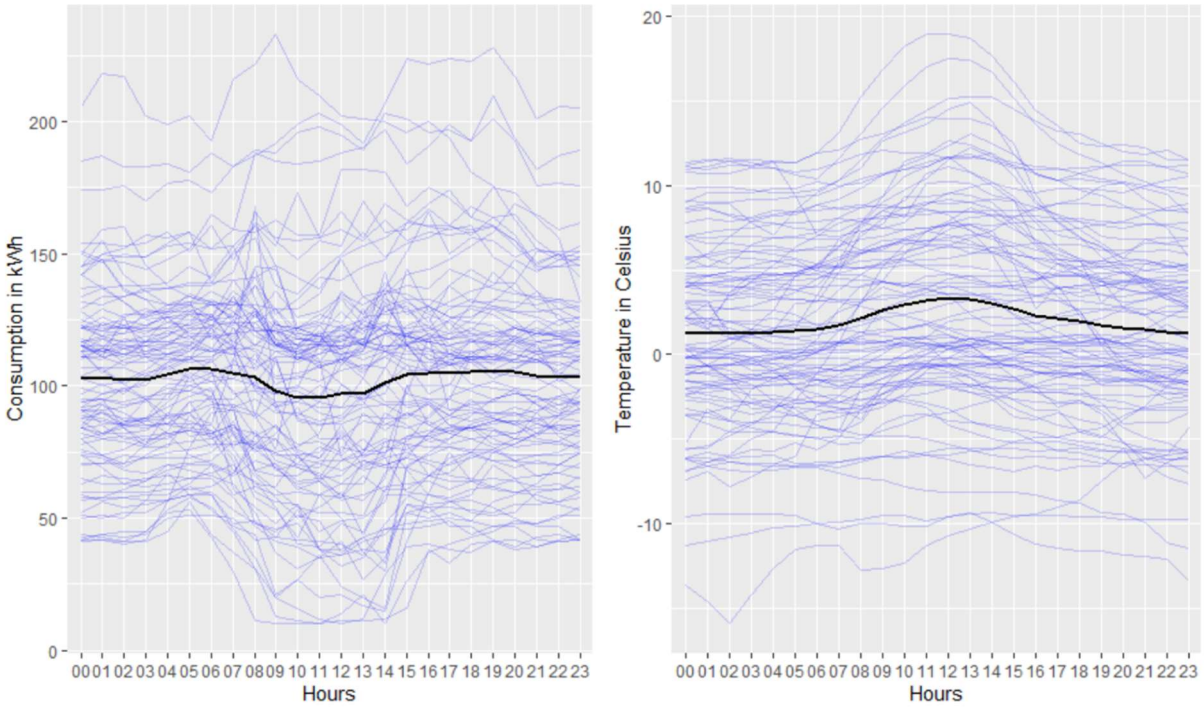

Figure 23. Heating consumption for one randomly selected anomalous location. On the right side is outside temperature measured from the nearest weather station in relation to the location. Mean values are shown as a black line. The individual blue lines represent one particular day within the given time interval of 3 months.

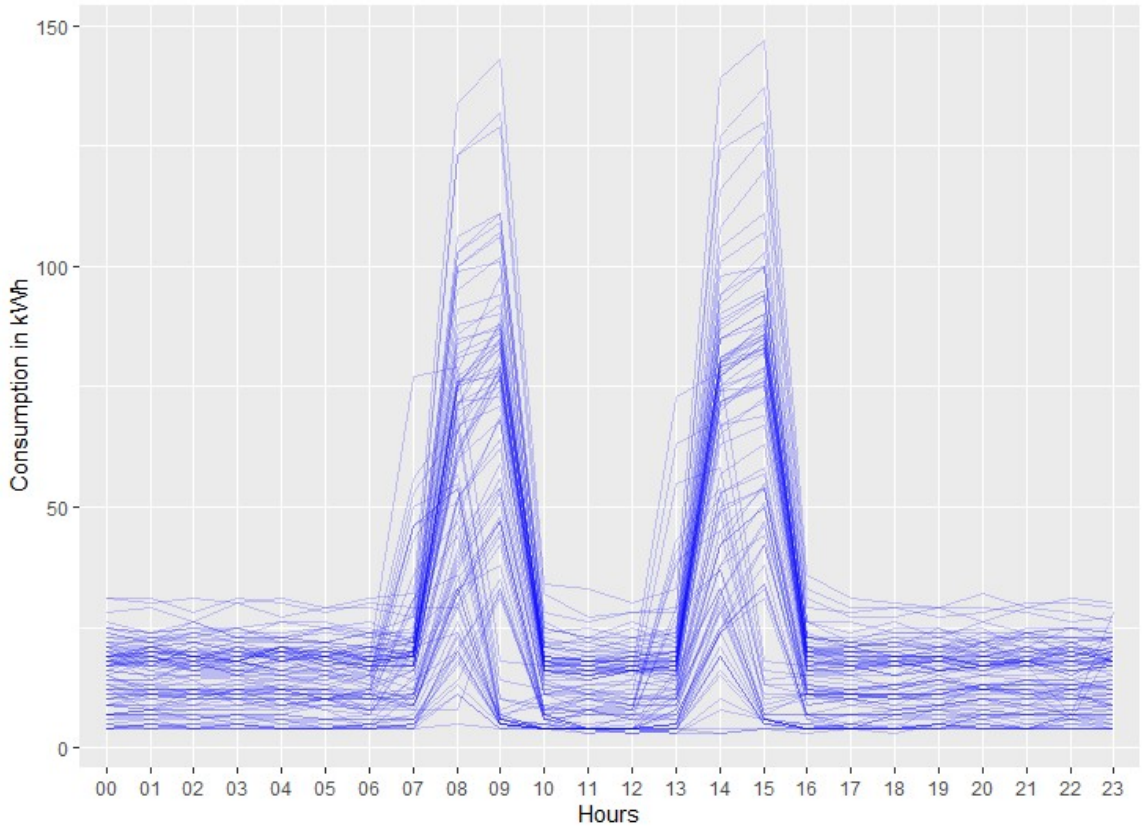

Figure 24. Heating consumption for one randomly selected anomalous location based on distance from the cluster center. The individual blue lines represent one particular day within the given time interval of 3 months.

5.3.5 The optimal number of clusters for data set Electricity

Three evaluation methods were used to find the optimal number of clusters for data set Electricity. These were Elbow method, Silhouette index, and Calinski-Harabasz index. Values for the number of clusters were predefined between 2 to 10. The results can be seen in Table 5 and Figure 25 below. Bold indicates the optimal value in the table.

Table 5. Results of evaluation methods for data set Electricity.

