# Master's Thesis

**Title:** Analysis of Residential Household Energy Consumption Using Smart Meter Data

**Programme/specialisation:** Spring semester, 2019

**Computer Science**

**Author:** Chibuzor Valentina Nwemambu

**Programme coordinator:** Open

**Supervisor(s):** Chunming Rong, Mina Farmanbar, Aida Mehdipourpirbazari

**Title of master's thesis:**

Analysis of Residential Household Energy Consumption Using Smart Meter Data

**Credits:** 30

**Keywords:**
- Smart Meter,
- Machine Learning,
- K-Means clustering,
- Hierarchical clustering,
- DBSCAN clustering,
- Power consumption

**Number of pages:** 66

**+ supplemental material/other:** 23

**Appendix A and Thesis_code.7z**

**Stavanger, 14/06/2019**
Analysis of Residential Household Energy Consumption Using Smart Meter Data

Master’s Thesis in Computer Science
by
CHIBUZOR VALENTINA NWEMAMBU

Main Supervisor
Chunming Rong

Co - Supervisor
Mina Farmanbar

Additional Co - Supervisor
Aida Mehdipourpirbazari

June 15, 2019
“I have tried fear, it didn’t seem to get me anywhere... I might as well try COURAGE”

- Chibuzor Nwemambu
Abstract

The need to change the source of electricity generation is apparent in the effect of climate change on the environment. Asides from the source change to renewable energy, the necessity for residents to understand their consumption rates and patterns is paramount to help reduce CO$_2$ emission and thereby reduce climate change.

This thesis discusses and implements the machine learning algorithm; K-Means clustering method, on a dataset derived from a town in Norway. The dataset is split into various features, to reveal the cluster and consumption patterns, their peak, off-peak, as well as mid-peak periods, in order to identify times where energy wastage can be minimized. It also experiments and compares two other algorithms; Hierarchical clustering and DBSCAN method, against the K-Means method, showing their differences and similarities, thereby deciding which algorithm is best suited for clustering the provided dataset.
Acknowledgements

Throughout the course of this thesis, I have been honored to receive tremendous help from lots of people around me, and I would love to say:

To my supervisors; Rong Chunming, Mina Farmanbar, and Aida Mehdipourpirbazari, thank you for all the time, effort, and encouragement you invested towards my achieving this thesis in the last six months. Your passion and dedication to its success are immeasurable and I truly appreciate you.

To my family; For your unwavering love and support throughout the course of this program, sharing in my stress and frustration, all the while encouraging and motivating me to push on. I cannot express how deeply grateful and fortunate I am to be yours and you, mine. I love you and this thesis is for you, WE DID IT!

To all my friends; Thank you for these two years of friendship(and counting), the trips, the laughs, the shared frustrations, and the pooled encouragement. Most especially Luca Tomasetti, Redjol Resulaj, and Ferdinand Rødne Tvedt, I couldn’t have asked for better. You are more than friends, you are my brothers.

to my love; Chisobem Nwokedi, You kept me focused through troubled times, reminding me of who I am, what my accomplishments are, and where I am headed. You are one in a million and I am so glad to have you in my life. Always and forever...
## Contents

Abstract vi

Acknowledgements viii

Abbreviations xi

1 Introduction 1
   1.1 Problem Definition ............................................. 2
   1.2 Scope of Experimentation .................................... 2
   1.3 Motivation ....................................................... 3
   1.4 Outline ......................................................... 3

2 Background and Related Work 5
   2.1 General Background ............................................. 5
      2.1.1 Electricity System and its History ....................... 5
      2.1.2 Grids, Meters, and their Data ........................... 6
         2.1.2.1 Conventional Electricity Grids and Meters ........ 6
         2.1.2.2 Smart Meter, Smart Grid ............................. 7
         2.1.2.3 Smart Grid and its Impact on Climate Change .... 7
      2.2 Technical Background ....................................... 8
         2.2.1 Machine Learning ....................................... 8
            2.2.1.1 K-Means Clustering ................................ 10
            2.2.1.2 Hierarchical Clustering ......................... 15
            2.2.1.3 DBSCAN ......................................... 17
      2.3 Related Work ................................................ 19

3 Solution Approach 23
   3.1 Machine Learning Platform ................................. 23
   3.2 Description of the Data ...................................... 24
   3.3 Data Pre-processing .......................................... 26
   3.4 Algorithm Implementation .................................... 29
      3.4.1 K-Means Clustering Implementation .................... 29
      3.4.2 Hierarchical Clustering Implementation ............. 30
      3.4.3 DBSCAN Implementation .................................. 30

4 Experimental Evaluation and Discussion 33
CONTENTS

4.1 Experimental Results ........................................ 33
  4.1.1 Daily Resolution ........................................ 33
  4.1.1.1 Similarities between Houses ....................... 35
  4.1.2 Hourly Resolution .................................... 36
    4.1.2.1 All Houses ........................................ 37
    4.1.2.2 All Dates - Four Seasons ....................... 38
    4.1.2.3 All Dates - Weekdays and Weekends .......... 41
    4.1.2.4 Individual Houses ............................... 44
    4.1.2.5 Winter and Summer for House gw_15 ........ 46
    4.1.2.6 Comparing Winter Seasons for Houses gw_15 and gw_33 47
    4.1.2.7 Winter and Summer for House gw_48 .......... 50
    4.1.2.8 Comparing the Winter Season for Houses gw_48 and gw_39 .................. 52
    4.1.2.9 Mondays and Sundays for House gw_33 .... 53
    4.1.2.10 Mondays and Sundays for House gw_39 ..... 55
  4.2 Comparing Algorithms .................................. 57
    4.2.1 House gw_13 ....................................... 57
    4.2.2 House gw_50 ....................................... 60
  4.3 Algorithm Discussion .................................. 62
  4.4 Result Summary ......................................... 63

5 Conclusion and Future Works  65
  5.1 Future Works .......................................... 66

List of Figures ............................................. 66

List of Tables .............................................. 71

A Appended Results for the Cluster Analysis  73
  A.1 Results for All Houses ................................. 73
  A.2 Results for Individual Houses ....................... 75
    A.2.1 House gw_14 ..................................... 76
    A.2.2 House gw_9 ....................................... 78
    A.2.3 House gw_1 ....................................... 80
  A.3 Results on Compared Algorithms ..................... 82
    A.3.1 K-Means clustering ............................... 82
    A.3.2 DBSCAN ........................................... 83
    A.3.3 Hierarchical clustering ......................... 84

Bibliography ................................................. 87
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>Alternating Current</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>CSO</td>
<td>Central Statistics Office</td>
</tr>
<tr>
<td>CCF</td>
<td>Cluster Classifying Forecast</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma Separated Value</td>
</tr>
<tr>
<td>CM</td>
<td>Conventional Meter</td>
</tr>
<tr>
<td>DQN</td>
<td>Deep Q Network</td>
</tr>
<tr>
<td>DR</td>
<td>Demand Response</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>Density - Based Spatial Clustering of Applications with Noise</td>
</tr>
<tr>
<td>DC</td>
<td>Direct Current</td>
</tr>
<tr>
<td>EDD</td>
<td>Energy Demand Distribution</td>
</tr>
<tr>
<td>eps</td>
<td>Epsilon</td>
</tr>
<tr>
<td>HDF5</td>
<td>Hierarchical Data Format 5</td>
</tr>
<tr>
<td>LAN</td>
<td>Local Area Network</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>minPts</td>
<td>minimum number of Points in eps</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Network</td>
</tr>
<tr>
<td>SG</td>
<td>Smart Grid</td>
</tr>
<tr>
<td>SM</td>
<td>Smart Meter</td>
</tr>
<tr>
<td>SMBM</td>
<td>Smart Meter Based Model</td>
</tr>
<tr>
<td>SS</td>
<td>Sum of Square</td>
</tr>
<tr>
<td>SSE</td>
<td>Sum of Square Error</td>
</tr>
<tr>
<td>SVM</td>
<td>Support of Vector Machine</td>
</tr>
<tr>
<td>WAN</td>
<td>Wide Area Network</td>
</tr>
<tr>
<td>WCSS</td>
<td>Within - Cluster Sum of Squares</td>
</tr>
</tbody>
</table>
Introduction

Humanity’s dependence on the consumption of electricity can be seen clearly in its everyday use. It is considered a basic amenity (in most countries) as it is required for charging mobile devices, operating electrical appliances in our homes, as well as powering electric vehicles among other functions. In most industries, it is needed for production and for some, to aid production. In fact, we can say that it is an intrinsic part of modern living.

Presently, the world’s leading energy source comes from coal, oil, and natural gases, which also accounts for a significant increase in climate change. For this reason, other renewable sources of energy generation (like hydropower, bioenergy, solar energy, geothermal energy, and others) have been researched and implemented to help reduce the dependence on fossil fuels. The introduction of renewable energy in electricity generation has seen an increase in the past couple of years, with hydropower as the leading source for electricity generation globally[1]. However, its growth is still considered slow, since the transition from the conventional method to renewable energy is financially expensive and time-consuming.

Buildings, with the inclusion of residential households, account for about 40% of the global energy consumption, and the energy demand is expected to grow in the coming years. For this reason, it is critical that more emphasis is placed on the use of renewable
energy, methods to improve efficiency as well as the sensitization of residents on energy wastage reduction. This use of renewable energy, was one of the driving forces for the introduction of the Smart Grid (SG) and Smart Meter (SM). Further discussion on this, is found in chapter 2.

The use of SMs is currently growing, as information retrieved from them is confirmed to be very useful in understanding daily energy consumption as well as strategizing and developing better ways to help in its reduction (through incentives and other means). For this information to be used strategically, Machine Learning (ML) techniques are applied to analyze the data. This analysis gives a window into the consumption patterns of the residence and can help create ways to reduce energy wastage and spot anomalies efficiently. The most common method of analysis using ML is through the clustering of similar information in such a way that a visible pattern emerges.

### 1.1 Problem Definition

Analyzing energy consumption derived through the implementation of SM data involves the gathering of energy readings from varied households, pre-processing these data to allow for optimal accuracy, clustering them and, through the result, determine the periods with significant energy consumption and periods with less. This analysis will also be able to help determine the best period to reduce or eliminate unnecessary use of electricity.

