Maria Portugal Queiroga Nogueira

Bachelor in Computer Science Engineering

Clustering Smart Metering Data for Energy Efficiency

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Computer Science and Engineering

Adviser:	Joaquim Ferreira da Silva, Assistant Professor, Faculty of Sciences and Technology NOVA University of Lisbon
Co-adviser:	Luís Tiago Ferreira, Engineer, EDP Distribuição

Examination Committee

Chairperson:	Doutor João Carlos Gomes Moura Pires
Raporteurs:	Doutor Nuno Miguel Soares Datia
Members:	Doutor Joaquim Francisco Ferreira da Silva

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

September, 2019

Clustering Smart Metering Data for Energy Efficiency

This document was created using the (pdf)LaTeX processor, based in the “novathesis” template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Acknowledgements

I would like to acknowledge the support of EDP Distribuição for access to the data set and to the support in the development of this project.

I am very grateful to my parents and sister, without their support in many ways I wouldn't have come this far. To my boyfriend that was my 24/7 rescue and handled my not so good moments. And to my friends that believed in me even when I didn't. Thank you all for always being there.

This work would not be possible without the help of my adviser professor Joaquim Silva that was always present and ready to help. Especially for making the effort of moving all along to EDP offices so that we could meet and that I could simultaneously maintain the best performance on the trainee program. Thank you for all the guidance on Machine Learning and all the suggestions and corrections on the dissertation.

I couldn't forget the help of Luis Tiago Ferreira wich was the big driver of this project and my guidance in all the work. Thank you to all the colleagues I came across during my journey within EDP. All the patience to explain the energy concepts and to make me understand the company operation. The emotional advice was also undoubtedly helpful. I have learned a little from each one of you. A very special gratitude to Pedro Ferreira that believe in me since the beginning and gave me the chance of staying at EDP.

Working and studying at the same time was not easy at all, but life isn't meant to be easy. I am who I am today thanks to the enriching experiences that I went through and for all the people who made them possible.

Abstract

Nowadays, huge quantities of metering data of each consumer are being taken from the electric distribution network through smart meters and stored in databases. These metering data are consumption readings useful for many data analysis applications as, for example, fraud detection. However, a large number of possible analysis on this data are still unexplored. In this dissertation, we explore smart meters as a way of improving Energy Efficiency. To do that, we need to understand more about the way clients consume and find behavioural patterns on their consumption. The identification of all these profiles of consumption is essential since it will allow EDP Distribuição (EDPD) to know more about its types of clients, providing focused feedback and consumption advice.

Clustering algorithms are useful to understand the distribution of patterns in large data sets. By creating several groups/clusters, we will be able to understand the profile of a specific client by the characteristics of its cluster. It makes it possible to categorise a customer from its group behaviour rather than expecting that each customer as it own specific profile. This allows to save a lot of time analysing the types of consumers.

In this dissertation, we intend to analyse clients from several perspectives in order to capture different types of behaviours. For example, we may want to analyse the clients based on their absolute consumption values, in order to compare their scales or compare them from the consumption regularity point of view. So, we use the clustering algorithms with the appropriate features as a data mining approach to find structure in our data.

The results that we present highlight the strengths and weaknesses of each clustering algorithm and validate their applicability to the EDP Distribuição (EDPD) use case. So, this dissertation will bring added knowledge about clustering techniques and analysis over smart metering data.

Keywords: Smart Metering Data, Clustering Algorithms, Unsupervised Learning, Electricity Consumption Clustering, Data Analysis

Resumo

Atualmente são retirados da rede de distribuição elétrica uma grande quantidade de dados de medição de cada cliente. Estas dados são obtidos através de smart meters e armazenados em bases de dados. As medições são leituras de consumo úteis para muitas aplicações de análise de dados como por exemplo, a deteção de fraude. No entanto, outras possíveis vertentes de análise destes dados estão ainda por explorar. Nesta dissertação, exploramos a utilização de smart meters como uma maneira de melhorar a eficiência energética. Para isso, precisamos compreender melhor a maneira como os clientes consomem e encontrar padrões comportamentais nos seus consumos. A identificação de todos esses perfis de consumo é essencial, pois permitirá à EDP Distribuição (EDPD) conhecer mais sobre os seus tipos de clientes, fornecendo feedback focado e conselhos de consumo.

Os algoritmos de Clustering são úteis para compreender a distribuição de padrões em grandes conjuntos de dados. Ao criar vários grupos / clusters, conseguimos compreender o perfil de um cliente específico pelas características do seu cluster. O que torna possível categorizar um cliente a partir do comportamento do seu grupo, em vez de esperar que cada cliente individual tenha um perfil de consumo específico, o que permite uma grande poupança de tempo na análise dos tipos de consumidores.

Neste projeto, pretendemos agrupar clientes de várias perspetivas, a fim de capturar diferentes tipos de comportamentos. Por exemplo, podemos querer analisar os clientes com base nos seus valores absolutos de consumo, para comparar suas escalas, ou compará-los do ponto de vista da regularidade do consumo. Assim, vamos utilizar como uma abordagem de Data Mining algoritmos de Clustering, com as features mais apropriadas, de modo a encontrar estruturas nos nossos dados.

Os resultados que apresentamos destacam ainda os pontos fortes e fracos de cada algoritmo de clustering e validam a sua aplicabilidade ao caso de uso da EDPD. Portanto, este trabalho traz um conhecimento adicional sobre técnicas de clustering e análise sobre dados de smart meters.

Palavras-chave: Dados de Smart Metering, Algoritmos de Clustering, Aprendizagem Não Supervisionada, Clustering de Consumos Elétricos, Análise de Dados

Contents

List of Figures

Acronyms

CTS	Client Transformation Substation.
DBSCAN	Density-based spatial clustering of applications with noise.
DTC	Distribution Transformer Controller.
DTS	Distribution Transformation Substation.
EDPD	EDP Distribuição.
EM	Expectation Maximisation.
HV	High Voltage.
LV	Low Voltage.
MBC	Model-Based Clustering.
MV	Medium Voltage.
NLV	Normal Low Voltage.
OECD	Organisation for Economic Co-operation and Development.
PCA	Principal Component Analysis.
PLC	Power Line Communication.
SS	Substation.
TS	Transformation Substation.
WCSS	Within-cluster sum of squares.

Chapter 1

Introduction

In the following chapter, the problem that motivates the study developed in this thesis and its framing in the real case of EDP Distribuição (EDPD) is explained.

1.1 Motivation

EDPD receives on a daily basis a large volume of data of its multiple network assets, explained in 2.2. They have data sets with historical measurements useful to several use cases are stored. This data is already being processed with algorithms which are essential to improve business processes. Fraud detection and prediction of consumption are some examples that allow better management of the electric grid's clients and resources. However, the usage of this data is not even halfway of reaching its full potential, and there is still much ignorance about the data itself and its faults. New methods of data analysis are developed every day to fulfil the business needs that go deeper in this data sets.

This project has two main motivations: one under the enterprise point of view and other under the scientific/technology point of view.

The first motivation is divided into the benefits to the clients and the benefits to the company of managing energy more efficiently. EDPD works with every single consumer of energy in Portugal regardless of traders. Therefore, it must encourage consumers to take efficient consumption measures for more sustainable energy use. EDPD intends to create a personal space within its website that allows the customer to have a greater perception of their consumption based on smart metering data. By giving more knowledge to the clients about their energy consumption, EDPD empowers them to manage their energy usage and to make smarter choices. More than just showing dashboards about its consumption, it is also interesting to give tailored Energy Efficiency Tips based on each type of client. This is done in order to make customers feel that they are not just receiving general tips, where half do not even apply to their type of consumption, but focused advice with which they identify. Finally, there is also the goal of creating a consumption comparison that can bring a component of *gamification* to the website. The goal is to encourage customers to be more efficient by comparing their consumption with those of their similar customers' group thus increasing the competitive spirit of reducing consumption. By doing this, they can reduce their consumption values allowing them to have smaller bills and avoiding excessive and unnecessary consumption. For EDPD, this approach will be useful for its civic fulfilment as national energy distributor. Moreover, it will be useful to get to know the users of the grid by for example understanding their distribution of consumption by location and the type of consumption to better management of the resources on the electricity grid.

On the other hand, we have the scientific motivation since this project allow us to build a study on unsupervised clustering of time series obtained from smart meters. In this project, we had the opportunity to explore up-to-date data with still an untapped potential.

1.2 Problem Definition

In this project, we intend to explore the data of smart meters in a not yet explored area: the energy efficiency. To give tailored advice on this, we need to know more about the clients' consumption details. However, there are more than 2 million clients with smart meters and growing every day, which makes it impossible to analyse them one by one. So, in order to avoid looking at each client individually expecting to find a specific profile, we will need to find a way to identify patterns of consumption and group them. First of all, it is necessary to do a more refined analysis of all consumers in order to identify how do they behave and what are the best features to group them. It is easy to create preconceptions about electrical consumption and customer types, but the question is: "Are these theories true?". Hence the need to analyse customers in the best possible way so that we can find out as much information as possible about customers. Thus, clustering results with non-biased features may allow to confirm or nullify the theories we have already conceived. After this first problem of getting to know more about the consumers and their characteristics, we will also need to group them so that we can produce a proper comparison of clients with their peers.

Thus, we are trying to answer two mains questions: What are the consumers characteristics? How can we group them?

1.3 Contributions

The expected contributions of this dissertation are:

- Analysis of smart metering data and its potential. A study over clustering smart metering data. Analysis of several approaches and comparison of their performances to understand which one provides the best solution.
- Identification of clients' consumption types. Creation of groups of clients under different perspectives to deliver focused energy efficiency tips and comparison of consumption dashboards to the clients in the EDPD website.

1.4 Structure of the Thesis

This document is structured with chapters as follows:

- Chapter 1 (Introduction) - This chapter describes the motivation to the development of this dissertation, what will be the contributions achieved by it and the structure of the document in order to guide its reading.
- Chapter 2 (Preliminary Concepts) - Definition of some preliminary concepts about energy to help the understanding of the overall thesis.
- Chapter 3 (Background and State of the Art) - In this chapter, an analysis is made over the clustering algorithms already used in similar studies, and some metrics that can be used to evaluate their performance.
- Chapter 4 (Methodologies for Solution) - Definition and explanation of the methodologies to use and that we identify as appropriate for solving our problem.
- Chapter 5 (Implementation and Discussion) - Describes all the implementations and work done as well as the analysis of results and their discussion.
- Chapter 6 (Conclusion) - In this chapter, we present the final conclusions and the guidelines for future work.

Chapter 2

Preliminary Concepts

The development of this project requires an initial understanding of the basic concepts of electricity and the components that are part of the Electricity Network. EDP is present in the three main activities of the electricity sector: generation, distribution and supply. Techniques like Machine Learning can be applied to the data produced by all these three stages by making use of its large quantities of metering data. It can also produce use cases that can be useful to the three different sectors. In this project, we will focus on the data from the distribution network.

2.1 Electricity Sector

Throughout history, the electricity sector experienced fundamental changes driven by technological, regulative and social transformations. Portugal faced one of the main changes related to the liberalisation of the market. In 2006 it changed from a regulated market to a free market that allows consumers to choose their electricity suppliers and where each supplier freely determines energy prices. EDP is present in the three main activities of the electricity sector: generation, distribution and supply. In the generation and supply, it works as a free market. However, the distribution sector is still regulated by the Regulatory Authority for Energy Services (ERSE). It implies that EDPD is a regulated company responsible for connecting all clients to the electricity grid and for providing consumption readings to suppliers regardless of the company. EDPD is the one responsible for all the data gathered from the Electricity Distribution Network. So it has access to all consumption data but cannot access the client's information since it belongs to the supplier, which in EDP group corresponds to EDP Comercial.

2.2 Electricity Grid

EDP Distribution is responsible for managing all the operations in the electricity grid and for ensuring that it works properly. The primary operations in the grid are the transport and reduction of energy voltage. The produced energy is delivered in High Voltage (HV), to the supply points that can be Medium Voltage (MV) or Low Voltage (LV). The transport is made through Electricity Distribution Networks which are constituted by cables and lines of all three voltages. Besides cables and lines, there are also assets of the grid: Substation (SS) and Transformation Substation (TS). The first one is responsible for transforming the voltage value that came as HV and the second for converting the network voltage from MV to LV and connect the network to consumer facilities. The TSs are divided in Client Transformation Substations (CTS), which are usually a group of clients, for example, a neighbourhood, and Distribution Transformation Substations (DTS), like industries that receive electricity in HV and MV. The first type of clients corresponds to the low voltage ones that are the great majority of clients connected to the grid. While the second type of clients exists in smaller quantities (approximately 60 000) than the LV ones (approximately 6 000 000), they still are very much relevant since they consume almost 50% of distributed energy. Nowadays, all these assets, from the SS and TS to the consumer facilities, have metering equipment that is monitoring everything that happens in the grid. The first measure the electricity that pass through them while the second measure the consumption of each client.

2.2.1 Smart Grids

EDPD is focused on transforming Electricity Distribution Networks in Smart Grids that meet future challenges arising from the so-called three D's: Digitization, Decarbonization and Decentralization. One of the reasons for Smart grids is the expected growth of global energy demand. Even though the expected decrease in the countries of the Organisation for Economic Co-operation and Development (OECD) it is foreseen a considerable growth in non OECD, pushed by China and India since there are still many people there in a situation of energy poverty [2]. Moreover, to tackle climate change, it is necessary to increase electrification and decarbonization. Besides this increasing, the main reason associated with smart grids is a shift in the paradigm, with more decentralized energy production, bi-directional energy transits, new types of consumption and more demanding consumers. These new challenges force the EDPD network to change in order to keep up with these changes.

Smart meters aim to promote sustainability and energy efficiency with technologies and initiatives that increase the predictability of network operation through real-time information transmission, data management, analysis and control. The traditional grids need to become smart by adding information and communication technologies [5]. The idea is that the grid is becoming a superposition of different layers more than just the physical components, Figure 2.1. Distribution networks need to be actively managed and monitored to adjust to the changing conditions. The installation of several sensors allows to measure and control the state of the grid in real-time, preventing interruptions of supply. They make it possible to focus the grid operation on pre-emptive rather than reactive actions.

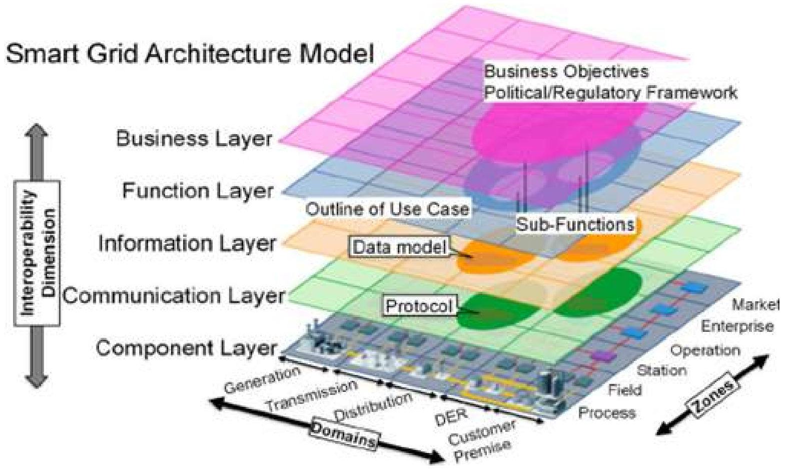


Figure 2.1: European smart grid architecture model framework [16].

2.2.2 Smart Meters

The smart meters are one of the main components of a smart grid [6]. Traditional meters and smart meters look similar but work in very different ways, Figure 2.2. In the past, the distributors could not see further than what happened at the substation level. With the installation of smart meters, it is possible to extend the smart grid until the end customer and to the low voltage. Smart meters provide daily readings and load diagram, wich are readings of every fifteen minutes. This measurements allow knowing more about the consumers and create tailored tariffs and products. Furthermore, it makes it possible to have a broader vision of what is happening on the grid. Though, this is a two-way communication since it also allows to remotely performs several operations:

- Change of tariff - the supplier charges the consumption depending on the chosen tariff option. The customers can be charged equal at any time or have different prices depending on the tariff periods of the day.

- Power changes - when there are changes in the contracted power it is necessary to update this information in the meter.

- Power on/off - enablement and disablement of supply. It can happen at customer's request or may be caused by payment failures or fraud.

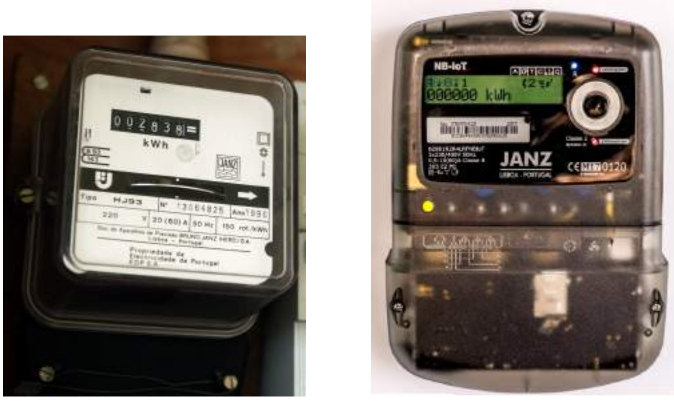


Figure 2.2: Traditional and Smart Electricity Meters.

Moreover, smart meters will allow customers to actively participate in the management of its consumption. They will be able to know their consumption almost in real-time, which allows them to budget better and to reduce their consumption.

The installation of smart meters by EDPD is part of the project InovGrid which started in 2007 with a pilot in Évora, Figure 2.3. This pilot project was the proof of concept and validation of the technology necessary to continue to the next phases. Then it was extended to six new locations between 2013 and 2015, and in 2016 it was defined that was time to start a progressive deployment and expand the installation of the smart meters to the whole country. The expansion is part of the road map 2016-2023, which implies that the works of installation of the smart meters are still being done and that more are installed every day.

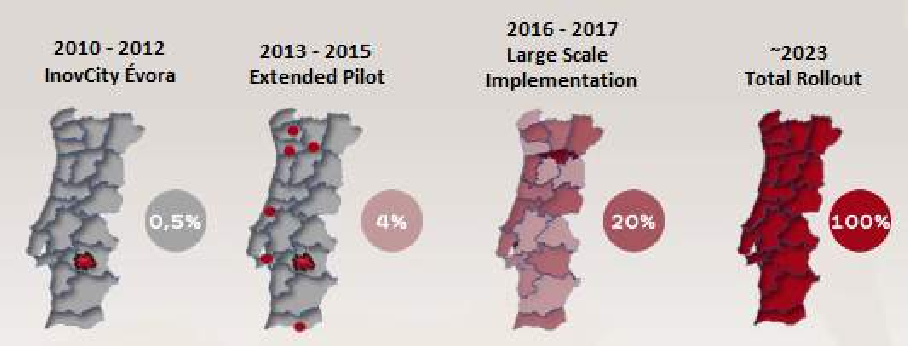


Figure 2.3: InovGrid project development and evolution of smart meters implementation in Portugal [33].

2.2.2.1 Data Communication Problems

EDPD already has smart meters installed in almost two million clients and emitting daily reading. However, not all of them already send load diagrams, only about four hundred thousand. It happens because of how the smart meters communicate the information. Smart meters communicate in two main ways via GPRS technology or TCP / IP. The first only applies to customers in very remote sites, where communications of another type are hampered, or small producers and other larger customers who need to send more reliable readings. In the second, the smart meters use Power Line Communication (PLC) to communicate through a Distribution Transformer Controller (DTC) placed in the nearest TS. There are DTCs connected to around 400 smart meters, in the most densely populated areas, and in those cases, it is impossible to get load diagrams to all smart meters or the DTC will be overloaded. Different manufacturers install the smart meters and other network components as per tender. Differences in component manufacturers affect data arrival performance.

Moreover, some operational problems affect the quality of the data arriving from the smart meters. One of the issues happens when there is a substitution of a meter. A meter can be changed for several reasons: if an equipment anomaly/breakdown is detected; if a change of supply from three-phase to single-phase occurs or if the counter reaches its limit. When this happens we face an abrupt change in the values of the totalizators since the new meter brings the counter with zero or is already an used meter containing measures. Another problem is the direct connections to the grid when, for example, a power failure happens, and the first objective is to get energy back. These are exceptions, but sometimes the client can have a smart meter, and the energy is not passing through it, so it measures zero consumption. Those are things that, in the past, would not bring significant problems, but now affects data quality.