No. of clusters	Elbow index	Silhouette index	Calinski-Harabasz index
2	4635.794	0.1869108	566.1432
3	**3564.343**	**0.2245251**	**602.4107**
4	3248.023	0.1741074	513.6818
5	2945.107	0.1526974	473.1157
6	2719.354	0.1647750	442.7331
7	2552.374	0.1667112	413.7030
8	2401.789	0.1639741	393.4930
9	2294.519	0.1489811	373.5618
10	2197.314	0.1558761	356.5328

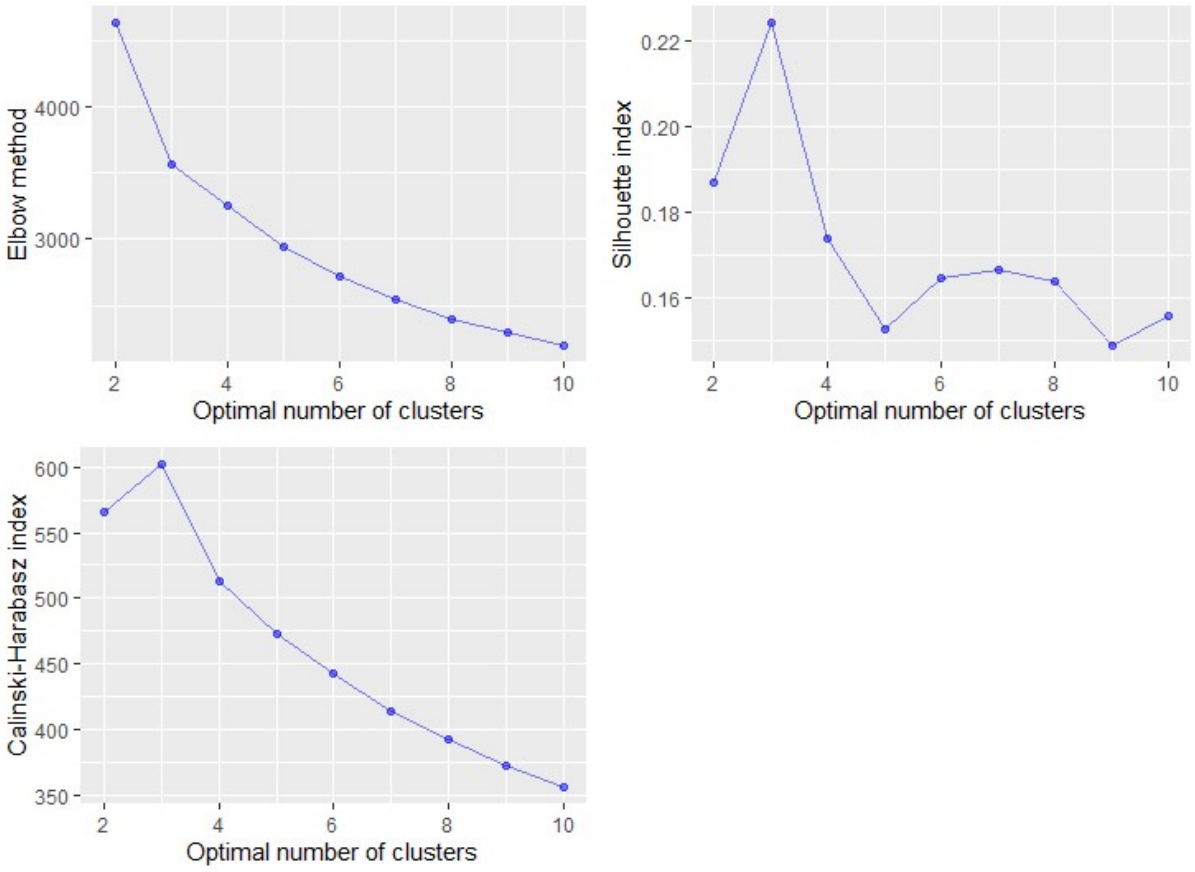

Figure 25. Evaluation metrics for data set Electricity.

5.3.6 Clustering results for data set Electricity

Results were unanimous when inspecting the three evaluation methods. All three evaluation methods suggested that the optimal number of clusters is 3. However, also this time the results with the elbow method were somewhat inconclusive and can be argued.

Again, by plotting cluster results and their respective centers, can be seen that 3 was not the optimal number to be able to extract detailed and meaningful profiles indicating possible high consumption during night time. Clustering results for k values of 2, 3, 5, 6 and 8 are represented in Figures 26, 27, 28, 29 and 30 below.

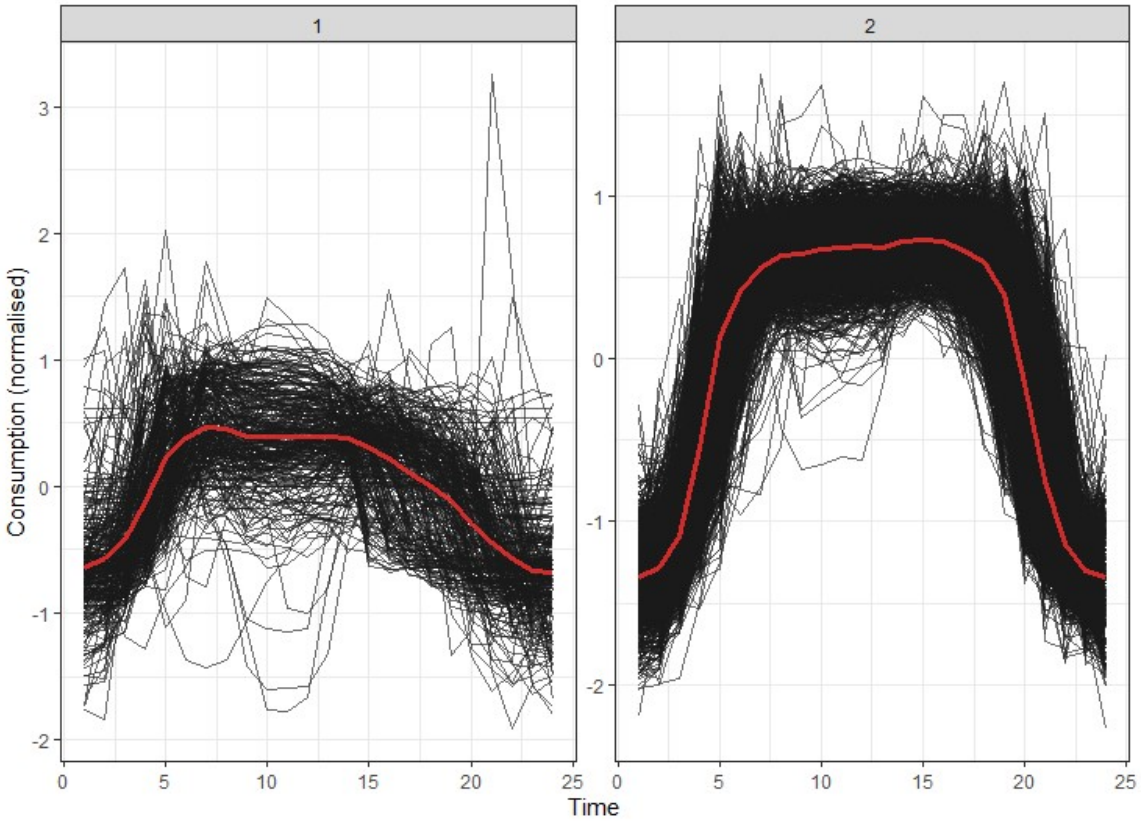

Figure 26. Clustering results with respective cluster centers with k = 2.