### 1.2 Scope of Experimentation

This thesis focuses on analyzing data retrieved from SMs of residential households in Stavanger, a town in southwestern Norway, and clustering them based on similarities in energy consumption, using extracted features from the dataset such as days of the week, seasons, holidays and others. In this thesis, the following points will be treated:

- description and understanding of the dataset,
- description of three clustering algorithms used (K-Means, Hierarchical, and DBSCAN),
- data pre-processing and filling of missing values,
- application of the K-Means algorithm on certain features,
- analysis of the results derived from K-Means clustering,
- application of the other algorithms on specific features,
- analysis of the results derived,
- comparison of the different algorithms tested.

1.3 Motivation

Electricity is generally produced for the end user’s consumption. Although that is the case, it is relatively connected to two other groups: electricity production, and distribution. All three, form a cycle for delivering the best and most efficient quality of service, which is hinged, as of recent, on the evaluation of the consumption pattern.

On the part of the consumer, one of the primary reasons for analysis is to understand how much electricity is required to run their household and help determine a more relatively efficient way to reduce wastage. On the part of the production and distribution, this analysis provides the opportunity to discover anomalies, which can be a partial solution to understanding why they occur in the first place, and how to eliminate them, as well as to develop better strategic packages which will help discourage energy consumption during peak hours. The Government and other regulatory bodies also require this analysis, as it provides them with the information required for tax reviews (if necessary), and a more detailed view of the cumulative effect of energy wastage on the economy. This information buttresses the convenience and necessity of using renewable energy as against the use of fossil fuels.

Ultimately, the ability to reduce energy wastage falls on the consumers and their knowledge on how much energy they can save with few precautions.

1.4 Outline

This thesis consists of five chapters, and a brief overview (with the exception of chapter 1) is given below:

Chapter 2 gives a general background into electricity, SGs, and SMs, the widespread of SGs and SMs, how data retrieval is conducted, how they are connected, their advantages, as well as effects on climate change. Also, it covers the technical background on ML and the different algorithms proposed for use in this thesis, its methods, and uses, concluding with works related to this thesis.
Chapter 3 focuses on the methods used, the ML platform used and the various libraries employed, the description of the dataset and its pre-processing as well as the implementation of the ML.

Chapter 4 outlines the experimentation and its algorithmic flow; the results achieved using the various features extracted, comparison of the results, and the different algorithms as well as discussions regarding the results.

Chapter 5 discusses future works as well as the conclusion on what the thesis covers, what methods were used, why certain methods were used, what the results and its accuracies are, and what its general effect would be, in its application on a bigger dataset.
2

Background and Related Work

2.1 General Background

2.1.1 Electricity System and its History

The first commercial electricity supply was generated in 1882 by Thomas Edison, who distributed Direct Current (DC) in residences and office buildings in Wall Street[2]. From that point, the evolution of the electrical industry rapidly changed, and further inventions and discoveries, like the Alternating Current (AC, which is currently used) were quickly revealed.

In more recent years, however, the electricity supply is considered to be an essential amenity in most countries, and substantial investment in this sector is typically owned by the government and is heavily regulated. Take Norway for instance, its municipalities and counties, along with central authorities own about 90% of its electricity production capacity [3].

Norway is considered one of the best worldwide, in regards to electricity production, producing over 149 terawatt-hours (TWh) in 2017, and consuming about 124 TWh (net consumption) [4], with the average energy consumption per household capped at 16,044 kilowatt-hours (kWh) in 2012 [5].
2.1.2 Grids, Meters, and their Data

In some countries, the importance or even the understanding of SMs, are not sufficiently emphasized as most people still use Conventional Meters (CM) whose readings are recorded monthly, and electricity consumption is difficult to manage due to the inaccessibility of the readings. Conventional meters are known to encourage wastage because of this, but SM eliminates most of these problems.

2.1.2.1 Conventional Electricity Grids and Meters

According to [6], a conventional grid is a complex interconnected electrical power system that generates, transmits and distributes power to the final consumers through the use of power generating plants, transmission lines, substations, transformers, distribution lines, and various other equipment. Its system structure is broadly divided into four sectors: power generation, transmission, distribution, and consumption(supply). For this thesis, the main focus will be on the electricity consumption sector.

One of the definite signs of a country’s GDP and social development, according to [2], is its energy consumption as most production processes require the use of power. Also, the percentage of people living with essential home appliances indicates its likely energy consumption.

Electric meters that are used to record the total energy consumption were introduced in the 19th century and have evolved over the years. The most commonly used is the Electromechanical Watt-Hour Meter. Data retrieval from a conventional meter is done monthly, with the current reading subtracted from that of the previous month to get the actual reading for the month. This method of functionality is quite unreliable and prone to errors because:

i. It requires the physical presence of a person to visit the location of the meter and take the readings;

ii. the billing derived from the meter readings are usually not the actual cost of energy consumed, but an estimated bill and;

iii. disconnection from the electricity grid is rather tedious and expensive as it requires the removal of the electrical cable connecting the building to the grid as against the touch of a button or the flick of a switch.
2.1.2.2 Smart Meter, Smart Grid

A Smart Grid (SG), on the other hand, is everything a conventional grid is, but with the advantage of advanced digital technology that allows the ability to interact with the consumers, producers, and distributors.

Smart Meters (SMs) are one of the new technologies which connect households to the Grid. On a fundamental scale, SM is installed and connected to the home, reads and retrieves electrical footprint of energy consumed (real-time energy consumption, usually on an hourly basis depending on its configuration), store this information and transmit them to the data collector through the Local Area Network (LAN), who in turn, transmits the data to the utility central collection points. The data is further processed by using the Wide Area Network (WAN) [7] to provide more accurate energy billings for the consumers, among other things.

One prominent feature of SM is that it allows for bi-directional communication between the consumer (through the meter) and the producer (the central electrical system). Another feature is the ability to disconnect and reconnect specific electrical loads. This feature is of relative importance, as it allows the consumer to manage their loads efficiently.

2.1.2.3 Smart Grid and its Impact on Climate Change

Many factors determine the changes made in the electricity system, and climate change is one of them. According to [8], the electricity and heat sector is the largest emitting sector of CO$_2$ in 2016, which accounted for about 42% of the global emission. Within the electricity sector, the second-largest emitter is buildings (including residential houses) with 27% CO$_2$ emission because of the firm reliance on electricity. Therefore, all avenue
to study the effect of SG on green gas and CO\textsubscript{2} emission is investigated and experimented by experts. Some of the main challenges faced are the rejection by some world leaders that climate change is real and also the additional cost for implementing SG. There is also the fear that if the right climate dimensions are not implemented in SG, it could help worsen climate risks.

Even with the risks involved, there is a great benefit attached to its proper implementation. The article [9] estimates that ‘a direct reduction of 6% in electricity consumption, with a range of 1% to 10%, can be achieved in the residential and small/medium commercial building sectors through the implementation of SG technologies’.

Some of the benefits of the SG can be actualized if:

- The SG allows for renewable energy generation like wind, solar and geothermal energies.
- Customers are made more aware of their energy consumption, real-time prices and actively participate in strategies to reduce wastage.
- System malfunction sensors are deployed to detect faults in real-time.
- Electricity generated during peak hours by resources using fossil fuels, can be shifted to resources with low carbon emission.

Lastly, it can be noted that renewable energy is a natural phenomenon and is rather unpredictable. Therefore this could pose a challenge to its implementation in the SG. However, implementing energy storage and discharge strategy by separating energy storage from demand will help reduce its unpredictability and enable its use during peak periods[10].

### 2.2 Technical Background

In order for the Machine Learning to be implemented for data analysis, it requires a large amount of data, as the quantity and quality of data determines the results generated.

#### 2.2.1 Machine Learning

The general phrase, ‘we learn from experience’ is the most fitting, regarding ML. Humans learn to perform tasks better and faster through constant practice, and ML involves a similar tactic.
It is a branch of Artificial Intelligence (AI) that involves the training of machines to learn, reason, understand, and eventually predict outputs/results from experience with previous data. That means machines get the ability to act without any constant influence or programming from an outsider.

Some daily applications which use machine learning are Apple’s "Siri", Google maps, social media services, email spam filtering, and many others, while broadly, its use in the medical industry, finance, pricing models, web search, pattern and image recognition and many more, have not gone unnoticed.

ML is divided into three types:

i. Supervised machine learning.

ii. Reinforced machine learning.

iii. Unsupervised machine learning.

**Supervised learning** - is categorized by the input and output pair of data already being provided. In a supervised learning algorithm, the data is divided into training and testing data. The training comprises of input and output pairs with the input specifying certain features paired up with the desired output based on those features. This training data becomes the experience/example which the machine implements to predict the desired output for the testing dataset.

The supervised learning algorithm is mostly used for prediction of future events, and some examples include Linear Regression, Naive Bayes, K-Nearest Neighbour algorithm, Logistic Regression, e.t.c.

**Reinforcement learning** - in reinforcement learning, the machine algorithm generates the desired output by learning from its experience. Through trial and error, the algorithm determines the optimal result by using its feedback from its actions and experiences. The primary goal for this algorithm is to quickly recognize actions that yield the desired output over some time. This algorithm is an iterative process that is especially useful in navigation, robotics, and gaming, and some examples include the Q-Learning algorithm, State Action Reward State Action algorithm (SARSA), Deep Q Network algorithm (DQN) and others[11].

**Unsupervised learning** on the other hand, has no pre-labeled or "training dataset," which enables it to determine the output. Since there are no labeled data, the algorithm gathers inference through its perception of what the related features(or lack thereof) are. The unsupervised learning algorithm is mostly used for exploratory analysis, and some
examples include anomaly detection and clustering, amongst others. For this thesis, the focus will be on clustering.

![Figure 2.2: Steps in Cluster Analysis](image)

Cluster Analysis divides data into clusters and is most effective with a large amount of data, as it becomes easier for the algorithm to see patterns. Figure 2.2 gives a general overview of the steps involved in this process.

The similarities between data in the same cluster are maximized, as much as the difference between each cluster group, which eventually generates an optimal result. However, some cluster groups are challenging to identify distinctively (without the color scheme) as data in these clusters may overlap due to their proximity.

