



UMEÅ UNIVERSITY

Finding Anomalous Energy Consumers Using Time Series Clustering in the Swedish Energy Market

Lukas Tonneman

Lukas Tonneman

Spring 2023

Submitted for the Degree of Master of Science in Computing Science and Engineering, 300ECTS

Supervisor: Juan Carlos Nives

Examiner: Henrik Björklund

Department of Computing Science, Umeå University, Sweden

Abstract

Improving the energy efficiency of buildings is important for many reasons. There is a large body of data detailing the hourly energy consumption of buildings. This work studies a large data set from the Swedish energy market. This thesis proposes a data analysis methodology for identifying abnormal consumption patterns using two steps of clustering.

First, typical weekly energy usage profiles are extracted from each building by clustering week-long segments of the building's lifetime consumption, and by extracting the medoids of the clusters. Second, all the typical weekly energy usage profiles are clustered using agglomerative hierarchical clustering. Large clusters are assumed to contain normal consumption patterns, and small clusters are assumed to have abnormal patterns. Buildings with a large presence in small clusters are said to be abnormal, and vice versa. The method employs Dynamic Time Warping distance for dissimilarity measure.

Using a set of 160 buildings, manually classified by domain experts, this thesis shows that the mean abnormality-score is higher for abnormal buildings compared to normal buildings with $p \approx 0.0036$.

Keywords: time-series analysis; clustering; electricity consumer clustering; anomaly detection; gaussian mixture model; hierarchical clustering

Acknowledgements

I would like to thank Juan Carlos Nives for his indispensable help, advice and support with the thesis. Furthermore, I would like to express my gratitude to Jonas Lundin for his guidance and support, and to all the individuals at Advania for providing me with the opportunity to become acquainted with you, and to do my thesis here. Next, I am grateful to Advania's client for facilitating the data that made this thesis possible. Finally, a special thanks to the domain experts for your efforts in the manual classification, and for sharing your expertise with me.

Contents

Abstract	i
Acknowledgments	ii
1 Introduction	1
1.1 Objective	1
1.2 Explainability	2
1.3 Data Background	2
1.3.1 Swedish Energy Market	2
1.3.2 Missing Data	3
1.3.3 What is a Favorable Consumption Pattern?	3
1.4 Readers Guide	4
2 Theoretical Framework	5
2.1 Dissimilarity Measures	5
2.1.1 Euclidean Distance	5
2.1.2 Dynamic Time Warping Distance	6
2.2 Clustering	6
2.2.1 Hierarchical Clustering	7
2.2.2 Gaussian Mixture Model-Based Clustering	7
2.3 Time Series Clustering	8
2.4 Normalization	9
2.5 Dimensionality Reduction	9
2.5.1 Multidimensional Scaling	10
2.6 Cluster Validity Indexes	10
2.6.1 Silhouette	10
2.6.2 Davies-Bouldin	11
2.6.3 Calinski-Harabasz	11
2.6.4 Dunn	11
2.7 Related work	12
2.7.1 Energy Consumption Clustering	12

2.7.2	Time Series Anomaly Detection	13
2.7.3	Energy Consumption Anomaly Detection	13
3	Method	15
3.1	Extracting TWEU Profiles	15
3.1.1	Dimensionality Reduction	16
3.1.2	Choosing K for Profile Extraction	16
3.1.3	Normalization	16
3.1.4	Differences Between Li et al.'s Methodology and This Thesis	17
3.2	Clustering the TWEU Profiles	17
3.2.1	Choosing K for Profile Clustering	17
3.2.2	Identifying Small Clusters	18
3.3	Assigning Abnormality-Score	18
3.4	Controlling for Temperature Differences	18
3.5	Dissimilarity Measurement	19
3.6	Evaluation	19
3.6.1	Evaluation Procedure	20
3.7	Addressing Missing Data	20
3.8	Tools	21
4	Results	22
4.1	TWEU Profile Extraction	22
4.2	Results from the Validation Set	22
5	Discussion	26
5.1	Selection of K for Profile Extraction	26
5.2	Clustering of the TWEU Profiles	26
5.3	Comparison to Domain Expert Classification	27
5.4	On the Mann-Whitney U Test	28
6	Conclusion	29
6.1	Remarks	29
6.2	Future Work	29
7	References	30

1 Introduction

Improving energy efficiency of Swedish buildings is vital for many purposes, including reducing the environmental impact of operating buildings, and the activity within them, and reducing the operating costs for building owners and stakeholders. The first step to improving energy efficiency is always to find areas of improvement, which for this thesis is buildings with suboptimal energy usage. Advania¹ is a software consulting firm with clients in the Swedish energy market. These clients are interested in finding energy consumption patterns indicative of wasteful energy consumption.

This thesis aims to explore methods to detect anomalous electricity consumption patterns by clustering time series energy consumption data. Clustering is an area of unsupervised machine learning, and the clustering of time series data is a well researched domain. In the area of energy consumption analysis, there is a substantial body of research dedicated to the clustering of time series energy consumption data. A number of key challenges have been identified:

- A high degree of domain expertise is necessary for valuable results. For instance, Kang and Lee (2015) begin their abstract by stating “The clustering of electricity customers might have an effective meaning if, and only if, it is verified by domain experts”.
- Cluster validation for unlabeled data is still an open problem. Tureczek and Nielsen (2017, p. 10) write that there “does not exist a single adequate index” for clustering energy consumption time series data.
- The data is often high dimensional, it is not uncommon to have dimensionality in the tens of thousands. Additionally, the data may be noisy and incomplete. It is often necessary to reduce the dimensionality of the data by extracting meaningful features, followed by selecting a suitable dissimilarity measure to compare them. Motlagh, Berry, and O’Neil (2019) describe how the key challenge in energy consumption time series clustering is addressing the extreme dimensionality of such data.

This project follows the methodology introduced by Li et al. (2018), who proposed a novel method for extracting typical usage profiles from energy consumption data, and clustering them. Furthermore, Li et al. showed that it is possible to discover abnormal consumption patterns using their method.

1.1 Objective

The objective of this thesis is to identify buildings with anomalous electricity consumption patterns. A subset of these outliers will have a potential for improvements in energy efficiency. The research question for this thesis is as follows.

How is it possible to identify buildings with anomalous energy consumption by clustering time series energy consumption data?

The approach taken by this problem is now outlined. First, extract typical weekly energy consumption profiles are extracted from every building’s consumption data. Second, all the profiles are

¹<https://www.advania.se/>

clustered to identify normal, and abnormal profiles. Third, by examining which clusters the profiles of a building are assigned to, an abnormality-score for all buildings is deduced.

In anomaly detection, it is common to classify objects as either an anomaly or not an anomaly in a binary fashion. In contrast, this thesis tries to place objects on an abnormality-scale. This is because the available resources for addressing inefficiencies are limited, and to make the best use of them it is best to start working on the buildings with the largest potential for improvement.

1.2 Explainability

As artificial intelligence models are increasingly used in decision-making processes, there is an emerging need for understanding these models when the decisions affect human lives. Model understandability and interpretability are important because it prevents partial and inaccurate results from the model from being taken at face value by users (Barredo Arrieta et al. 2020). Nevertheless, model explainability is not a key focus of this thesis. The provided analysis is not intended to be used directly in decision- or policymaking. Rather, it will highlight anomalous buildings from a larger data set which will be manually reviewed.

1.3 Data Background

Since this thesis has been done in collaboration with industry, it is necessary to describe the data this project has worked with, how it has been collected, along with the relevant context of the Swedish energy market. The data is facilitated by a client of Advania for this thesis. The identity of the client is confidential. The contents of this section is largely based on an interview with a domain expert from the client firm².

The data describes the energy consumption of 3 694 buildings spread out all over Sweden. It is in the form of megawatt consumption in hourly granularity. The buildings are heterogeneous, being a mix of residential, and a wide range of commercial buildings. For many buildings, the data stretches back several years, but the lengths of the series varies between buildings. Some buildings are currently reporting their consumption, while others have ceased their reporting.

1.3.1 Swedish Energy Market

Sweden is divided into four electricity areas. As illustrated by Figure 1, the four areas are

- Area 1: Far northern Sweden
- Area 2: Northern Sweden
- Area 3: Middle Sweden (this area includes Sweden's two largest cities, Stockholm and Gothenburg)
- Area 4: Southern Sweden (where Malmö is located, Sweden's third largest city)

The electricity price does not vary within areas, but may vary between them. From the dataset, the electricity area of every building is known.

The Swedish energy market consists of four types of actors

1. Electricity producers

²Conducted remotely on May 10th, 2023.

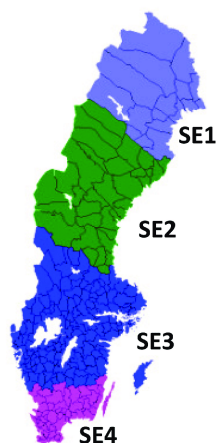


FIGURE 1: The four electricity areas of Sweden (Hansson et al. 2017).

2. Grid owners
3. Electricity consumers
4. Third-parties (Swedish term: Ombud)

Electricity consumers has a contract with the local grid owner, an electricity producer of their choice and optionally a contract with a third party. The grid owner is responsible for measuring the consumption and reporting it to the relevant electricity producing company. The quality of the reporting is strictly regulated, and the grid owner is also responsible for finding and fixing errors in the measured data, including retroactively. At the request of the consumer, the grid owner may also share the data with a third party (for example, Advania's client). This is data however, not subject to the same strict regulations as the data sent to the energy producers.

There is usually one single meter for each building, but apartment buildings may have separate meters for each apartment. When this thesis uses the term *building*, it refers to a distinct electricity meter, which may or may not be a separate building.

1.3.2 Missing Data

In the data available to this thesis (as well as energy consumption data generally), periods of missing data are common. There are several reasons for these gaps, three of which are outlined below.

First, errors in the measurement process might not be adjusted in the data provided to the client (and by extension, this thesis), however these errors will be retroactively addressed in the data supplied to the electricity producer. Second, gaps may form when a building changes ownership. This is because the third party only receives data on behalf of the consumer who owns the building, rather than the building itself. Hence, when ownership changes, a new contract between the new owners and the third party needs to be negotiated before data collection may resume. This gap does not occur for grid owners or electricity producers. Third, gaps may occur as a result of random failures of the data pipeline.