Finally, the faults in the data or strange values of consumption could imply situations of fraud, i.e. people using energy from the network without paying. Those situations are already being monitored by the fraud detection department that also uses the data sets of smart meters' readings to do data analysis and find these cases.

2.3 Energy Efficiency

Environmental concerns are taken into account in everything EDPD does. More than ensuring sustainable energy distribution, it has a very active role in the improvement of energy efficiency in end-use, promoting behavioural and technological changes. Energy efficiency is a crucial practice for reducing global warming greenhouse gas emissions and contributing to the preservation of the environment. The work of this dissertation is included in the measures implemented within the company to promote efficient energy usage.

Energy efficiency is the rational and responsible use of energy avoiding lost: using less to perform the same task without compromising the comfort of consumers. EDPD intends to teach the consumers good energy efficiency practices that allows them to reduce their energy bill. There are three main ways to achieve energy efficiency (Figure 2.4): load reduction, load shifting and peak shaving. The first one is a general decreasing of consumption over the time which means that a consumer reduces his size/magnitude of consumption. The second one is when a client relocates his consumption in peak to the other periods allowing a better distribution of the consumption, for example, by turning on the washing machines at off-peak hours. Finally the peak shaving is when a client reduces his consumption on peak by maintaining the same activities but with a better management of energy.

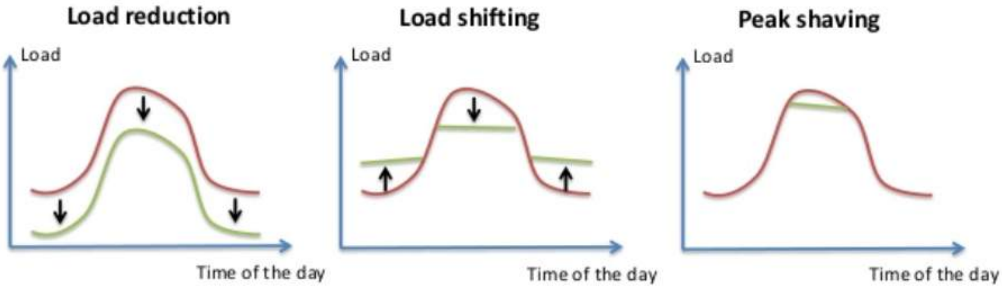


Figure 2.4: Three main ways of achieving energy efficiency [20].

The energy efficiency measures are divided into two main aspects, the technological one, for example, using more efficient equipment, and in a behavioural aspect like leaving lights on in empty rooms. The technological is hard to advise since the clients already have their equipment, and it will be expensive and a waste to change all the appliances still working. In this situation, EDPD can only advise being careful when buying new ones by looking at the energy label. The energy label is the same in all of the state state members of the UE-27 and allows to identify the class/level of energy efficiency of an equipment, Figure 2.5. Behavioral measures are part of the need to educate people to use energy more efficiently. EDPD intends to encourage its customers to adopt good habits of consumption by giving focused tips and comparisons of consumption. "Efficiency is not a fad, it is a way of being"[10].

There several simple measures that we can include in our daily routine to reduce unnecessary consumption [10]:

1. Make the most of sunlight. In winter, take advantage of the sun to warm the house by opening the blinds and curtains. On the contrary, in summer avoid direct sunlight during the day and promote natural ventilation at night by opening the windows on opposite sides of the house;

2. Instead of using the elevator, choose, whenever possible, the stairs;

3. Turn off the lamps when not in use;

4. Do not leave the equipment in standby;

5. Avoid turning on the climate in areas of the house that are not being used;

6. When you have climate control equipment turned on, keep the doors and windows closed;

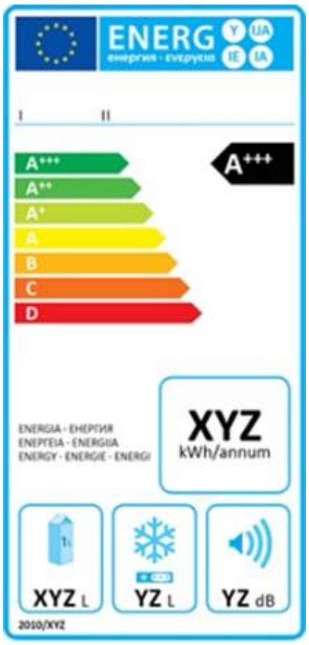


Figure 2.5: EU energy label to rate appliances in terms of energy efficiency [10].

7. Reduce the number of times you open the fridge door and optimize the opening time;

8. Let the food cool before putting it in the fridge;

9. Use the microwave preferably for small and easily prepared meals;

10. Use the sun and wind to dry clothes whenever possible;

11. Avoid printing documents.

More than this general tips the good habits of consumption could be divided by the type of use 2.1. To know more, you could find it on [12].

Table 2.1: Distribution of spending on energy at home, by type of use [1].

Type of Use	Value%
Cooking	40%
Water heating	27%
Electrical Equipment	15%
Room Heating	11%
Lighting	6%
Room Cooling	1%

Chapter 3

Background and State of the Art

This chapter describes the work made by several researchers in similar studies and several techniques already used in this type of projects.

3.1 Clustering Smart Metering Data

Tureczek et al., 2018 [32], study the clustering of smart metering data by applying several techniques of feature extraction. McLoughlin et al., 2015 [26], focuses on the clustering of load profiles. Most of the work we have found in this area uses data granularity of 15-minute interval, mainly due to the level of operational development of smart meters. Kwac et al., 2013 [22], proposes to segment customers based on their lifestyle by analysing 15-minute interval energy consumption data. Gajowniczek et al., 2015 [15], uses hierarchical clustering as a data mining technique to detect household characteristics and to understand the correlation of the time of usage of home appliances. Figueiredo, et al., 2005 [14], focuses on the characterisation of electricity consumers by combining supervised and unsupervised learning. These and other researchers use different clustering algorithms, but none of them is close to the specific topic contained in the smart meter's data of this dissertation. Despite that, these approaches were taken into account as they may include parts or details that can be useful to enrich the analysis we have developed in this dissertation. Beckel et al., 2014 [7], studies an approach that enables tailored energy efficiency advice for private households. However, it is based on supervised learning and socio-economic characteristics of each client wich EDPD don't have.

Clustering is useful when we want to describe and classify data, but the classes and their cardinality aren't previously known (Unsupervised Learning). Clustering Algorithms group sets of objects such that similar objects belong to the same cluster and are as similar as possible, while members of different clusters are as dissimilar as possible

[21]. Clustering is useful to extract knowledge of the data and to understand its structure and its relations. The term refers to the task of finding clusters in a data set and not to the algorithm itself since there are many possible Clustering techniques. Clustering is especially significant when we are dealing with huge quantities of data since it's almost impossible to analyse the data one by one to understand it. In these cases, clustering is truly useful since it organises unlabelled data in groups such that analysing and characterising each cluster by looking at a representative data point is enough, and all other data points inherit the characteristics of its group. Sometimes clustering is used as a pre-processing technique before other data analysis such as forecasting.

In this state of the art and in the whole dissertation we mainly focus on electricity consumption clustering since it is our real use case.

In the next Section, we do an analysis over multiple data representation techniques and clustering algorithms already used by other scientists in similar studies.

3.2 Data Representation

One of the main steps in clustering is the representation method and the choice of features which affects the efficiency and accuracy of the solution [13]. This method is a way of transforming the raw data set into a vector of another dimensionality by feature extraction or dimensionality reduction techniques. It is helpful to reduce the complexity and memory requirements of the data in order to speed-up the clustering process. Moreover, this avoids the loss of meaning during the analysis when we are dealing with several dimensions, the so-called *curse of dimensionality* [21]. For example, in the case of Euclidean distance if we use many coordinates the final result may end up having no significant differences between the different samples.

3.2.1 Features

The feature extraction is a way of choosing the best attributes to represent the consumption data to the clustering. The choice of features is usually application dependent [28], affecting the way the clusters are created, so we should use metrics that detect discriminant singularities of the groups. This means that we can choose several feature sets depending on what we want to detect and analyze in each of the project phases. We can extract specific features of the readings, instead of using the raw consumption data, allowing a dimensionality reduction and less sensibility to outliers. Moreover, using a smaller data set and less features makes it possible to use more complex and sophisticated clustering algorithms with lower computational efficiency.

The quality of the choice of features to use always need to be assessed by evaluating the results of the clustering algorithm where they are used. First, we analyse some statistical measures that allow to represent the global nature of consumption data:

- **Mean** - Mean is an important feature in most of the applications. For example, the mean of the consumption of electricity of an asset throughout the time will help to group assets with similar load magnitude. This measure is also used indirectly in other more complex metrics. The absolute sum of consumption ends up reflecting the same as the mean for equal time intervals.

$$\mu = \frac{1}{n}\sum_{i=1}^{i=n} x_i \tag{3.1}$$

- **Standard Deviation** - The standard deviation reflects the square root of the variance, which is average of the squared deviation of individual values to the mean. It is a measure of data dispersion around the mean. In our case it may inform how variant is the client consumption, for example.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(x_i - \mu)^2} \tag{3.2}$$

- **Coefficient of variation** - The coefficient of variation is also a measure of dispersion but standardised since it measures relative variability by dividing the standard deviation by the mean. It turns out to be better than the variance or the standard deviation for absolute values as these take very much into account the consumer's scale since they do not divide by the mean.

$$Cv(X) = \frac{\sigma}{\mu} \tag{3.3}$$

- **Normalised Skewness** - It is a measure of symmetry that evaluates the distribution of the values of a variable concerning the equilibrium of its concentration over the range of these values. If the result of this metric is positive, this means that there is greater weight in the distribution of values higher than the average. If it is negative, there is greater weight in the values less than the average value. The information about the balancing of the consumption values may be important to distinguish them and organise groups. This Skewness is normalised by the cube of the standard deviation in order to be similarly important as the other used metrics without having a different scale/weight.

$$Skew(X) = \frac{\frac{1}{n}\sum_{i=1}^{i=n}(x_i - \mu)^3}{(\sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(x_i - \mu)^2})^3} \tag{3.4}$$

- **Normalised Kurtosis** - It measures of shape that characterises the flatness of the concentration of values. It allows us to see if the values are more or less concentrated around the mean. It may be useful for detecting outliers that in this case corresponds to consumption peaks or breaks. Like Skewness this metric is also normalised.

$$Kurt(X) = \frac{\frac{1}{n}\sum_{i=1}^{i=n}(x_i - \mu)^4}{(\sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(x_i - \mu)^2})^4} \tag{3.5}$$

The values of consumption that we can get naturally from the data seem quite informative, and so we can probably use them as the data for clustering. However, due to granularity issues and computational efficiency, we can transform the raw consumption values in the absolute sum or percentage of consumption over a time interval in order to capture different behaviours. In these cases we can end having many features which for some algorithms could be challenging to deal, besides that, we do not want to have correlated features that do not differentiate the clients. Due to this, we can use dimensionality reduction techniques to reduce the number of features and obtain the most significant ones, as Principal Component Analysis (PCA), Subsection 3.2.2, and Autoencoders, Subsection 3.2.3.

3.2.2 Principal Component Analysis

The PCA is used to eliminate self-correlated features that are redundant or partially redundant leading to an excessive computational weight preventing the operation of some algorithms. This approach transforms the data set in a set of orthogonal coordinates. It is a dimension reduction technique that finds the directions in which the data has the most significant variance to find the principal components. The principal components are sorted by the proportion of explained variance of the original data set, so by getting the first components we are only obtaining the more informative ones. With PCA, we can get a high percentage of the variance of the data set only with some components, usually less than the number of original features, maintaining or even increasing the quality of the clustering.

3.2.3 Autoencoders

Autoencoders are a type of artificial neural network used to learn a way to represent a data set. They are usually a multi-layer perceptron since they have hidden layers where the data is coded more than just the output layer. The information in this type of neural networks moves in only one direction, in a feed-forward way, which means that the output of one layer is always the input of the next layer. The Autoencoders are made up of the encoder, the code and the decoder independently of the number of layers (Figure 3.1). The code is the representation of the input data that is used to compress our data set but also used to reconstruct the original data. The results of the output layer are used to compare with the input ones and calculate the data loss with an error-function.

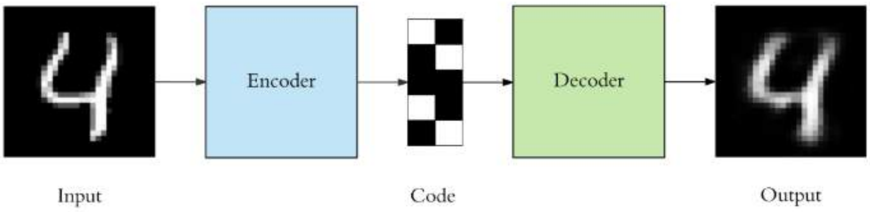


Figure 3.1: Example of the several parts of an Autoencoder and how it works [4].

To build a neural network we need to specify the activation function which determines the output of a layer based on a mathematical function. There are several different activation functions that change the way our data is coded and thus the result. They are divided in binary step function, linear activation function and non-linear activation functions. The last ones are more complex and allow to extract more complex information of the data. In Chart 3.2 we can see the difference between two of the more common non-linear activation functions. See [30] for details on Autoencoders activation functions.

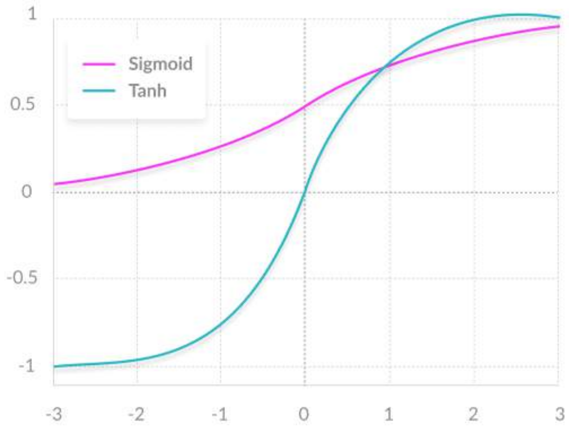


Figure 3.2: Example of two different activation functions - Sigmoid and Tanh [27].

Autoencoders have “demonstrate a promising ability to learn meaningful features from data” [34], allowing one to find a structure in data that is not immediately visible. When using the Autoencoders as a dimensionality reduction technique, we use the output data of the middle layers, the encoder, since we want a more compressed representation of our data [18]. We still build the decoder part to calculate the performance of the Autoencoder but do not use the final output of the neural network since we only intend to reduce the number of features.

3.3 Sample Size Determination

A question that arises a lot in statistical analysis is how to determine the appropriate sample size for the phenomenon we are studying. A sample with a size close to the actual population can yield more accurate study results. However, it is not always possible to obtain this quantity of data for several reasons. In a census, for example, it is expensive to interview many individuals. In our case, many customers do not have smart meters installed yet, and even the smart meters in communication sometimes have flaws and problems. So it is necessary to use a sample of our population and choose wisely its size to obtain a statistical study the more accurate possible.

The needed sample size can be calculated through the formula 3.6, obtained by the equation of the estimation of a proportion. We will consider a confidence level of 95%, equivalent to a Z-Score of 1.96, which is more than enough to extrapolate the conclusions to the rest of the population and a margin of error of 5%. Standard deviation in the formula is related to the maximum variance of Bernoulli distribution, which is approximate to normal distribution when sample sizes are over some tens, and is obtained by $\sqrt{p(1-p)}$. This value is set to the most conservative case: the case where the probability of an event carries the greatest uncertainty, that is, 0.5 [17].

$$n = \frac{(Z\text{-}Score)^2 \times StdDev \times (1 - StdDev)}{(Margin\ of\ error)^2} = \frac{(1.96)^2 \times \sqrt{0.25} \times (1 - \sqrt{0.25})}{(0.05)^2} = 425 \qquad (3.6)$$

By doing the calculus, we obtain that a sample size of approximately 425 would be enough to get accurate conclusions about the population in study.

3.4 Prototype Clustering

Prototype Clustering algorithms assign each data point (instance) to the cluster represented by the closest prototype (centroid, medoid, etc.). The prototype is a data instance that is representative of all the data in the cluster. These type of algorithms are largely heuristic which means that they do not rely on a formal model (model-free). Due to this, they are not useful to understand the nature of data and for representing the relationship between the features and clustering outcome in the best way. In this type of models, the data is clustered with the assumption of a specific distribution of the data which led to errors in the cases when the distribution of our data is different from the one assumed by the algorithm as we can see in Figure 3.3.

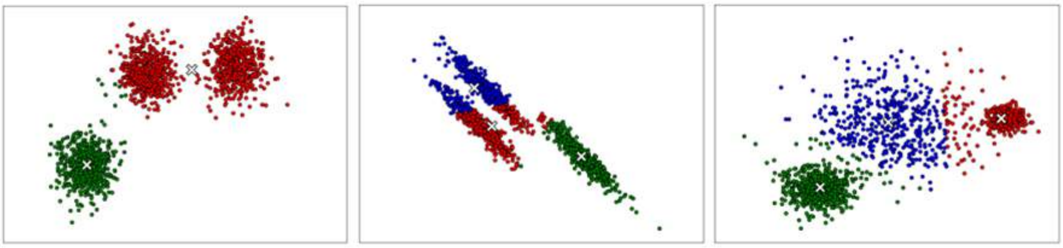


Figure 3.3: Prototype clustering problems when the number of clusters chosen is incorrect (left), the clusters are not spherical (middle) or they have different variances (right) [21].

3.4.1 K-Means Clustering

The K-Means Clustering starts with a random set of k prototypes and assigns each point to one of this prototypes, creating the initial clusters, based on the distance to the nearest prototype [21]. Then the algorithm recalculates the prototypes as the centre of each cluster and form new clusters. The k-means process is repeated until convergence or some stopping criterion. It requires the number of clusters as input which can be a problem since in many cases we cannot know it a priori and it's difficult to predict. The algorithm forces the number of requested clusters even if there were more or less real natural clusters in data. Moreover, it assumes that the clusters are spherical and have the same shape and variance leading to errors as in Figure 3.3.

However, k-means is computationally very efficient for large data sets producing quite good results [3][32]. Being able to choose the number of clusters can present itself as a good thing in specific uses of this algorithm. The k-means is a widely used clustering algorithm besides some disadvantages, so it was taken as a clustering method which was worthy of analysis.

3.4.2 Hierarchical Clustering

Hierarchical clustering solves the k-means problem of not knowing the number of clusters since its output is a hierarchy of the relation between clusters, a dendrogram. So we can understand the best number of clusters with the help of the visual information given by the dendrogram. This structure can be built in an agglomerative or divisive way. In the first one, we start with clusters of single point that are successively merged while in the second we start with a larger cluster with all data points and split them.

This clustering algorithm may present some disadvantages since it still is highly heuristic when Euclidean distance is used and real clusters in our data set are not spherical and of similar volume, introducing some errors in the clustering. However, if other distances such as Mahalanobis distance can be considered, this algorithm may produce good results. Considering the nature of our data, we may be interested in grouping clients by different levels, for example by splitting on a first level where we can immediately distinguish the most common load profiles from the others.

3.5 Density-Based Clustering

Density-based spatial clustering of applications with noise (DBSCAN) is a density-based clustering algorithm in which points close to each other are grouped based on a distance measure and a minimum number of neighbours. This approach forms groups of elements whose neighbourhood has a density of objects above a certain value. However, objects that are away from any other object (reflecting the low density of their neighbourhood) are discarded and will not be integrated into any group. The algorithm requires two parameters: the ϵ which is the minimum distance for two points to be considered as neighbours and *minPts* the minimum number of points in the neighbourhood of a point *p* so that *p* is considered a core point and for a cluster with all its neighbours. It works by iteratively expanding the cluster to all reachable point from *p*. If the neighbourhood population of some point *q* is fewer than *minPts* then *q* is considered as noise.