Various clustering technique exists, but only three are discussed in this thesis, and they are:

- K-Means Clustering.
- Hierarchical Clustering and.
- Density -Based Spatial Clustering of Applications with Noise.

### 2.2.1.1 K-Means Clustering

K-Means clustering is a centroid-based unsupervised learning algorithm were ’k’ denotes the number of clusters required for optimal grouping of the dataset. Centroid points are
the designated location for a cluster, and the closer a data point is to a centroid, the more likely it is to be in that cluster.

According to [12], k-means can be expressed statistically as finding the number of cluster; $C := c_1, ..., c_k$ that minimizes the Within-Cluster Sum of Squares (WCSS) also referred to as the Variance using:

$$WCSS_1 := \sum_{c_i \in C} \sum_{j=1}^{d} 2|c_i| \sum_{x \in C_i} (x_{ij} - \mu_{ij})^2$$

where $\mu_{ij}$ is the mean coordinate of cluster $i$ and dimension $j$.

In order to use k-means, the number of clusters must be pre-determined. K-means is centroid-based which means based on the number of clusters selected, $k$, the number of center points (centroids) is strategically located to assign each datum to the cluster nearest to it. That means that the determination of $k$ is paramount to the application and output of the algorithm.

The determination of $k$ is implemented as:

![Figure 2.3: Algorithm for $k$ selection](image)

After $k$ has been selected, the k-means algorithm becomes iterative, by:
i. determining the distance between a data and the closest centroid to it, using the Euclidean Distance (straight line distance),

ii. assigning the data to the closest centroid,

iii. determining the mean of all data clustered at each centroid and using that value as the new centroid point.

That means each centroid is randomly placed far from one another and each datum is associated with the centroid closest to it. Once all data are assigned, the centroids are re-adjusted until they are equidistant. This process is repeated multiple times until the centroid can no longer be adjusted.

K-means clustering is an easy but powerful algorithm used in ML. Figure 2.4 is a visualization of how an un-clustered dataset can be clustered. One of its significant challenges, however, is determining the optimal number of clusters required for the dataset. Using brute force to determine $k$ is possible, but as the dataset grows, it becomes rather unrealistic to take that approach.

Two groups of methods are available to determine the optimal number of clusters, the statistical testing methods, and direct methods. The statistical testing methods allow for making inferences about the data by observing its pattern and coming to a conclusion based on the observation, while the direct method uses the optimization of specific criteria to achieve its aim. In this thesis, the focus was on the use of direct methods Elbow method and the Silhouette method to calculate $k$.  

\[ \text{Figure 2.4: An Example of an Un-clustered and Clustered Dataset} \]
K-Means is considered to be an NP-Hard problem when applied to a $n$-dimensional dataset.

**Elbow Method** - This method helps determine the number of clusters by calculating k-means for a range of times, (for example between 1 to 10 (assuming range = $i$)). For $i$, it calculates the Sum of Squared Error (SSE). SSE measures the differences, in the distance, between each point and the mean of the group. It shows the compactness between data points in a cluster. It is given statistically as:

$$SSE := \sum_{i=0}^{n} (x_i - \bar{x})^2$$

where $x_i = \text{data point in the cluster}$,

$\bar{x} = \text{group mean}$ and,

$(x_i - \bar{x})^2 = \text{deviation of data point from the group mean}$

Using the elbow method, the algorithm to determine the optimal number of $k$ is given below as:

i. Pick a range of $k$ (for example, 1 to 10),

ii. calculate the SSE for each $k$,

iii. plot a curve from the result gotten in ii.,

iv. an elbow shape is formed from the plot.

SSEs can then be plotted on a graph, and the optimal cluster identified as the point just before the SSE flattens out, looking somewhat like an elbow. In most cases, a sharp curve is very distinct, and the number of clusters can easily be identified. However, there exist other cases where a smooth curve is visible, and thus, the optimal number of clusters cannot be readily determined.

The SSE values take a range of between -1 and 1, where a value approaching 0 indicates its very nearness to a neighboring cluster, a value approaching 1 indicates a further distance from a neighboring cluster, and all negative values indicate erroneous clustering.
Figure 2.5 shows an example of how the SSE can look like an elbow, with a distinctive curve point at $k=2$.

**Silhouette Method** - This method analyses the separation distance between clusters by calculating the average distance from all: the data points in the cluster $x$; data points in the next closest cluster to $x$, $y$; and then calculates the coefficient. Simply put, this method calculates how well each data fits in its cluster.

$$\frac{y^i - x^i}{\max(x^i, y^i)}$$

The algorithm for this method is given as follows:

i. Pick a range of $k$ (for example, 1 to 10),

ii. calculate the silhouette of observation for each $k$,

iii. plot a curve from the result gotten in ii.,

iv. the maximum point is considered to be the optimal number of clusters.

With the value of $k$ determined, the dataset based on the applied feature, is then fed into the k-means algorithm.
Some advantages of K-Means clustering are\cite{13}:

- it is straightforward to understand and easy to implement,
- depending on the $k$, sub-clusters can also be determined,
- it is scalable,
- it is computationally fast when dealing with large datasets.

Disadvantages are:

- it is sensitive to density, size and non-spherical shapes,
- re-computation might produce slightly different clusters,
- it cannot identify outliers.

2.2.1.2 Hierarchical Clustering

Hierarchical clustering is an algorithm that has a tree-like structure and are of two types: *Divisive* and *Agglomerative*. Divisive clustering is built with a top-down approach, with all data grouped in a single cluster and subsequently split, step by step, into separate clusters based on their dissimilarities. Agglomerative clustering, on the other hand, is built in a bottom-up approach. In Agglomerative clustering, every data is an individual cluster, and with each progression, any two data with enough similarities are grouped to form a new cluster. In this thesis, the Agglomerative Hierarchical clustering was employed.

Hierarchical clustering is represented using a ’dendrogram’ as seen below.
When analyzing a dendrogram, the $x$ axis represents the data, while the $y$ axis represents the distance between clusters. Better explained, the height of the connecting cluster determines the similarities between them. The more the cluster height, the more dissimilar they are from each other.

In hierarchical clustering, the number of clusters cannot be pre-determined like in the k-means clustering. However, to determine the most appropriate number of clusters, a distance value is picked from the $y$ axis. The determination can be done using either the Euclidean distance or the Manhattan distance. In this thesis, however, only the Euclidean distance is implemented. The Euclidean distance is the most popular distance used, and is calculated as the square root of the sum of two data points.

$$EuclideanDistance := \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

When using either method aforementioned, the objects are grouped into clusters based on their similarities. Each cluster is further merged with other clusters to form a more significant cluster, and this is done repeatedly until no further cluster can be generated.

Similarities between clusters are quite crucial for understanding the reasons why particular clusters are merged. There are various methods (linkage parameters) to determine these similarities, but only the applied method in this thesis will be discussed.
**Ward’s Minimum Variance** - This method groups clusters by minimizing the Sum of Square. This is simplified as:

i. calculating the mean of each cluster,

ii. within each cluster, calculate the difference in distance between each object and the mean of the cluster,

iii. calculate the square of the value from ii. for each object,

iv. add all squared values in each cluster,

v. finally, add all SS values for all clusters.

This method allows clusters with the minimum linkage (within-cluster) distance to merge.

There are other methods/techniques which can be used to calculate similarities between clustering. For a further and well-structured read on this topic, please see [14].

Some advantages of Hierarchical clustering are [13]:

- it is straightforward to understand,
- it works well for small datasets,
- the number of clusters can be determined by observing the dendrogram and deciding on the most appropriate,
- it is less sensitive to noise.

Disadvantages are:

- Once clusters are merged, it becomes permanent. Therefore errors made during clustering cannot be undone,
- it is not scalable when applied on a large dataset,
- it cannot identify outliers.

### 2.2.1.3 DBSCAN

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an algorithm that segments clusters based on how dense a region is. For a simpler explanation, let us assume a dataset $x$. DBSCAN functions by:
i. dividing \( x \) into different sections\((n \text{ sections})\),

ii. within each \( n \), it takes a point \( \alpha \) and calculates the distance to other points,

iii. all neighbouring points within distance of \( \alpha \) is grouped as a cluster,

iv. these steps are iterated until all points within each \( n \) are visited,

v. points that do not get clustered are considered outliers.

For DBSCAN to be accurately implemented, two parameters are required: Epsilon, \('eps'\) and Minimum Points, \('minPts'\). \( eps \) determines the radius of the section \( n \) and \( minPts \) is the minimum number of points within \( eps \). This means a cluster is formed if a minimum number of points are within a pre-determined radius.

In DBSCAN, there are three characterization of points, Corepoints, Border points and Noise points. Core points are those points which fall within \( eps \) and have more than the specified \( minPts \). These points are usually at the densest region of its cluster. Border points are points that do not have the required \( minPts \) to form its core point, but fall within the \( eps \) of another core point and are therefore part of that cluster. They usually form the border of clusters, as the name suggests. Noise points however, are neither core points nor border points and are considered as outliers.

Some advantages of DBSCAN are [13] it:

- identifies both spherical and non-spherical shapes.
- can find clusters within clusters\((\text{so long as the clusters are not connected})\).
- is scalable,
- can identify outliers.

Disadvantages are:

- it is sensitive to the density of the dataset.
- it is sensitive to \( eps \) and \( minPts \).
- outliers might be part of a real dataset but will still not be considered.
### Difference between K-Means, Hierarchical and DBSCAN Clustering

<table>
<thead>
<tr>
<th>Criteria</th>
<th>K-Means</th>
<th>Hierarchical</th>
<th>DBSCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Algorithm Type</strong></td>
<td>Partitioning spatial</td>
<td>Hierarchical spatial</td>
<td>Density-based spatial</td>
</tr>
<tr>
<td><strong>K-Selection</strong></td>
<td>Done prior to clustering</td>
<td>Done after dendrogram is observed</td>
<td>Determined by the algorithm</td>
</tr>
<tr>
<td><strong>Outlier Detection</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Cluster Pattern</strong></td>
<td>Splits dataset into sections and cluster based on centroid position</td>
<td>Starts with each individual data as its own cluster and builds upwards</td>
<td>Splits dataset into sections and cluster based on ( eps ) and ( minPts )</td>
</tr>
<tr>
<td><strong>Time Complexity</strong></td>
<td>( O(n) )</td>
<td>( O(n^2) )</td>
<td>( O(n^2) ) without index structure and ( O(n \log n) ) with index structure</td>
</tr>
<tr>
<td><strong>Outputs</strong></td>
<td>Might differ if run multiple times</td>
<td>Same result every time</td>
<td>Same results every time</td>
</tr>
<tr>
<td><strong>Affected by Density</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 2.1: Difference between the three selected algorithms

### 2.3 Related Work

Related topics have been broadly discussed in recent years using both global and local data to reinforce several points regarding SM data analysis and energy consumption analysis. Yi Wang et al. [15] were able to establish (using data from Web of Science) that there has been a rapid increase in the number of publications regarding this topic between the years 2010 and 2017, with a significant increase noticed in 2012.