1.3.3 What is a Favorable Consumption Pattern?

In the aforementioned interview, the expert talked about how to discern between “good” and “bad” consumption patterns. The expert stressed the difficulty of objectively classifying a consumption pattern as favorable or not, as it is highly nuanced and context dependent. Generally, a consumption pattern is good when it can be explained by the operation. Meaning that, for instance, a period of high consumption can be explained by a corresponding high intensity process necessary for the

operation. Regretfully, this kind of context is rarely available for energy consumption time series analysis. One indicator of good consumption that can be used without context information, is strong cyclic trends, as opposed to erratic consumption.

Moving on to indicators of poor usage, sudden spikes in consumption is often bad, and worse if the spikes are frequent. Spikes are bad because they often occur during peak electricity price. Moreover, erratic consumption poses a challenge for grid-level electricity capacity planning. Furthermore, frequent and sudden drops in consumption are also suspicions. However, a certain degree of sudden drops in consumption is expected from random electricity outages, as they are is normal.

Another reason to avoid consumption peaks overlapping with pricing peaks, in addition to the direct financial cost, is environmental consideration. This is because the energy mix becomes less environmentally favorable as the price increases. The reason for this is that oil and coal powered reserve/peak power plants may come online to supplement the electricity grid at peak consumption. Additionally, high Swedish prices elevate the chance of electricity import from countries with poor energy mix, which in turn sullies the energy mix for Swedish consumers.

Finally, the expert pointed out that there may not be a direct relationship between a consumption pattern being common and being good. This runs counter to a fundamental premise of this thesis: Poor energy usage is overrepresented in abnormal consumption patterns.

1.4 Readers Guide

This thesis begins with Chapter 2 on the theoretical framework for the thesis. Next, is Chapter 3 on the methodology of the thesis, where method selections and decisions are motivated. Chapter 4 details the results of the project, followed by Chapter 5 which provides a discussion on the project. Lastly, Chapter 6 shows what conclusions may be drawn.

2 Theoretical Framework

This chapter briefly goes over some theoretical concepts central to the understanding of this thesis. Furthermore, it will introduce the specific technologies that have been used.

2.1 Dissimilarity Measures

In the field of time series clustering, it is important to use a suitable dissimilarity measure. Liao (2005) writes that selecting an appropriate dissimilarity measure with regard to the characteristics of the data is *the* key to satisfactory results. Batista et al. (2014) even suggest that for time series clustering, the choice of a dissimilarity measure is much more important than even the choice of a clustering algorithm. Paparrizos and Gravano (2017) describe time series analysis as critically depending upon the choice of the dissimilarity measure. Comparing time series is challenging due to the often immense dimensionality of the data. Moreover, many distance measures produce unintuitive results and are highly sensitive to noise in the data (Aghabozorgi, Seyed Shirshorshidi, and Ying Wah 2015).

Agner (2019) compared Euclidean to Dynamic Time Warping (DTW) for whole time series clustering in an energy consumption context, and found that DTW did not provide any observable improvement. Kim and Kim (2020) compared ten distance measures to find the most suitable for power consumption pattern clustering, but found there was no one best measure. Iglesias and Kastner (2013) evaluated four distance measures and found that Euclidean and DTW were the two most suitable for energy consumption data. However, Kang and Lee (2015) are critical of Iglesias and Kastner's results, ruling out Euclidean distance completely and discuss undesirable quirks in DTW. Kang and Lee go on to propose a novel distance measure, k-sliding, that they argue possesses more favourable properties than DTW. They motivate their findings by comparing clustering results with expert evaluations.

Another paper introducing a novel measure is Paparrizos and Gravano (2017). The new measure is called shape-based distance (in conjunction, the paper also proposes two clustering algorithms, k-Shape and k-MultiShapes). Paparrizos and Gravano found their new method to be better than, or similar to, the state of the art for generic time series clustering and classification. Specifically, they found that constrained DTW (cDTW) yielded similar results, but that their method required significantly less computational resources. However, they did not compare shape-based distance to the k-sliding measure. Moreover, Agner (2019) did not find Paparrizos and Gravano's methodology superior to cDTW for energy consumption data.

2.1.1 Euclidean Distance

The Euclidean distance is a fundamental concept in mathematics and statistics that is widely used in various fields, including data science and machine learning.

The Euclidean distance between two n-dimensional points p and q is defined as the square root of the sum of the squares of the differences between their corresponding coordinates. Mathematically, the formula for Euclidean distance is given by

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2},$$

where $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$ are the n -dimensional points.

2.1.2 Dynamic Time Warping Distance

Dynamic Time Warping (Bellman and Kalaba 1959) is a technique used for comparing time series data in the temporal domain (as opposed to the frequency domain). The Dynamic Time Warping (DTW) algorithm is a distance measure that allows for the calculation of the dissimilarity between two time series by warping one series to align with the other. DTW has been widely used in fields such as speech recognition, finance, and bioinformatics.

Intuitively, the algorithm tries to find the minimum total distance between two time series by not comparing them point by point, but instead by comparing a point on the first series, with the nearest point on the other series. It systematically tries to find the optimal pairing of points, under some constraints, which forms an optimization problem. Mathematically, the problem may be described as shown by

$$DTW(x, y) = \min_{\pi} \sum_{(i, j) \in \pi} d'(x_i, y_j)$$

where x and y are two sequences of data, π is a set of point pairings, called a *path*, for the sequences (subject to constraints detailed below) and d' is the base metric used to compare the points. Typically, d' is Euclidean distance.

These are the constraints enforced on the path:

1. No point in either series may remain un-paired.
2. The starting and ending points of the path, must be the first and the last points of the aligned sequences.
3. The indices in the first sequence must be mapped to monotonically increasing indices in the second sequence, and vice versa. This condition preserves the time-ordering of points.
4. Optionally, more constraints may be placed on the path limiting its shape. A common constraint is to limit the size of the window for the search for optimal pairing, this is called the Sakoe-Chiba band constraint.

Note that one point in one series may be paired with several points in the other series in a one-to-many relationship. This stands in contrast to Euclidean distance where points are paired one-to-one.

When the fourth constraint is used, the measure is sometimes denoted cDTW (constrained DTW), as opposed to DTW for the unconstrained configuration.

2.2 Clustering

Clustering is an area of unsupervised machine learning (Hastie et al. 2009). Intuitively, the goal of clustering is to partition a data set into a number of groups (named “clusters”), in such a way that minimizes within-group variance and maximizes between-group variance. The number of clusters is denoted by K .

Within a cluster, the distance between all members is defined. Additionally, for each member, the sum of the distances to all other members, is also defined. The member with the lowest sum of distances is called the *medoid* of the cluster. The medoid of a cluster is not to be confused with the centroid of a cluster, which is the average point in the cluster, the centroid is not necessarily a member. However, in a well-formed cluster, the medoid is expected to be in the vicinity of the centroid.

2.2.1 Hierarchical Clustering

Hierarchical clustering is a way of greedily clustering data without necessarily knowing how many clusters there are beforehand. As described by Hastie et al. (2009), a hierarchical clustering algorithm will follow one of two paradigms:

1. Agglomerative (bottom-up)
2. Divisive (top-down)

Under the agglomerative paradigm, the initial state consist of one cluster for each data point. The algorithm will then recursively merge the two most similar clusters until only one cluster remain. Conversely, when employing the divisive variant the initial state consist of every data point in one single cluster. Then, the algorithm recursively finds and splits the cluster that, when divided, yields the maximum between-group dissimilarity. When no more cluster can be split, the algorithm terminates.

There are several methods for evaluating the dissimilarity between two clusters. First, a dissimilarity measure between the *elements* of the clusters has to be determined, Euclidean distance is a popular choice. Second, a strategy for comparing *clusters* is needed. There are many variants for this, the three most simple being the following(Hastie et al. 2009).

1. Single linkage – The distance between two clusters is the distance between the nearest pair of elements from both clusters.
2. Average linkage – The distance between two clusters is the average distance between all points within them.
3. Complete linkage – The distance between two clusters is the distance between the least similar pair of elements from both clusters.

2.2.2 Gaussian Mixture Model-Based Clustering

Gaussian mixture model-based clustering, or just GMM clustering, is a soft clustering algorithm which means it does not assign each point to a group outright like a hard clustering algorithm would (e.g. hierarchical clustering). Instead, each point is assigned a probability for each class. A high probability means the point is likely to belong to that class, and vice versa.

Bouveyron et al. (2019) describe the GMM algorithm as centered around the Gaussian distribution (also known as the normal distribution). In the n dimensional space of the data points exist K Gaussian distributions in n dimensions, where K is given. Each distribution has a probability, moreover they have a mean and a variance for each dimension. It is possible to calculate to which proportion each point belongs to each cluster. All of this is illustrated by Figure 2, which shows a probability density function of a Gaussian mixture model using two distributions ($K = 2$) in one dimension. Notice that while the two distributions have different means, variance and probability (illustrated here as the “height” of the bell curves), overlap in the center.

The black line in this theoretical example is the sum of both distributions. However, in the case of trying to fit a model to some real world data, the black line would be the raw data, and the GMM algorithm would try to find a mixture of two Gaussian distributions (if the algorithm was given $K = 2$) that best fits the data.

By using Maximum Likelihood it is possible to fit a GMM model to some data. This is commonly done using the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 1997). The algorithm consists of two steps that are repeated until convergence.

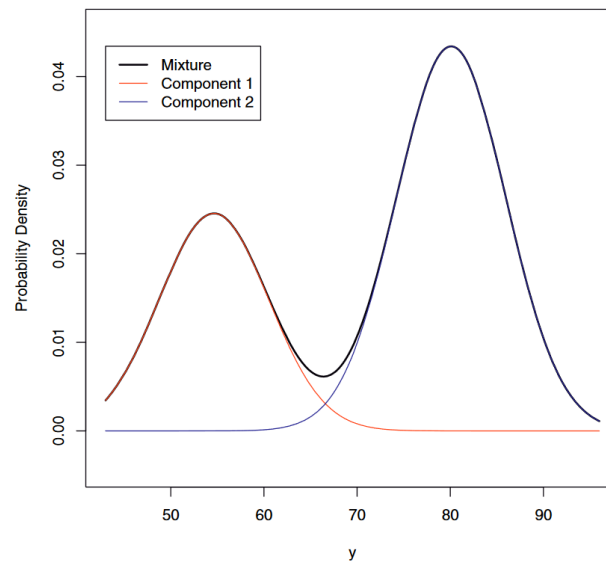


FIGURE 2: Probability density function for a one-dimensional univariate finite normal mixture with two mixture components (Bouveyron et al. 2019, p. 16).