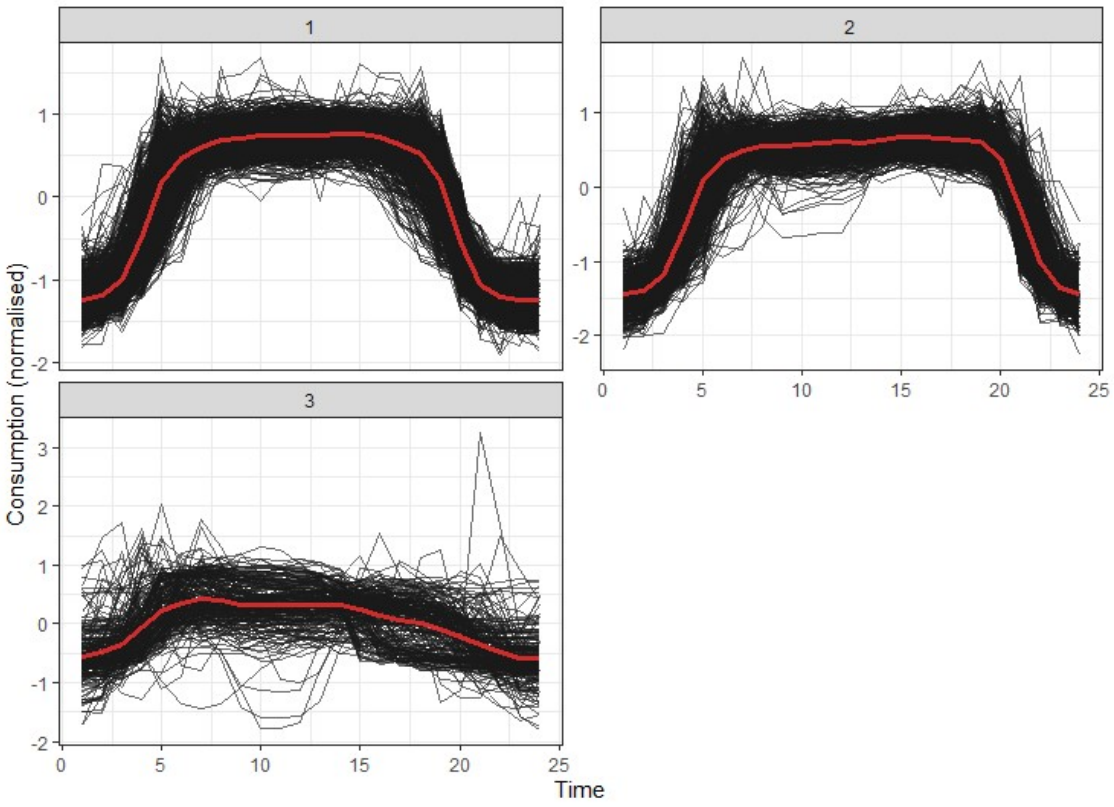

Figure 27. Clustering results with respective cluster centers with k = 3.

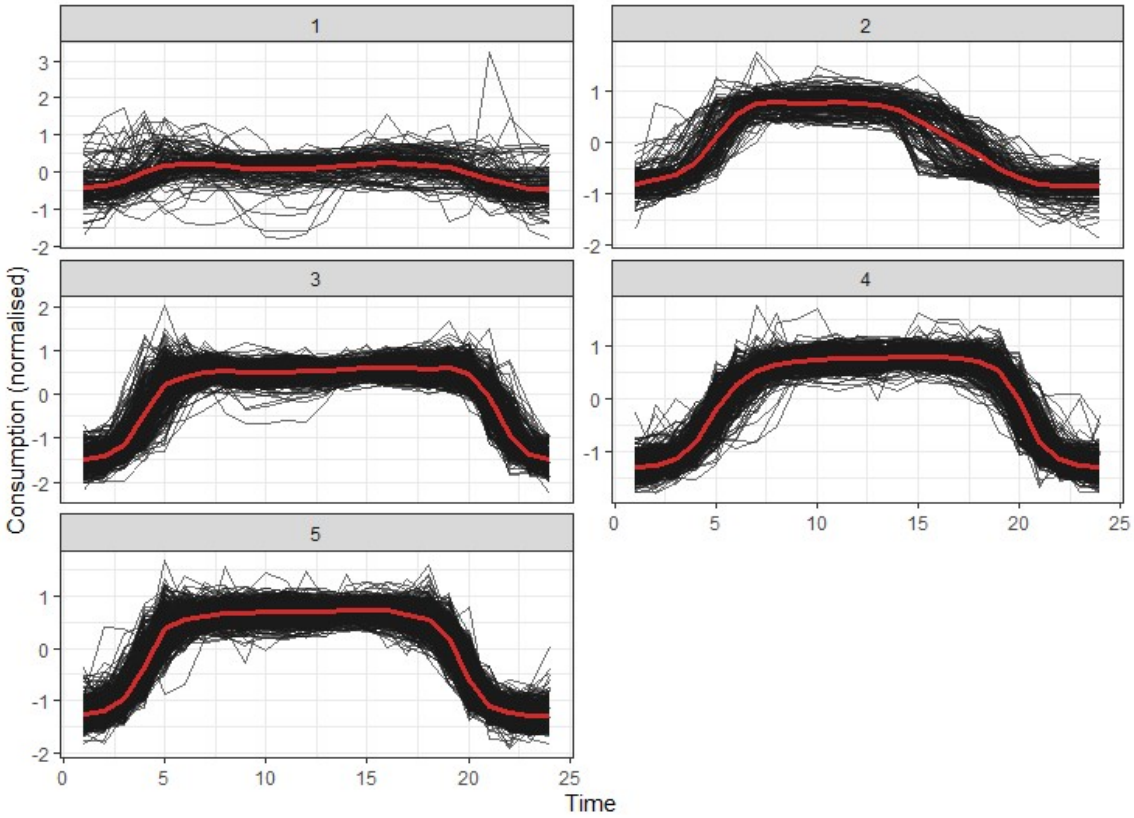

Figure 28. Clustering results with respective cluster centers with k = 5.