Publications span across topics like load analysis and forecasting, to topics about energy fraud, the security of customer data as well as smart meter data analysis and its effect on climate change.

Data gotten from SMs has proven to be very useful for various analyses, and many researchers are creating positive use for their analysis. Alexander Lavin and Diego Klabjan [16] used this data to identify and group similar energy accounts of customers in both commercial and industrial buildings. The hourly-interval data were normalized, and the \( k\)-means clustering algorithm was applied to achieve the desired result.

Datong Zhou et al.[17] uses different machine learning methods to develop estimators for gauging individual treatment effects for residual Demand Response(DR), by incorporating latent variables that allow for improved prediction accuracy. Jungsuk Kwac et al. [18],
on the other hand, uses a three-step approach to help determine residents who will most benefit from DR programs.

Ning Lu et al. [19] discusses methodologies to extract specific data signatures with varying time resolutions, allowing for the possibility of building a database using SM data management system. This system can then be accessed for various information like anomalies in grid operations, consumer energy consumption, and lots more. Using three methods: K-Nearest Neighbors, Support Vector Machine (SVM) and Artificial Neural Network (ANN), Jesse Eisse [20] predicts the energy consumption of individual buildings and these predictions were further tested to determine if they presented any anomalies. On the other hand, Baran Yildiz et al. [21] determined an optimal forecasting method for the individual household, by comparing its developed method ‘Cluster Classify Forecast’ (CCF) and a more popular method ‘Smart Meter Based Model’ (SMBM). It was determined that CCF outperformed its counterpart because of some factors, one of which is that CCF applies the training and testing phase, which are generally applied to the entire dataset, to each household, thereby achieving a better output.

Paula Carroll et al. [22] focuses on using Neural Networks (NN) and Elastic Net Logistic Regression to determine the number of occupants in residence, using data retrieved from the Central Statistics Office of Ireland (CSO) which also collaborated in this project.

Alexander Turecze et al. [23], much like this thesis, works with SM data obtained from domestic/residential buildings in the small Danish city of Esbjerg. Methods and wavelets were applied to extract significant features (autocorrelation coefficients) which eventually aided the production of smaller, better-defined clusters with less ‘within-cluster’ variance. Adrian Albert and Ram Rajagopal [24], on the other hand, propose and develop a new
methodology called Energy Demand Distribution (EDD) to group a large population of users by their variability signatures. Factors that determine the EDD choices were also analyzed to reveal certain large appliances that showed a stronger effect on driving variability compared to conscious efforts by residents on energy usage.

Vitaly Ford and Ambareen Siraj [25] address privacy concerns, by disaggregating energy consumption using Fuzzy C-Means clustering to deduce consumers’ energy usage profile and also currently investigating solutions for energy fraud, data theft and other illicit activities relating to SM data. Another report which deals with security and privacy is [26], which surveyed current developments in big data analytics regarding energy consumption and its security, providing a broad coverage over certain aspects of energy big data analytics and security/privacy as well as real-time and tight cyber-physical coupling.

Nathaniel J Williams et al. [27] disaggregates microgrid customers in Tanzania according to their energy consumption behavior. While using the k-means clustering algorithm, customers were divided into distinct clusters based on the mean daily electricity consumption and mean normalized load profiles. The conclusion can be used to the advantage of optimizing either the microgrid system design or the load profile for the system by the microgrid developers.
This thesis aims to determine and understand energy consumption patterns within each household; between different households and, in comparing them, find their similarities. In order for these to be achieved, clustering is done at various levels. Clustering was performed based on the selected features, which will be discussed later in this chapter.

This chapter focuses on the methods used, data description, pre-processing techniques of the data set, existing approaches, clustering architecture, techniques, and proposed solutions.

### 3.1 Machine Learning Platform

In this thesis, Python programming language was used through the ‘Jupyter Notebook’ open-source web application, and several libraries for machine learning were employed.

The significant libraries imported and used are briefly described below:

- **Numpy** is fundamentally designed for scientific computing with Python and is used to provide support for the multi-dimensional dataset, and sophisticated (broadcasting) functions as well as mathematical function for various calculations [28].
Chapter 3 Solution Approach

- **Pandas** is an open-source, BSD-licensed library that provides high-performance data structure and operations of the data, allowing it to be manipulated and analyzed, as well as being viewed in structured tables among other things.

- **Matplotlib** is a Python 2D plotting library used for data visualization by plotting data in interactive environments.

- **SciPy** is a Python-based ecosystem of open-source software for mathematics, science, and engineering [29], and is used in this thesis to perform hierarchical clustering.

- **Scikit-learn** is an efficient machine learning tool used for data mining and analysis and contains k-means and the DBSCAN algorithms used in this thesis.

All libraries used in this thesis are open-source libraries and are used in combination with one another.

### 3.2 Description of the Data

The data presented for this experiment was obtained from 50 residential houses located in Stavanger, a southwestern city in Norway. Each house is equipped with an SM which invariably takes data readings in 10 seconds intervals. That should give every house, a total of 8640 readings/values per day. The dataset was organized using a calendar-based approach, which means, power is recorded every day of the year. The data gathered spans for a little over one year, beginning from the 7th of February 2017 to the 9th of April 2018.

The unprocessed data was initially stored in Hierarchical Data Format 5 (HDF5) but was later converted to a Comma Separated Value (CSV) format and all queries to the database were made from CSV.

The unprocessed data received, contains names of each houses represented as an alias "gw_" and a digit attached to it, for example, "gw_0, gw_22, ..." etc. These houses contain load demand data for multiple days, with the entire dataset saved in a nested dictionary. The key of the dictionary is the house alias, while its value is another dictionary with its key as the date and its value being two lists, one containing the time (expressed in Epoch) and the other containing the power consumed at the related time. That means each house has multiple dates, and each date has multiple values of time and power.

A couple of information are worthy of note in regards to the dataset:
- The data used in this research are private and therefore, specific details, like the identity of the residents or the exact location of the residence, cannot be discussed.

- No form of energy-saving tips were discussed, with the residents; therefore, all power consumption is based on their regular ‘every day’ schedule.

- The number of occupants living in each residence is unknown as well as their age demographics, occupation, and any other information that might suggest the level of energy consumption.

- Although each house is expected to have 8640 paired values of time and power, some houses had missing data, which were eventually filled during pre-processing.

Another detail worthy of note is that a significant amount of data was missing in the dataset for July 2017 as well as from 25th November 2017 to 12th December 2017. This was due to some technical problems involving the servers which affected the recording process of the data. This period is, therefore, not considered in the analysis.

Asides from the significant data gap mentioned above, the minor missing data were used as the basis for the selection of houses regarding pre-processing. The final number of houses selected was 30 and were chosen based on the number of entries per day, and the number of days available with information.

From the dataset, it was determined that a total of 426 days worth of information was recorded, each house is ideally expected to have 8640 value entries (10 seconds resolution) per day, which amounts to each house having 3,680,640 values entered.
3.3 Data Pre-processing

Data pre-processing is broken down into data importation and selection, filling missing data and time conversion. In order to successfully import the dataset, specific libraries were imported into Jupyter Notebook.
One of the first hurdles tackled in this thesis was the sorting and organization of the data into daily and hourly resolutions. The hourly resolution was further broken down into weekly, monthly, and seasonal profiles.

To select the dataset, first, the houses with more recorded values in more days were selected. Next, to get the houses with the highest number of entries, the number of entries for all days of each house was counted and calculated, to check if it had more or less than 80% missing data of the ideal number of entries (8640). If a house had more than 80% missing data, we assume that this is a result of power loss, equipment failures, or communication issues, and so these houses are dropped. However, if it had less missing data, then it was retained and filled. Days with null values were deleted from the set.


**Filling Missing Data**

After 30 houses were selected, the missing data in these houses had to be filled with specific values. Usually, missing values could be filled using the previous day’s value, but due to the nature of this thesis, the values chosen were those from the corresponding days of the previous week. That means missing values from Tuesday of week 12 were filled with values from Tuesday of week 11 (never the proceeding week). If the value for the previous week is not available, then, the values from 2 weeks prior were chosen. That gave a total of 30 houses with a maximum number of entries for the maximum number of days.

**Time Conversion**

The time presented in the original entries by the SMs were recorded in Unix timestamp (Also called Epoch timestamp), the number of seconds which have elapsed since January 1st, 1970. Each time was saved as a 13 digit value, representing time in milliseconds. The `strptime()` method was used to convert the timestamp to human readable time as a `datetime` object, allowing for proper processing of the data when clustering.

**Data Resampling and Time Resolution**

Due to the volume of the data processed, the latency of the algorithms was observed to rise while the throughput dropped. In total, the dataset contained 2.9 million rows and 26 columns. Run-time was about 2 hours, and any slight error made during implementation would result in running the algorithm all over again; therefore, it was considered unwise to process the information using the 10 seconds resolution. Thus, resampling was implemented using the `resample('H').sum()` to reshape the time records to more efficient representation of hourly resolution. This transformation was also applied to the power values.

**Feature Extraction**

After the data had been formatted to reflect a readable and understandable representation, the next task was to extract particular features which are needed for clustering. Features were extracted based on the type of data available, the amount of data available, consistency of the output after clustering multiple times, and the importance of the feature to the thesis. With these criteria, the following features were decided:

- **Daily Mean**: This is the power consumption for each day represented as a single figure, and is derived by adding all values registered for a day and dividing by the number of entries. This feature was mainly used for daily resolution.
- **Number of Days**: This is the difference between the date examined and the 1st of January 2017.