1. Expectation Step – Given the current parameters and data, calculate how well the model fits the data.
2. Maximization Step – Determine how the model parameters can be tweaked to improve fit.

To illustrate how data points may be assigned to clusters, imagine a point in Figure 2 being in a spot where the blue line is “above” the orange line. This can be interpreted as the point having higher probability of belonging to the blue cluster compared to the orange one. If the lines are near each other, it means the model is not confident about which component the point belongs to. If the algorithm is still fitting the model to the data, the point will influence the parameters of both components.

The asymptotic time complexity of GMM clustering is $O(NKD^3)$ where N , K and D are the number of data points, number of Gaussian components and dimensionality of the data, respectively (Pinto and Engel 2015).

In the GMM-based clustering implementation used for this thesis, the soft clustering produced by the algorithm is converted to a hard clustering by assigning the points to their highest probability cluster. This means that it is possible to have empty (zero members) clusters.

2.3 Time Series Clustering

Data from energy consumption comes in different forms. The granularity (or sample rate) may vary, from one reading per day to one reading every 60 seconds, to even more extreme granularity. Consumption is recorded in Watts. The data may come from different sources, e.g. residential, office, or other types of buildings. When a data set contains data from multiple types of buildings, it is said to be *mixed use*.

Aghabozorgi, Seyed Shirshorshidi, and Ying Wah (2015) partition time series clustering into three categories.

1. Whole time series clustering – A set of discrete time series are clustered, usually by conven-

tional clustering methods.

2. Subsequence clustering – Subsequences of long time-series are extracted and clustered.
3. Time point clustering – This can be viewed as an extension of subsequence clustering, where the subsequences consist of one single time point. But one important distinction is that in time point clustering some objects may be considered as noise and not be assigned to any cluster.

This project falls under whole time series and subsequence clustering.

Aghabozorgi, Seyed Shirkhorshidi, and Ying Wah (2015) go on to describe three general approaches to whole time series clustering.

1. Raw-data approach
2. Feature-based approach
3. Model-based approach

In a raw-data approach (also known as shape-based approach), little to no preprocessing of the data is performed. Instead, the clustering algorithms are applied directly to the raw data. These approaches often experiment with various dissimilarity measures.

In a feature-based approach, features are extracted or engineered from the raw data. These features have a lower dimensionality than the raw data, and are typically compared using Euclidean distance. This is the approach chosen in this project.

In a model-based approach, a model is used to describe the time series. The raw data is transformed into parameters for the model. The model parameters are then compared using some dissimilarity measure and clustered by some method.

When clustering energy consumption data, outliers tend to congregate in smaller clusters (Al-Jarrah et al. 2017).

2.4 Normalization

Normalization is a common preprocessing technique used in data analysis and machine learning to scale and transform data to improve its performance and interpretability. Normalization involves adjusting the values of the features or variables in a data set to a common scale or range.

Z-score standardization is a technique used to transform the features of a data set to a standard normal distribution with a mean of 0 and a standard deviation of 1. It involves subtracting the mean of the feature from each value and then dividing by the standard deviation of the feature as described by

$$n = \frac{x - \mu}{\sigma}$$

where the normalized value is denoted by n , the original value x , the mean μ and the standard deviation is denoted by σ .

2.5 Dimensionality Reduction

In this thesis, a method of dimensionality reduction called Multidimensional Scaling (MDS) is employed. Explaining this is the subject of this subsection.

2.5.1 Multidimensional Scaling

Multidimensional Scaling (MDS) is a method of reducing the dimensionality of a data set, while preserving the dissimilarities between all elements of the set (Hastie et al. 2009).

Formally MDS can be described as a minimization problem. Let $x_1, x_2, \dots, x_N \in \mathbb{R}^p$ be a set of N data points in p dimensions of real values, and let d_{ij} be the dissimilarity or distance between elements i and j (Euclidean distance is a popular choice for this). Next, let k be the number of dimensions that we want our data to be in. Now we may define a *stress function* as

$$S_M(z_1, z_2, \dots, z_N) = \sum_{i \neq i'} (d_{ii'} - \|z_i - z_{i'}\|)^2 \quad (2.1)$$

where $i, i' \in 1, \dots, N$. The goal of MDS is to minimize the stress function by optimizing a set of data points $z_1, z_2, \dots, z_N \in \mathbb{R}^k$. When the minimization is complete, the final set of z_1, z_2, \dots, z_N can be considered the output of the MDS algorithm.

There are many ways to minimize the stress function. The MDS implementation used for this thesis uses the SMACOF (Scaling by Majorizing a Complicated Function) algorithm (Pedregosa et al. 2011).

One interesting aspect of MDS is that the elements themselves do not need to be defined. As shown by Equation (2.1), it is sufficient to only have the distances between the elements defined.

2.6 Cluster Validity Indexes

It is important to be able to evaluate the “goodness” of the clusterings produced by clustering algorithms. For this purpose, several cluster validity indexes have been developed. There are two fundamental types of validity indexes, external and internal. External measures may be used when the data is labeled, meaning there is a ground truth to compare a clustering to. Internal measures by contrast do not use any outside information, instead relying simply on the structure and shape of the clustering. In this application there is no ground truth, consequently external measures can not be utilized (Javed, Lee, and Rizzo 2020).

A common application of cluster validation is to determine the number of clusters in a data set. This is difficult without domain expertise, Tureczek and Nielsen (2017) writes that there does not exist any adequate index for cluster validation.

The following internal cluster validity indexes are used in this thesis.

- Silhouette index
- Davies-Bouldin index
- Calinski-Harabasz index
- Dunn index

In the following subsections, these measures are introduced.

2.6.1 Silhouette

The Silhouette cluster validity index (Rousseeuw 1987) is a commonly used metric for evaluating the quality of clustering results. It is based on the idea of measuring the cohesion of points within a cluster, and the separation between clusters. Mathematically, it is defined as

$$\frac{a - b}{\max(a, b)}$$

where a is the average intra-cluster distance, and b is the average inter-cluster distance. It ranges from -1 to 1 where a higher score indicates better clustering.

2.6.2 Davies-Bouldin

The Davies-Bouldin cluster validity index (Davies and Bouldin 1979) can intuitively be understood as the ratio of the within-cluster variation, to the between-cluster variation. Consequently, a good clustering will be assigned a low score.

$$DB_k = \frac{1}{k} \sum_{i=1}^k \max_{j=1, \dots, k, i \neq j} \left(\frac{\text{diam}(c_i) + \text{diam}(c_j)}{d(z_i, z_j)} \right)$$

describes the formal definition of the index. Where k is the number of clusters, z_i is the centroid of the i^{th} cluster, function d is some distance measure (e.g. Euclidean distance), and the function for the diameter of the i^{th} cluster $\text{diam}(c_i)$ is defined as

$$\text{diam}(c_i) = \sqrt{\frac{1}{n_i} \sum_{x \in c_i} d(x, z_i)^2} \quad (2.2)$$

where n_i is the number of elements in the i^{th} cluster.

2.6.3 Calinski-Harabasz

The Calinski-Harabasz cluster validity index (Caliński and Harabasz 1974) can be roughly understood as the ratio of the between-cluster variation and the within-cluster variation. As it is similar to the inverse of the Davies-Bouldin index, high scores are awarded good to clusterings. Mathematically, the index is defined as

$$CH_k = \frac{BCSM}{k-1} \cdot \frac{n-k}{WCSM}$$

where n is the total number of elements in all clusters, and k is the number of clusters. $BCSM$ (between cluster scatter matrix) is defined as

$$BCSM = \sum_{i=1}^k n_i \cdot d(z_i, z_{tot})^2$$

and $WCSM$ (within cluster scatter matrix) is defined as

$$WCSM = \sum_{i=1}^k \sum_{x \in c_i} d(x, z_i)^2$$

In these equations, z_i is the centroid of, and n_i is the number of elements in, cluster c_i . z_{tot} is the centroid of the entire data set. The function d is some distance measure, for instance Euclidean.

2.6.4 Dunn

Dunn's cluster validity index (Dunn 1974) is similar to the Calinski-Harabasz index in the sense that it also involves a ratio of the between-cluster variation mean and the within-cluster variation. However, it is not identical, being more sensitive to outliers. The formal definition of the index is

$$DU_k = \min_{i=1, \dots, k} \left\{ \min_{j=1+1, \dots, k} \left(\frac{\text{diss}(c_i, c_j)}{\max_{m=1, \dots, k} \text{diam}(c_m)} \right) \right\}$$

where k is the number of clusters. The function diss is defined as

$$\text{diss}(c_i, c_j) = \min_{x \in c_i, y \in c_j} d(x, y)$$

and $diam$ is defined as

$$diam(c) = \max_{x,y \in c} d(x,y) \quad (2.3)$$

d is some distance measure, for example Euclidean. The function $diam$ defined in Equation (2.3) is not to be confused with the function defined in Equation (2.2) by the same name, as they are different functions.

2.7 Related work

This section will go over some similar works in related fields.

2.7.1 Energy Consumption Clustering

First, four papers on the clustering of energy consumption time series data are presented.

Damayanti et al. (2017) clustered 2014 energy consumption data from the West Java province of Indonesia. They took a raw-data approach and clustered with various partition-based clustering techniques and found that the K-Harmonic Means algorithm was suitable. They only used a single cluster validity index (Davies-Bouldin Index, see Section 2.6.2) and found the optimal $K = 2$.

Li et al. (2018) clustered 40 buildings from the campus of the University of Wollongong, Australia with data from 2014–2015. A novel and complex strategy is used which includes a type of feature extraction where they extracted 24-dimensional typical daily energy usage (TDEU) profiles from the buildings, which were then clustered through agglomerative hierarchical clustering into nine clusters. They claim to have successfully discovered abnormal electricity consumption patterns, but do not elaborate on the nature of the anomalies.

The extraction of the TDEU profiles was performed by splitting the energy consumption time series of a building into distinct days, which were reduced in dimension by MDS. The resulting vectors were clustered by GMM clustering. The medians of each cluster were selected as the TDEU profiles for that building.