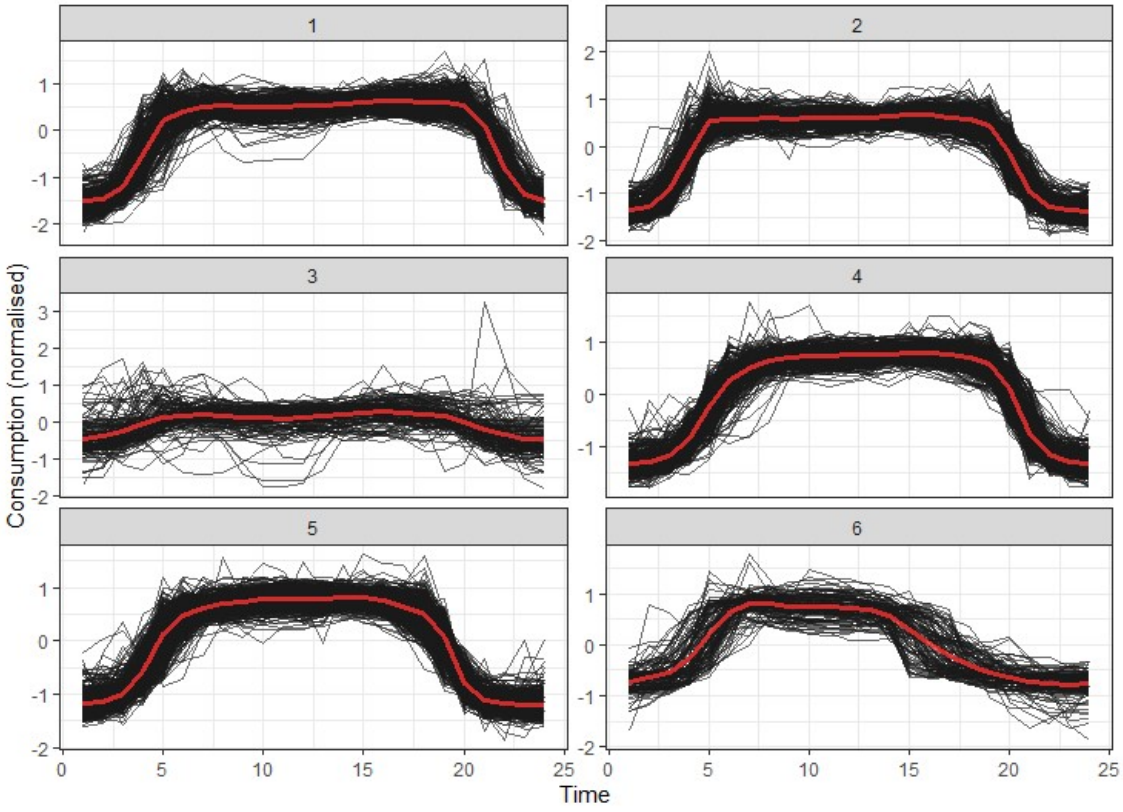

Figure 29. Clustering results with respective cluster centers with k = 6.

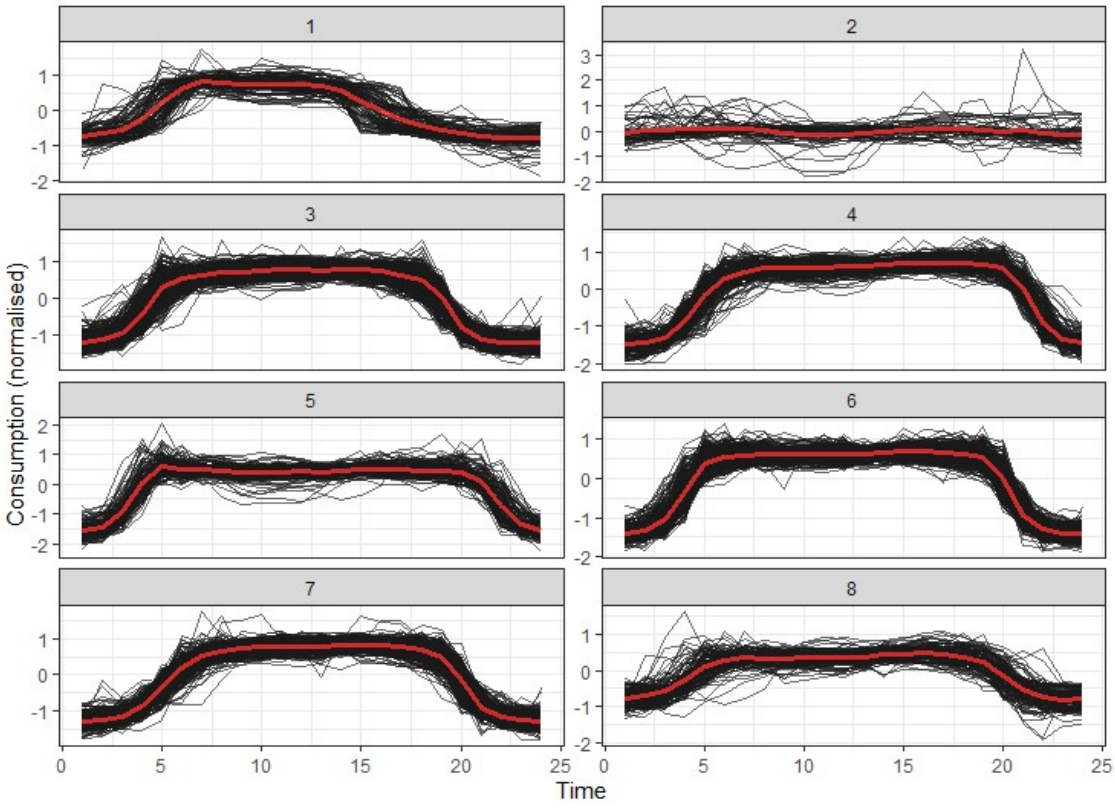

Figure 30. Clustering results with respective cluster centers with k = 8.

To find possible anomalous locations where night time consumption was high, this study continued with 8 clusters. With 2 clusters in Figure 26, 3 clusters in Figure 27 or 5 clusters in Figure 28 the shapes were deviating too little to be able to extract ‘flat’ profiles that can be seen in cluster 2 in Figure 30 when dealing with a k value of 8.

When dealing with 6 clusters in Figure 29, cluster 3 represented the one with ‘flat’ profiles. However, it still had too many locations not suitable for our desired outcome since it showed a minor up peak for the day time consumption. Cluster size distribution in Table 6 was also indicating that with 8 clusters the cluster with possible anomalous locations has 56 locations compared to the one with 6 clusters that has 139 locations, although both indicated characteristics of an outlier cluster compared with the rest of the distributions of members.

Profiles of the locations in clusters 3, 4, 5, 6 and 7 all showed somewhat similar consumption patterns. Incline peak in the morning and decline peak in the evening with quite steep slopes. The consumption during the day also remained very steady. Cluster 8 was a bit different concerning the steepness of the slope towards the peaks but other than that very similar to the others. Cluster 1 was the only one besides cluster 2 that deviated from the majority with a gradual decline peak starting at 15:00. These peaks represented the opening and closing times of these locations. Distribution of size was also different for cluster 1 and 8 related to others. However, cluster 5 also had a smaller size of members but probably because the declining peak was occurring a bit later than with others which indicated later closing times for these locations. Locations with unusual consumption patterns concerning the cluster center existed clearly in every cluster and mostly in cluster 2 in Figure 30 with 8 total clusters.

Table 6. Cluster size distribution with data set Electricity.

Cluster	Members with 2 clusters	Members with 3 clusters	Members with 6 clusters	Members with 8 clusters
1	305	1059	350	111
2	1732	741	432	56
3		237	139	518
4			451	264
5			546	156
6			119	448
7				345
8				139

5.3.7 Extracting 24-hour stores for data set Electricity

Before clustering, a subset of locations operating 24 hours was extracted. This was achieved by choosing stores that were labeled open at 02:00 during the night by using ‘open’ as a value of the status column. This approach was adopted since there was no better information available and it should be noted that there is a possibility for false statuses. However, when manually investigating the opening hours for 10 random stores, they all verified to be open at 24 hours. This subset of data contained 41 locations.