- **Days of the week**: This identifies all weekdays in the dataset using the numbers 1 through 7 to represent each day of the week, with day 1 representing Monday and day 7 representing Sunday. This identification was extracted using `tempDT.isoweekday()`. Within this extraction, two sub-features were also extracted, namely, Weekdays and Weekends.

- **Holidays**: This is all Norwegian holidays identified in the dataset. Importing the `holidays` library gave access to all global and local holidays, but for this thesis, the focus was placed strictly on local holidays pertaining to Norway. This was extracted using `holidays.NO()`, and each day was represented with either `True` for a holiday or `False` for other days.

- **Seasons**: Each season (Winter, Spring, Summer, and Autumn), were extracted from the dataset with the standard seasonal date used as a reference. This means Spring is considered as being from 20th of March to 19th of June, Summer is from 20th of June to 21st of September, Autumn is from 22nd of September to 20th of December, while Winter is considered as being from 21st of December to 19th of March.

All features were split into two groups, one group for individual houses and another group for all houses as a unit. This split allowed for the clustering of individual houses for each season, each day of the week, weekdays and weekends but also allowed for clustering of all houses as well, for the features mentioned above.

It is worthy of note that other features were also extracted but were quickly observed to be of negligible importance as a better representation of them was already highlighted in the major features listed above.

### 3.4 Algorithm Implementation

In this thesis, three algorithms are evaluated and compared; **K-Means Clustering, Hierarchical Clustering and DBSCAN**. All three have been discussed extensively in section 2.2.1.

In the feature extraction process, dates relating to each feature is stored in a list and is used during the implementation of the algorithms.

#### 3.4.1 K-Means Clustering Implementation

Before the implementation of the algorithm, $k$ had to be determined. Testing the silhouette method, continuous false $k$ was generated, sometimes suggesting the optimal
number of k to be 0, and sometimes up to 26; therefore, the silhouette method was
disregarded. The elbow method turned out to be more reliable and thus was used in the
experimentation.

As expected, each house had a different $k$ but ranged between 2 and 4. For the individual
house clustering, the exact $k$ for each house was used; however, for the implementation
of the group clustering, since the $k$ varied, the highest number of $k$, was chosen to avoid
certain data being under clustered, and to allow for uniform correspondence.

For implementation, the saved dates are passed through the algorithm with specific
criteria, i.e., the date, and a boolean flag, inquiring if the clustering is done for a single
house, or all the houses as a group. Once the criteria are met, the data is then scaled
and translated individually between the values of zero and one using the MinMaxScaler().

Once the output and the determined number of clusters are combined, the algorithm is
performed on the dataset, displaying the clustered set both as a line plot and a scattered
plot. The cluster label attached to each input is also displayed, showing which house, for
instance, fall into which cluster.

3.4.2 Hierarchical Clustering Implementation

For this algorithm, the daily resolution was taken as input and passed, using the
AgglomerativeClustering() function with parameters set for affinity at 'Euclidean' and linkage
at 'Ward'. The Euclidean distance was chosen because it is much more understandable
and quite intuitive, while the Ward minimum variance was chosen because its accuracy
is, somewhat, based on the use of Euclidean distance measure since its function is to
minimize the pooled 'within-cluster'. Also, Ward is more precise at uncovering clusters
with a difference in sizes or clusters irregularly located. The number of clusters was
changed to see varying results in the scatter plot, but was, ultimately, left at 3.

The output is displayed both as a dendrogram and a scattered plot.

3.4.3 DBSCAN Implementation

As with the hierarchical clustering implementation, the daily resolution is taken as
input for this algorithm. The input data was standardized and transformed using
StandardScaler().fit_transform() and used as input into the DBSCAN function.

The DBSCAN() function was implemented using varying values of $eps$ and $minPts$, but was
finally adjusted to $eps$ at 0.3 and $minPts$ at 10 samples as the satisfactory parameters.
The output was displayed as a scattered plot as well as the number of clusters for each house and the estimated number of outliers.
Experimental Evaluation and Discussion

In this chapter, the results derived from the implementations in chapter 3 are much dependent on how well the data is pre-processed. With the features earlier discussed, results are achieved, visual outputs are displayed, and comparisons are made to interpret the findings. This chapter mainly focuses on two aspects. The first is the interpretation of the output, using only the K-Means algorithm to determining the 'Peak', 'Off-Peak' and 'Mid-Peak' periods, for two groups (i.e., the daily resolution and the hourly resolution, and individual houses and all houses as a unit), based on the features. The other aspect is comparing the output of all three algorithms using only the daily resolution as input. These two aspects broadly cover the objective function of this thesis.

4.1 Experimental Results

4.1.1 Daily Resolution

With the daily resolution as input, all days of each house were combined to determine the mean per hour, giving each house a single line representation on the plotted graph. This combination was done to discover which houses would be clustered together based on their similarities in hourly consumption.
Below is the clustered generated:

![Figure 4.1: Clusters of all houses an individual line, based on hourly consumption, using daily resolution dataset.](image)

For this specific dataset, three clusters were used, because only thirty lines were available to be clustered and using a higher amount of cluster (four), resulted in a constant change of the cluster pattern. The use of precisely three clusters gave a consistent result.

The table below shows each house, their total consumption, and which cluster they belong.

From the table, it can be observed that houses under cluster 1 have relatively lower electricity consumption compared to houses in cluster 2, with significantly higher electricity consumption. Houses in cluster 0, form the intermediate.
### Table 4.1: Cluster Category for Each House and its Total Consumption

<table>
<thead>
<tr>
<th>House ID</th>
<th>Cluster Number</th>
<th>Total Consumption (ln kWh)</th>
<th>House ID</th>
<th>Cluster Number</th>
<th>Total Consumption (ln kWh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gw.29</td>
<td>0</td>
<td>247.107</td>
<td>gw.26</td>
<td>1</td>
<td>155.833</td>
</tr>
<tr>
<td>gw.13</td>
<td>0</td>
<td>212.896</td>
<td>gw.14</td>
<td>1</td>
<td>169.764</td>
</tr>
<tr>
<td>gw.12</td>
<td>0</td>
<td>251.704</td>
<td>gw.27</td>
<td>1</td>
<td>148.230</td>
</tr>
<tr>
<td>gw.52</td>
<td>0</td>
<td>304.800</td>
<td>gw.53</td>
<td>1</td>
<td>159.076</td>
</tr>
<tr>
<td>gw.22</td>
<td>0</td>
<td>209.475</td>
<td>gw.31</td>
<td>1</td>
<td>108.812</td>
</tr>
<tr>
<td>gw.9</td>
<td>0</td>
<td>211.943</td>
<td>gw.10</td>
<td>1</td>
<td>114.305</td>
</tr>
<tr>
<td>gw.17</td>
<td>0</td>
<td>282.331</td>
<td>gw.55</td>
<td>1</td>
<td>174.800</td>
</tr>
<tr>
<td>gw.25</td>
<td>1</td>
<td>158.181</td>
<td>gw.37</td>
<td>2</td>
<td>362.907</td>
</tr>
<tr>
<td>gw.33</td>
<td>1</td>
<td>72.636</td>
<td>gw.8</td>
<td>2</td>
<td>348.948</td>
</tr>
<tr>
<td>gw.16</td>
<td>1</td>
<td>174.752</td>
<td>gw.1</td>
<td>2</td>
<td>457.105</td>
</tr>
<tr>
<td>gw.28</td>
<td>1</td>
<td>178.684</td>
<td>gw.39</td>
<td>2</td>
<td>341.434</td>
</tr>
<tr>
<td>gw.50</td>
<td>1</td>
<td>71.096</td>
<td>gw.0</td>
<td>2</td>
<td>309.829</td>
</tr>
<tr>
<td>gw.47</td>
<td>1</td>
<td>145.684</td>
<td>gw.18</td>
<td>2</td>
<td>386.773</td>
</tr>
<tr>
<td>gw.32</td>
<td>1</td>
<td>118.738</td>
<td>gw.36</td>
<td>2</td>
<td>424.158</td>
</tr>
<tr>
<td>gw.48</td>
<td>1</td>
<td>99.103</td>
<td>gw.15</td>
<td>2</td>
<td>364.748</td>
</tr>
</tbody>
</table>

### 4.1.1.1 Similarities between Houses

Determining the relative similarities between each house based on the k-means cluster was possible using the `adjust_rand_score()` function in Python, with identical clusters recorded as 1.0 while very dissimilar clusters tend towards 0.0. This determination is achieved by first calculating the Rand Index, (the number of pairs of data that either belongs in the same group or different groups divided by the total number of all the pairs of data). It is represented statistically as:

\[
ARI = \frac{RI - E|RI|}{Max(RI) - E|RI|}
\]

where ARI = Adjusted Rand Index, RI = Rand Index, E|RI| = Expected, Rand Index, and Max(RI) = Maximum Rand Index.

Using the above method, similarities between a few houses are displayed below;
Table 4.2: Similarities between few houses using Adjustment Rand Index method

<table>
<thead>
<tr>
<th>Houses</th>
<th>House_14</th>
<th>House_32</th>
<th>House_9</th>
<th>House_16</th>
<th>House_31</th>
</tr>
</thead>
<tbody>
<tr>
<td>House_52</td>
<td>0.7603</td>
<td>0.8792</td>
<td>0.9721</td>
<td>0.9186</td>
<td>0.9394</td>
</tr>
<tr>
<td>House_0</td>
<td>0.7348</td>
<td>0.8330</td>
<td>0.9277</td>
<td>0.8793</td>
<td>0.8909</td>
</tr>
<tr>
<td>House_25</td>
<td>0.9681</td>
<td>0.8681</td>
<td>0.9593</td>
<td>0.9071</td>
<td>0.9196</td>
</tr>
<tr>
<td>House_26</td>
<td>0.7565</td>
<td>0.8718</td>
<td>0.9689</td>
<td>0.9161</td>
<td>0.9247</td>
</tr>
<tr>
<td>House_50</td>
<td>0.7373</td>
<td>0.8470</td>
<td>0.9348</td>
<td>0.8851</td>
<td>0.8970</td>
</tr>
</tbody>
</table>

From the table, houses within the same cluster shown in table 4.1 are quite similar compared to houses in other clusters. The most dissimilar pair of houses noticed in this comparison are houses gw_33 and gw_28 with its relative similarity recorded at 0.6344.