This type of clustering solves some of the problems stated about the prototype-based clustering. For example, it allows different shapes of clusters without any prior assumption [24]. This technique has the advantage of easily recognising the shape of high-density clusters even if they have a non-spheroid or non-ellipsoid shape (just contiguous). However, when considering all elements whose neighbourhood are low density, it does not form groups with them since it ignores this elements. Nevertheless, in the case of our problem we may use this technique to identify the scattered elements by filtering the high density clusters.

3.6 Model-Based Clustering

Model-Based Clustering (MBC), contrary to prototype-based clustering, relies on a formal model and does not use a heuristic approach to build clusters. It tries to recover the distribution from the data assuming that a generative model produced the data. It can be thought as generalising k-means clustering to incorporate information about the co-variance structure of the data and the Gaussian's latent variables. It starts with the assumption that some Gaussian mixture generates the data and initialises the algorithm using the results obtained from agglomerative hierarchical clustering with several distributions. Then it attempts to optimise the fit between the data and some of the Gaussian models [29]. This is done by means of an Expectation Maximisation (EM) algorithm which is an iterative method to maximise the probability of the data being generated by a given model. In the Expectation phase it uses a function as hypothesis and in the Maximisation phase the parameters of this function are computed maximising the probability of the data being generated by it. In the Figure 3.4 we can see the geometric characteristics of the clusters depending on the 14 different Gaussian Models considered by this approach.

This algorithm is available in several packages as the *Mclust* [29] in R. The *Mclust* algorithm of this package returns the best model that fits the data and also the optimal number of clusters which is one of the major needs in the other algorithms. Moreover, it also assigns the data points to the distinct clusters.

3.7 Chi-Square Test

Chi-square test is used to test how expectations compare to observed values. It may be used to test the independence between two categories of values. Thus, if the expected frequencies of a category in case of independence are significantly different from the observed frequencies, then the hypothesis of independence must be rejected. The chi-square distribution is used for these tests. The samples used must be random and large

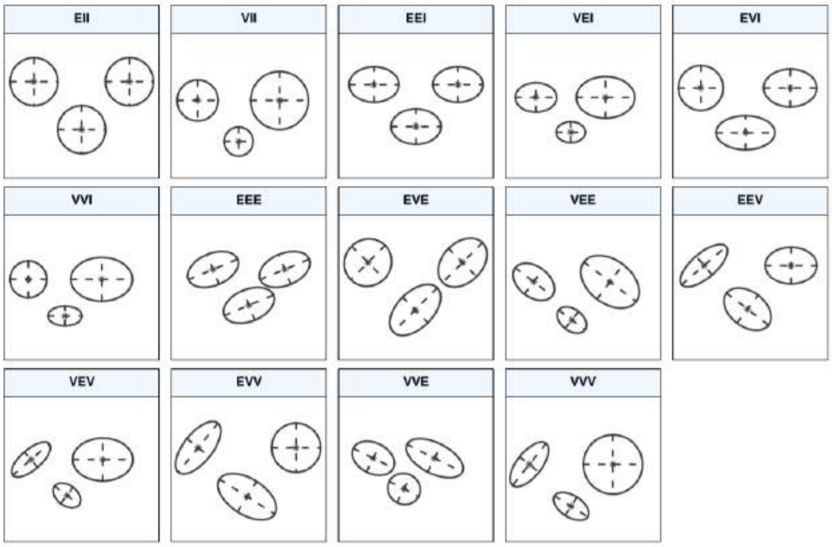


Figure 3.4: 14 possible Gaussian models with different volume, shape, orientation, and the associated model names [29].

enough in order to be representative of the data. In practical terms, the X^2 is obtained by

$$X^2 = \sum_{i=1,j=1}^{i=m,j=n} \frac{(x_{i,j} - m_{i,j})^2}{m_{i,j}} \quad . \tag{3.7}$$

Where $x_{i,j}$ is the observed value for the instance i of one of the categories and instance j of the other and $m_{i,j}$ is the expected value in case of independency for the same pair of instances. Then if X^2 is higher than the value in chi-square table for $(m-1)\times(n-1)$ degrees of freedom and for a level of significance of α (typically 0.05), the hypothesis of independence must be rejected. Available software usually uses the p-value to indicate how to evaluate the hypothesis of independence. After the input of the contingency table, it reads the degrees of freedom and a usual practical rule is followed: if p-value is less than 0.05, the hypothesis of independence must be rejected.

3.8 Clustering Validation

Clustering algorithm's quality can be measured with two main analyses: mathematical analysis of the clusters without taking into account the final application which is usually called cluster validity; and analysis of the clusters in the environment of its application, the cluster evaluation [19].

Cluster validity can be based on its cohesiveness that reflects two primary objectives, high intra-class similarity and low inter-class similarity. The determination of the optimal number of clusters in the clustering is essential since it affects the results and performance of the algorithm.

3.8.1 Elbow Method

Elbow Method is one of the methods that helps to avoid the problem of not knowing the number of clusters. It compares the Within-cluster sum of squares (WCSS) according to the choice of k (number of clusters), Chart 3.5. The WCSS calculates the sum of the quadratic distance between the points and the centroid within each cluster, measuring its cohesion.

$$WCSS = \sum_{k} \sum_{x \in C_k}^{n} (X_i - c_k)^2 \tag{3.8}$$

The purpose of this procedure is to minimise the WCSS until we find the elbow zone, where the curvature of the graph is the most significant and it starts to stabilise, to obtain the appropriate and more consistent number of clusters. However, this analysis doesn't allow to check if the created clusters have meaning in our data set.

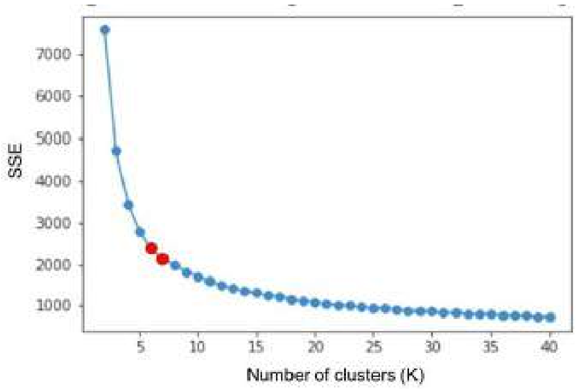


Figure 3.5: Definition of the number of clusters with the elbow method [8].

3.8.2 Silhouette Score

The Silhouette Score evaluates the consistency within clusters by taking into account their tightness and separation. It is a value ranging from -1 to 1 calculated by averaging the following formula over all data points i considering all clusters:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{3.9}$$

Where *a(i)* is the average distance between some point i and every point within its cluster and *b(i)* the average distance of i to all points of its neighbouring cluster, which is the cluster with a smallest average distance to i. A negative *s(i)* means that the point is assigned to a wrong cluster weighing negatively on the average. while a larger *s(i)* means that the point is well clustered.

A Silhouette Score closer to 1 means more distance between clusters and more tightly clusters. So the highest value of this score will suggest the best performance for the data. By testing the clustering algorithm for several numbers of clusters we can get the highest score to obtain the optimal number of clusters.

3.8.3 Data Visualisation

Sometimes Data Visualisation is the only way of perceiving some characteristics of the data set. To evaluate the results of the clustering algorithm we can use several types of visualisations to perform better analysis. It can helps to see things that are not possible by only using the numeric values obtained by the cluster validity. Usually Data Visualisation techniques are application based methods which means that their use depends on the problem that we intend to solve. Choosing the right type of visualisation has a big impact on the conclusions we can draw from it.

3.8.3.1 Hairball plots

Hairball plots is a data visualisation technique in which the several energy consumption profiles are plotted on a single graph with lines of reduced opacity to detect patterns [25]. To evaluate the results of the clustering algorithm, it is possible to plot the several time series within the same cluster in the same graph. Thus we can check the consumption profiles that are similar and also the ones that are deviations [23]. This seems to be one of the best ways to analyse the results of the clusters based on the application since it allows to understand if the clients of one clusters have similar profiles. If the clustering algorithm is grouping clients well, a pattern in which the lines of reduced opacity overlap is observed, we can see this both on Figure 3.6 and Figure 3.7.

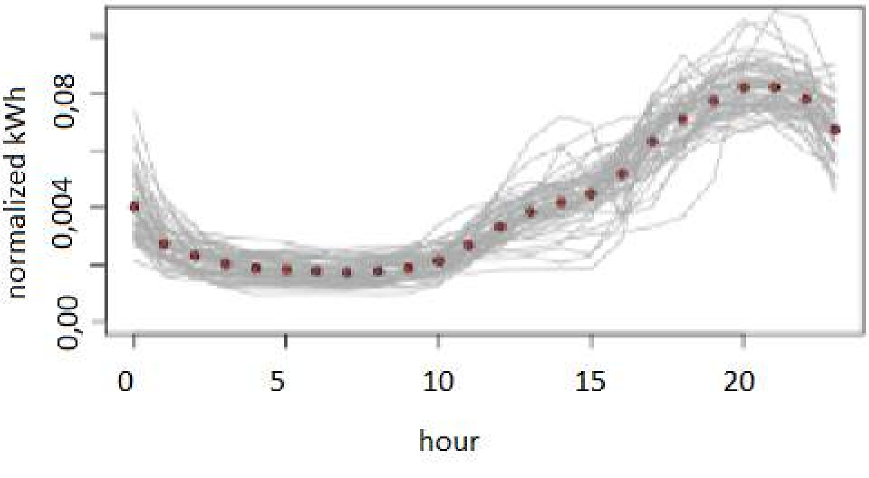


Figure 3.6: Hairball plot of several consumption profiles of the same cluster. Each line represent the consumption profile of one client and the lines that are furthest from the common pattern are the outliers of this cluster. Adapted from [23].

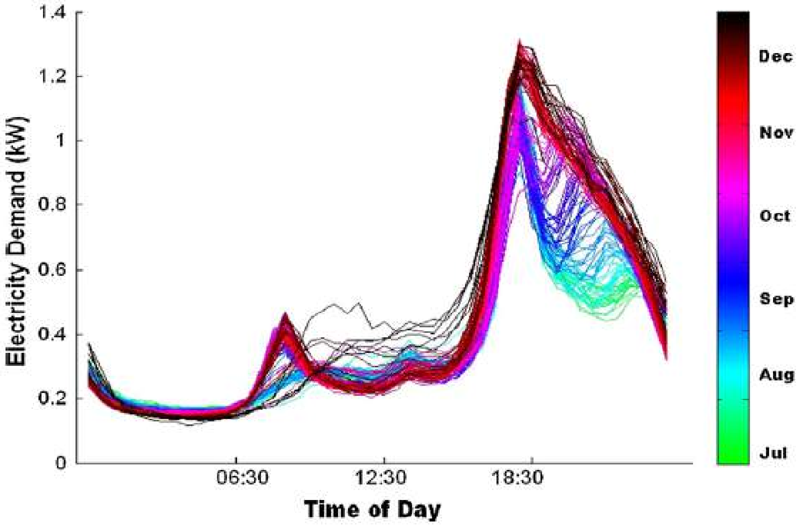


Figure 3.7: Hairball plot of one of the profile classes obtained by the classification project of [26].

Chapter 4

Methodologies for Solution

This chapter explains the methodologies defined to achieve our objective. Firstly, Section 4.1 gives an overview of data exploration, which includes data definition, collection and pre-processing. Then, in Section 4.2, we explain what are the techniques used to solve our problem, which will be explored in Chapter 5.

When dealing with the results of clustering, there isn't a correct division but rather a more or less adequate division for the data set depending on our purpose. Due to this, one of the primary difficulties of clustering is choosing the adequate algorithm, from a range of algorithms, that better explains the data. Since the clustering can be based on different aspects of the data, another difficulty is understanding how to represent the data set to feed the algorithm. In most cases, real data doesn't come with an adequate format for clustering, so we need to choose which features of the data to use. So, when working with clustering, the one of the main steps is the pre-processing of the data since this choice defines in what characteristics of data the algorithm is focusing on and how it affects the forms of the clusters. Moreover, the choice of the algorithm is also an important step since even with the right features we can end up in mistakes caused by the selection of an incorrect algorithm.

4.1 Data Exploration

Machine learning complexity increases exponentially when dealing with raw data. Moreover, when we deal with real cases, we do not know what to expect as opposed to when we use, for example, the Iris data set [9] or other known data set. In most of these data sets, we already know what they contain and what are the expected classes among them. When dealing with real data, we can end facing the garbage in - garbage out paradigm, Figure 4.1. It happens when we do not know our data and make the mistake of feeding the Machine Learning algorithms with wrong or meaningless inputs. In these cases, even if we have a perfect algorithm, we will end up with the meaningless output, "The analysis are as good as the data". This data exploration phase is what delays most projects with machine learning analysis. In our project, we need to do the complete work of data definition and collection from EDPD databases. Moreover, the pre-processing is necessary for the preparation of the raw data to the algorithms.

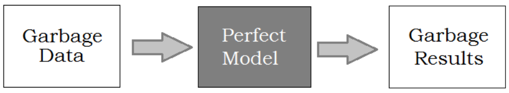


Figure 4.1: Garbage Paradigm.

4.1.1 Data Definition and Collection

All the data used in this dissertation is a subset extracted from the original data set with the anonymization of the individual clients and does not reflects any personal information.

The work made in this thesis started by the analysis of which would be the data from EDPD databases, that can be used to solve our problem. As we already explain in the 2.2.2, load diagrams only exist for a small number of customers which would imply that we would discard a large number of customers who could not receive focused tips and the comparative of consumption. For these reasons, we choose to use daily consumption since this is a study with an application to a real use case and EDPD intends to group and analyse as many customers as possible and not be restricted to such specific cases. Moreover, an analysis of consumption at every 15 minutes is too fine-grained, which would imply that the algorithm would take into account all the small changes and outliers. In the future we may want to go to that level of detail but for now we want to understand the high-level behavior of the clients. We chose to define the temporal space of our data set as all the readings of 2018, since this is the most recent year with full data readings and therefore the year in which there is a higher quality of data, both because it has more smart meters installed and communication errors are being corrected and improved every year, as described in Section 2.2.2. Another data definition agreed with EDPD was that initially, this analysis would only make sense for the Normal Low Voltage (NLV) clients to which the general energy efficiency tips apply. The choice is made because clients with tension levels over the NLV are already industrial customers that need different advice on Energy Efficiency, and that does not make sense to compare with smaller clients.

In order to carry out this project, it was necessary to extract our data set. To do so, we need to create programs in SAS, using SQL queries, that return the daily consumption of each client. The queries need to contain the filters that we defined, like the one of only wanting the NLV clients, together with other filters that restrict our sample to the most reliable possible, i.e. with the fewest faults possible. At first, we start by thinking that adding several theoretical filters to our queries would allow us to find the most reliable and random sample. We limited our sample to the smart meters of the most trusted manufacturers and to those installed longer and to the clients with currently active contracts. We build queries based on those filters and check if these restrictions would not be biasing our data for analysis. In the first case, we have to ensure that the selection of the meter manufacturer does not skew the type/location of the customer. Which means, we must ensure that there is no correlation between the manufacturer and other customer characteristics, for example, if all the smart meters of the manufacturer with best performance are located in Lisbon. We investigate the business and conclude that the manufacturer does not affect the type of clients or locations. The production of the components operates with a tender for lower prices where the manufacturer with a lower bid wins a percentage of components to produce. Components of the meters are produced without any specific end location and then assembled without any location criteria, which means that filtering smart meters by manufacturers is not biasing our sample. On the other hand, smart meters issuing data via GPRS are mostly coming from large industries or valuable customers located in remote areas, as we already explain in Section 2.2.2.1. So using only clients with smart meters issued by GPRS biases the data and therefore is not a valid option. With this in mind, we build our queries with the restrictions that we only want measurements of energy consumed and not produced, smart meters from the most trusted manufacturers and the other conditions already described.

By running the queries to extract data of several clients, we notice that for different months, a different number of clients and readings was returned , Figure 4.2, as we can see there are consumption values that seem zero. It happens not because the consumption were effectively zero, but because in these months there is no consumption data. These missing values are caused by several issues explained in 2.2.2.1. Obtaining these results made us conclude that the filters were not enough to get the most reliable data, and so we start thinking about other approaches.

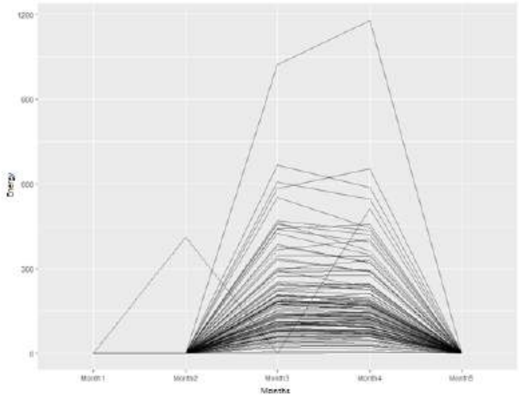


Figure 4.2: Consumptions over five months of clients with missing data. The y axis represents the value of consumption in a period of time (kWh).

The new approach to obtain our sample is to create a ratio of the quality of the readings. Since constraining the data set with theoretical filters were not enough. We understand that we had to evaluate the quality of the data set data in practice. So we create a query that for each smart meter it calculates the volume of days in which there were readings of the three tariff periods over one year and divides that for the supposed number of measures that should exist multiplying by one hundred, formula 4.1. It creates a percentage of availability of communication of each smart meter. A smart meter with a ratio of 100% means that it has communicated all the supposed readings of the year.

$$\frac{Volume}{31Dec2018 - 1Jan2018 + 1} \times 100 \tag{4.1}$$

The result of this query was a table with the top 2018 of clients with the best data quality, Figure 4.1, which means that choosing the first ones implies choosing the most reliable sample.

Table 4.1: Example of the top 2018 of smart meters with best data quality.

Instalation	MinDate	MaxDate	Volume	Days	Ratio
Client A	01/01/2018	31/12/2018	365	365	100%
Client D	01/01/2018	31/12/2018	364	365	99,725%
Client F	01/01/2018	31/12/2018	362	365	99,178%
Client E	01/01/2018	31/12/2018	359	365	98,56%
Client B	01/01/2018	31/12/2018	307	365	84,109%
Client G	01/01/2018	31/12/2018	294	365	80,548%
Client C	01/01/2018	31/12/2018	279	365	76,438%
Client H	01/01/2018	31/12/2018	232	365	63,562%
...	...	...	...	...	...

As explained before, EDP wants to imply the Use Case of the Energy Efficiency to as many customers as possible, but the quantity of data can be challenging to analyze, and to get tangible results. Besides, vast quantities of data entail a much higher computational cost, and so we would have to use more computationally efficient algorithms that are often not so accurate because they work based on brute force. So to start and especially for the part of data mining, we think it would make sense to use a smaller sample, but with a size that would be enough to make reliable conclusions about our data, as explained in Section 3.3. We ended up choosing a sample of 999 clients since it is larger than the appropriate minimum size and it is still not too large to have a high computational cost.

4.1.2 Data Preparation

The daily consumption data arrive on the form of totalizators, so it is necessary to calculate the difference of each measurement minus the previous one to obtain the energy consumed in that day-time interval. Besides, these differences must be made for each time of use period (*peak*, *half-peak and off-peak*) since there are three different measurements every 24h. As already mentioned in a real-world scenario, some data problems hamper the performance of the algorithms that need this information. In the beginning, our data seems clean and ready to analyse, but by starting to explore it, we face several errors. So it became necessary to understand the problems and develop methodologies that still maintain the quality of the data for the analysis, even when facing temporary data errors. The first detected error was related to the substitution of a meter where we face an abrupt change in the values of the totalizators. In these cases, we substitute the wrong measured values by the weekly average consumption in order to find an approximation of what would be the real consumption value.

4.1.3 Features Extraction

After having the data set extracted and prepared to be analysed, we need to define how to represent the data. The choice of features affects what the algorithm will focus on and consequently affects the results. As we have already said, the definition of features depends largely on the approach we want to take to the data and the final application.