Motlagh, Berry, and O’Neil (2019) had access to the load curves of some 20,000 Australian residential buildings from around 2011–2014, and used that data to evaluate two clustering approaches, one feature-based and one model-based. They found the latter to be superior, in addition, the model-based approach could also handle poor data quality better than the feature-based. The optimal K was found to be 12 clusters.

Iliev (2022) did his master’s thesis on the analysis of electricity usage time series with K-means clustering. At his disposal was the 2018 electricity consumption from 360 Swedish buildings (mixed use), which the author successfully classified as either “residential” or “office” ($K = 2$). For further work Iliev suggests finding anomalous time series.

Most of the papers so far state their goal as classifying the individual consumers into two or more distinct categories, attempt to learn more about consumption patterns for the purposes of grid-level planning, or consumption forecasting.

Compared to related works, this project has a comparatively large set of data, which is also more heterogeneous than many works. Both in terms of the buildings’ usage (office, residential, e.t.c.), but also in the geographic spread of the sites and how much the data is dispersed in time, with the earliest data from 2017, and the latest data being just hours old (at the time of the analysis) as new data is being continuously ingested into the client’s database.

2.7.2 Time Series Anomaly Detection

The next two works write about time series anomaly detection generally, as opposed to energy consumption data clustering specifically. Shaukat et al. (2021) details three types of anomaly detection:

1. Supervised Anomaly Detection – Where labeled data is available.
2. Semi-supervised Anomaly Detection – Where the data is highly homogeneous permitting the data to be modeled by a single concept. An anomaly is something deviating from this concept.
3. Unsupervised Anomaly Detection – No labeled data. The paper exemplifies this type with clustering-based anomaly detection coupled with some assumptions about the data, like assuming data in small clusters is prone to be anomalous.

This thesis belongs to the third type. Shaukat et al. go on to detail several challenges related to anomaly detection such as, how to define normality, detecting malicious anomalies trying to appear normal, normal behaviour changing over time, the difficulty of generic anomaly detection, lack of labeled data and finally some challenges related to scaling and deploying software employing anomaly detection. The paper concludes that the detection of anomalies is a challenging task, and that consequently, most techniques are tailored to specific applications.

Blázquez-García et al. (2021) presents a taxonomy based on the main aspects of outlier detection techniques: The input data, the type of outlier, and the nature of the method. The paper identified four types of outliers.

1. Point outliers – Where a single point may be considered an outlier.
2. Subsequence outliers – Where a subsequence of a time series may be considered an outlier. Blázquez-García et al. divides this type into five subtypes: Discord detection, dissimilarity-based, prediction model-based, frequency-based, and information theory. They note that the most straightforward approach is to calculate the distances between the subsequences, and clustering them.
3. Outlier time series – Where an entire time series may be considered an outlier. They state that the most common approach in this category is clustering. The authors go on to find two types of approaches: Dimensionality reduction and dissimilarity-based approaches.

For future research, the paper raises the issue of determining a threshold for outliers, remarking that they could not find a single instance of dynamic or adaptive threshold selection, in subsequence and entire time series outlier analysis. Furthermore, Blázquez-García et al. notes that Euclidean distance is used in the majority of the works, and calls for a study examining how using different dissimilarity measures can improve outlier detection.

2.7.3 Energy Consumption Anomaly Detection

Lastly, four papers specifically on energy consumption anomaly detection are described.

Pan, Yin, and Jiang (2022) developed a method for detecting point outliers in real time using a combination of ARIMA (Autoregressive Integrated Moving Average model), a convolutional neural network(CNN) and bi-directional long and short-term memory neural networks (Bi-LSTMs). They used five months of consumption data from houses in the USA, including indoor temperature, and a range of climate data. Points above 3σ from the model prediction were classified as anomalies. Unfortunately, the quality of the outlier detection is not quantitatively evaluated.

Cui and Wang (2017) developed a point outlier detection method by using historical consumption of Norwegian schools to create a Gaussian model exhibiting normal consumption, and flagging consumption moments when the consumption exceeds some multiple of σ . At the request of domain

experts, the method has zero tolerance for false positives. The model was evaluated by comparing results to expert evaluations, and it acquired zero false positives and very few false negatives. The method does require manual parameter tuning for each building, and does not account for seasonal variations.

Oprea et al. (2021) presents a complex method for fraud detection in household energy consumption. The method consists of using unsupervised machine learning to detect point outliers in consumption patterns, buildings with more than 15% outliers are considered suspicious. Using the labeling of the aforementioned step, a supervised machine learning model is trained, acquiring an accuracy of 90%.

Liu et al. (2021) proposes a data mining-based framework for the identification of daily electricity usage patterns. Their methodology is similar to Li et al. (2018), but they expand it by engineering five statistical features from the profiles, and improving the interpretability of the results, among other things. Liu et al. differentiates between the terms outliers and anomalies. Outliers are odd daily electricity usage patterns and are identified via clustering, and excluded. Anomalies are buildings that deviate from expected values in the five statistical features, and are of interest. The anomaly detection is not evaluated.

Comparing these papers to this thesis, none of the papers mention dissimilarity measure, whereas this is a central problem for this thesis. Furthermore, this project is differentiated by having access to expert evaluations, which is rare, and by having a highly heterogeneous and unusually large data set.

3 Method

This section describes the methodology of this thesis. It follows the methodology introduced by Li et al. (2018) who detected abnormal energy consumption by a novel approach. The methodology of this thesis can be divided into three distinct steps.

1. Extract typical weekly energy usage profiles (TWEU profiles) from all buildings.
2. Cluster the profiles and identify small clusters.
3. Assign buildings an abnormality-score based on how prevalent they are in small clusters. Then, rank all buildings by the score in order to identify the most abnormal.

3.1 Extracting TWEU Profiles

For every building, the consumption data was partitioned into weekly (Monday through Sunday) segments. Segments with incomplete or missing data were excluded from the analysis. Then, MDS was used to reduce the dimensionality of the weeks of data from 168 (one for each hour of the week) down to 14 (see Section 3.1.1 for motivation).

The low-dimensional representations of the weeks are then clustered by GMM clustering, into $K = 8$ clusters (see Section 3.1.2 for motivation on choosing $K = 8$). For each cluster, the medoid is calculated (in the reduced dimension space), the medoid profile is selected as the TWEU profile of that cluster. The amount of profiles that were assigned to that cluster, relative to how many segments the building had, is recorded as the relative weight of that cluster. All extracted TWEU profiles were subsequently normalized according to Section 3.1.3. Recall that in GMM clustering it is possible to have empty clusters. These clusters were ignored.

GMM clustering is selected because this thesis follows the methodology introduced by Li et al. (2018), who used GMM in this fashion. Li et al. motivate their choice of GMM by performing a comparison of four clustering methods: GMM-based, K-means, K-medoid, self-organizing map, and agglomerative hierarchical clustering. They were compared using three indexes: Silhouette, Dunn, and Davies-Bouldin index. Li et al. found GMM-based clustering to be the most suitable.

It is important to extract these profiles because it is not suitable to cluster raw data directly. Previous research has shown it is possible to extract features from raw energy consumption time series data and to use the features to cluster the data, instead of clustering on the raw data (Chicco, Napoli, and Piglion 2006). Moreover, careful preprocessing of the data can improve clustering (Tureczek, Nielsen, and Madsen 2018). Finding a way to represent time series data in reduced dimensionality is a key challenge (Motlagh, Berry, and O'Neil 2019). Dimensionality reduction or feature extraction is important because it may not be computationally feasible to cluster raw, high dimensional time series. Moreover, high dimensional fine granularity time series data may suffer from noise, which can negatively impact dissimilarity measures (Aghabozorgi, Seyed Shirshorshidi, and Ying Wah 2015)

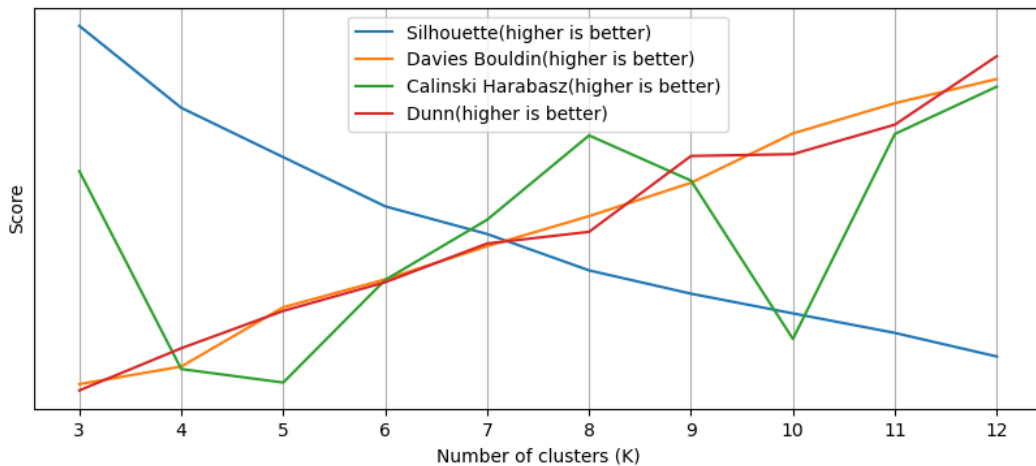


FIGURE 3: The average index scores for the evaluated range of K s. The Davies-Bouldin index has been inverted so that it can be interpreted as higher is better, like the other indexes.

3.1.1 Dimensionality Reduction

For this project, MDS was selected for dimensionality reduction. This is because this project follows the methodology introduced by Li et al. (2018), where MDS was used in the same fashion. They motivate their choice by stressing MDS's ability to retain information about the pairwise distances among the observations (Li et al. 2016).

Two differences between the original paper and this thesis is that while they reduced their 1-day segments down to two dimensions, this thesis reduces the 7-day segments down to $2 \cdot 7 = 14$ dimensions. Second, this project uses DTW, replacing Euclidean distance. Note that while DTW was used for the dimensionality reduction, the resulting reduced dimensionality representations are compared using Euclidean distance.

3.1.2 Choosing K for Profile Extraction

It is important to select an appropriate K , selecting a low K risks hiding minority patterns in large clusters. While selecting a high K risks producing empty or duplicate clusters (Benítez et al. 2016). K was determined by evaluating a range of possible values on 100 randomly sampled buildings from the entire dataset. K s in the range 3–12 were evaluated with the cluster validity indexes of Silhouette, Davies-Bouldin, Calinski-Harabasz and Dunn. The segments from each building were clustered five times for each prospective K to stabilize the scores. Figure 3 shows the average scores for each index for each K . The scores from each index were normalized to fit into the same plot.