4.1.2 Hourly Resolution

During the experimentation, several features were considered that developed different clustering results. With these features, all available dates for all houses were clustered in a line plot, to visualize their output. The x axis represents the hours of the day from 0 to 23, and the y axis represents the amount of power consumed in Watts(W).
4.1.2.1 All Houses

Figure 4.2: Clusters of all houses based on hourly consumption using the hourly resolution dataset.

The line plot illustrates clustering based on dates and shows the output as 4 clusters with similar consumption patterns. Each line represents each day of a house; all clustered relatively close to each other and no proper distinction for where each cluster begins and ends or which dates fall into what cluster. With the values from power consumed, it was possible to determine the 'Peak', 'Mid-Peak' and 'Off-Peak' periods of consumption.

Figure 4.3: Consumption chart showing Peak, Mid-Peak and Off-Peak periods for all dates in the dataset.
With the intensity going from the highest to the lowest point of each color section, \textit{8pm} is noticed to have the highest power consumption while \textit{1am} has the lowest power consumption in the above chart.

### 4.1.2.2 All Dates - Four Seasons

With the season’s features, the following results were obtained:

![Figure 4.4: Clusters of all houses in Spring](image)

*Figure 4.4: Clusters of all houses in Spring*

![Figure 4.5: Clusters of all houses in Summer](image)

*Figure 4.5: Clusters of all houses in Summer*
With the clustering results achieved, it is observed that Winter and Spring have the highest value of power consumed within the hour, as well as all the seasons, reaching close to 6,000kW between the hours of 3 pm and 12 am cumulatively, while Summer and Autumn have their consumption capped at under 5,000kW as the highest values for the seasons.
Here different peak times apply to the different seasons but, fall still, within the same range of 7 pm and 8 pm. The same can be said for off-peak periods which look to be between the range of 12 am to 4 am with the exception of Summer that has one of its lowest consumption at about 2 pm in the afternoon. The mid-peak period varies for all four seasons.
4.1.2.3 All Dates - Weekdays and Weekends

**Figure 4.9:** Clusters of all houses during weekdays

**Figure 4.10:** Clusters of all houses during weekends
Figure 4.11: Clusters of all houses on Sundays

Figure 4.12: Clusters for all houses on Mondays
Studying the line plot and the consumption chart for days like Sundays and Mondays, the expectation is to see much difference in consumption rate and pattern. However, not a lot is revealed as they both have similar consumption patterns with the only major difference being that more power is consumed during the week than on the weekends, with some consumption reaching above 5,500kW within the hour. Not many differences are seen in the peak to off-peak period as well. Looking further into individual houses, however, reveals obvious differences between them, and more coherent analysis can be observed.
4.1.2.4 Individual Houses

Going into clustering for individual houses, a more clear pattern is observed. Better distinction (from the group plots) can be made between each cluster within a house and analyzing this together with the consumption chart allows for the understanding of power consumption as well as the living pattern of the residents. Due to the volume of the data analyzed, only a few results are shared, due to the limitation on space. Randomly selecting four houses, the line plots and consumption charts of houses, gw_39, gw_48, gw_15, and gw_33 are discussed below.

With individual houses plotted, a clear pattern highlighted is noticed, showing four most likely pattern the household experiences.

![Figure 4.14: Clusters over house gw_15 based on the hourly consumption using hourly resolution dataset.](image)

For house gw_15, it is quite obvious that an increase in daily consumption usually starts from between the hours of 5 am and 6 am, and the lowest consumption appears between the hours of midnight and 4 am. Asides from the consumption times, it is also clear that house gw_15 consumes much power on a daily basis as compared to other houses, due to its almost constant elevated state.
House gw_33 has a distinctively low daily consumption, as most of the cluster is quite close to 0kW. Even with that, two clusters show a group of high power consumption between the hours of 10 am and 3 pm as well as 6 pm to 10 pm.

House gw_39 shows a clear pattern of increased power consumption in the early mornings and a drop between the hours of 8 am and 12noon thereafter, another increase.
Chapter 4 Experimental Evaluation and Discussion

House gw_48 does not give any clear indication of elevated power consumption within certain periods of the day.

By viewing the consumption chart, it is much easier to understand the line plot. Take, for instance, house gw_15; the line plot shows increased power consumption between 6 am and 11 am. With the chart as a guide, the peak period is recorded between the hours of 8 am, and 10 am, with the maximum peak attained around 10 am, and this corresponds with the line plot. The off-peak period can also be observed around the same time, as shown in the line plot, which is between 12 am and 5 am. Take house gw_48 as another example; its consumption pattern is not visibly determined in the line plot as there is no distinct pattern of increase or decrease in consumption. This uncertainty is also reflected in the chart, as high power consumption is observed at varying times of the day.

Analyzing the individual cluster based on seasons, Summer and Winter for houses gw_15 and gw_48 will be discussed as well as comparison with seasons of other houses. For all four seasons, Summer has the least consumption for every house.

4.1.2.5 Winter and Summer for House gw_15

House gw_15 has one of the highest consumption rates in the dataset, and this is most noticed during the Winter period, very likely due to the use of electric heaters. With the comparison of this to its Summer consumption, a clear difference emerges in the amount
of power consumed for the periods, but not so much difference is noticed in its pattern of consumption as well as the peak, off-peak and mid-peak periods.

During Winter, power is consumed mostly in the evenings between 6 pm and, but its peak is noticed at about 10 am in the mornings. For the Summer, much power is consumed in the mornings between the hours of 8 am, and 10 am, and also in the evenings between 9 pm and 10 pm. The off-peak period is relatively the same, with a little adjustment, so also the mid-peak period.

**4.1.2.6 Comparing Winter Seasons for Houses gw_15 and gw_33**

House gw_33 consumes less power than gw_15 during the Winter and in its overall consumption.

Haven compared the consumption chart to the line plot; house gw_15 shows a clear interpretation of rising power consumption from a dormant state (off-peak period) between the hours of 12 am and 5 am, where it begins to increase till 10 am and drops to its mid-peak period.
Figure 4.19: Clusters over house gw_15 during Winter using the hourly resolution dataset.

Figure 4.20: Clusters over house gw_15 during Summer using the hourly resolution dataset.

This change is illustrated in the chart, which shows the off-peak period between 12 am and 4 am, and then part of the peak period is recorded between 9 am, and 10 am. The mid-peak period is observed between the hours of 12 noon and 5 pm in the evenings. From then on, power consumption increases again between the hours of 6 pm and 8 pm and eventually starts declining around 10 pm.

House gw_33, on the other hand, has one of the lowest consumption rates. Even with
such low rates, its Winter consumption is the highest at 27,869kWh all of its four seasons compared to its Summer consumption of 11,409kWh.

Its cluster pattern shows peak periods to be in the evenings while its off-peak periods are in the mornings. The consumption chart agrees with the clustered plot with regards to the peak periods which occurs between the hours of 5 pm and 9 pm, but the off-peak and mid-peak period seem scattered across time from midnight to 1 pm without any apparent pattern.
4.1.2.7 Winter and Summer for House gw_48

Another result illustrated is for the Winter and Summer season for house gw_48. Its Winter consumption shows a higher rate than its Summer consumption (39,475kWh, at Winter and 16,089kWh, at Summer). The chart shows its peak period between 2 pm to 3 pm in the afternoons and also within the hours of 9 pm, 4 am, and 6 am in the Winter. Its off-peak period is shown between 12 am to 2 am as well as 9 am to 10 am. Summer consumption shows a different consumption pattern with peak periods between 9 am and 10 am (which is the off-peak periods for the Winter). The off-peak period in the Summer is mostly noticed within the hours of 2 pm and 4 pm as well as from 10 pm to 11 pm.
Figure 4.24: Clusters over house gw_48 during Winter using the hourly resolution dataset.

Figure 4.25: Clusters over house gw_48 during Summer using the hourly resolution dataset.
4.1.2.8 Comparing the Winter Season for Houses gw_48 and gw_39

House gw_39 consumes more power than gw_48 in the Summer even though its Summer consumption is the least amongst the four seasons.

Figure 4.27: Clusters over house gw_39 during Summer using the hourly resolution dataset.

House gw_39 has one of the highest consumption rates within the dataset. Comparing its Summer consumption with that of its Winter, Summer is recorded at about 49,027kWh while its Winter consumption is at 141,146kWh. Although a generally high consumption for both seasons, there is a significant drop in the Summer season by about 92,000kWh.
Figure 4.28: Consumption chart showing Peak, Mid-Peak and Off-Peak periods of Summer for Houses gw_39 and gw_48

Analysis of the line plot during the Summer shows that power is usually consumed in the early hours of the morning and then declines between 9 am and mid-day only to increase again from about 3 pm till late evening. The consumption chart shows that there are peak periods in the Summer at around 4 am and 6 am in the mornings and also by 2 pm and 3 pm in the afternoons. The off-peak periods are from 12 am to 2 am as well as 9 am, and 10 am, and this corresponds with the cluster generated by k-means.

Houses gw_48 has a low power consumption rate with its Summer consumption totaled at 16,089kWh while its Winter consumption is capped at 39,475kWh. Like most of the other houses, the Summer season has the lowest consumption rate. Analysis of the cluster graph shows, several periods stand out, but overall, it is a bit difficult to determine a very distinct pattern in consumption. With the consumption chart, however, 6 am, as well as 9 am to 10 am are the peak periods while its off-peak period is mainly between 10 pm and 11 pm. The mid-peak period is noticed between the hours of 1 pm and 3 pm, which can also be seen in the clustered line plot.

4.1.2.9 Mondays and Sundays for House gw_33

Taking Mondays and Sundays as the example for the feature ‘days of the week’, house gw_33’s Mondays show an elevation in power consumption during mid-afternoons towards evenings and all other period is mostly clustered towards 0kW showing that electricity is usually not consumed during the mornings and early afternoons but towards the evenings. Whereas, for Sundays, although similar to Mondays with cluster close to 0kW, a slight elevation is observed around mid-afternoon. This elevation suggests that more power is
consumed during the weekends than the weekdays, but generally, electricity is conserved when not in use for this residence.