It was expected that at least some of the stores operating 24 hours a day would be clustered into cluster 2 in Figure 30. This was based on an assumption that electricity consumption for the locations operating night hours was similar or higher to the usage of day time. However, this assumption can’t be generalized since some 24-hour stores probably have a lot more traffic during the day time compared with night time at least in bigger cities. More people and higher frequency of opening cold storage doors, for example, might have an impact on the total electricity consumption.

11 locations out of 56 in cluster 2 were labeled and verified to be 24-hour stores. This indicated that at least for some locations the assumptions were correct and 24-hour stores had a “flat”

profile. However, 24-hour stores seemed to be distributed evenly into every cluster, which was surprising. Reasons for this phenomenon would be another interesting topic to investigate but it's outside of the scope for this study. The majority of them were distributed to clusters 2, 5, 7 and 8. In Figure 31 below we can see a histogram representing a distribution of 24-hour stores within clusters from 1 to 8.

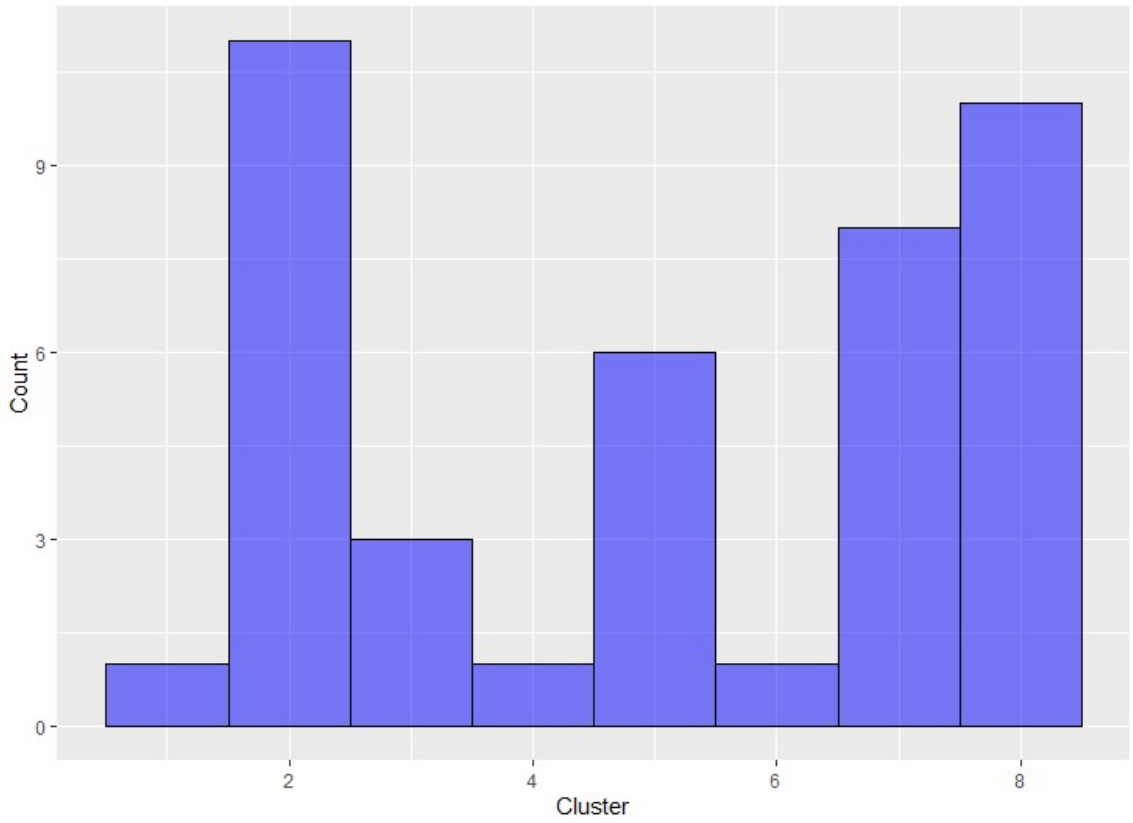

Figure 31. Histogram of the distribution of 24-hour stores within clusters from 1 to 8.

One randomly selected store from cluster 2 that was verified to be operating on a 24-hour basis is represented in Figure 32 below. The selected location is showing underlying steady consumption throughout the day, but variance and peaks can also be detected. However, a major peak in consumption occurs during the 14:00 to 18:00 and another minor one occurs from 0:00 to 02:00.

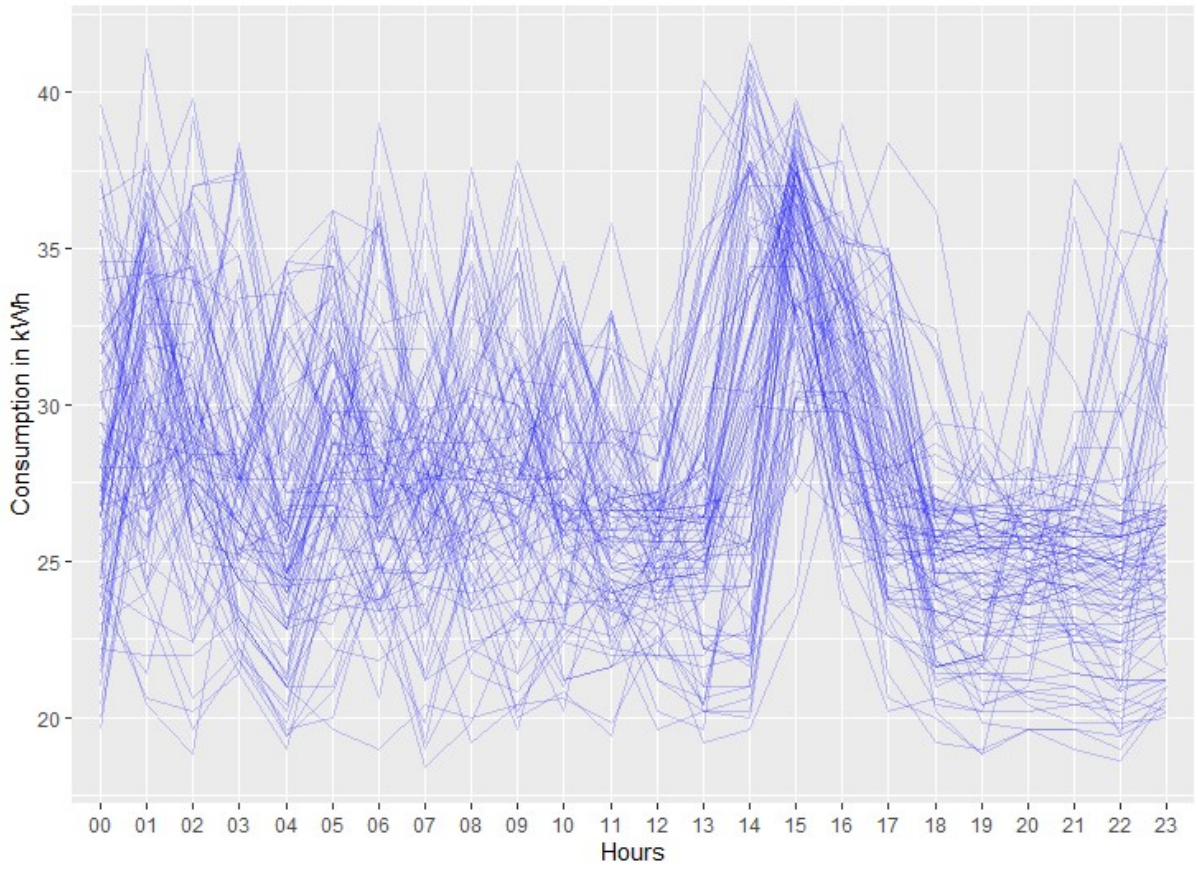

Figure 32. Representation of a randomly selected 24-hour location by the hour.