The figure shows how the Silhouette score is monotonically decreasing, while the Davies-Bouldin and Calinski-Harabasz are monotonically increasing. These stand in contrast to the Calinski-Harabasz index which has three distinct local maxima at K equals 3, 8, and 12. Of these 8 is selected as the optimal K .

3.1.3 Normalization

The reason for using normalization is that it enables effective comparison of buildings of varying size, shifting the focus from the absolute values of the consumption, to the shape of the consumption pattern. Normalization is applied to TWEU profiles after their extraction. Tardioli et al. (2018) investigated various normalization methods and found that Z-Score normalization is the most suitable for data with mixed use buildings. Iliev (2022) compared three methods for normalization in a similar context and found that Z-Score along with Min-max normalization is adequate. With this

background, Z-Score normalization is selected.

3.1.4 Differences Between Li et al.’s Methodology and This Thesis

This thesis builds upon Li et al. (2018), but some key changes have been made. This subsection will explain these differences.

This implementation uses weekly segments instead of daily segments as the paper suggests. This is because long term patterns may be detected when taking a weekly perspective, which that may not be visible on a daily granularity.

Furthermore, the paper selected the typical daily energy usage profiles (TDEU profiles) of a cluster by taking the median consumption for each of the 24 hours of the day. Since this method splices data from different days, it constructs a consumption pattern that, possibly, never occurred in the real world. Moreover, averaging the consumption risks smoothing out abnormal patterns, which runs contrary to the objective of this thesis. The selection process for this thesis is instead to select the medoid of the cluster as the TWEU profile. This ensures that the selected pattern actually occurred.

Additionally, while the original paper tested a range of 2–14 K s for each site, selecting the best K for each, this implementation instead extracts 8 profiles from every site. The reasons for this is computational complexity. Li et al. (2018) had access to 40 buildings, while this thesis has access to data from 3 694 buildings. Replicating the optimization of K would induce a further twelve-fold increase in computational load, which is beyond the resources allocated to the thesis.

Finally, this project uses DTW distance for dissimilarity measure, as opposed to Euclidean distance used by Li et al. (2018). See Section 3.5 for the motivation of this decision.

3.2 Clustering the TWEU Profiles

From the TWEU profile extraction process, 22 730 profiles were extracted. Notice that the number of profiles is less than expected, if all buildings have eight TWEU profiles. This is because the GMM-based clustering produced some empty clusters. The next step was to cluster the profiles using agglomerative hierarchical clustering, following the methodology introduced by Li et al. (2018).

In the particular implementation of hierarchical clustering used for this thesis, the algorithm is initiated with a dissimilarity matrix of all profiles. As calculating a dissimilarity matrix has complexity $O(n^2)$ this was significant computational task requiring ≈ 260 million profile comparisons, the endeavour was not simplified by the use of DTW as dissimilarity measure. The next subsections will discuss how K was selected, along with how small clusters were identified.

One difference between Li et al. (2018) and this thesis is that Li et al. use Ward’s linkage whereas this thesis uses average linkage. This is because Ward’s linkage is not compatible with DTW. Average linkage is chosen instead because it is the most suitable for time series clustering (Chicco, Napoli, and Piglion 2006).

3.2.1 Choosing K for Profile Clustering

The selection of K for this clustering was difficult as evaluating any range of K s requires significant computational resources and time. Recall from Section 2.2 that outliers tend to congregate in smaller clusters. Choosing a suitable K is important because choosing a too small K risks forcing outliers into large clusters which goes against the assumption that abnormal patterns congregate in small clusters. On the contrary, choosing a too large K risks forcing normal patterns into small clusters, which similarly runs contrary to the aforementioned assumption. For the selection of K , the thesis continues to follow the methodology introduced by Li et al. (2018) who selected $K = 9$ for their $n = 40$ buildings, a clusters to buildings ratio of about 1:4. This ratio is maintained for the selection

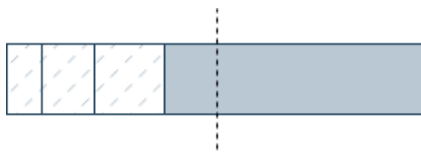


FIGURE 4: Four clusters ordered by size. The width of the boxes represent cluster size and the total length represents the total number of elements. The dashed line shows the 50th percentile. The three smallest clusters are below the 50th percentile, and are hence small clusters (represented by dashed boxes). Solid-colour boxes represent large clusters.

of K for this thesis.

Having 9 clusters is unreasonable because it is in the interest of this thesis to find outliers, the goal is not to group the consumers into a handful of meaningful classes, which is the objective of many similar works. It is unreasonable to assume that the outliers will be similar to each other, instead, most outliers will be abnormal in their own way. This is why K must scale with n , and can not remain constant. The specific ratio of 1:4 is reasonable because it was the optimal K Li et al. found for 40 buildings was $K = 9$.

3.2.2 Identifying Small Clusters

When using agglomerative hierarchical clustering on energy consumption data, it has been shown that clustering produces one dominant cluster, along with many small clusters (Kang and Lee 2015). Given this, the dominant cluster may be named the normal cluster, and the many small ones abnormal. For this thesis, a small cluster is a cluster below the 50th percentile in size. The cluster on the 50th percentile is not included.

This is illustrated by Figure 4 showing four clusters, three small and one large cluster. The three smallest clusters combined make up less than 50% of all members, but adding the fourth cluster would bring the total above 50%. Hence, the three smallest are designated small clusters, and the fourth is designated large.

3.3 Assigning Abnormality-Score

The “abnormality-score” of a building is a measure of how abnormal the building’s consumption is, the higher the score, the more abnormal. The score is assigned based on how large of the proportion of the buildings profiles are in small clusters. Recall that each building has 8 TWEU profiles (except for the buildings that ended up with some empty profile clusters, which effectively have fewer than 8 profiles).

Next, recall that each of these profiles have a weight decided by how many weekly segments they represent. If the TWEU profile of a building is assigned to a small cluster, that buildings abnormality-score will increase with respect to the weight of that profile. This creates a scale starting from having every TWEU profile in the normal cluster, to having no TWEU profile in the normal cluster, which will be awarded the maximum abnormality-score.

3.4 Controlling for Temperature Differences

Ambient temperature affects energy consumption in buildings (Riihimäki and Koponen 2012). Since the buildings in the data set are spread out over time and space, the varying temperature adds noise to the data. This noise can mask fine differences in consumption patterns. This problem is addressed by partitioning the data into week-long segments, which are normalized. Normalization is effective

because it removes the effect of a raised base-consumption during winter. It does not remove the effect of fast (within one week) temperature changes.

Out of the papers reviewed for this thesis, only one takes into account the issue of temperature induced noise. In his master's thesis, Agner (2019) took a model-based approach which took into account, among other things, the ambient temperature. The approach of this thesis is different because Agner's model-based solution is not applicable as this thesis takes a feature-based approach.

3.5 Dissimilarity Measurement

DTW was selected for the dissimilarity measure of this project. This is because it can adequately quantify the differences between two time series (Iglesias and Kastner 2013). The novel dissimilarity measure *k*-sliding proposed by Kang and Lee (2015) was not selected because there is no highly optimized open source implementation of the *k*-sliding distance, while there is such implementations of DTW. A naive implementation of *k*-sliding was developed for this thesis, but proved several times slower than DTW, rendering it an unsuitable choice.

This thesis has used unconstrained DTW, meaning the algorithm has searched for the optimal path among all possible paths. This stands in contrast to Agner (2019) who used a Sakoe-Chiba band of two hours. Agner went on to find that DTW, in their configuration, was not better than Euclidean distance. This is in opposition to Kang and Lee (2015) who are dismissive of Euclidean distance. Going back to the choice of constrained or unconstrained DTW, Paparrizos and Gravano (2017) found that while carefully tuning the constraints can improve classification accuracy, but that unconstrained DTW is superior in a clustering context.

Finally, some papers have expressed concern for the computational complexity of DTW, preferring other measures to DTW with this motivation (Öhman 2022; Paparrizos and Gravano 2017). But, with respect to the resources allocated to this project, the computational complexity of DTW is deemed acceptable.

While DTW is used for the clustering algorithms and dimensionality reduction, Euclidean distance is used for the cluster validity indexes. The reason for this is that the open implementations of the indexes that have been employed do not permit altering the dissimilarity measure. Moreover, it was deemed outside the scope of this thesis to implement new versions of the indexes with the aforementioned functionality.

3.6 Evaluation

A domain expert from the client firm has manually classified 150 sites as either normal or abnormal energy usage. The 150 sites were randomly sampled from electricity areas three and four. The models created during this project have been evaluated by having the model classify these 150 sites, and comparing the results to the expert classification.

Finally, the model was validated with another 160 randomly selected buildings, disjoint from the set of 150 buildings used for testing. The 160 buildings were split into four sets of sizes 45, 45, 45 and 25. Afterwards they were packaged into three packages consisting of one 45-size set joined with the 25-size set, as illustrated by Figure 5. The packages were then assigned to three distinct domain experts from the client, who manually and independently, classified all buildings in their package. The reason for the partial overlap of the packages is to evaluate the quality of the expert classification, or more specifically, to evaluate how well the experts judgements align with the judgement of other experts.

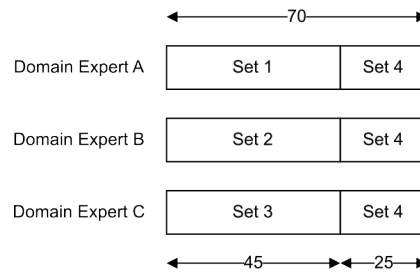


FIGURE 5: The three packages compiled from the four sets of validation buildings. Each package was assigned to a distinct domain expert.