Firstly, we should start by defining the temporal dimensionality of our features. The first hypothesis is using the raw data, which means to use the daily readings directly. We discard this option since the clustering of daily readings are fine-grained, which implies being too susceptible to outliers and to small differences in consumption that we are not interested in analyzing. Moreover, to the EDP's specific use case of the Energy Efficiency tips and the comparison with similar clients, we aim to get the general behaviour of the customers and not the behaviour of specific days. On the other hand, we have weekly clustering, which makes much more sense since there are differences in festivity weeks, vacation weeks, workweeks, among others. It also makes sense to consider the monthly cluster since it will allow finding differences in consumption over the months of the year. For example, we could find that most of the residential customers consume more in winter than in summer and that they probably need to improve their house's isolation and improve energy efficiency related to heating. Finally, we also think that it could be useful to consider yearly information despite it being coarse-grained. Yearly features allow understanding the customer's overall consumption picture regardless of the time of the year, seasons and other external factors that may affect consumption. Using different time frames allows us to analyse customers from several perspectives and better understand their behavior.

Another essential factor in extracting features from data is defining what they should take into account. The main goal of this project is to find similar clients. However, the question is similar to what level? Clients with similar magnitudes of consumption? Which means to compare clients according to average consumption values. For example, in this plot of weekly consumption, Figure 4.3, we can separate, just by looking, the customers with higher consumption compared to the others. On the other hand, we may also want to compare customers according to their consumption profiles by comparing the data on the same scale, allowing us to compare customer profiles of consumption even if they have different consumption magnitudes.

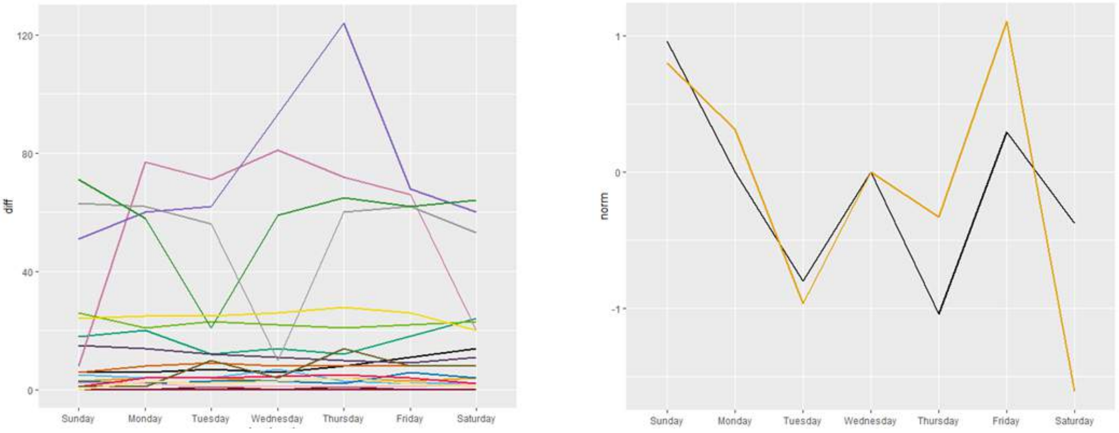


Figure 4.3: Example of features based on magnitude and on format of consumption by each day of the week. The left Figure represents the value of consumption in kWh (y axis) for each day of the week (x axis), while the right Figure represents the normalized consumption (y axis) for each day of the week (x axis).

By looking to absolute values, we are comparing consumption scales and customers' magnitude, which is useful since the contracted power seems to be little informative. At the same time, we may want to normalize the features, by dividing the consumption of a specific period for the total consumption of each client, so that we are comparing similar scales and looking at the format of consumption. Therefore, we conclude that in our case, we intend to use more than one type of features in order to fit several perspectives as we want to get a description of a customer as specific as possible.

To take into account the magnitude we will try to use as a feature the absolute sum of consumption, whether for a week, a month or the whole year. On the other hand, to take into account the pattern of consumption, we will try to use as feature the absolute sum of consumption by the time-of-day tariff periods since it is the temporal division we have in our data that allows seeing the different formats of consumption. In this case, we will start by analysing what happens when using non-normalized values, but it will probably make sense to normalize them so that it does not take into account the scales. So another set of features to consider are the relative weights (percentages) of consumption according to the tariff period (*peak*, *half-peak and off-peak*). To better understand visually what these features represent see Figure 4.4. From now on, whenever it is convenient we refer *magnitude* and *time-of-day tariff periods* analysis as the analysis of the features explained above.

Finally, we are also interested in studying the regularity of consumption and their outliers. For this, we decided to experiment using the coefficient of variation that allows us to realize which customers that vary more their consumption regardless of the scale. Also, we hypothesized to use the Skewness since it allows us to understand if the customer has more or fewer consumption peaks or valleys compared to the average.

We will try to use these different features in several ways doing different combinations

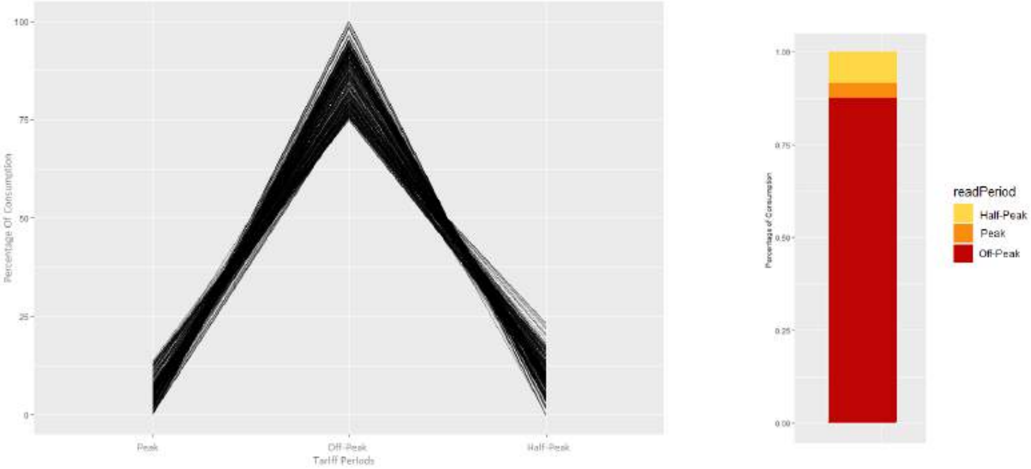


Figure 4.4: Example of clients represented by the percentages of consumption according to the tariff period. On the left side we can observe the hairball plot of the pattern of consumption of one cluster based on the percentage of consumption (x axis) depending on the tariff period (y axis). The right side is a stacked bar plot that represents the percentage of consumption that this type of profile has in each tariff period.

of them and evaluating the results obtained to understand which options are best. In addition to the features themselves, we will also test two features reduction techniques, PCA and Autoencoders, to realize how far we are not incurring errors by excess of features or repeated information.

4.2 Clustering Algorithms

The solution that we propose is an approach to identify groups of clients with similar electricity consumption according to different perspectives. This approach is part of a class of problems known as Clustering, which has already several studied and implemented algorithms, and that groups similar clients and separate different ones in different clusters. The distribution network has a large quantity of smart meters, making it impracticable to analyse the consumption of each client one by one and so we use clustering, more than just to group the clients, to understand what is happening in our data. So in the first part of the project, we intend to use the clustering algorithms as a data mining technique. The goal is trying to understand the universe of clients of low voltage with smart meters by analysing its habits and behaviour. By grouping, similar clients it is easier to characterise each one since we can easily understand the behaviour of all the clients by its group behaviour. Moreover, it will allow to create tailored advice and suitable groups to deliver a comparison of the consumption of each client with its group consumption.

By the analysis of state of the art and considering the nature of the data set, we choose 3 clustering algorithms to consider in our dissertation: Model-Based Clustering (MBC) by using the Mclust package, Density-based spatial clustering of applications with noise (DBSCAN) and Hierarchical Clustering with the R package hclust. The first one makes a general analysis of the data distribution without forcing many parameters like the number of clusters or the clusters format. It seems a reasonable hypothesis since we know very little about our data set, and we want to do an unbiased exploration. Moreover, the MBC seems appropriate due to the fact that it is expected that the consumption values have a Gaussian distribution as consumption tends to be distributed around a mean more or less equally decaying on its left and right sides. Many clients don't consume equally through time; for example, clients usually consume more during the day than at dawn when they are sleeping. The MBC already tests spherical clusters, so we thought that would not make sense to use K-means because if clusters had this format, MBC would probably identify them. Therefore K-means will only be used if we see that clusters are indeed spherical and that K-means is computationally more efficient. We will also try the Hierarchical Clustering, with hclust package on R. Although it also uses Euclidean distance, it creates a visible hierarchy (dendrogram) that may allow to obtain different and interesting conclusions. Finally, we consider the possibility of using the DBSCAN since it allows to detect clusters with different densities and peculiar formats. Without testing, we do not know whether or not our data have these characteristics and if the DBSCAN could find something that others can't.

The approach taken in this dissertation is very exploratory, and we intend to make several analyses using different algorithms and techniques that allow drawing different conclusions.

4.3 Evaluation

In order to evaluate the algorithms and compare their performances we are going to use both mathematical methods, to analyse the cluster validity, and application based methods that do the cluster evaluation.

As validity method we will consider Silhouette Score since it allows to understand if points are being well attributed to clusters. However, this score only evaluates the cohesion of clusters and not exactly its result given the specific use case. So in order to complement it we will use several data visualisation techniques that allow us to gain more insights from the results. We are going to consider the hairball plots that validates the pattern detected by each cluster. Moreover, we will also consider box plots to understand the distribution of values within a cluster, the stacked bar plots which allow an aggregated visualisation of the consumption per tariff, among others. With the combination of data visualisation and numerical scores we can more safely analyze the quality of the results and draw conclusions.

Chapter 5

Implementation and Discussion

In this chapter, we use several clustering algorithms with different features to understand what is the best approach to solve our problem. The implementation and results analysis were made iteratively. In each iteration, we analyse a different set of features and clustering algorithms, since they depend on each other, and discuss the results in order to evaluate its performance.

In the Subsection 5.7 we do an aggregation of the several experiences made in this dissertation in order to guide the reader. Moreover, for reading convenience, some complementary figures are available in appendix A.

5.1 Exploratory Introduction

Since we need to understand our data and how the clients consume, we begin by doing small weekly analysis. For these analysis, we started using the mclust library, which provides a MBC algorithm in R. We choose this approach due to the fact of being the only algorithm that does not need the number of clusters as input. In the case of the *magnitude* analysis, we could use the contracted power levels as the number of clusters, but we do not want to bias the analysis by any existing prejudices. Moreover, this contracted power does not always reflect the actual consumption of the customers. MBC can help us to find out what is the best division of absolute consumption or what are the different formats that exist among the clients. Moreover, it already tests several distributions of the data, so for example, when it tests spherical clusters, it will test a distribution similar to k-means. The idea is to start with the algorithm that requires less prior knowledge in order to understand our data and what information can be taken. After defining the best approach to this data set, we can define the best features and algorithms, evaluate their performance and compare them to each other.

The exploratory introduction was the first phase of data analysis that took place simultaneously with data set definition and data extraction. For this reason, at that time we only had available a small data set of the consumption data for March 2018. We started to use only one week as an experiment and to begin to understand what data we had and what kind of approaches we could take. So in the Subsection 5.1.1 and in the beginning of the Subsection 5.1.2 we use as data input the week from third to ninth of March. Although the Section 5.1 is not very long, this was one of the most complex phases of the project due to the lack of knowledge of the structure and content of the data. For the sake of relevance here, we present only the leading exploratory introduction analysis that led us to the most interesting conclusions.

5.1.1 Magnitude Analysis

As explained before in Section 4.1.3, we want to analyse the client's consumption over several perspectives. First, we analyse the magnitude of consumption, which means to separate the clients based on their scale of consumption, obtaining the clients that consume more or less during a week. If we do not take this into account, we risk telling a customer that he is consuming 200% more than his group and is consuming too much even though we may be comparing a store with an empty house all day. To cluster based on magnitudes, we use as feature the total sum of consumption for one week.

The algorithm divides the clients into three clusters of small, medium and big consumers which allows us to categorise the customers according to consumption levels to build a better comparative with their group of similar customers, Figure A.1.

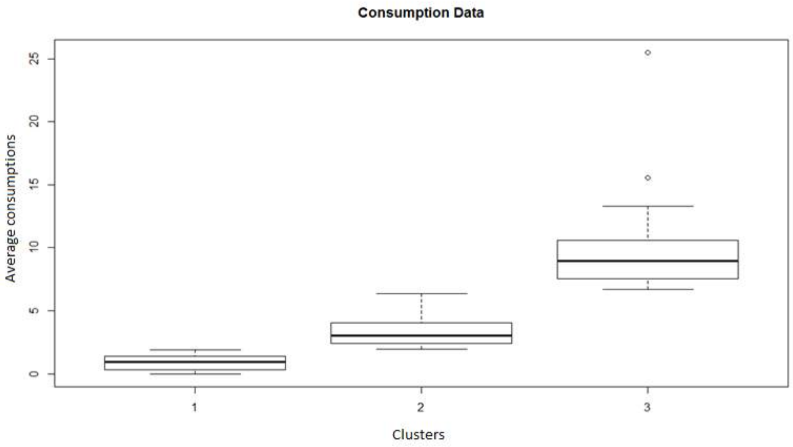


Figure 5.1: Box plot of the average consumption of the clients of each cluster obtained by running the MBC over the absolute sum of consumption of one week.

5.1.2 Time-of-Day Tariff Period Analysis

At the same time, we test the same type of algorithm but taking into account the different time-of-day tariff periods i.e. three features each one with the consumption during the period of Peak, Half-Peak and Off-Peak.

First, we test it with the absolute sum of consumption for each tariff period in each week as features and it returns few clusters that are pretty much equal to the ones produced by the algorithm based on *magnitude*. It happens because the difference in scale between clients is much more significant than the difference between tariff periods. So we need to ensure that all values of consumption of this three last features fall into similar scales since we want to focus in similar patterns/shape of consumption rather than in magnitude.

In the analysis of this Section we mainly want to detect and understand the differences of consumption taking into account the time-of-day tariff periods. So, we normalise the three features (consumption during the period of Peak, Half-Peak and Off-Peak), as explained in Section 4.1.3, obtaining the percentages of consumption by tariff period during the week for each client, regardless of consumption scales. This algorithm separates our data into nine different clusters - see the nine ellipses and colours of the Figure 5.2 - depending on if the client consumes more at peak, half-peak or off-peak. The MBC algorithm creates clusters of different sizes and formats based on the distribution of the data and creates cohesive clusters as we can confirm in their hairball plots in Figure A.2.

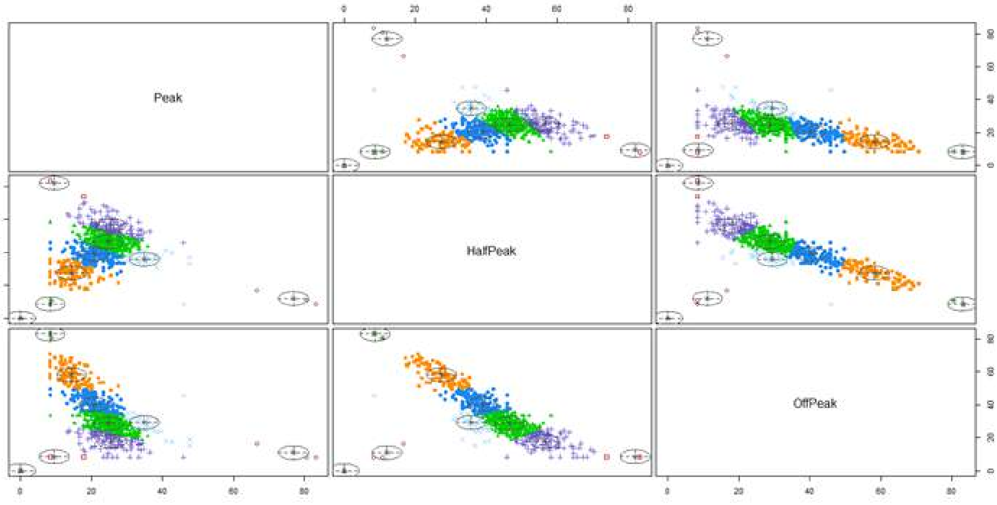


Figure 5.2: Clusters obtained by running the MBC over the percentage of consumption by tariff period. Axis of the figure are internal values produced by MBC for the features in the data set.

The algorithm identifies nine different patterns of consumption among our sample of one week of each client, as shown in Figure 5.3, and shows us that some patterns are more common than others, based on the size of the clusters that we can see in Figure A.3, i.e. bigger clusters means more common patterns among the clients. For example, clients that mostly have their consumption on off-peak period (Cluster 8) are less than the ones who consume mostly on peak and half-peak (Cluster 3). It allows us to identify the most common patterns but at the same time, the ones that occur less often.

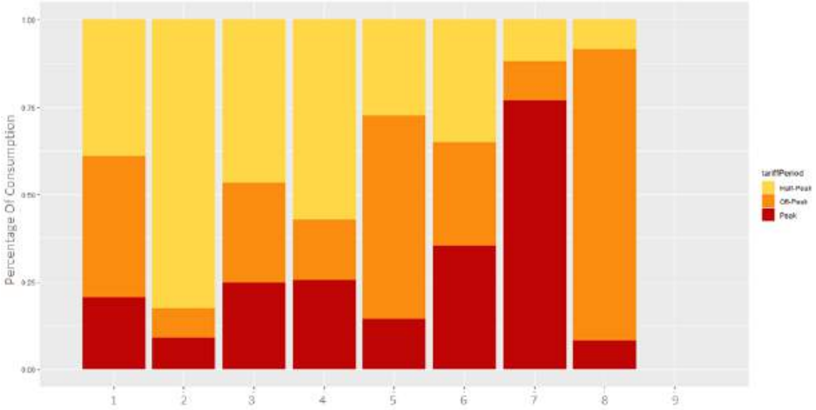
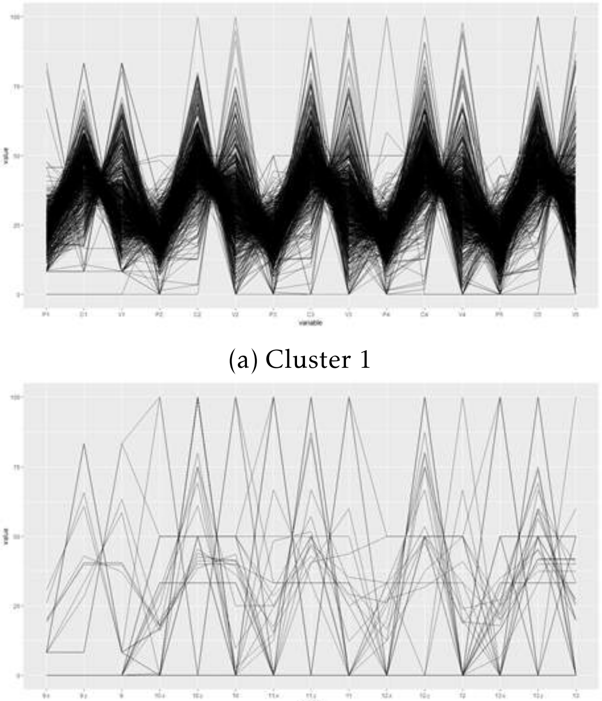


Figure 5.3: Percentage Stacked Bar Chart of the nine different patterns of consumption obtained by each cluster in A.2.

It was an interesting conclusion to realize that even in a small sample, it was possible to detect several patterns of consumption. These first analysis eventually helped to define the path chosen in the development of this dissertation. After this small analysis, to start digging in our data set, we decided to scale our analysis to more than one week and understand what happens in this case. We set several hypotheses: "Do the clients maintain in the same clusters over the weeks? Do clusters maintain their number or increment with the number of weeks? Does it make sense to analyse the clients weekly, or are they stable enough that we can only do a monthly analysis?"