5.3.8 Possible anomalous locations for data set Electricity

The final stage was to extract possible anomalous locations by conditions given at the end of chapter 5.3.2. This part was divided into two different approaches for clarification. In the first approach, possible anomalous locations in cluster 2 in Figure 30 contained 56 unique locations. These locations had a “flat” profile and were showing only a little variance between day and night time consumption.

For the second approach, IQR was again used when extracting locations based on their distances from the cluster centers. This yielded 84 unique locations. Results for IQR are presented in Table 7 and Figure 33 below in the form of a boxplot. Anomalous locations are shown as black dots. The threshold for anomalous locations was calculated to be greater than 1.82 according to Maximum in Table 7.

Table 7. IQR results for data set Electricity.

Minimum	Q1	Q2	Q3	Maximum
0.255	0.748	0.927	1.177	1.815

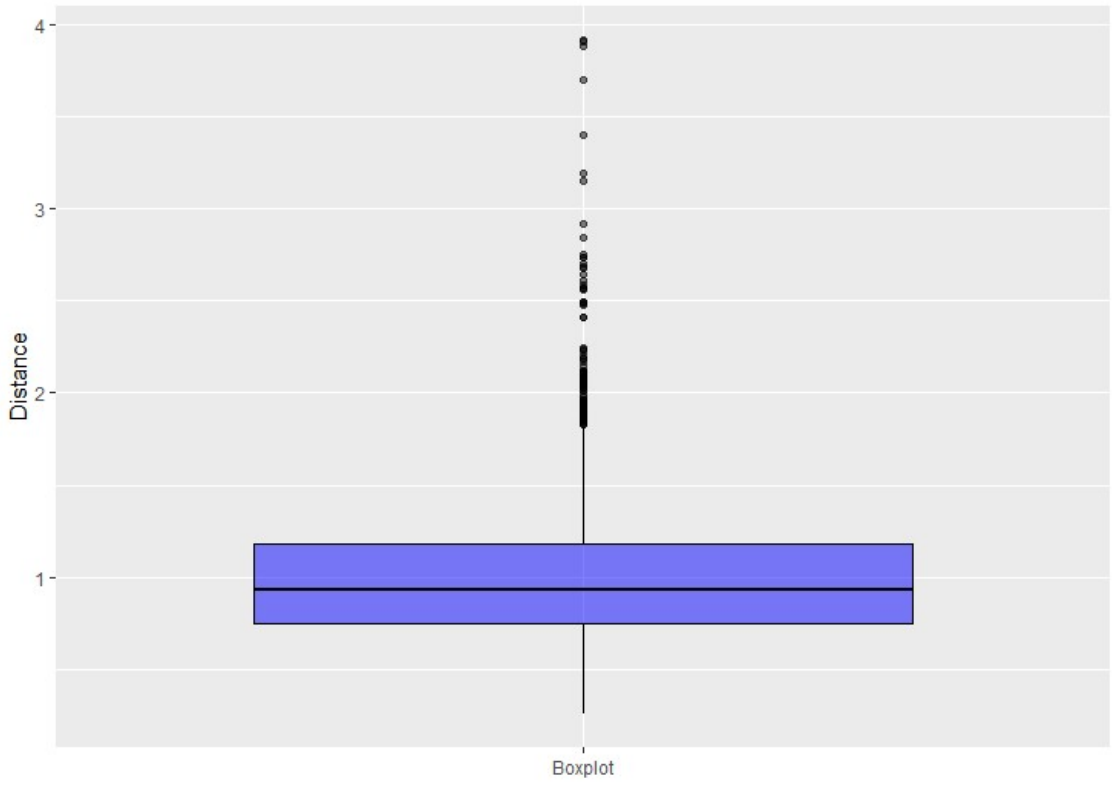

Figure 33. Boxplot representing IQR results for data set Electricity.

Cross-checking locations of both results indicated 14 duplicate values. The first approach resulted in 45 locations after removal of 11 24-hour locations extracted earlier and described in chapter 5.4.7. The second one resulted 84 locations. After removing 14 duplicate values, the final data set contained 115 unique locations.

Visualizations of the result set are represented in Figures 34 and 35 by two randomly selected possible anomalous locations. The first location shown in Figure 34 represents a member of cluster 2 from figure 30. This specific location is showing no difference during the day and night regarding consumption. The second location shown in Figure 35 represents a member

from cluster 5 and it is considered to be a shape-based anomaly due to its distance from the cluster center. The location has a very unique consumption profile and it deviates a lot from the cluster mean represented in Figure 30 in cluster 5.