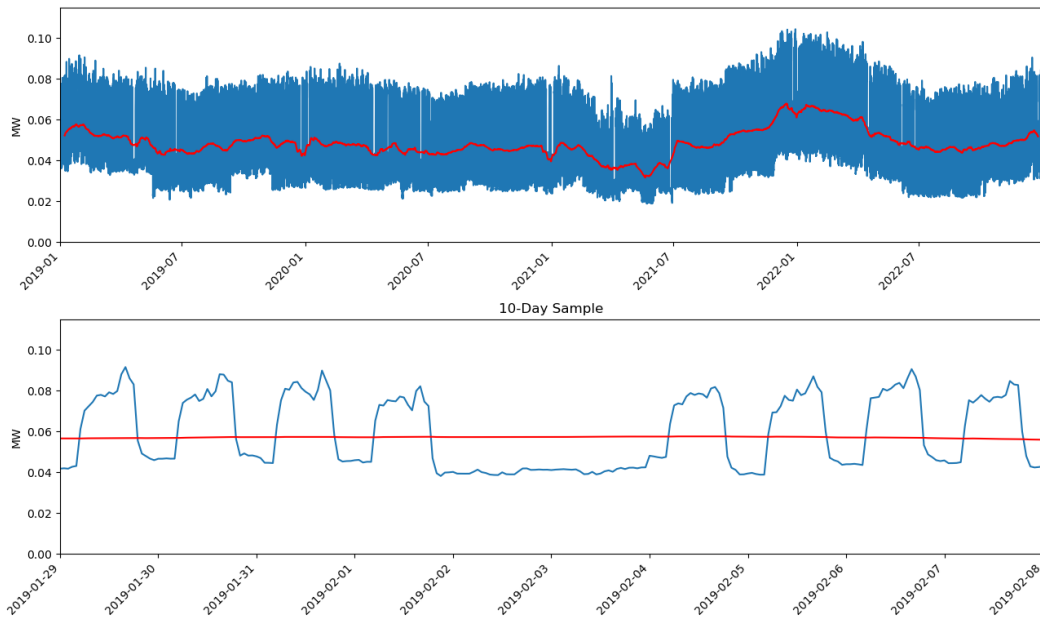


FIGURE 6: An example of a consumption pattern the manual reviewers faced. This particular specimen was classified as normal.

3.6.1 Evaluation Procedure

This subsection describes how the domain experts classified the consumption patterns as either normal or abnormal. The process of classifying one building consisted of viewing two plots, one of the lifetime consumption of the evaluated building, and one plot of the consumption over 10 days for a more detailed view. In addition to the raw data, the plots also showed a two-week centred rolling average to reveal trends. Figure 6 shows one example of this. After viewing the plots, the expert would then classify the building as either normal, or abnormal.

3.7 Addressing Missing Data

The quality of the available data varies a lot. First, there is high variation in how much data is available for the different buildings, buildings with less than 180 days of data have been excluded because this has been deemed to be too little data. Second, it is common to have gaps in the data. These gaps range from just a single missing reading, to months of missing data. Data gaps of one hour (a single missing reading) have been repaired by linear interpolation. Gaps of exactly 24 hours are repaired

by replaying the consumption from the previous day. Other gaps are not addressed.

3.8 Tools

The Python programming language has been used for this project. Many packages have had a vital role for the project, including NumPy (Harris et al. 2020), Pandas (McKinney 2010), SciPy (Virtanen et al. 2020), Scikit-learn (Pedregosa et al. 2011), Statsmodels (Seabold and Perktold 2010) and tslearn (Tavenard et al. 2020).

4 Results

This section goes over some results of the thesis. First, it details the results of the TWEU profile extraction of one building. Next, the abnormality scores of the validation set are presented, and the scores of normal versus abnormal buildings are compared.

4.1 TWEU Profile Extraction

Figure 7 shows the results of the TWEU profile extraction process for the consumption data shown in Figure 6. The plots exemplify how the GMM clustering algorithm successfully grouped similar patterns together, moreover it also exemplifies how abnormal patterns congregate in small clusters. The two smallest clusters are the sixth (5 members) and seventh (4 members) clusters. The sixth cluster is abnormal because it has low consumption on Mondays, and the seventh cluster, similarly, has an abnormally low consumption on Fridays. Conversely, the two largest clusters are the first (71 members) and fifth cluster (43 members). Both of which exhibit a highly cyclic pattern, the consumption being nearly identical for all weekdays.

4.2 Results from the Validation Set

Recall that a set of 25 buildings was evaluated by all three experts. Table 1 shows the results of this evaluation. The rightmost column shows the majority vote of the experts, and is interpreted as the ground truth when evaluating the model. In order to measure the inter-rater agreement, Fleiss' kappa is used (Fleiss 1971). Given the data in Table 1, we have $\kappa \approx 0.50$ on a scale from 0–1 where 1 is complete consensus, and 0 is a random result. Against the null hypothesis of the experts classifying randomly, we have $p \approx 0.00094$, meaning that there is a $> 99\%$ chance that the experts did not classify the buildings randomly.

Recall that 160 buildings were manually classified by three domain experts for the purpose of validating the thesis. From them, 1 002 TWEU profiles were extracted, and clustered into $K = 40$ clusters. Figure 8 illustrates the sizes of these 40 clusters, there are many small clusters, including 12 single member clusters, and one dominant cluster of 686 clusters.

Figure 9 shows the score distribution for both the normal and abnormal sites as classified by the domain experts. In the bin for the highest scores, abnormal sites make up the majority. Conversely, in the bin for the smallest scores, normal sites more than two-fold outnumber the abnormal. Simultaneously, the plots show a not insignificant level of both false positives and false negatives.

A one-sided Mann-Whitney U test (Mann and Whitney 1947) showed that the abnormal set is stochastically larger than the normal set of buildings, with $p \approx 0.0036$. This means that there is a 0.36% chance the abnormal and normal buildings have the same distribution. To rephrase, there is a $> 99\%$ chance that abnormal sites have a larger scores than normal sites. Mann-Whitney U test is selected because it nonparametric test, meaning it does not require assumptions regarding the normality, mean, or variance of the examined populations. One-sided is selected because it is of interest if the abnormal sites have larger scores than normal, as opposed to if they are just different.

Because the sample sizes are not equal, and because there exists no natural pairing, no pairing test

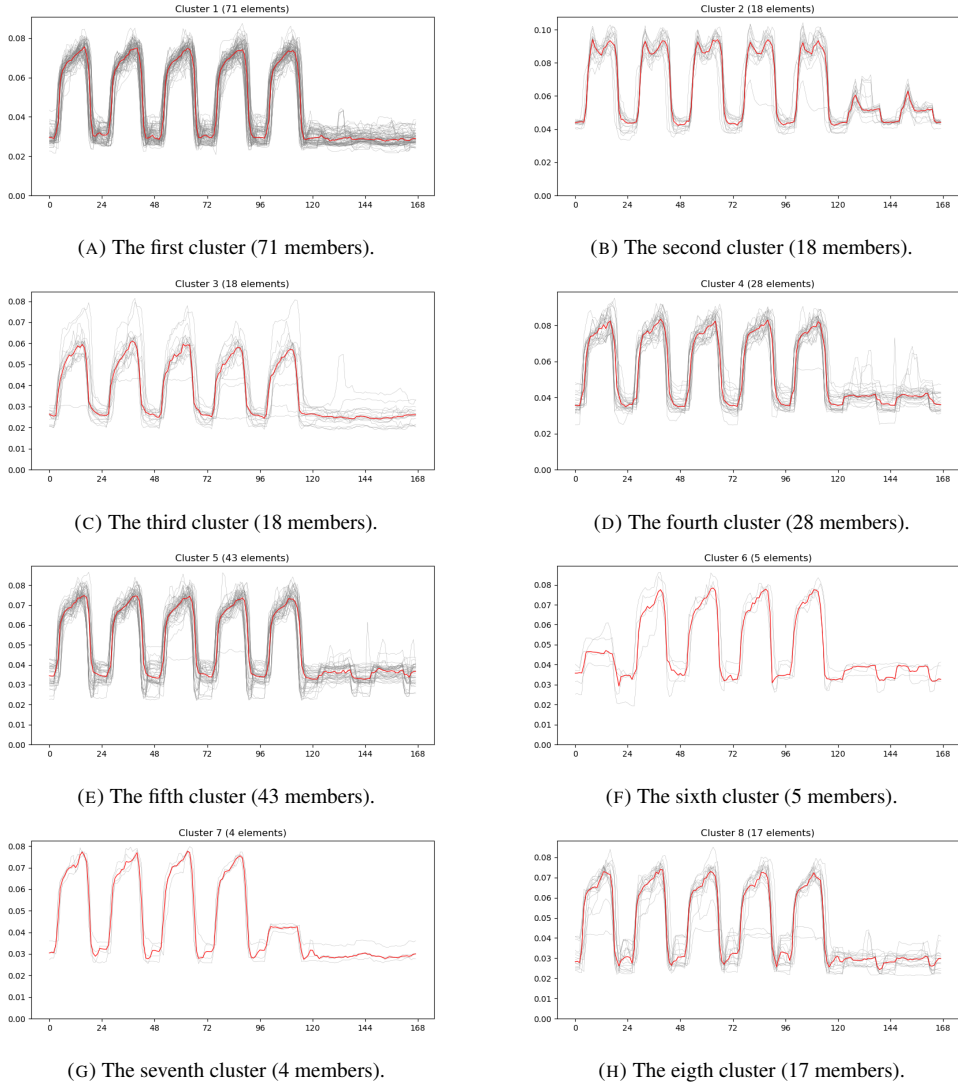


FIGURE 7: The eight clusters extracted from the site shown in Figure 6. The red lines shows the cluster medoids, which are selected as the TWEU profile the corresponding cluster. The thin grey lines show each member of the cluster. The Y-axis is energy consumption in megawatts, and the X-axis is time (in hours) starting at Monday 00:00 and ending at Sunday 23:00.

Table 1 Three experts opinion on 25 different buildings, including the final verdict (the majority result). A zero means the building was classified as normal by the corresponding expert, and vice versa.