We analyse the consumption of the clients over five weeks, of the end of February and March of 2018, by also using as features the total sum of consumption of each tariff period but for each week, so 5x3=12 features. It results in two clusters in Figure 5.4, which allows us to draw some conclusions. The cluster 1 contains 901 clients of the whole sample and has a hairball plot where we can understand a pattern of similar average consumption over the weeks. It confirms the initial hypothesis that humans are beings of habits that usually consume the same way. However, we can detect some outliers to this pattern, which shows us that even a client that usually follows the same habits can have weeks when it comes out of it. These outliers can be related to vacation weeks, weeks with holidays, like Christmas, or even just because of a change of the daily routine like a dinner out. On the other hand, cluster 2 (Figure 5.4) detects mainly clients with significant differences in consumption between time-of-day tariff periods. It detects clients with zero consumption. As already explained in Section 4.1.2, we are using the most accurate sample possible and dealing with faults, so these values are not invalid measurements but real zeros measurements in the data. These zeros could mean several things: the customer did not actually consume during that time, operational problems with the smart meters or we can even be facing a fraud scenario (Section 2.2.2). However, in order to better understand the behaviour of these clients, we need to do long-term analysis.



(a) Cluster 1

(b) Cluster 2

Figure 5.4: Pattern of consumption by week obtained by running the MBC over the percentage of consumption by tariff period over five weeks. Plot ordered by weeks and based on A.4. In the x axis P, V and C represent respectively the three tariff periods Peak, Off-Peak and Half-Peak of the different weeks. The y axis represents the consumption value for each period (kWh).

5.2 Clustering Composite weeks

We extract data of daily readings of 2018 for the clients of our sample. First, we start by doing similar analysis to the already made but now over the fifty-two weeks of the year for each client. We begin by using the raw data and lots of features of the consumption of all weeks of the year to understand what is the general behaviour of the clients.

5.2.1 Magnitude Analysis

Following the methodology already used, we begin by analyzing the magnitude of customer consumption over 2018 with MBC. For this, we use for each customer fifty-two features of the absolute consumption per week. Similarly to the *magnitude* analysis of only one week, we obtained three clusters as we can see in Figure 5.5. The clustering separates the clients by its general magnitudes as we can see in the hairball plots on Figure 5.6, where the cluster 1 to 3, respectively, contains the customers with the highest to the lowest consumption. However, in this analysis, we can observe that the clients do not have the same magnitude on all weeks. Even if a client has five of the fifty two weeks in a different cluster of its normal one, this does not affect clustering because in the combined features these weeks do not have enough weight. We can see, for example, in cluster 2 (Figure 5.6) that some clients consume considerably less on summer weeks comparatively with winter weeks. It would be interesting to be able to assign different groups to customers in these different weeks.

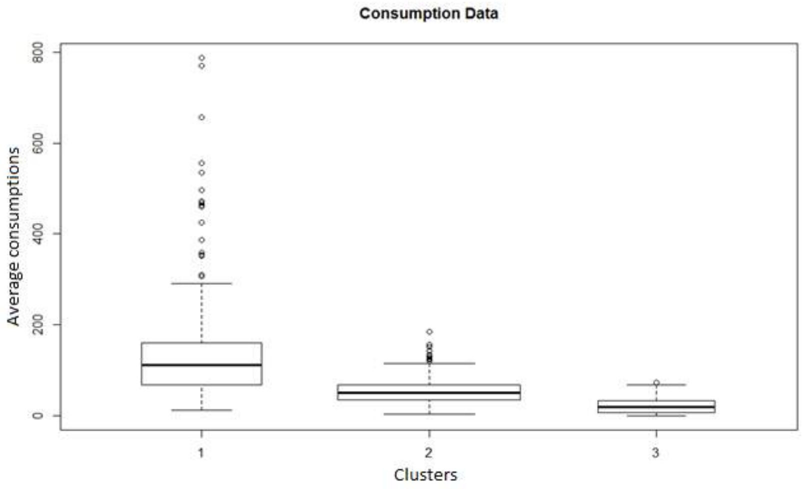
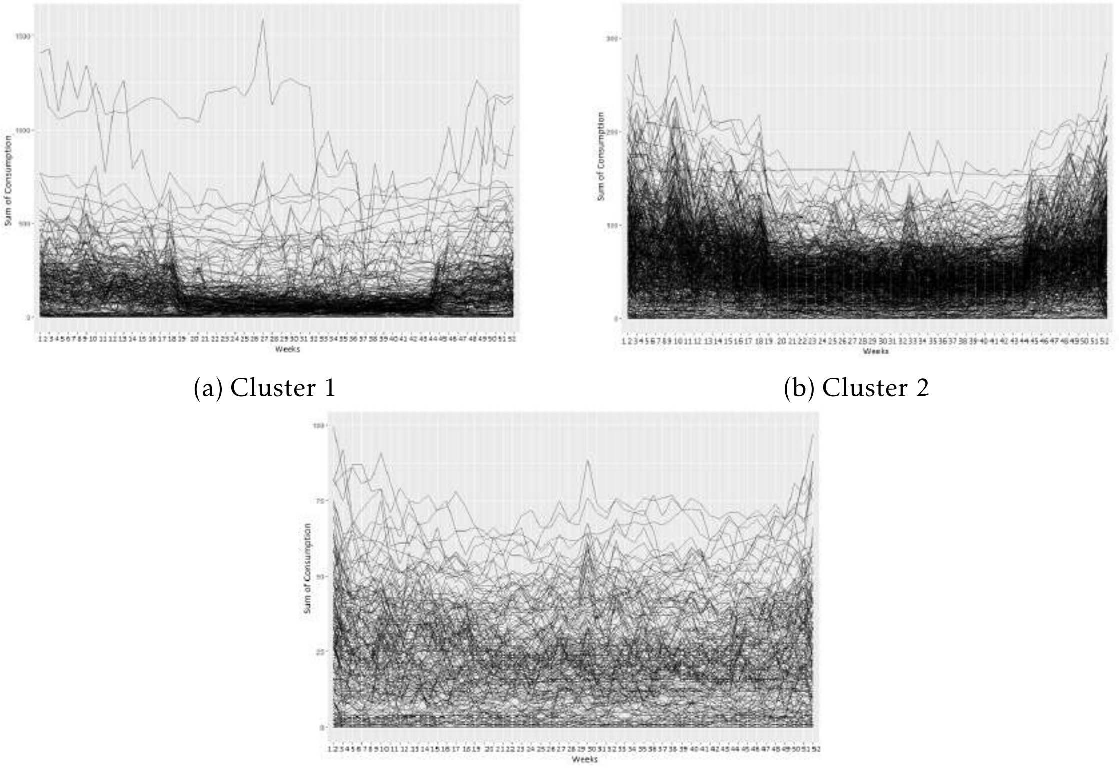


Figure 5.5: Box plot of the average consumption of the clients of each cluster obtained by running the MBC over the fifty two features of the absolute sum of consumption of each week of 2018.

The hairball plot already let us understand that this approach did not have good results, however as explained in Section 4.3 we will check the accuracy of our results with the help of a numerical evaluation metric, the silhouette score. In this case, Table 5.1, the silhouette score is almost near zero supporting the fact that the clusters are not appropriate. When facing situations like this they may be due to several causes: a wrong choice of algorithm, a wrong choice of features or even to a wrong choice of data set. In this case, the clients' weeks are distinct between them, so using all these features together with such big differences may cause the algorithm to fail to find significant differences.

Table 5.1: Silhouette score of the clusters obtained by running the MBC over the absolute sum of consumption of the fifty two weeks of 2018.

	Average Silhouette widths	Cluster sizes
Cluster 1	-0.27423	243
Cluster 2	-0.03328	543
Cluster 3	0.49848	222
Mean	**0.02845**	



(a) Cluster 1

(b) Cluster 2

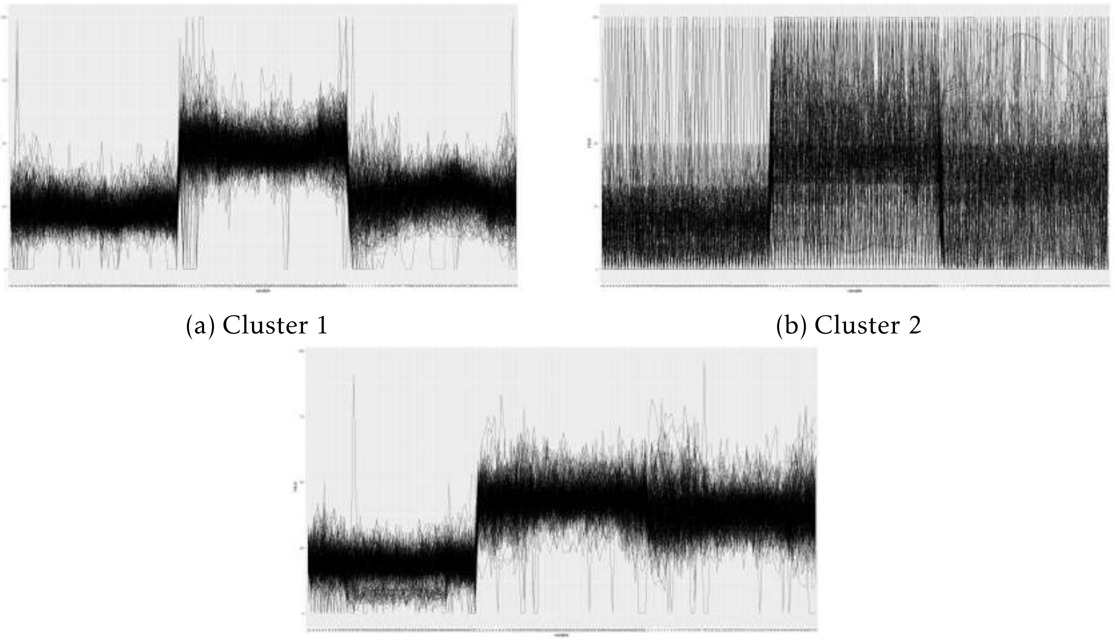
(c) Cluster 3

Figure 5.6: Hairball plot of the clusters obtained by running the MBC over the absolute sum of consumption of the fifty two weeks of 2018. The x axis contains the fifty two weeks and the y represents the consumption value for each period (kWh).

5.2.2 Time-Of-Day Tariff Periods Analysis

To this analysis we used the percentage of consumption for each time-of-day tariff period throughout the week, which leads to 52x3=156 features. It allows to study the differences of consumption between tariff periods and understand when customers consume more or less. It results in three similarly sized clusters two of them with a more specific pattern and the other with the less constant clients which do not immediately emphasise any pattern, Figure 5.7. The cluster without pattern (Figure 5.7 - cluster 3) groups the clients who do not have a similar trend of consumption over the year. The other two clusters show the two main patterns of consumption over a year: the clients that consume less in off-peak (cluster 1) and the ones that equally consume on off-peak and half-peak (cluster 2). This clusters mainly happen because composite features imply the same cluster for all weeks over the year. Similarly to the magnitude here we also have a similar result of silhouette plot as we can see in Table 5.2. Probably this happens because composite features imply the same cluster for all weeks over the year. Another possible reason may be the use of a large number of features.

By using the features of the total sum of consumption of each tariff period over every week of one year, as we do in this approach, we get 156 features, as we can see in Formula



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

Figure 5.7: Hairball plot of the clusters obtained by running the MBC over the percentage of consumption by tariff period over the fifty two weeks of 2018. In the x axis it is represented the fifty two weeks of each tariff period in the order peak, half-peak and off-peak. The y axis represents the absolute sum of consumption.

Table 5.2: Silhouette score of the clusters obtained by running the MBC over the percentage of consumption by time-of-day tariff period of each of the fifty two weeks of 2018.

	Average Silhouette widths	Cluster sizes
Cluster 1	0.19856	364
Cluster 2	-0.28223	256
Cluster 3	0.2156121	379
Mean	**0.08183**	

5.1. A large number of features implies a greater complexity that can make the algorithm computationally unbearable. Even when the system hangs this complexity, as in this project, it may be accumulating errors. Moreover, the algorithm can fall in a curse of dimensionality, as explained in 3.2, and features may end up losing their utility. Due to this, we try to apply some techniques of dimensionality reduction to understand if we can find more specific clusters than the three we found using the raw data.

$$\{w|w \text{ is a Week in the year}\}|\times|\{Peak, HalfPeak, OffPeak\}| = 52\times3 = 156 \text{ features per client} \tag{5.1}$$

5.2.2.1 Reducing Features by Principal Component Analysis

First, we start by applying the PCA to the 156 features. To explain 99% of the data variance, Figure 5.8, we need 125 components and to explain 95% it requires 86 components. It shows that we still need many components in order to represent our data without losing too much variance. However, we test the algorithm with 86 components in order to understand if the resulted clusters are better than the ones obtained by using the whole set of features. It also returns three clusters, but by analysing their hairball plots, Figure A.5, we can see that the cluster are less cohesive and with more outliers than the ones produced by the clustering of the 156 features, Figure A.4. One hypothesis to explain this result is the fact that PCA may sometimes distort the information contained in the initial data since it creates a matrix that gives different weights to the original features. In our case, PCA does not seem to be the best option since its representation of the data does not bring significant changes to the result and still creates clusters that appear to be less suitable than the originals. This is confirmed by the silhouette score in Table 5.3 as it shows a reduction in its value compared to the previous analysis.

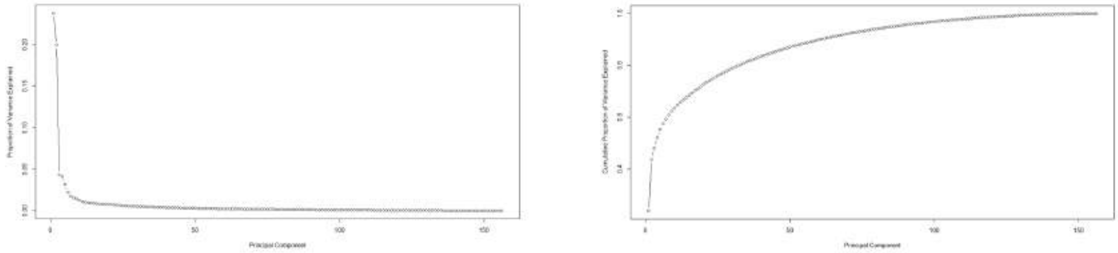


Figure 5.8: Proportion of variance explained (left) and cumulative proportion of variance explained (right). Plots obtained by running PCA over the sample of the percentage of consumption by tariff period over the fifty two weeks of 2018.

Table 5.3: Silhouette score of the clusters obtained by running the MBC over the matrix of the 86 principal components of the percentage of consumption by tariff period over the fifty two weeks of 2018.

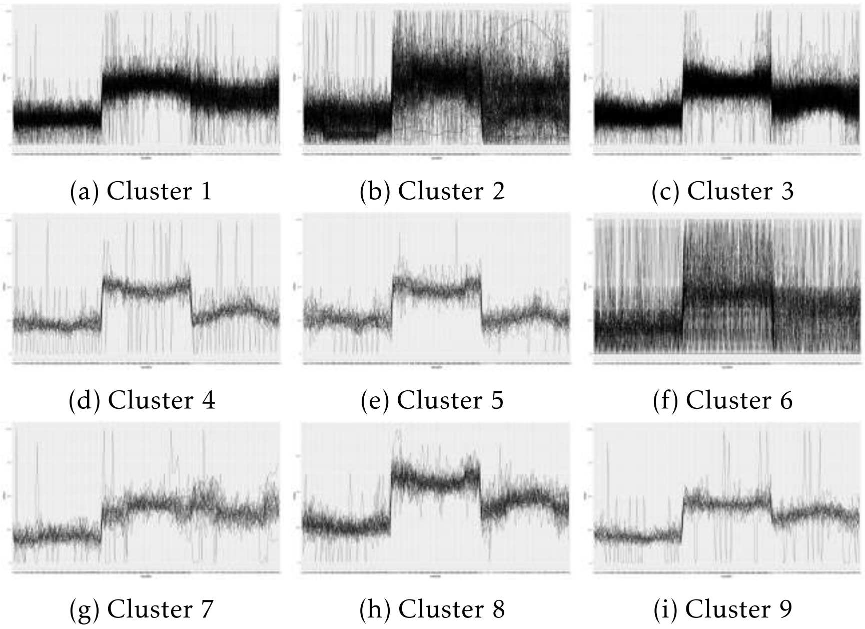
	Average Silhouette widths	Cluster sizes
Cluster 1	-0.17226	223
Cluster 2	-0.23115	208
Cluster 3	0.23026	568
Mean	**0.04434**	

5.2.2.2 Reducing Features by Autoencoders

As an alternative to PCA, we apply several types of autoencoders to understand if they could have a better performance on our data set and this specific application.

First, we start by building a neuronal network based on sigmoidal activation function with just three layers the input and output layers and a middle layer that represent our

156 initial features (Eq. 5.1) in only fifty new one. The output data of that middle layer, with smaller dimensionality, was the one used as input to the MBC algorithm. By running it, we obtained nine clusters, Figure 5.9, which are much more than the ones obtained by the algorithm with the non-reduced data and with the data reduced with PCA. Besides the immediate difference in the number of clusters we can observe that in this case, the algorithm could distinguish more specific patterns in a small number of clients. It happens because *sigmoid* smooths the data, i.e. it represents the data giving similar values to an interval of close values. For this reason, our algorithm can distinguish more easily some of the existing patterns by immediately separating those similar values. However, since it smooths a lot, our data set ends up putting lots of distinguishing clients all in the same cluster and creating clusters with small differences. This approach as a silhouette score -0.16920 which is worst than the ones obtained with PCA and even without the dimensionality reduction. Nevertheless, this is a case where data visualisation allows us to draw far more informative conclusions than evaluation metrics. Because as explained earlier for specific applications, it might make sense to be able to discover more specific patterns such as those found by autoencoders, even that overall the algorithm does not classify all clients as well.


(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

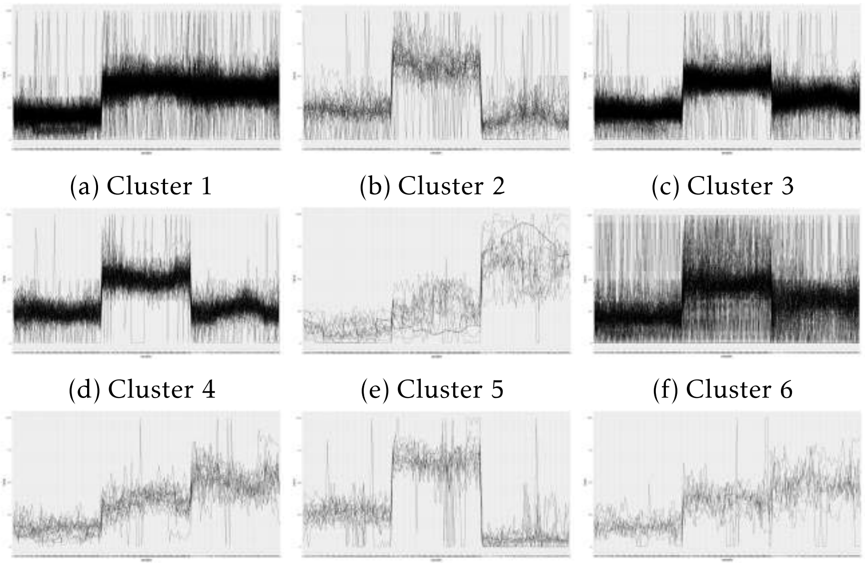
(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

(g) Cluster 7 (h) Cluster 8 (i) Cluster 9

Figure 5.9: Hairball plot of the clusters obtained by running the MBC over the percentage of consumption by tariff period over the fifty two weeks of the year reduced with the neuronal network based on the logistic function (*sigmoid*). In the *x* axis it is represented the fifty two weeks of each tariff period in the order peak, half-peak and off-peak.

Secondly, we also test the dimensionality reduction with autoencoders but by using the hyperbolic tangent function (*tanh*) which is *zero centred*, and so it is better to model *strongly negative*, *neutral* and *positive values*, see [27] for details. We prove this idea by looking to the results, Figure 5.10, that show how the clustering algorithm could obtain much more particular patterns of the clients with higher differences of consumption between tariff periods, which could be useful since we want to understand the different patterns of consumption and aggregate the similar ones. The fact that this approach could find more specific patterns is reflected in its silhouette score of -0.05594 that is higher than the one of *sigmoid* but still negative. So, the same conclusions taken in the *sigmoid* could be applied to *tanh*, which is the fact that the evaluation metrics are not enough to conclude if a clustering algorithm is having the right/appropriate results.