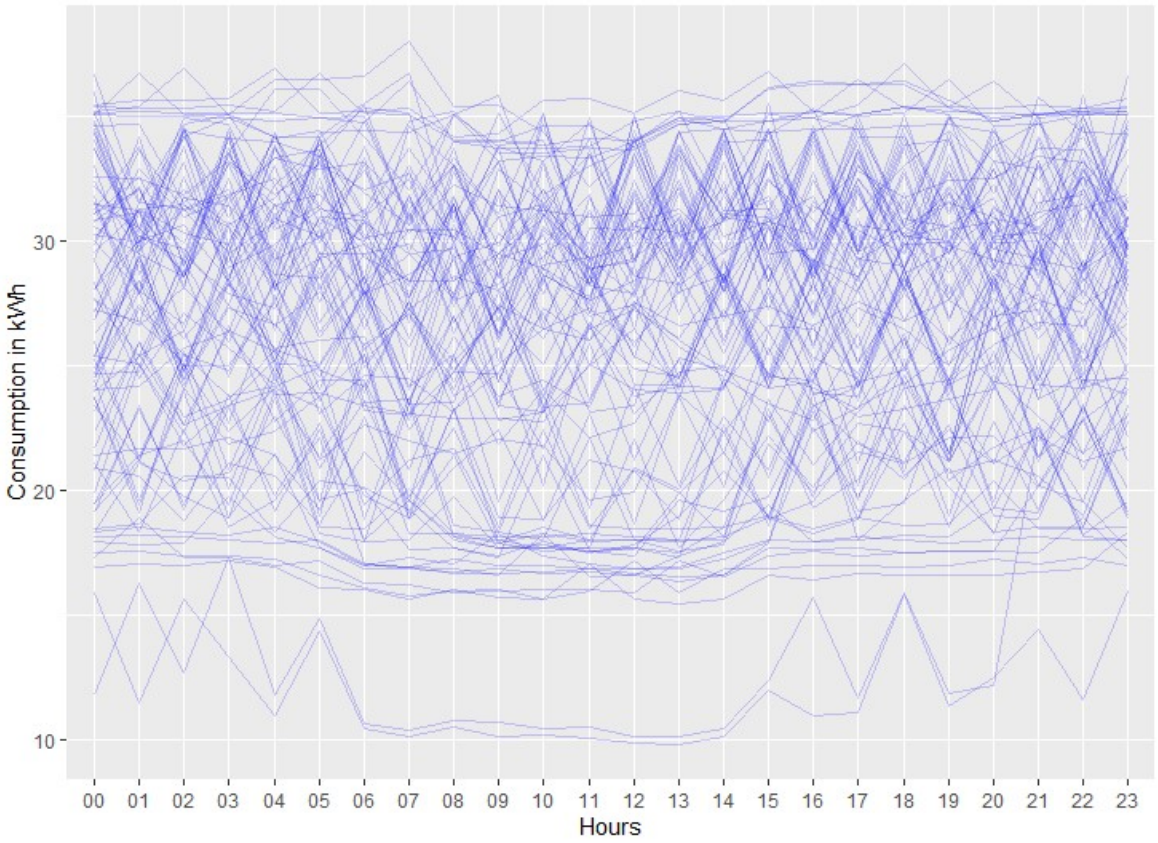

Figure 34. Electricity consumption for one randomly selected anomalous location by the hour.

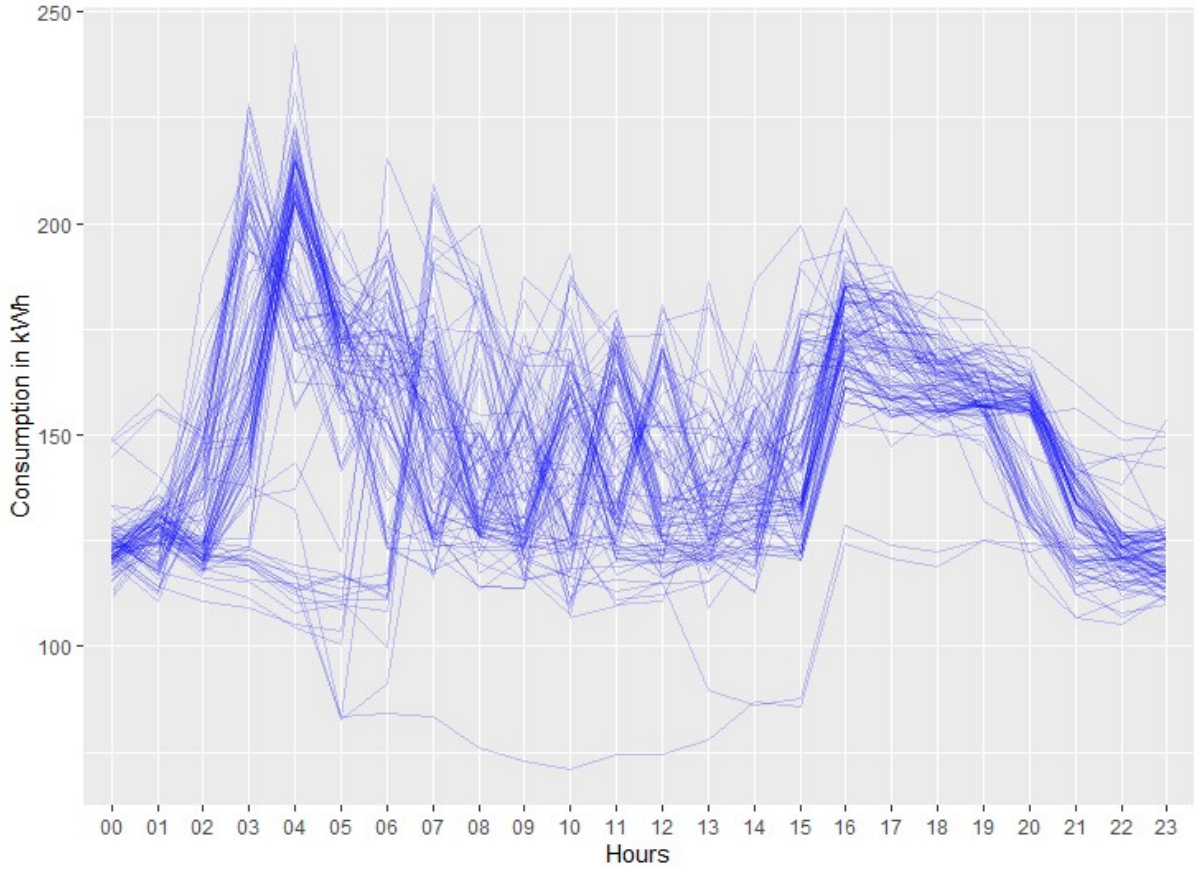

Figure 35. Electricity consumption for one randomly selected anomalous location based on distance from the cluster center by the hour.

5.3.9 Summary of the results

For both data sets Electricity and Heating possible anomalies were found. Heating data set contained 531 locations before clustering and extraction of possible anomalies. By implementing the methods described in the workflow in chapter 5.3.1 resulted in 64 possible anomalies.

Similarly for Electricity data set contained 2037 locations before clustering and extraction of possible anomalies. After performing steps described in the workflow in chapter 5.3.2 resulted in 115 possible anomalies. After performing cross-checking between Heating and Electricity data set anomalies 8 duplicate values were found that were anomalous both in Heating and Electricity. Overall out of 2037 locations 179 (8.8%) were possible anomalies and 171 of them were unique. This was considered to be an indication that there is no clear relation between

anomalous heating and electricity consumption through 8 locations that appear to have possible anomalous behavior with both. Further investigation on possible anomalies found and their verification was left for the company experts since this task can’t be performed to obtain reasonable and verified conclusions without domain knowledge.

6 DISCUSSION AND CONCLUSIONS

Inspiration for this study was the fact that by finding possible anomalous locations may have an impact on saving natural resources by reducing and optimizing energy consumption.
The first objective of this thesis was to cluster heating consumption data using the K-means algorithm to obtain load profiles and then extract possible anomalous locations by using correlation with outside temperature and variance. The second objective for the thesis was similar to the first one and K-means clustering was used to obtain load profiles for electricity consumption data set. Profiles with high baseload during the night were extracted as possible anomalies. In both cases also locations that were deviating a lot from their respective cluster centers were considered as possible anomalies by using IQR.

These load profiles received as a side product are already valuable and will help the company to understand its customers better and serve as a base for possible future research. A deeper investigation of these profiles is desirable and can inspire or reveal valuable information to work with.