Building	Expert 1	Expert 2	Expert 3	Verdict
1	0	0	0	Normal
2	0	0	0	Normal
3	0	0	0	Normal
4	0	0	0	Normal
5	0	0	0	Normal
6	0	0	0	Normal
7	1	1	1	Abnormal
8	0	0	1	Normal
9	0	0	1	Normal
10	0	0	0	Normal
11	1	1	1	Abnormal
12	0	0	1	Normal
13	0	0	0	Normal
14	0	1	0	Normal
15	0	1	0	Normal
16	1	1	1	Abnormal
17	0	0	0	Normal
18	1	0	0	Normal
19	1	0	0	Normal
20	1	1	1	Abnormal
21	0	0	1	Normal
22	0	0	0	Normal
23	0	0	0	Normal
24	1	1	1	Abnormal
25	0	0	0	Normal

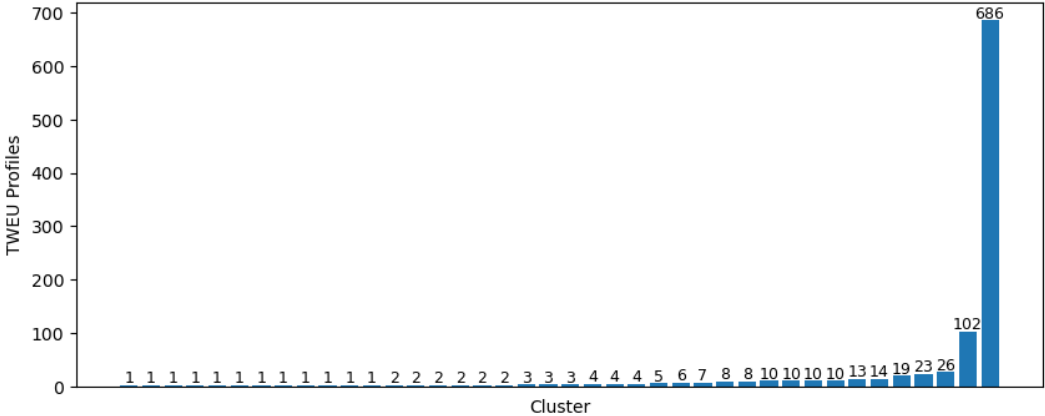


FIGURE 8: The cluster sizes of the 40 clusters produced from clustering your the validation set (1 002 TWEU profiles).

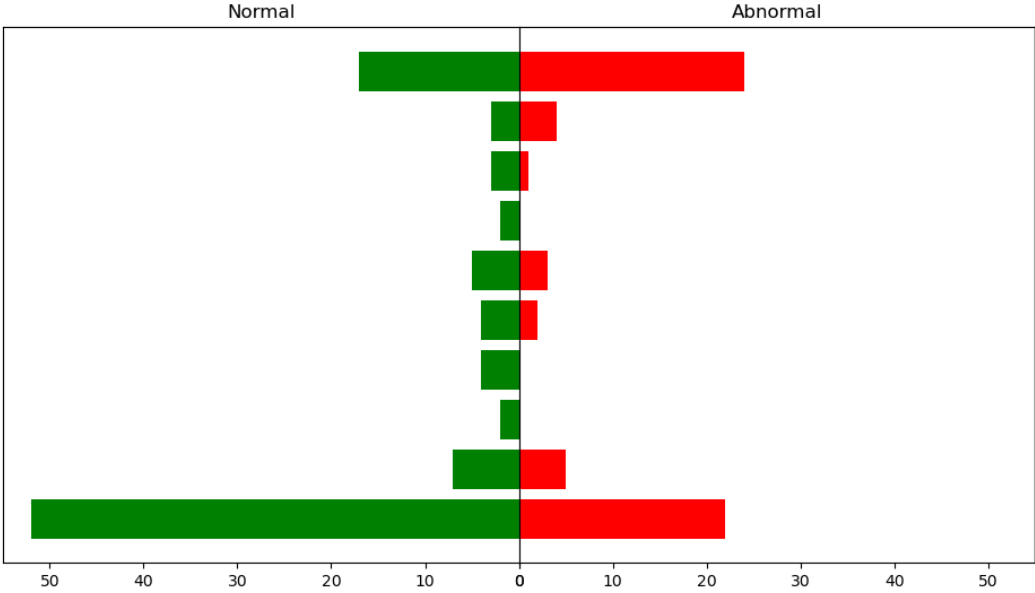


FIGURE 9: A double histogram of the abnormality-score assigned to the buildings in the validation set, the highest scores are at the top, and vice versa. The green histogram (left) is shows the score distribution of the buildings classified as normal by domain experts. The red histogram (right) shows the corresponding histogram for the buildings classified as abnormal by the experts.

is suitable. Moreover, as Figure 9 shows, the distributions are not Gaussian, rendering tests that assume normality unsuitable.

5 Discussion

This section features a discussion on the project, examining weaknesses and comparing the thesis to the state of the art.

Recall from Section 3.4 almost no other papers seemingly addressed the problem of noise introduced by temperature differences. This statement is supported in findings by Tureczek and Nielsen (2017). This is believed to be because no other paper had access to such a diverse set of data, this is supported by the fact that Agner (2019), who also used data from the entirety of Sweden and discovered the same problem, did address it.

Recall that the first step of the methodology presented in this thesis is to partition the lifetime consumption of buildings into week-long segments. This imposes a hard limit on the periodicity of detectable patterns, any patterns over 7 days, such as seasonal patterns, can not be unveiled using this methodology. For example, imagine a consumption pattern that highly correlates with ambient temperature compared to other buildings. This type of anomaly could indicate lackluster insulation and would be of interest for a stakeholder looking to reduce costs. But using week-long segments inherently prevents this type of anomaly detection.

5.1 Selection of K for Profile Extraction

Recall that Section 3.1.2 explains how $K = 8$ was selected for the extraction of TWEU profiles from buildings. Figure 3 showed that the Silhouette score was monotonically decreasing, while the Davies-Bouldin and Calinski-Harabasz were monotonically increasing.

The behaviour of the Davies-Bouldin and Calinski-Harabasz scores may be explained as follows: Adding more clusters will inevitably shrink the average size of the clusters, which, since far apart data points are no longer forced into the same cluster, will decrease the within-cluster variation.

The reason the Silhouette score is monotonically decreasing may be because it rewards having high average between-cluster distance. When having few clusters, outliers forced into large clusters may shift the cluster centroid, thus increasing the distances between the clusters. However, when increasing the number of clusters, the outliers can be put into their own small clusters, leaving the other clusters to appear closer to each-other, as the outliers are no longer pulling them apart. This could lower the average between-cluster distance, reducing the score as more clusters are added.

The meaningless results of these indexes supports the conclusion drawn by Kang and Lee (2015): What constitutes good clustering with real world electricity consumption data, is ambiguous.

5.2 Clustering of the TWEU Profiles

For the clustering of the TWEU profiles, a dissimilarity matrix was computed with complexity $O(n^2)$, which is not scalable. One possible solution to this would be to partition the profile set. Given that the sizes of the partitions are not trivially small, and are randomly sampled to prevent bias, the consequent hierarchical clustering performed on the partitions should still be able to identify abnormal patterns.

For example let us explore the effect of partitioning the profile set in two. The number of comparisons required to create a dissimilarity matrix for 22 730 objects is

$$\binom{22\,730}{2}$$

Assuming the dissimilarity measure is symmetric, which DTW is. If we partition the set into two sets of $\frac{22\,730}{2} = 11\,365$, the number of comparisons is

$$\binom{11\,365}{2} \cdot 2$$

Comparing them we have that

$$\frac{\binom{11\,365}{2} \cdot 2}{\binom{22\,730}{2}} \approx 0.50$$

This means that partitioning the set in two would yield a reduction in computational load of $\approx 50\%$ for the calculation of the dissimilarity matrix.

Moving on to the sizes of the clusters, recall Figure 8 which shows the cluster sizes from the validation set, it shows many small clusters and one dominant. This supports Kang and Lee (2015), who found the same to be true for agglomerate hierarchical clustering of electricity consumption data.

A possible improvement of the TWEU profile clustering is to divide the data set according to its use, as opposed to having mixed use clustering. By clustering more homogeneous data, for instance, only residential buildings, a narrower and more suitable view of normal consumption can be created. Consequently, abnormal usage is easier to detect.

5.3 Comparison to Domain Expert Classification

Regarding evaluation of the precision of the expert classifications in Section 4.2, a Fleiss' kappa of ≈ 0.50 is acceptable, but not great. The mediocre result could be explained by simple plots the expert viewed (exemplified by Figure 6), not giving them an understanding of the consumption pattern. This could have been improved by allowing the experts to interactively explore the time series data via some medium such as Grafana¹.

Recall that Figure 9 shows the score distributions for both normal and abnormal buildings, as classified by domain experts. The plots show a significant level of both false positives and false negatives. Moreover, recall that the interviewed domain expert stressed the difficulty in evaluating a consumption pattern without context, and that what constitutes good and bad consumption is highly nuanced and context dependent.

Moreover, recall that the interviewed domain expert noted that common consumption patterns might not be good consumption. This problem is also described by the literature (Himeur et al. 2021; Pan, Yin, and Jiang 2022).

Given the reasons outlined above, there is an upper limit to how well anomalies can be detected by only clustering the energy consumption. It must be admitted that this approach is fundamentally limited.

Compared to similar works in electricity consumption anomaly detection, the accuracy of this method is low. Several reasons can explain this.

1. The data used for this project is unusually heterogeneous. It is common to work with homogeneous data, such as only residential buildings or only schools (Cui and Wang 2017; Oprea et al. 2021). It is more challenging to find anomalies in mixed use data.

¹<https://grafana.com/>

2. The results are compared to expert evaluations. It is very rare to have access to domain experts, for this thesis, only two such papers have been found (Kang and Lee 2015; Cui and Wang 2017). It is often more difficult to optimize for alignment with experts, than optimizing for some statistical metric.
3. No contextual information has been taken into account. The literature shows this is important (Himeur et al. 2021; Oprea et al. 2021).

5.4 On the Mann-Whitney U Test

Regarding the choice of the Mann-Whitney U test for comparing the scores of the abnormal and normal buildings, the selection can be questioned. The test, along with several similar tests, assumes the two samples to be independent. This assumption may be questioned. As all buildings are clustered together, they inevitably impact the clustering of each-other, and the sizes of the clusters. These two metrics determine the scores of the buildings. Ergo, it can be argued that the two populations are not independent. However, the link between the consumption pattern of one building, and the score of another is very weak on an individual level. And because the samples are randomly sampled reflecting the real world, the assumption is deemed sufficiently met.

6 Conclusion

6.1 Remarks

In conclusion, this thesis has shown that it is possible to identify buildings with anomalous energy consumption by clustering time series energy consumption data using the methodology laid out in this thesis. However, the suggested methodology is plagued by large numbers of both false positives and false negatives. Better dissimilarity measures, clustering algorithms, more homogeneous data, and more contextual information on the data could improve this. Nevertheless, determining what constitutes anomalous energy consumption is a highly nuanced and context dependent question that is difficult to answer given only the raw consumption data. Even domain experts find the task to be difficult when manually reviewing the data.