For the consumption chart, Mondays tend to have their peak periods in the evenings (as suggested by the line plot) between the hours of 8 pm and 11 pm as well as 7 am in the mornings. Off-peak is registered in the early hours of the mornings as well as between 8 am, and 9 am. Sundays, on the other hand, show a stack contrast from Mondays,
Figure 4.31: Consumption chart showing Peak, Mid-Peak and Off-Peak periods of ‘Mondays’ and ‘Sundays’ for Houses gw_33 with the peak period between 12noon and 3 pm as well as 7 pm in the evenings, while the off-peak period is noticed in the early hours of the mornings and mid-peak periods towards late evenings (10 pm, 11 pm, and 12 pm).

4.1.2.10 Mondays and Sundays for House gw_39

Figure 4.32: Clusters over house gw_39 for all Mondays using the hourly resolution dataset.
Figure 4.33: Clusters over house gw_39 for all Sundays using the hourly resolution dataset.

Figure 4.34: Consumption chart showing Peak, Mid-Peak and Off-Peak periods of 'Mondays' and 'Sundays' for Houses gw_39.

House gw_39 shows the day starting with approximately 2,500kW of power in the morning for both Mondays and Sundays, which suggests this as a routine practice of having certain appliances turned on overnight. Viewing the consumption chart, with guidance from the line plot, Mondays show peak periods to be between the hours of 2 pm and 5 pm as well as 6 am, and its off-peak period mostly occurring towards midday and midnight. For Sundays, however, the peak period is between 8 am, and 9 am as well as 3 pm to 5 pm. Off-peak hours are concentrated towards midnight to the early mornings.
From the analyses above, a clear distinction can be seen between the different houses, showing their consumption patterns as well as their routine peak, off-peak, and mid-peak periods. This difference can also be observed within each house (however small) based on the clustered features, highlighting that no two days or periods are the same, but similar patterns can be recognized within households.

Due to the large amount of output generated, only a few are displayed in this section. However, other generated results are attached as an appendix.

4.2 Comparing Algorithms

As earlier discussed, three types of unsupervised learning algorithms were tested to compare their respective output. The dataset used for this test is the daily resolution for individual houses using a scattered plot for easier analysis. After running the algorithms, the following outputs were derived.

Taking the output for houses gw_13 and gw_50 as examples, the difference between each algorithm is illustrated.

4.2.1 House gw_13

Using K-Means, the optimal number of clusters is determined by the elbow method to be four. DBSCAN algorithm determined the number of clusters as six and estimated 28 outliers present in the dataset, which falls beyond the boundaries of the clusters. For the hierarchical algorithm, determining the ideal number of clusters was done by studying the dendrogram and then deciding on a constant distance for all houses (in this case, 5000) which gave four clusters. The following outputs were generated for house gw_13 for the different algorithms.
Figure 4.35: K-Means scatter plot for House gw_13 with four cluster points

Figure 4.36: DBSCAN scatter plot for House gw_13 with six clusters
Figure 4.37: Dendrogram for House gw_13

Figure 4.38: Hierarchical scatter plot for House gw_13 based on dendrogram
4.2.2 House gw_50

Another example is the house gw_50. After applying all three algorithms, the output is shown in 4.39 and below are derived.

![Figure 4.39: K-Means scatter plot for House gw_50 with three cluster points](image1)

![Figure 4.40: DBSCAN scatter plot for House gw_50 with three clusters](image2)
Figure 4.41: Dendrogram for House gw_50

Figure 4.42: Hierarchical scatter plot for House gw_50 based on dendrogram
Using the K-Means algorithm, the ideal number of clusters was determined as three, and the dataset is grouped showing the centroid points. DBSCAN also decides on three clusters, same as K-Means, which are also similar in placements but it estimates that 38 outliers exist in the dataset, which falls beyond the boundaries of the clusters. Hierarchical, on the other hand, decides one cluster since no distance of 5000 is attained.

These results clearly show the difference between the three algorithms, their weaknesses, their strengths, and why, ultimately, the k-means algorithm was chosen for the thesis.

4.3 Algorithm Discussion

Following the analysis of all three algorithms, the K-Means algorithm is regarded as the algorithm with the best output for the dataset. This primarily is because, it considers all data points and gives a more structured and reliable output which can help understand the dataset better and thus, its analysis.

The only distinct disadvantage to using the K-Means algorithm in the thesis is deciding the optimal number of clusters based on the features, and this is specific to the hourly resolution. Since specific dates within each house are grouped based on features, determining the optimal number of clusters was a bit challenging as it is usually determined in relation to the house and not to the date. That means each house is taken as a unit (notwithstanding the multiple days and hours entry) and the elbow method determines the cluster for that house irrespective of features. Therefore the number of clusters was decided as the maximum number of clusters generated throughout the search (in this case, four). However, for the comparison between algorithms, The elbow method, which was calculated for each house, was applied since the dataset used was that of the daily resolution.

Another disadvantage of k-means is the fact that it clusters in a spherical manner, which means some data points which should fit better in a different cluster might get pushed to another group because it falls within its SS.

DBSCAN was one of the most favorite method used and a significant challenge to the k-means method, as it takes into consideration, the amount of data within an area and groups them based on their closeness, among other factors. The downside to this method, though, is the representation of some data as outliers. This setback can be limited by increasing the $\text{eps}$ or $\text{minPts}$ to accommodate for the inclusion of these outliers, but this has a significant impact because it dramatically reduces the number of clusters expected and still, not all data points are captured in each cluster. Take house gw_13 again as an example, increasing the $\text{eps}$ from 0.3 to 0.5, results in a decrease in clusters from six.
to three clusters and an increase only in the \textit{minPts} from 10 to 15, results in a cluster decreased to four and even in both cases, not all outliers were encompassed.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{DBSCAN_cluster}
\caption{DBSCAN cluster showing increased \textit{eps} to 0.5 and increased \textit{minPts} to 15 respectively.}
\end{figure}

Applying this method on an arbitrary dataset might be the best, but when dealing with actual data, no value is considered an outlier since they are all recorded by the SM. This means not all data is considered in the clustering process, and therefore, the output is not an accurate representation of the dataset.

The hierarchical clustering was the least productive algorithm, as its disadvantages significantly repressed its advantages. First, it was grossly inadequate for the dataset (due to its size), both in time complexity (run time), and in the display on a dendrogram as well as the amount of memory required for processing. Secondly, the determination of the optimal number of clusters involves a physical examination of each house’s dendrogram to see the best fit. This process is not ideal for a large dataset and frankly removes the notion of machine learning from the entire process. The latter point, however, can be overcome by using different techniques like the \textit{Dynamic Cut Method}\cite{30} or \textit{Hubert’s Gamma Statistics}, among others. Unfortunately, these techniques were not tested due to time constraints. The final disadvantage of using this algorithm was observed in the display of the clusters in a scattered plot. Each group seemed to be clustered vertically even though some data points could be seen to not fit naturally into its cluster. Ultimately, results from this algorithm failed to show distinct consumption patterns and related data points within the dataset.

\section{4.4 Result Summary}

This thesis has produced several conclusions based on the analysis performed

- Shows three distinct profiles of houses based on their consumption 4.1,
• Shows consumption pattern and rate (on an hourly basis) of all the houses as a group using the various features extracted from the dataset.

• Shows the consumption pattern and rate for individual houses using the various features extracted from the dataset.

• Shows difference in pattern and rate between results derived from each feature, taking all houses as a group.

• Shows difference in pattern and rate between results derived from each feature based on individual houses.

• Clusters the dataset, using three different machine learning techniques and,

• Compares these three techniques to buttress the decision for analysis using the k-means algorithm.
Conclusion and Future Works

Daily, the importance of data information and interpretation is increasing as every person, from the producers to the consumers, the distributors and even the government understands the need to analyze these data whether for personal reasons, for financial or economic reasons, or on a broader perspective, for tackling climate change.

In this thesis, two areas of focus were analyzed, with each having sub-layers. One was to cluster the given dataset using the decided k-means clustering by subdividing the dataset into various features, while the other was to compare k-means with two other algorithms (DBSCAN and Hierarchical clustering) using the daily resolution of the dataset to determine which algorithm gave the preferable result.

The application of k-means to the various features yielded precise and reliable results showing consumption patterns as well as peak, mid-peak and off-peak periods for all houses taken as a group and each house as an individual unit. The comparison between the three algorithms gave experiment-based evaluations, showing the difference in each of the algorithm’s cluster style and ultimately deciding on the most appropriate for the dataset.

From the output derived, each house profile can distinctly show the total consumption per feature and the particular consumption periods on an hourly basis. This output allows for analysis on an hourly basis, thereby allowing residents(or whoever views the
results), to pinpoint times in which wastage occurs and, going forward, proffer solutions to curtail it. Analysis and interpretation of such results can also encourage residents to utilize electricity during off-peak periods, which could be less expensive than the peak periods, thereby reducing the overall amount incurred in electricity bills. This benefit can also help in the reduction of CO$_2$ emission, as the reduction in energy consumption during peak periods, means avoidance of the use of power plants, which encourages pollution.

5.1 Future Works

Further improvement can be made to varying parts of this thesis, beginning with the features.

The inclusion of further details like; the number of residents in a house, the occupational status of the residents and the age demographics and the types of appliances in the household, can help to understand better why specific houses consume more electricity than others and also help to detect anomalies if there exists any.

Furthermore, works regarding load disaggregation can be performed in order to study and identify individual appliance’s consumption patterns, providing residents with a clear indication of its energy usage, and how it can be reduced.

Another aspect would be to increase the amount of data used for the experimentation and also diversify the origin of the dataset, taking a large amount of SM readings from very different regions in order to get better comparable outputs.