We also did some experiences with more layers, but we conclude that more layers only decreases the number of clusters since it is smoothing so much the data that it is becoming increasingly similar. Finally, we also test the *Leaky ReLu* activation function that is more similar to the linear function, and that has results similar to the first ones obtained by running the clustering algorithm without using dimensionality reduction techniques. We can conclude that the autoencoders, although cannot properly classify all customers, they could find a structure on the data that the PCA could not, so for this specific application autoenconderes seem more suitable.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

(g) Cluster 7 (h) Cluster 8 (i) Cluster 9

Figure 5.10: Hairball plot of the clusters obtained by running the MBC over the percentage of consumption by tariff period over the fifty two weeks of the year reduced with the neuronal network based on the hyperbolic tangent function (*tanh*). In the x axis it is represented the fifty two weeks of each tariff period in the order peak, half-peak and off-peak.

The analysis of these multiple clustering algorithms using the composite features of all weeks over one year shows that, besides having a trend, some clients do not consume equally in all weeks. This type of clustering forces that a client only has one cluster in all weeks, hiding some peculiarities of the consumption pattern that we could observe when analysing one week only. Moreover, it is impossible to detect the most appropriate groups for customers who have very different consumption between weeks. The clustering of all weeks as features is useful when we want to have a broad idea of the consumption of the client over the year, but it is not when we want to give specific energy efficiency tips for each type of week. These conclusions support the fact that it also makes sense to analyse the clients' consumption weekly separately.

5.2.3 Regular and Irregular Consumption

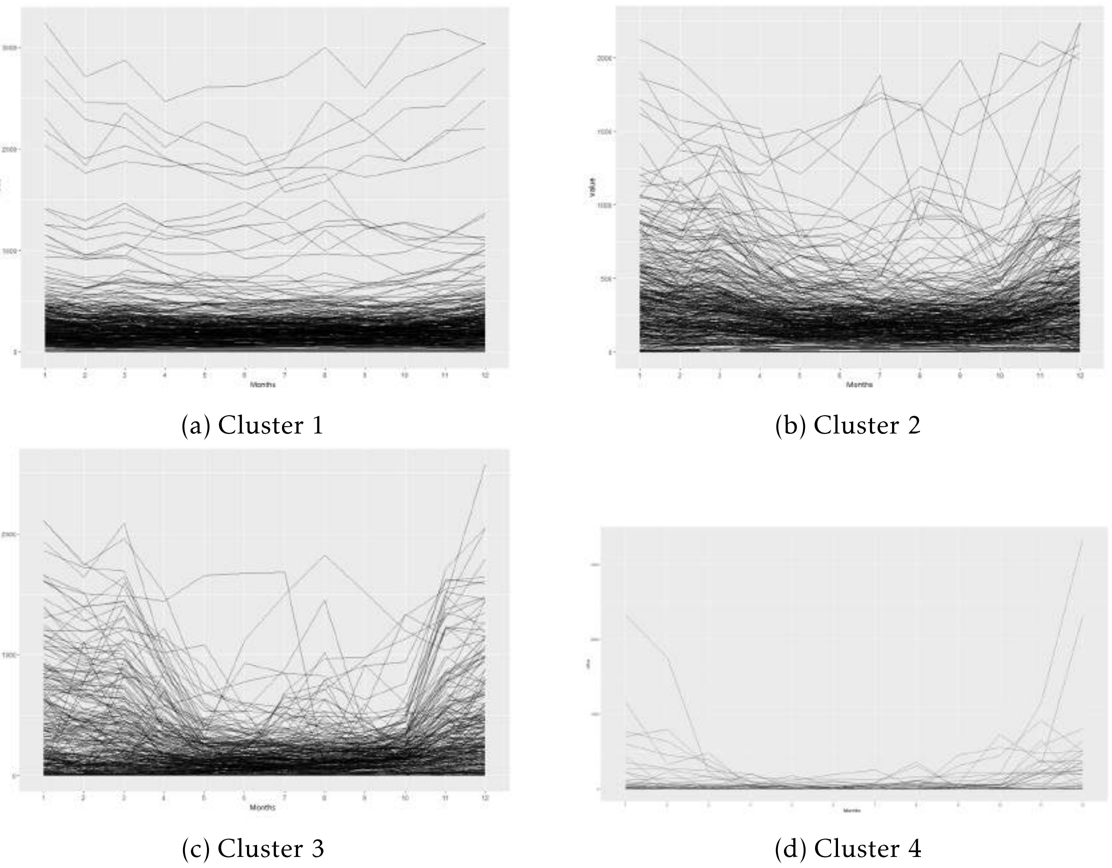
Realising if customers are generally stable over the different weeks or months over a year means that the customer is a regular customer who probably does not need to receive weekly or monthly different tips. On the other hand, if they are irregular customers, that means the tips have to adapt to their different ways of consuming.

The analysis of the results obtained by the clustering of composed features makes us understand that this kind of analysis will be good if we want to compare the regularity of consumption of the clients. However, by using the absolute values or percentages of consumption, we are focusing on the similarity of change more than just on their regularity. So we explored other hypotheses of features that would be capable of capturing the regularity and irregularity of a time series.

We try to use the coefficient of variation over the consumption of the twelve months of the year as a unique feature in order to understand which customers maintain similar consumption values over the months and which tends to change. The MBC divides our data into four different clusters, Figure 5.11, where cluster 1 represent the client with more stable consumption along all month of the year and the others (cluster 2, 3 and 4) contain the irregular clients that have significant differences of consumption from one month to another. We can confirm with the silhouette score, in Table 5.4, that this approach has significantly better results. The coefficient of variation seems to be a feature with significant differences to clearly distinguish the clients. Moreover in a more application based analysis it could be really helpful to separate the regular and irregular clients.

Table 5.4: Silhouette score of the clusters obtained by running the MBC over the coefficient of variation of the months of one year (2018). In the x axis it is represented the twelve months.

	Average Silhouette widths	Cluster sizes
Cluster 1	0.66426	339
Cluster 2	0.48275	344
Cluster 3	0.39169	261
Cluster 4	0.38614	55
Mean	**0.5152**	



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

Figure 5.11: Monthly consumption's patterns of the clusters obtained by running the MBC over the coefficient of variation of each month of one year (2018). The y axis represents the sum of consumption in kWh and the x the twelve months of the year.

5.3 Clustering separated weeks

Based on the conclusions obtained from the Section 5.2 and on the needs defined for the EDP's use case, arose the need to analyse the fifty two weeks of the clients separately in order to capture their particularities, for example, the cluster reflecting mostly the vacation weeks. It allows a more accurate analysis of customers' week types and the understanding of their behaviour over time.

To do this, we need to analyse the similar weeks independently of its time in the year. So we split the consumption data of the clients according to the week of the year, creating less features and so obviously there is no need of feature reduction. However, it creates a bigger data set with fifty-two lines of data per client since we consider only one week per line (*week of client*) as the example given in Table 5.5. It implies that different weeks of the same client can belong to different clusters as shown in Figure 5.12.

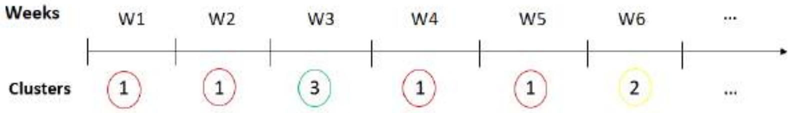


Figure 5.12: Example of the weeks and its respective clusters for one client.

5.3.1 Magnitude Analysis

Firstly, we test this new approach with the magnitude information by running the MBC over the total sum of consumption of each week per client. Like in the clustering of only one week we have only one feature, but in these case, we have a different line for each week, Table 5.5.

Table 5.5: Example of the input data to the magnitude clustering of separated weeks.

Client	Week	Sum
Client A	1	240.000
Client A	2	350.000
Client A	3	274.000
Client A	4	255.000
Client A	5	247.000
...	...	...
Client A	52	247.000

The results are pretty similar to the ones obtained by testing the MBC over only one week, which makes sense since the data set contains independent weeks grouped by their characteristics and not composite weeks of one year that are evaluated together.We obtain three clusters, Figure A.6, with different sizes and different magnitudes of consumption in its weeks. The cluster 3 has the weeks of more consumption, while cluster 1 has the weeks of less consumption. The size of clusters, Figure A.8, gradually increases from clusters 3 to 1, which shows that weeks with small magnitude are much more frequent while higher consumption are not so typical. At the same time, we realise that most clients have their weeks in more than one cluster, i.e. per year, clients tend to be in at least two different clusters of magnitude, Figure A.7. Then we create a stacked bar plot, Figure 5.13, that allows understanding the percentage of customers according to the number of weeks in a cluster. For example, in percentage and ignoring the fact that clusters have different sizes, how many clients in cluster 1 have their fifty-two weeks there? These are the questions we want to see answered with these graphs. It is possible to see that cluster 1, which is the lowest average consumption cluster, is a cluster where the customers typically have the fifty-two weeks in the same cluster, i.e. they are customers with similar consumption magnitudes over the weeks. While in cluster 3, the vast majority of customers fall there only in one week or so, which shows us that this is a cluster of outlier weeks rather than typical weeks. This analysis allows us to conclude that there are few customers with very high consumption over the whole year. Those high consumption weeks can be outlier weeks of a client that consume less or weeks of an irregular client that is always changing its consumption over the year.

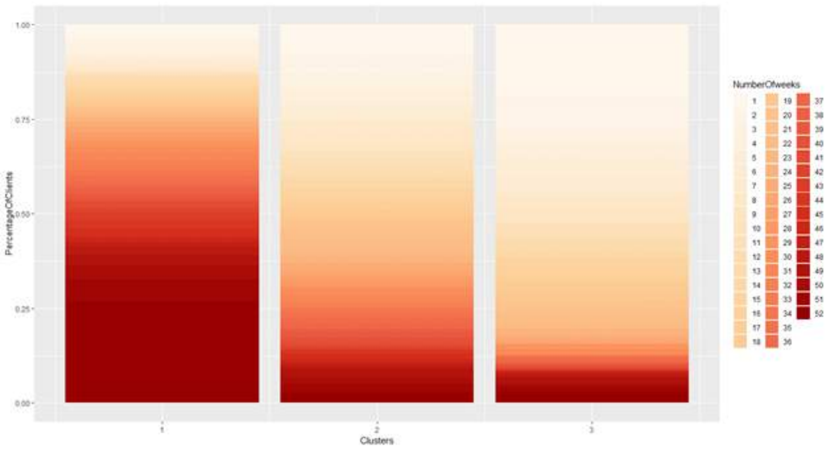


Figure 5.13: Stacked bar plot of the percentage of customers by the number of weeks per cluster. Clusters obtained by running the MBC over the absolute sum of consumption per week per client.

Finally, we create a table that allows the analysis of the changes from cluster to cluster for the weeks of a same client. To do this, we implement a small code that increments a counter when a client falls simultaneously on cluster x and y, Table 5.6. With this analysis, we can conclude that in terms of magnitudes, customers jump mainly between adjacent clusters; it is less common for customers with lowest-consumption (cluster 1) to have weeks outliers with the highest cluster consumption (cluster 3). Even the ones were weeks are in both cluster 1 and 3 they usually do not pass directly from one to the other without passing through the intermediate cluster (cluster 2).

Table 5.6: Number of clients sharing different pairs of clusters.

	1	**2**	**3**
1	-	667	196
2	667	-	241
3	196	241	-

The analysis of the Silhouette score in Table 5.7 shows that this approach has indeed better results than the approach with the composite features of each week. From values almost close to zero, we move to more than 0.5, remembering that this metric has a range from -1 to 1. It confirms the idea that this will be a better approach to solve our problem since customers have similar weeks between them, but customers with all their weeks with equal magnitudes are rare.

Table 5.7: Silhouette score of the clusters obtained by running the MBC over the absolute sum of consumption per week per client.

	Average Silhouette widths	Cluster sizes
Cluster 1	0.65367	33328
Cluster 2	0.48409	15022
Cluster 3	0.19464	3598
Mean	**0.5728**	

5.3.2 Time-Of-Day Tariff Periods Analysis

Secondly, we test the approach of *clustering separated weeks* with the time-of-day tariff periods features, i.e. the percentages of consumption by tariff period of each week per client, Table 5.8. As in the magnitude clustering, we also obtain nine clusters, Figure 5.14, similar to the ones we obtained when we ran the MBC over one week, Subsection 5.1.2. However, the patterns of consumption, Figure 5.15, have slightly different formats since we are looking into much more weeks in this case. As we can see it was able to find non spherical clusters of the huge amount of data. Moreover, MBC could even separate the clusters with the smallest differences which are the nearest clusters (red, green and blue) and the ones with most similar patterns (cluster 1, 2 and 3), Figure 5.15. In this stacked bar plot we can also observe apparently stranger formats like the ones of cluster 6, 7, 8 and 9 where the clients consume mostly on only one tariff period (cluster 6, 7 and 9) or where they don't have consume at all at any period (cluster 8).

Table 5.8: Example of the input data to the *time-of-day tariff periods* clustering of separated week.

Installation	**Week**	**Percentage of Consumption**		
		Peak	**Off-Peak**	**Half-Peak**
Client A	1	28.75%	25.00%	46.25%
Client A	2	23.64%	39.34%	37.00%
Client A	3	24.81%	36.86%	38.33%
Client A	4	25.88%	33.73%	40.39%
Client A	5	23.48%	31.98%	44.54%
Client A	6	21.48%	37.59%	40.93%
...	...	...	...	...

This approach resulted in the creation of more specific patterns in the clusters opposed to the three ones obtained in the *composite weeks* clustering in Subsection 5.2.2, where all weeks of a client were forced to belong to the same cluster and having the same profile of consumption, which might not match the correct representation. In the *magnitude analysis* (Subsection 5.3.1) some customers already jump from cluster to cluster and in the *time-of-day tariff periods* analysis we also found this behaviour. It happens since any small change in the routine of the week is enough to the client consume more or less at different tariff periods. We can observe, Figure A.11, that most of the clients have weeks

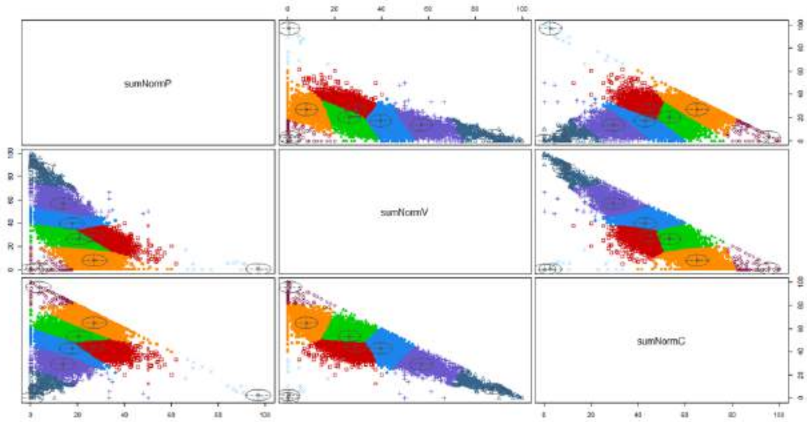


Figure 5.14: Clusters obtained by running the MBC over the percentage of consumption by tariff period per week per client.

in at least three different clusters and that it is uncommon that a client has the same pattern over the whole year, even if only changing between similar clusters as one and two.

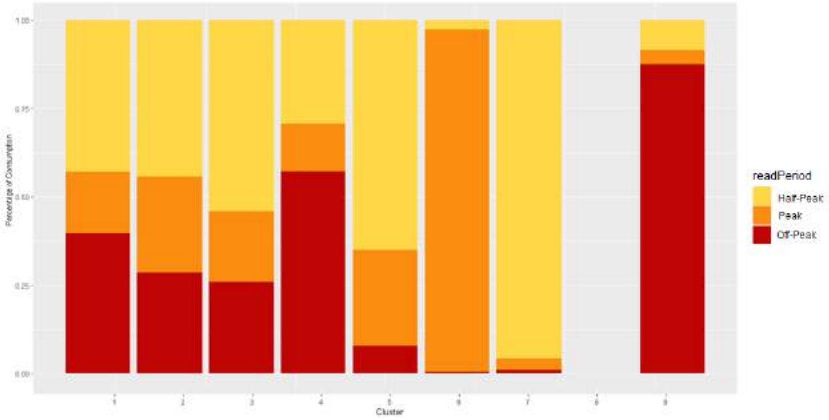


Figure 5.15: Percentage Stacked Bar Chart of the nine different patterns of consumption obtained by each cluster in A.9.

Moreover, analysing the results allow us to understand that there are patterns of consumption much more common than others. The bar plot, Figure A.10, shows that the patterns as the ones of clusters 1, 2 and 3 are much more common than the others. Typically the clients follow those types of patterns, with well-distributed consumption over the tariff periods, much more than the ones of the other clusters that have much less common patterns with higher consumption on only one period or with zero consumption in all periods.

Finally, we also analyse the percentage of customers according to the number of weeks in a cluster, Figure 5.16, as we have done to the magnitude. It allows an understanding that some clients of clusters 1, 5 and 8 have their fifty-two weeks in the same cluster. It shows us that those are the clusters of regularity while for example, cluster 6 typically contain the outlier weeks as it only has around one week per customer.

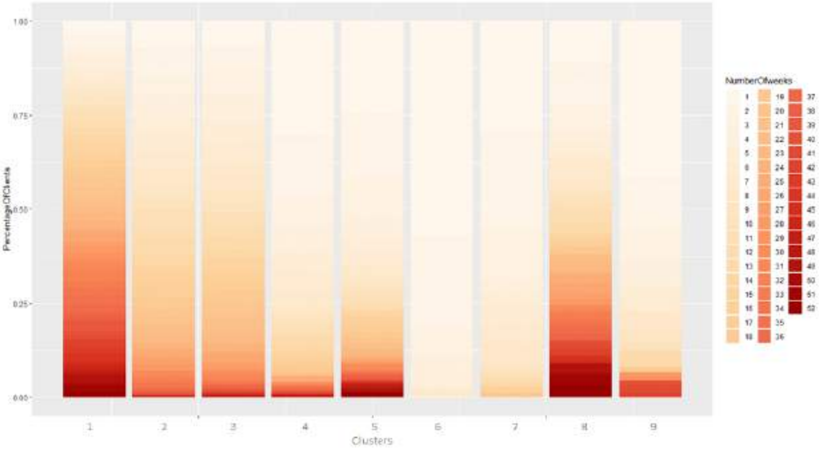


Figure 5.16: Stacked Bar Chart of the percentage of customers by the number of weeks per cluster. Clusters obtained by running the MBC over the percentage of consumption by tariff period per week per client.

In the *time-of-day tariff periods* analysis we also found a significant rise in the value of the silhouette score when using the separate weeks approach compared with the composite one. This analysis allows us to find the specific patterns of consumption of each week without forcing that all weeks of a customer belong to the same cluster.

Table 5.9: Silhouette score of the clusters obtained by running the MBC over the percentage of consumption by tariff period per week per client.

	Average Silhouette widths	Cluster sizes
Cluster 1	0.37036	23139
Cluster 2	0.42779	1920
Cluster 3	0.33656	16877
Cluster 4	0.39918	2937
Cluster 5	0.37177	2617
Cluster 6	0.74566	523
Cluster 7	0.53778	575
Cluster 8	1.00000	3234
Cluster 9	0.89812	126
Mean	**0.4093**	

In conclusion to this analysis, we confirm that we are effectively able to obtain much more specific information about the clients by using their weekly consumption independently rather than creating composite features over several weeks. This approach is more useful for our specific use case because it allows changing the type of energy efficiency tips and the consumption comparison itself throughout the year depending on the customer's consumption changes.

5.3.3 Comparison with other algorithms

After obtaining positive results by running the model-based over the features of *separeted weeks* of the *time-of-day tariff periods* (Subsection 5.3.2), we choose to try other algorithms since we cannot exclude other hypotheses only for getting results that seem to be good at the outset but may have some hidden problems. We start by comparing it with the DBSCAN, but the first problem we faced was that it was very manual since there is no general way of choosing the value of *minPts* neither *epsilon*. It is usually used to discover clusters with different formats and the parameters depends on what we want to find. We do several experiences in order to try to understand which are the best options.