Another conclusion is that validating clusterings without access to ground truth remains an open problem. The meaningless results of the indexes used for the TWEU profile extraction illustrated in Figure 3, support the conclusion draw by Kang and Lee (2015), that what constitutes good clustering with real world electricity consumption data is ambiguous.

6.2 Future Work

No combination of feature extraction method and dissimilarity measure exists that adequately mirrors the view of a domain expert. It might be possible to engineer such a combination.

The cost of having domain experts manually classifying hundreds of consumption patterns is not insignificant, but the ability to evaluate models relative to a ground truth is nothing less than vital. Having even more manually classified data would increase the chances of success for any future project. Other possible remedies this project has not utilized are resampling and cross-validation methods which can improve performance in spite of small testing data sets.

As previously suggested by Pérez-Chacón et al. (2018), including electricity price could provide a better analysis. As previously stated, one of the indicators of poor energy usage is having consumption spikes overlap with grid peak consumption (which is very highly correlated with price). Interestingly, Liu et al. (2021) found that two of the most important contextual data types are the ambient temperature, and distinguishing working days from non-working days. Including these three variables into the analysis could significantly improve accuracy.

Finally, using the K-sliding distance (Kang and Lee 2015) could prove helpful. Many papers stress the importance of considering domain expertise, and the K-sliding distance is developed specifically to address this concern. Developing a highly optimized (possibly GPU accelerated) open implementation of K-sliding would be of great use to the area of electricity consumption time series analysis.

7 References

- Kang, Jimyung and Jee-Hyong Lee (2015). “Electricity Customer Clustering Following Experts’ Principle for Demand Response Applications”. In: *Energies* 8.10, pp. 12242–12265. ISSN: 1996-1073. DOI: 10.3390/en81012242.
- Tureczek, Alexander Martin and Per Sieverts Nielsen (2017). “Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data”. In: *Energies* 10.5. ISSN: 1996-1073. DOI: 10.3390/en10050584.
- Motlagh, Omid, Adam Berry, and Lachlan O’Neil (2019). “Clustering of residential electricity customers using load time series”. In: *Applied Energy* 237, pp. 11–24. ISSN: 0306-2619. DOI: <https://doi.org/10.1016/j.apenergy.2018.12.063>.
- Li, Kehua et al. (2018). “Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering”. In: *Applied Energy* 231, pp. 331–342. ISSN: 0306-2619. DOI: <https://doi.org/10.1016/j.apenergy.2018.09.050>.
- Barredo Arrieta, Alejandro et al. (2020). “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58, pp. 82–115. ISSN: 1566-2535. DOI: <https://doi.org/10.1016/j.inffus.2019.12.012>.
- Hansson, Julia et al. (Mar. 2017). “The Potential for Electrofuels Production in Sweden Utilizing Fossil and Biogenic CO₂ Point Sources”. In: *Frontiers in Energy Research* 5. DOI: 10.3389/fenrg.2017.00004.
- Liao, T Warren (2005). “Clustering of time series data—a survey”. In: *Pattern recognition* 38.11, pp. 1857–1874.
- Batista, Gustavo EAPA et al. (2014). “CID: an efficient complexity-invariant distance for time series”. In: *Data Mining and Knowledge Discovery* 28, pp. 634–669.
- Paparrizos, John and Luis Gravano (2017). “Fast and accurate time-series clustering”. In: *ACM Transactions on Database Systems (TODS)* 42.2, pp. 1–49.
- Aghabozorgi, Saeed, Ali Seyed Shirkhorshidi, and Teh Ying Wah (2015). “Time-series clustering – A decade review”. In: *Information Systems* 53, pp. 16–38. ISSN: 0306-4379. DOI: <https://doi.org/10.1016/j.is.2015.04.007>.
- Agner, Felix (2019). “Creating Electrical Load Profiles Through Time Series Clustering”. MA thesis. Sweden: Lund University.
- Kim, Jaehwi and Jaehee Kim (2020). “Comparison of time series clustering methods and application to power consumption pattern clustering”. In: *Communications for Statistical Applications and Methods* 27.6, pp. 589–602.
- Iglesias, Félix and Wolfgang Kastner (2013). “Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns”. In: *Energies* 6.2, pp. 579–597. ISSN: 1996-1073. DOI: 10.3390/en6020579.
- Bellman, Richard and Robert Kalaba (1959). “On adaptive control processes”. In: *IRE Transactions on Automatic Control* 4.2, pp. 1–9.

- Hastie, Trevor et al. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.
- Bouveyron, Charles et al. (2019). *Model-based clustering and classification for data science: with applications in R*. Vol. 50. Cambridge University Press.
- Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the royal statistical society: series B (methodological)* 39.1, pp. 1–22.
- McLachlan, Geoffrey J. and Thriyambakam Krishnan (1997). *The EM Algorithm and Extensions*. Wiley.
- Pinto, Rafael Coimbra and Paulo Martins Engel (2015). “A fast incremental gaussian mixture model”. In: *PloS one* 10.10, e0139931.
- Al-Jarrah, Omar Y et al. (2017). “Multi-layered clustering for power consumption profiling in smart grids”. In: *IEEE Access* 5, pp. 18459–18468.
- Pedregosa, F. et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
- Javed, Ali, Byung Suk Lee, and Donna M. Rizzo (2020). “A benchmark study on time series clustering”. In: *Machine Learning with Applications* 1, p. 100001. ISSN: 2666-8270. DOI: <https://doi.org/10.1016/j.mlwa.2020.100001>.
- Rousseeuw, Peter J. (1987). “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- Davies, David and Don Bouldin (May 1979). “A Cluster Separation Measure”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-1*, pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.
- Caliński, Tadeusz and Jerzy Harabasz (1974). “A dendrite method for cluster analysis”. In: *Communications in Statistics-theory and Methods* 3.1, pp. 1–27.
- Dunn, Joseph C (1974). “Well-separated clusters and optimal fuzzy partitions”. In: *Journal of cybernetics* 4.1, pp. 95–104.
- Damayanti, R et al. (Mar. 2017). “Electrical Load Profile Analysis Using Clustering Techniques”. In: *IOP Conference Series: Materials Science and Engineering* 180.1, p. 012081. DOI: 10.1088/1757-899X/180/1/012081.
- Iliev, Göran (2022). “Analysis of Electricity Usage Time Series with K-means Clustering”. MA thesis. Sweden: Uppsala University.
- Shaukat, Kamran et al. (2021). “A review of time-series anomaly detection techniques: A step to future perspectives”. In: *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 1*. Springer, pp. 865–877.
- Blázquez-García, Ane et al. (2021). “A review on outlier/anomaly detection in time series data”. In: *ACM Computing Surveys (CSUR)* 54.3, pp. 1–33.
- Pan, Haipeng, Zhongqian Yin, and Xianzhi Jiang (2022). “High-Dimensional Energy Consumption Anomaly Detection: A Deep Learning-Based Method for Detecting Anomalies”. In: *Energies* 15.17, p. 6139.
- Cui, Wenqiang and Hao Wang (2017). “A new anomaly detection system for school electricity consumption data”. In: *Information* 8.4, p. 151.
- Oprea, Simona-Vasilica et al. (2021). “Anomaly detection with machine learning algorithms and big data in electricity consumption”. In: *Sustainability* 13.19, p. 10963.

- Liu, Xue et al. (2021). “A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data”. In: *Energy and Buildings* 231, p. 110601.
- Chicco, G., R. Napoli, and F. Piglion (2006). “Comparisons among clustering techniques for electricity customer classification”. In: *IEEE Transactions on Power Systems* 21.2, pp. 933–940. DOI: 10.1109/TPWRS.2006.873122.
- Tureczek, Alexander, Per Sieverts Nielsen, and Henrik Madsen (2018). “Electricity Consumption Clustering Using Smart Meter Data”. In: *Energies* 11.4. ISSN: 1996-1073. DOI: 10.3390/en11040859.
- Li, Ran et al. (2016). “A novel time-of-use tariff design based on Gaussian Mixture Model”. In: *Applied Energy* 162, pp. 1530–1536. ISSN: 0306-2619. DOI: <https://doi.org/10.1016/j.apenergy.2015.02.063>.
- Benítez, Ignacio et al. (2016). “Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance”. In: *Electric Power Systems Research* 140, pp. 517–526. ISSN: 0378-7796. DOI: <https://doi.org/10.1016/j.epsr.2016.05.023>.
- Tardioli, Giovanni et al. (2018). “Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach”. In: *Building and Environment* 140, pp. 90–106. ISSN: 0360-1323. DOI: <https://doi.org/10.1016/j.buildenv.2018.05.035>.
- Riihimäki, Henri and Pekka Koponen (2012). “Prediction of energy consumption from outdoor temperature for houses electrically heated via heat storage”. In: *VTT, Helsinki, Research Report VTT-R-02882-12*.
- Öhman, Albin (2022). “Machine learning for cable shoe press classification on embedded systems”. MA thesis. Sweden: Umeå University.
- Harris, Charles R. et al. (Sept. 2020). “Array programming with NumPy”. In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2.
- McKinney, Wes (2010). “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.
- Virtanen, Pauli et al. (2020). “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.
- Seabold, Skipper and Josef Perktold (2010). “statsmodels: Econometric and statistical modeling with python”. In: *9th Python in Science Conference*.
- Tavenard, Romain et al. (2020). “Tslern, A Machine Learning Toolkit for Time Series Data”. In: *Journal of Machine Learning Research* 21.118, pp. 1–6.
- Fleiss, Joseph L (1971). “Measuring nominal scale agreement among many raters”. In: *Psychological Bulletin* 76.5, pp. 378–382.
- Mann, Henry B and Donald R Whitney (Mar. 1947). “On a test of whether one of two random variables is stochastically larger than the other”. In: *The annals of mathematical statistics*, pp. 50–60.
- Himeur, Yassine et al. (2021). “Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives”. In: *Applied Energy* 287, p. 116601.
- Pérez-Chacón, Rubén et al. (2018). “Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities”. In: *Energies* 11.3. ISSN: 1996-1073. DOI: 10.3390/en11030683.