As pointed out by previous studies on the clustering of load profiles deciding the optimal number of clusters was considered to be a difficult task. Validity indices used did not provide the best optimal number of clusters for the desired outcome, which was to obtain meaningful and informative profiles. This might be due to the nature of these validity indices, which were not created to evaluate whether a cluster is informative for the user. Still, it was considered to be a good approach to use some validation for reference and this was something also done in all previous studies with load profiling. However, it should be noted that without any validation measures used to decide which clusters are important and provide the information needed is a matter of one's opinion. From this perspective, a better understanding of energy systems, meters, and devices used in the field of grocery and retail would have been valuable.
Challenges for consumption data sets were the big size of the data and limitations of computational power. For the additional metadata, the main challenges were zero values or missing ones. This was not considered to affect finding anomalous locations since the main task for this data in this study was to provide the necessary building type of 11 to extract the subset of data needed.

Overall two types of different possible anomalies were found with the total number of 171. These were distance or shaped-based anomalies and the locations from the small anomalous clusters with low variance or high negative correlation with the outside temperature. Inspection of a possible relation between these anomalies or some common factors causing them was considered to be outside of the scope of this study.

A deeper investigation of the anomalies received from this study would be intriguing. Usually for anomalies to occur there is a reason causing them. It could be possible that behind some groups of anomalies a common factor or factors are causing them. Improved additional metadata with more detailed features such as the number of windows, visitor count or detailed description of cooling systems used with manufacturing year could be used for example. More detailed weather data with more features such as solar radiation, wind speed, etc. could be used. However, it could also be that there is no common abstract factor causing the anomalies and it's due to human error of incorrect setup for ventilation or cooling systems for example. These types would be harder to understand since it might be that not everything is marked as it should be in the paper or in the data stored.

In the future different kinds of clustering algorithms such as K-medoid, Hierarchical clustering, etc. could be tested by using different distance measures and data reduction techniques to see whether the results are similar to the ones obtained from this study. It would be an interesting topic also to compare results with different time intervals and seasons to see whether possible anomalous locations are occurring all the time or just a specific time.

If true anomalies exist, they can be labeled. This labeled data can then be used as training data for supervised classification models such as neural networks, decision trees, etc. This future model would then be able to classify possible anomalous behavior more effectively with the chosen method. As an example, this kind of solution could be implemented for some energy monitoring system and it could provide instant feedback when raw consumption data is fed into it.

7 REFERENCES

[1] European Comission 2019, Available: https://ec.europa.eu/energy/en/topics/energy-efficiency/energy-performance-of-buildings/overview

[2] Energiavirasto 2019, Available: https://energiavirasto.fi/documents/11120570/13026619/Raportti-National-report-2019-Finland/5f0408b2-5903-11cf-29a3-3a6d0ed0d2a5/Raportti-National-report-2019-Finland.pdf

[3] Motiva 2012, *Kaupan kylmälaitteiden ja -järjestelmien lauhdelämmön käyttöönotto*. Available: https://www.motiva.fi/files/7973/Kaupan_kylmalaitteiden_ja_-jarjestelmien_lauhdelammon_talteenotto_Laskentaohje.pdf.

[4] S Zhong & K Tam 2015, *Hierarchical Classification of Load Profiles Based on Their Characteristic Attributes in Frequency Domain*.

[5] Lundström, L. 2017, *Adaptive Weather Correction of Energy Consumption Data*.

[6] Hayn, M., Bertsch, V. & Fichtner, W. 2014, "Electricity load profiles in Europe: The importance of household segmentation", *Energy Research & Social Science*, vol. 3, pp. 30-45.

[7] Abreu, J., Pereira, F. & Ferrão, P. 2012, "Using pattern recognition to identify habitual behavior in residential electricity consumption", *Energy and Buildings*, vol. 49, pp. 479-487.

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

[9] Haben, S., Singleton, C. & Grindrod, P. 2016, "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data", *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 136-144.

[10] Elexon 2012, *Load Profiles and their use in Electricity Settlement*. Available: https://www.elexon.co.uk/wp-content/uploads/2013/11/load_profiles_v2.0_cgi.pdf.

[11] Zhou, K., Yang, S. & Shen, C. 2013, *A review of electric load classification in smart grid environment*.

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

[13] Räsänen, Teemu & Voukantsis, Dimitrios & Niska, Harri & Karatzas, Kostas & Kolehmainen, Mikko. 2010. ”Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data.” *Applied Energy*. 87

[14] Laurinec, P. & Lucka, M. 2016, ”Comparison of Representations of Time Series for Clustering Smart Meter Data.”

[15] Laurinec, P. & Lucka, M. 2018, "Clustering-based forecasting method for individual consumers electricity load using time series representations", *Open Computer Science,* vol. 8, pp. 38-50.

[16] Cui, W. & Wang, H. 2017, "A New Anomaly Detection System for School Electricity Consumption Data", *Information,* vol. 8, pp. 151.

[17] Chahla, C., Snoussi, H., Merghem, L. & Esseghir, M. 2019, *A Novel Approach for Anomaly Detection in Power Consumption Data*.

[18] Chandola, V., Banerjee, A. & Kumar, V. 2009, "Anomaly Detection: A Survey", *ACM Comput.Surv.,* vol. 41.

[19] Abraham, B. & Chuang, A. 1989, "Outlier Detection and Time Series Modeling", *Technometrics*, vol. 31, no. 2, pp. 241-248.

[20] Aggarwal, C.C. (ed) 2017, *Outlier Analysis*, Second edition, Springer.

[21] Aghabozorgi, S., Seyed Shirkhorshidi, A. & Ying Wah, T. 2015, *Time-series clustering – A decade review*.

[22] Wu, J. 2012 "K-means Clustering: An Ageless Algorithm" in *Advances in K-means clustering* Springer, pp. 7.

[23] Dent, I., Craig, T., Aicklen, U.& Rodden, T., 2014. ”Variability of behavior in electricity load profile clustering; Who does things at the same time each day?”

[24] Starczewski A., Krzyżak A. 2015, *Performance Evaluation of the Silhouette Index*. Lecture Notes in Computer Science, vol 9120. Springer.

[25] Caliński, Tadeusz & JA, Harabasz. 1974, *A Dendrite Method for Cluster Analysis*. Communications in Statistics - Theory and Methods.

[26] Ünlü, R. & Xanthopoulos, P. 2019, Estimating the number of clusters in a dataset via consensus clustering.

[27] Krawczak, M. & Szkatuła, G. 2014, *An approach to dimensionality reduction in time series*.

[28] Bagnall, A., Ratanamahatana, C., Keogh, E., Lonardi, S. & Janacek, g. 2006, "A Bit Level Representation for Time Series Data Mining with Shape Based Similarity", *Data Min.Knowl.Discov.,* vol. 13, pp. 11-40.

[29] Fu, T. 2011, *A review on time series data mining.*

[30] Jain, S., Shukla, S. & Wadhvani, R. 2018, *Dynamic selection of normalization techniques using data complexity measures*.

[31] Investopedia 2020. Available: https://www.investopedia.com/terms/z/zscore.asp

[32] Project R. 2019, About. Available: https://www.r-project.org/about.html.

[33] Hartigan, J. A. and Wong, M. A. 1979, "Algorithm AS 136: A K-means clustering algorithm.", *Applied Statistics*, 2

[34] Statistics Dictionary, StatTrek 2020. Available: https://stattrek.com/statistics/dictionary.aspx?definition=interquartile%20range