We start with small values for these parameters, a *minPts* of 5 and an *epsilon* of 0.5. The results, Figure 5.17, were not good since it results in a vast quantity of clusters and lots of clients considered as noise. It happens because the size of *epsilon* seems too small for our data set, implying that all the points outside the central density were considered as noise. In our case, we want to have groups for all clients, even if it is in the group of the client's exception that has very different consumption. At the same time, since the *minPts* is also reduced, it creates more than three hundred small clusters without meaning.

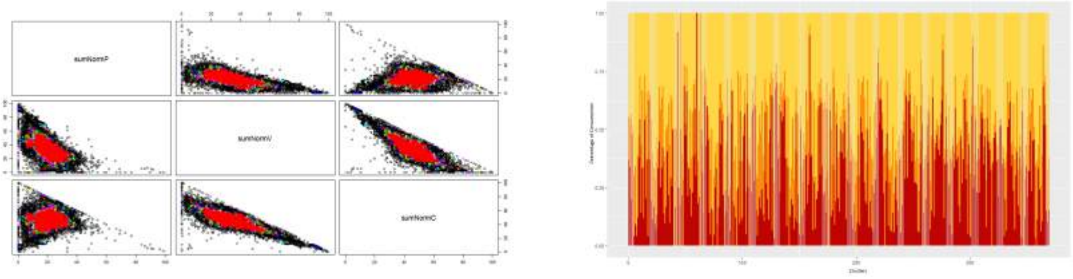


Figure 5.17: Clusters (left) and percentage Stacked Bar Chart of the different patterns of consumption (right) obtained by running DBSCAN, with *epsilon* of 0,5 and *minPts* of 5, over the percentage of consumption by tariff period per week per client.

Secondly, we increment the value of *epsilon* gradually in order to obtain a distribution of clusters without weeks considered as noise, we maintain a *minPts* of 5 but use a value of 15 for the *epsilon*. In this case, Figure 5.18, it only splits our data into the two most different clusters of the weeks with zero consumption and the others. This hypothesis is also not what we wanted since we want a more in-depth analysis than just identifying weeks of zero consumption. However, it can be helpful if EDP wants to capture clients in fraud or with problems in the smart meters since usually it is normal that all houses have a standby consumption.

By analysing the previous results, we understand that with DBSCAN we will always need to have noise, so we choose to use a smaller *epsilon* with a value of 2 but also a bigger *minPts* of 15 so that it does not build so much meaningless clusters of noise. The results, Figure A.12, are much better with these parameters since it considers much fewer points as noise than the initial approach, moreover, it also creates a number of clusters much

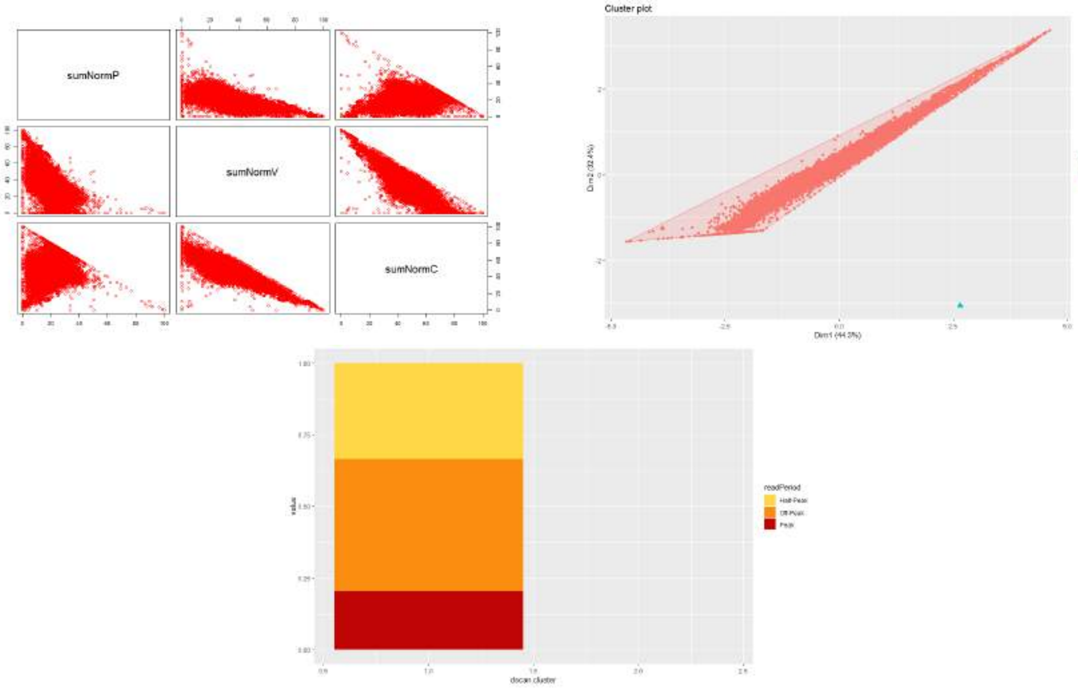
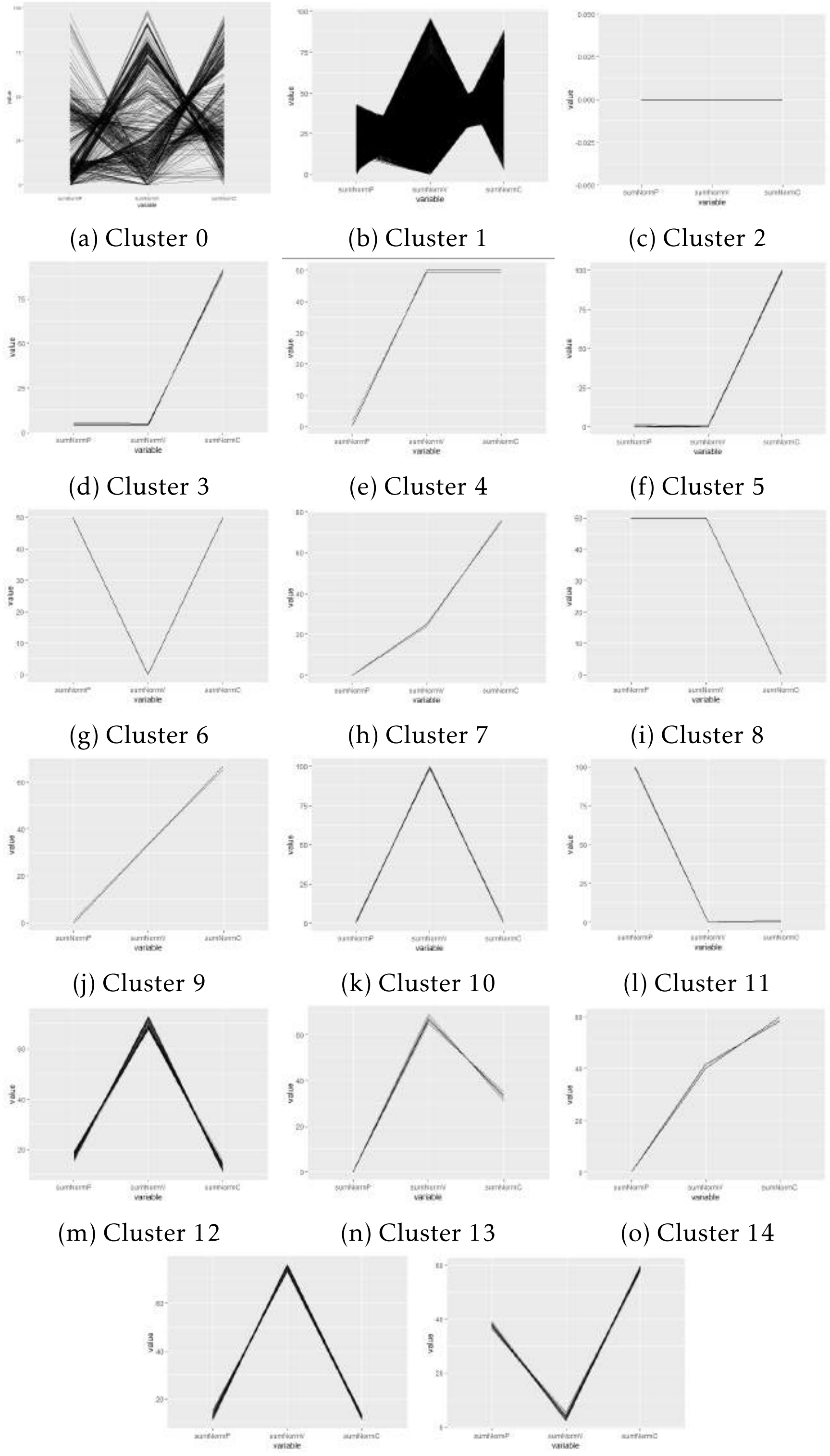


Figure 5.18: Clusters (top) and percentage Stacked Bar Chart of the different patterns of consumption (down) obtained by running the DBSCAN, with *epsilon* of 15 and *minPts* of 5, over the percentage of consumption by tariff period per week per client.

more acceptable. So it shows that this parameters are more suitable to our data set.

The advantages of DBSCAN are that they allow finding similar weeks with very particular patterns, Figure 5.19. The disadvantages are that besides the too small and specific clusters, it creates a massive cluster with all the other weeks. In our case, we want to use clustering in order to find groups to which we can generalize tips and at the same time to compare clients to the ones in their groups. So it does not make sense to have a perfect division and classification for less than one-quarter of the weeks and then in the other weeks have a huge generalist group without a specific pattern. Moreover, it is still to manual which mean that we need to adjust the parameter manually until we found something that make sense. The problem is that a lot of times, and like in our case, we don't know exactly what we are looking for. So DBSCAN is indeed good when looking structure in data with different densities and strange formats which do not happen in this application. These conclusions are supported by the Silhouette score, which suggests a reduction in clustering quality when we use this algorithm instead of MBC.

As explained in Section 4.2 we also experiment to use hierarchical clustering with the Euclidean distance and with the features of the percentages of consumption by tariff period of each clients' week. The first detected problem was the fact that it is computationally very complex to run the algorithm over so many rows of data and to draw the dendrogram. After running it, we obtained the dendrogram on Figure 5.20 where we can detect some hierarchies. The first problem is how to choose where to cut it to obtain the right number of clusters. First, we experiment the cut in the level where the dendrogram



(a) Cluster 0

(b) Cluster 1

(c) Cluster 2

(d) Cluster 3

(e) Cluster 4

(f) Cluster 5

(g) Cluster 6

(h) Cluster 7

(i) Cluster 8

(j) Cluster 9

(k) Cluster 10

(l) Cluster 11

(m) Cluster 12

(n) Cluster 13

(o) Cluster 14

(p) Cluster 15

(q) Cluster 16

Figure 5.19: Hairball plot of the seventeen clusters obtained by running the DBSCAN, with *epsilon* of 2 and *minPts* of 15, over the percentage of consumption by tariff period per week per client. In the x axis we can observe the time-of-day tariff period Peak, Off-Peak and Half-Peak respectively denoted by *sumNormP*, *sumNormV* and *sumNormC*. The y axis is the percentage of consumption.

Table 5.10: Silhouette score of the clusters obtained by running the DBSCAN, with *epsilon* of 2 and *minPts* of 15, over the percentage of consumption by tariff period per week per client.

	Average Silhouette widths	Cluster sizes
Cluster 1	-0.73717	562
Cluster 2	0.20689	46697
Cluster 3	1.00000	3234
Cluster 4	0.91558	20
Cluster 5	0.99942	415
Cluster 6	0.99709	375
Cluster 7	1.00000	117
Cluster 8	0.98426	17
Cluster 9	1.00000	33
Cluster 10	0.99569	88
Cluster 11	0.99752	141
Cluster 12	0.99951	107
Cluster 13	0.55231	40
Cluster 14	0.98104	42
Cluster 15	0.94442	20
Cluster 16	0.67192	23
Mean	**0.26763**	

seems to have the most significant differences, so we choose 5 clusters. However, it cannot find any specific patterns, as we can see on Figure 5.21, mainly because clusters are not naturally spherical as we already observed with the MBC. Secondly we experiment to use 9 clusters since it supposedly is a good number by the result of the MBC but it still results in clusters without any specific patterns, Figure A.13.

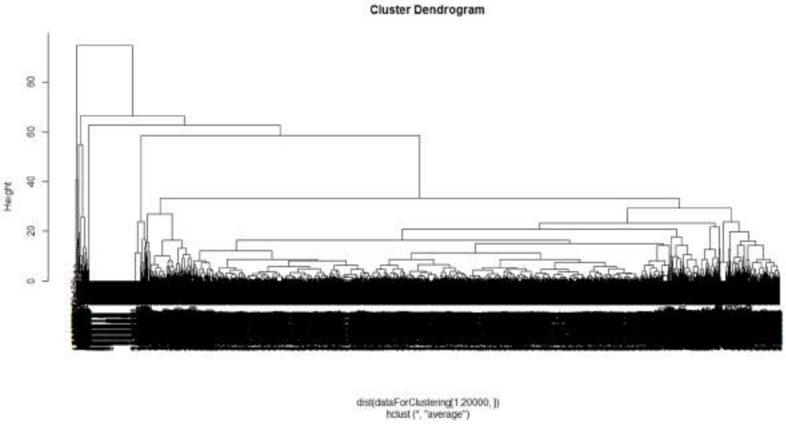
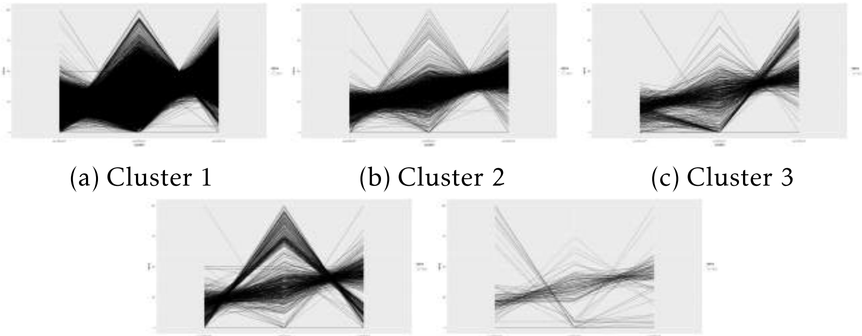


Figure 5.20: Dendrogram obtained by running the Hierarchical Clustering Algorithm over the percentage of consumption by tariff period per week per client.

By the analysis made, we can conclude that the MBC is the most suitable one to our application since it does a correct division of the similar patterns of consumption and allows us to create types of weeks. It shows us that our data is not naturally well-divided or have spherical clusters that could be easily found by using a prototype-based clustering. MBC can infer the distribution of the data and then divide them into groups which were



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

Figure 5.21: Hairball plot of the five clusters obtained by running the Hierarchical Clustering Algorithm over the percentage of consumption by tariff period per week per client.

what we need since we do not know a priori what the behaviour of the data of our project was.

5.4 Monthly and Yearly Clustering

By realizing that we obtained more complete results to this application by using the MBC with *separated weeks* in 5.3 rather than with composed in 5.2, we decided to apply this approach to a temporal granularity of months and the whole year. We applied both *magnitude* and *time-of-day tariff periods* analysis as we had done in the weekly analysis.

5.4.1 Magnitude Analysis

To perform the magnitude analysis, we consider the absolute sum of consumption (one feature), as showed in Table 5.11, for each *month of client* and *year of client*. More than knowing the scale of the clients in the several weeks we also want to have a coarse-grained analysis that allows seeing a broader vision of how the client consume through the year by the monthly analysis. At the same time, we also want to split the clients by the sum of consumption of the whole year, which allows understanding their general scale of consumption. It is a more similar categorization to the one made by the contracted power since it puts one client in only one cluster for the whole year.

The results of the monthly analysis are similar to the ones obtained by the weekly analysis since it also produces three clusters, Figure A.14, with the months with smaller, medium and large absolute consumption. Even the silhouette score, in Table 5.12, is similar to the one obtained by running the MBC over the weekly data, Section 5.3. The first two clusters contain almost the same quantity of months while the third contains fewer months, Figure A.15.

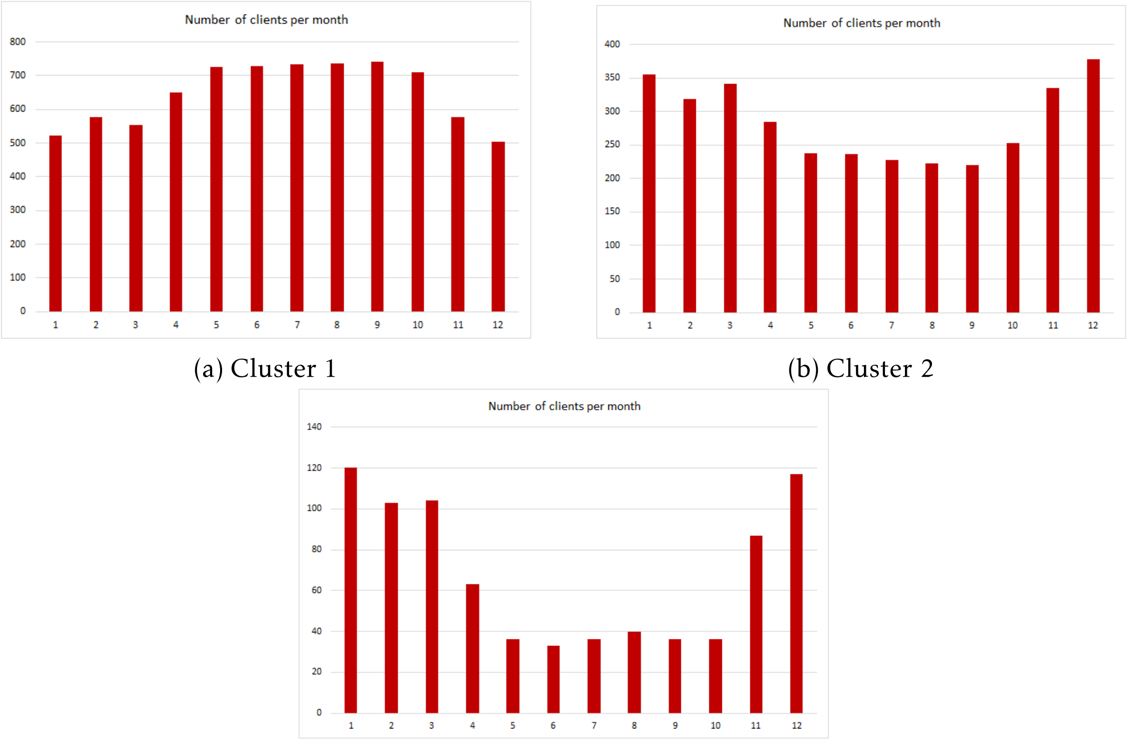
By analysing in more detail the number of clients per month of each cluster, Figure 5.22, we confirm the idea that a significant part of the clients consumes more than their normal level of consumption on the winter months. The cluster of higher consumption,

Table 5.11: Example of the input data to the magnitude clustering of separated month and years. The sum values in this case are artificial and just to exemplify. The structure of the table is ready for several years although we currently analyze only one year.

Client	Month	Sum
Client A	1	440.000
Client A	2	550.000
Client A	3	474.000
Client A	4	455.000
Client A	5	447.000
...	...	...
Client A	12	447.000

Client	Year	Sum
Client A	2018	7440.000
Client A	2017	7800.000
Client A	2016	7474.000
Client A	2015	7455.000
...	...	...

cluster 3 , contains much more clients in the winter months while in the smaller, we observe the opposite. It shows that the clients tend to consume less in the summer months and jump to the higher consumption clusters in winter.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

Figure 5.22: Number of clients per month obtained by running the MBC over the absolute sum of consumption of each month of 2018.

By applying the algorithm to the absolute sum of the consumption of one year, it results in six clusters, Figure 5.23. It happens because this value takes into account much more data condensed in only one feature which implies that each client only belongs to one cluster all year long even when it has irregular consumption with changes of scale

Table 5.12: Silhouette score of the clusters obtained by running the MBC over the absolute sum of consumption of each month of 2018.