Finally, the algorithms can be improved, especially the hierarchical clustering, by implementing other linkage parameters and distances to observe if its outcome is more comparable than that of Ward and Euclidean. Also, the testing of other unsupervised learning algorithms is encouraged to see if there is any significant improvement in the accuracy of the output.
List of Figures

2.1 An Electro-Mechanical Meter and a Smart Meter. ........................................ 7  
2.2 Steps in Cluster Analysis ................................................................. 10  
2.3 Algorithm for k selection ................................................................. 11  
2.4 An Example of an Un-clustered and Clustered Dataset ................................. 12  
2.5 Example of SSEs’ plotted, showing the optimal k as 2 ................................. 14  
2.6 Example of a Hierarchical Dendrogram ................................................... 16  
2.7 Number of Publication indexed by WoS [15] ........................................... 20  
3.1 Hourly-based electricity consumption for House gw_13 showing all 426 days; each day represented as a line. ................................................................. 26  
3.2 All imported Libraries ........................................................................... 27  
4.1 Clusters of all houses an individual line, based on hourly consumption, using daily resolution dataset ............................................................. 34  
4.2 Clusters of all houses based on hourly consumption using the hourly resolution dataset ............................................................. 37  
4.3 Consumption chart showing Peak, Mid-Peak and Off-Peak periods for all dates in the dataset ................................................................. 37  
4.4 Clusters of all houses in Spring .................................................................. 38  
4.5 Clusters of all houses in Summer ................................................................ 38  
4.6 Clusters of all houses in Autumn ............................................................... 39  
4.7 Clusters of all houses in Winter ............................................................... 39  
4.8 Consumption chart showing Peak, Mid-Peak and Off-Peak periods for all four seasons ....................................................................................... 40  
4.9 Clusters of all houses during weekdays ....................................................... 41  
4.10 Clusters of all houses during weekends ..................................................... 41  
4.11 Clusters of all houses on Sundays ............................................................ 42  
4.12 Clusters for all houses on Mondays ........................................................... 42  
4.13 Consumption chart showing Peak, Mid-Peak and Off-Peak periods for Weekdays, Weekends, Mondays and Sundays ............................................ 43  
4.14 Clusters over house gw_15 based on the hourly consumption using hourly resolution dataset ................................................................. 44  
4.15 Clusters over house gw_33 based on the hourly consumption using hourly resolution dataset ................................................................. 45  
4.16 Clusters over house gw_39 based on the hourly consumption using hourly resolution dataset ................................................................. 45  
4.17 Clusters over house gw_48 based on the hourly consumption using hourly resolution dataset ................................................................. 46  
4.18 Consumption chart showing Peak, Mid-Peak and Off-Peak periods for all four houses using all dates ............................................................... 47
4.19 Clusters over house gw_15 during Winter using the hourly resolution dataset. ................................................. 48
4.20 Clusters over house gw_15 during Summer using the hourly resolution dataset. ................................................. 48
4.21 Consumption chart showing Peak, Mid-Peak and Off-Peak periods of Winter and Summer for house gw_15 ................. 49
4.22 Clusters over house gw_33 during Winter using the hourly resolution dataset. ................................................. 49
4.23 Consumption chart showing Peak, Mid-Peak and Off-Peak periods of Winter for Houses gw_33 and gw_15 ................. 50
4.24 Clusters over house gw_48 during Winter using the hourly resolution dataset. ................................................. 51
4.25 Clusters over house gw_48 during Summer using the hourly resolution dataset. ................................................. 51
4.26 Consumption chart showing Peak, Mid-Peak and Off-Peak periods of Winter and Summer for House gw_48 ................. 52
4.27 Clusters over house gw_39 during Summer using the hourly resolution dataset. ................................................. 52
4.28 Consumption chart showing Peak, Mid-Peak and Off-Peak periods of Summer for Houses gw_39 and gw_48 ................. 53
4.29 Clusters over house gw_33 for all Mondays using the hourly resolution dataset. ................................................. 54
4.30 Clusters over house gw_33 for all Sundays using the hourly resolution dataset. ................................................. 54
4.31 Consumption chart showing Peak, Mid-Peak and Off-Peak periods of 'Mondays' and 'Sundays' for Houses gw_33 ................. 55
4.32 Clusters over house gw_39 for all Mondays using the hourly resolution dataset. ................................................. 55
4.33 Clusters over house gw_39 for all Sundays using the hourly resolution dataset. ................................................. 56
4.34 Consumption chart showing Peak, Mid-Peak and Off-Peak periods of 'Mondays' and 'Sundays' for Houses gw_39 ................. 56
4.35 K-Means scatter plot for House gw_13 with four cluster points ................................................................. 58
4.36 DBSCAN scatter plot for House gw_13 with six clusters ................................................................. 58
4.37 Dendrogram for House gw_13 ................................................................. 59
4.38 Hierarchical scatter plot for House gw_13 based on dendrogram ................................................................. 59
4.39 K-Means scatter plot for House gw_50 with three cluster points ................................................................. 60
4.40 DBSCAN scatter plot for House gw_50 with three clusters ................................................................. 60
4.41 Dendrogram for House gw_50 ................................................................. 61
4.42 Hierarchical scatter plot for House gw_50 based on dendrogram ................................................................. 61
4.43 DBSCAN cluster showing increased eps to 0.5 and increased minPts to 15 respectively. ................................................................. 63

A.1 Clusters of all houses during the holidays ................................................................. 74
A.2 Clusters of all houses on Tuesdays ................................................................. 74
A.3 Clusters of all houses on Wednesdays ................................................................. 75
A.4 Clusters of all houses on Saturdays ................................................................. 75
A.5 Clusters over house gw_14 during holidays using the hourly resolution dataset. ........................................... 76
A.6 Clusters over house gw_14 for all Tuesdays using the hourly resolution dataset. ........................................... 76
A.7 Clusters over house gw_14 for all Wednesdays using the hourly resolution dataset. ......................................... 77
A.8 Clusters over house gw_14 for all Saturdays using the hourly resolution dataset. ........................................... 77
A.9 Clusters over house gw_9 during holidays using the hourly resolution dataset. ........................................... 78
A.10 Clusters over house gw_9 for all Tuesdays using the hourly resolution dataset. ........................................... 78
A.11 Clusters over house gw_9 for all Wednesdays using the hourly resolution dataset. ......................................... 79
A.12 Clusters over house gw_9 for all Saturdays using the hourly resolution dataset. ........................................... 79
A.13 Clusters over house gw_1 during holidays using the hourly resolution dataset. ........................................... 80
A.14 Clusters over house gw_1 for all Tuesdays using the hourly resolution dataset. ........................................... 80
A.15 Clusters over house gw_1 for all Wednesdays using the hourly resolution dataset. ......................................... 81
A.16 Clusters over house gw_1 for all Saturdays using the hourly resolution dataset. ........................................... 81
A.17 K-Means scatter plot for house gw_9 with three cluster points .............. 82
A.18 K-Means scatter plot for house gw_14 with five cluster points ............... 83
A.19 DBSCAN scatter plot for gw_9 with two clusters and thirteen outliers ....... 83
A.20 DBSCAN scatter plot for gw_14 with four clusters and twenty outliers .... 84
A.21 Dendrogram for gw_9 ................................................................. 84
A.22 Hierarchical scatter plot for gw_9 based on dendrogram ....................... 85
A.23 Dendrogram for gw_14 ................................................................. 85
A.24 Hierarchical scatter plot for gw_14 based on dendrogram ....................... 86
List of Tables

2.1 Difference between the three selected algorithms .......................... 19
4.1 Cluster Category for Each House and its Total Consumption ............ 35
4.2 Similarities between few houses using Adjustment Rand Index method .. 36
Due to the volume of output generated, only a few results are displayed for ease of comprehension. In this section however, a few more outputs are displayed to show their clustering pattern. This section is divided into two, results generated for; all houses as a unit, and individual houses. Results relating to comparison of algorithms are also included.

For the unit clusters, only holidays, Tuesdays, Wednesdays and Saturdays will be displayed. For individual houses, information regarding only three randomly chosen houses; gw_14, gw_9, and gw_1, will be displayed, and the features covered include holidays, Tuesdays, Wednesdays and Saturdays.

For comparing algorithms, the results generated for houses gw_14 and gw_9, using k-means, hierarchical and DBSCAN algorithms will be displayed.

**A.1 Results for All Houses**

This section displays generated results for holidays, Tuesdays, Wednesdays and Saturdays of all houses as a unit, relative to 4.1.2.3
Figure A.1: Clusters of all houses during the holidays

Figure A.2: Clusters of all houses on Tuesdays
A.2 Results for Individual Houses

This section displays generated results for houses gw_14, gw_9, and gw_1 holidays, Tuesdays, Wednesdays and Saturdays of all houses as a unit, relative to 4.1.2.4
A.2.1 House gw_14

Figure A.5: Clusters over house gw_14 during holidays using the hourly resolution dataset.

Figure A.6: Clusters over house gw_14 for all Tuesdays using the hourly resolution dataset.
Figure A.7: Clusters over house gw_14 for all Wednesdays using the hourly resolution dataset.

Figure A.8: Clusters over house gw_14 for all Saturdays using the hourly resolution dataset.
A.2.2 House gw_9

**Figure A.9:** Clusters over house gw_9 during holidays using the hourly resolution dataset.

**Figure A.10:** Clusters over house gw_9 for all Tuesdays using the hourly resolution dataset.
Figure A.11: Clusters over house gw_9 for all Wednesdays using the hourly resolution dataset.

Figure A.12: Clusters over house gw_9 for all Saturdays using the hourly resolution dataset.
A.2.3 House gw_1

Figure A.13: Clusters over house gw_1 during holidays using the hourly resolution dataset.

Figure A.14: Clusters over house gw_1 for all Tuesdays using the hourly resolution dataset.
Figure A.15: Clusters over house gw_1 for all Wednesdays using the hourly resolution dataset.

Figure A.16: Clusters over house gw_1 for all Saturdays using the hourly resolution dataset.
A.3 Results on Compared Algorithms

This section displays generated results for k-means, hierarchical and DBSCAN algorithms, relative to 4.2

A.3.1 K-Means clustering

![K-Means scatter plot for house gw_9 with three cluster points](image)

**Figure A.17:** K-Means scatter plot for house gw_9 with three cluster points
A.3.2 DBSCAN

Figure A.18: K-Means scatter plot for house gw_14 with five cluster points

Figure A.19: DBSCAN scatter plot for gw_9 with two clusters and thirteen outliers
Figure A.20: DBSCAN scatter plot for gw_14 with four clusters and twenty outliers

A.3.3 Hierarchical clustering

Figure A.21: Dendrogram for gw_9
Figure A.22: Hierarchical scatter plot for gw_9 based on dendrogram

Figure A.23: Dendrogram for gw_14
Figure A.24: Hierarchical scatter plot for gw_14 based on dendrogram
Bibliography