	Average Silhouette widths	Cluster sizes
Cluster 1	0.64601	7737
Cluster 2	0.49066	3437
Cluster 3	0.18811	814
Mean	**0.5704**	

over time. By looking to the silhouette score in Table 5.13 we observe a decreasing of the clustering performance. Caused by being coarse-grained, this turns out not being an accurate enough approach, however, in our case it may be interesting since it allows dividing customers into general annual groups similar to the contracted power groups, which allows customers to realize when they consume much more than their group's average annual consumption. To understand if the contracted power could be used as a measure of the magnitude of the clients or if we can confirm that it is not informative enough and that using it as a feature would bias our results, we produce the graphic on Figure 5.24. In it, we can observe that the lowest contracted powers are indeed more present in the lowest clusters and the higher on the highest ones but that the contracted powers are distributed across all clusters. It means that they are not informative enough to categorize clients' consumption since most clients have a contracted power based on their peak of consumption even if it happens only during one-hour of only one day of the whole year.

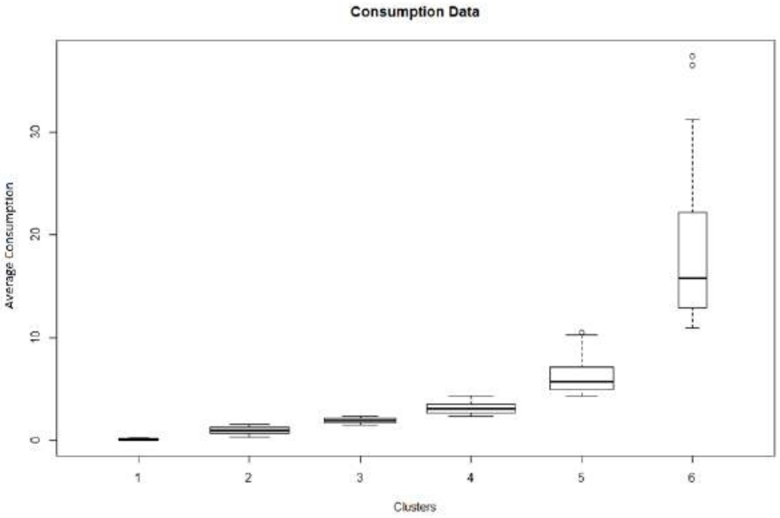


Figure 5.23: Box plot of the daily average consumption by year of the clients of each cluster obtained by running the MBC over the absolute sum of consumption of one year (2018).

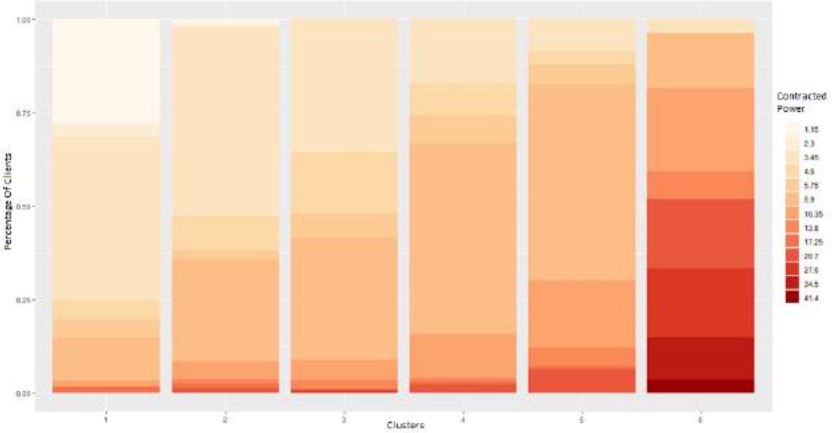


Figure 5.24: Percentage Stacked Bar Chart of the distribution of the contracted power per cluster.

Table 5.13: Silhouette score of the clusters obtained by running the MBC over the absolute sum of consumption of one year (2018).

	Average Silhouette widths	Cluster sizes
Cluster 1	0.87790	61
Cluster 2	0.33103	242
Cluster 3	0.64956	204
Cluster 4	0.39209	306
Cluster 5	0.35725	156
Cluster 6	0.21924	30
Mean	**0.4489**	

5.4.2 Time-Of-Day Tariff Periods Analysis

To analyse the different formats of consumption for each month and year, we use the percentage of consumption of each tariff period. First, we calculated these values for each month and ran the MBC over it, obtaining nine different clusters with the different formats of consumption of the months, Figure 5.25. It creates clusters similar to the ones obtained in the weekly analysis, which makes sense since they are aggregating the weekly information. It can be interesting for understanding, for example, if the clients consume later in vacancy months because they do not do the usual routine. Besides, having higher interval of data it can still find the detail of energy consumed in each tariff period as we can see by the high silhouette score in Table 5.14. The monthly analysis is useful to detect differences in the times of the year. Moreover, due to its greater granularity, it allows tips not focused on weeks, that may contain more outliers, but rather on monthly behavior. This can also be interesting for EDPD since it allows to not overload customers with loads of tips for every week.

The yearly analysis, with the percentage of consumption of each tariff period, becomes too broad to analyse profiles of consumption. This analysis produces clusters, Figure 5.26, with patterns much more similar between them since the difference of consumption are lost in the aggregation of the values. The silhouette score in Table 5.15 shows a decreasing

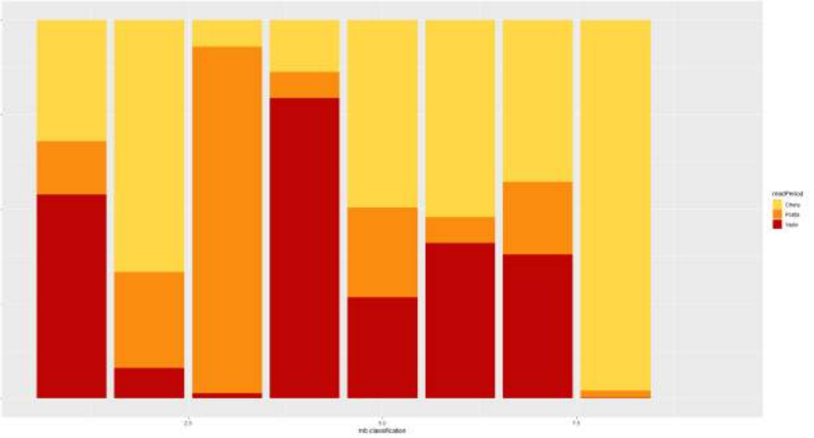


Figure 5.25: Percentage Stacked Bar Chart of the nine different patterns of consumption obtained by each cluster in A.16.

Table 5.14: Silhouette score of the clusters obtained by running the MBC over the percentage of consumption by tariff period of each month of 2018.

	Average Silhouette widths	Cluster sizes
Cluster 1	0.42874	464
Cluster 2	0.35215	1697
Cluster 3	0.76861	30
Cluster 4	0.46063	132
Cluster 5	0.38709	3397
Cluster 6	0.39210	517
Cluster 7	0.29807	5178
Cluster 8	0.86443	75
Cluster 9	1.00000	498
Mean	**0.3757**	

in its value similar to the one of the yearly *magnitude* analysis. However, in the magnitude, these aggregated analysis are still desirable to understand the quantity of consumption even for significant periods, but in the format, we lose some specific format information by putting everything in one single variable for each month. After this analysis, we can conclude that the magnitude analysis could be made in several time-intervals allowing to categorise the client's consumption over several perspectives but that the *time-of-day tariff periods* analysis loses meaning with higher granularity.

5.5 Chi-Square Test

In our specific case, we think of the chi-square test as a way to help EDPD get to know its customers and characteristics better. As mentioned earlier, it is quite easy to conceive a priori some ideas about customer consumption, for example, that customers in the city consume differently from customers in the more rural area. So this project aimed to be able to do an unbiased consumer analysis that would allow us to understand which socio-economic characteristics of customers affect their consumption. Moreover, we want

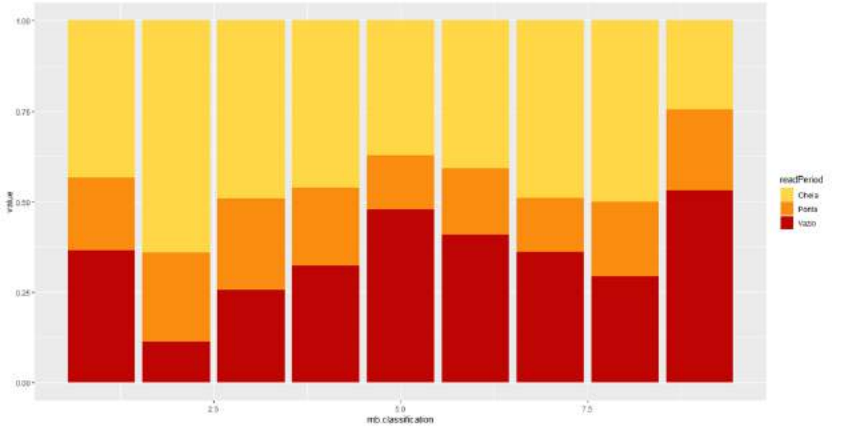


Figure 5.26: Percentage Stacked Bar Chart of the nine different patterns obtained by running the MBC over the percentage of consumption by tariff period of one year (2018).

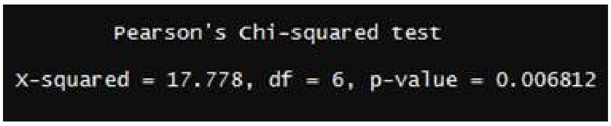
Table 5.15: Silhouette score of the clusters obtained by running the MBC over the percentage of consumption by tariff period of one year (2018).

	Average Silhouette widths	Cluster sizes
Cluster 1	0.27151	174
Cluster 2	0.35222	74
Cluster 3	0.03799	196
Cluster 4	0.39329	165
Cluster 5	0.11469	89
Cluster 6	0.51003	117
Cluster 7	0.19395	59
Cluster 8	0.48279	90
Cluster 9	-0.32508	35
Mean	**0.25931**	

to understand how consumption relates to the groups already used by EDPD as the contracted power.

These analysis would be more interesting with more specific socio-economic attributes such as the family income, the social class, the household constitution and others. This information would allow us to understand even more the customer characteristics to better-focused tips. However, as explained in Section due to energy sector laws, EDPD only has access to consumption data and geographical location. Therefore we can only perform analysis based on these two types of variables which already help us a lot in understanding the distribution of customers.

First, we start by analysing how the topology of urban areas could affect the regularity of consumption, i.e. if, in the rural areas, the consumers are more or less regular compared to urban areas. So we test whether the coefficient of variation concerning the consumption, (5.2.3) depends on the type of areas (rural, semi-urban and urban) or if these two variables are independent. To define the different population zones, we use the report [31]. Then we use these zones and the results of the MBC with the coefficient of variation as input to the chi-square test. The obtained results, in Figure 5.27, return a p-value less than 0.05, and therefore we can reject the null hypothesis that the variables are independent. In other words, we can conclude that the type of urban areas and consumption depend on each other. By analyzing the contingency table in more detail (not shown here due to its uncomfortable reading), we observe that consumption is more irregular in rural areas. It probably happens because in the city people have much more routine while the country life is more irregular since it is affected by the periods of the year and weather conditions.

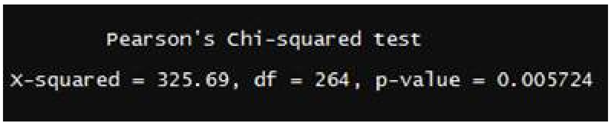


```
        Pearson's Chi-squared test

X-squared = 17.778, df = 6, p-value = 0.006812
```

Figure 5.27: Chi-square test for the coefficient of variation and topology of urban areas (rural, semi-urban and urban).

Then, since the population density must depend on the municipalities, we decided to confirm the hypothesis that the division by municipalities would also affect the regularity of consumption. So we test the hypothesis of independence of the 88 existing municipalities in our sample, of the 308 in Portugal, and the groups of regularity resulting from the clustering with the coefficient of variation. The *p*-value, in Figure 5.28, is also less than 0.05, and so we confirm that the regularity of consumption depends on municipalities.

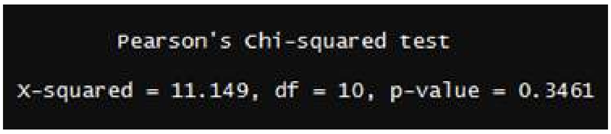


```
        Pearson's Chi-squared test

X-squared = 325.69, df = 264, p-value = 0.005724
```

Figure 5.28: Chi-square test for the coefficient of variation and municipalities.

We then decided to test other consumption results, and so we ran the chi-square test with the topology of urban areas and the resulted clusters on annual consumption magnitudes analysis (Subsection 5.4.1). Opposite to the regularity, in this case, we cannot conclude that the hypothesis of independence can be rejected since the *p*-value is superior to 0.05 as we can see in Figure 5.29. It shows us that the magnitudes of consumption do not depend on the type of region (urban, semi-urban and rural).

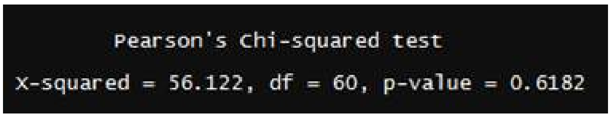


```
        Pearson's Chi-squared test

X-squared = 11.149, df = 10, p-value = 0.3461
```

Figure 5.29: Chi square test for the annual magnitude of consumption and population density.

Finally, we think that at the level of magnitude, it will be interesting to understand its relationship with the contracted power. The contracted power is a way of grouping clients widely used by EDPD but which does not always correspond to the actual consumption taken by the customers as we see in Figure 5.24. Hence the interest of confirming whether the contracted power depends on the annual magnitude clustering results or if they are independent. It returns a *p*-value of approximately 0.6, Figure 5.30, and so, since 0.6 is clearly greater than 0.05 we strongly suspect that the variables are independent which confirms that the contracted power is not a good indication of customer consumption magnitudes.



```
        Pearson's Chi-squared test
X-squared = 56.122, df = 60, p-value = 0.6182
```

Figure 5.30: Chi square test for the coefficient of variation and population density.

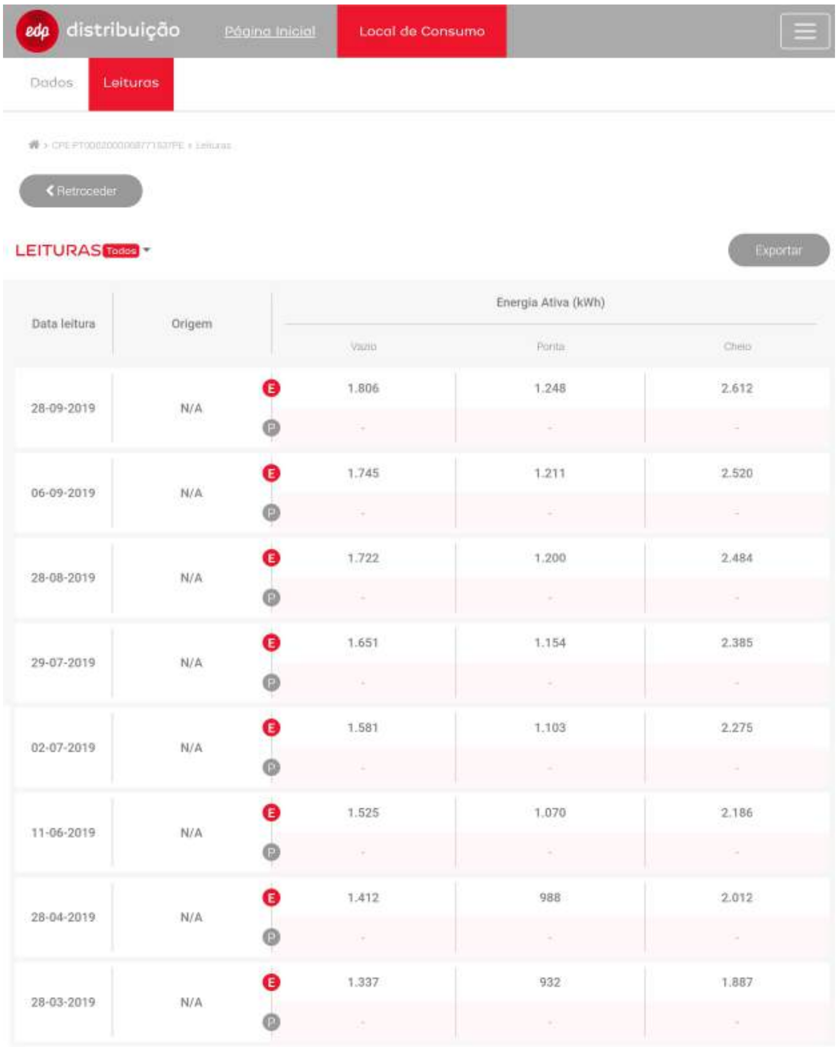
5.6 Practical application to EDP's use case

This project contributes to a comparative analysis of several clustering techniques and different approaches to the consumption data set obtained from smart meters. However, more than the scientific study, the project fits into a real use case within the scope of EDPD's work that will bring benefits to both clients and the company.

The first applicability of this dissertation is that it allows EDPD to understand better the types of consumption of its customers and the way they are distributed. This new knowledge allows better management of the distribution network, which was previously made only based on fixed values as the contracted power. As we can see from the chi-square analysis, the contracted power is not a rating that reflects the actual consumption of the customers and is therefore not enough indicative of the magnitude of consumption. This results may indicate that many customers do not have the most appropriate contracted power, which can reflect in *circuit breaker trip* or an oversized of EDPDnetwork caused by customers with high contracted powers but with low consumption.

Secondly, besides not having a business relationship with the clients, since it is done through the traders, EDPD as distribution network operator is responsible for delivering energy to all clients and for contributing to higher energy awareness. With this in mind, they created a personal area on the website where customers can request network connection, report readings and report anomalies and malfunctions. Moreover, it already allows seeing the history of the reported readings, Figure 5.31. However, in the future EDPD intends also to show consumption history, based on smart metering data, consumption comparisons and focused energy efficiency tips.

In this Section, we show the practical application of the clusters obtained in the previous Sections to the real use case. To do so, we choose random specific clients so that we can analyze which clusters they belong to in the different approaches and understand what information could be made available for this type of clients. For the sake of interest we show only the main conclusions that we could take of this analysis. Currently, this information is not yet being made available on the EDPD website, but this is an analysis



Data leitura	Origem		Energia Ativa (kWh)		
			Vazio	Ponta	Cheio
28-09-2019	N/A	E	1.806	1.248	2.612
		P	-	-	-
06-09-2019	N/A	E	1.745	1.211	2.520
		P	-	-	-
28-08-2019	N/A	E	1.722	1.200	2.484
		P	-	-	-
29-07-2019	N/A	E	1.651	1.154	2.385
		P	-	-	-
02-07-2019	N/A	E	1.581	1.103	2.275
		P	-	-	-
11-06-2019	N/A	E	1.525	1.070	2.186
		P	-	-	-
28-04-2019	N/A	E	1.412	988	2.012
		P	-	-	-
28-03-2019	N/A	E	1.337	932	1.887
		P	-	-	-

Figure 5.31: Visual of the website as it is right now in the Section of the historical reported readings.

of several ways of showing the comparative of consumption and the possible energy efficiency tips.

The first analysed client belongs to the cluster 1 of the coefficients of variation, which means that it is a regular client through the months. Clients with a low coefficient of variation do not need tips focused on the season. Moreover, they could have the same type of tips over the whole year without the need to review their group and type of consumption seasonally. This client falls in the same cluster of magnitude through all months of the year since it is a client with a small daily average consumption. So besides being always possible to increase energy efficiency, this is a client with small and consistent consumption over the year and not the most worrying case. To a client like this we could show, Figure 5.32, its average daily consumption by month compared to the annual average consumption of the clients in its annual magnitude cluster. Moreover, we can conclude he has a right contracted power since it is in the smallest clusters and contracting the smallest power possible, 3.45. In terms of format the months of the client change between two main clusters the 5 and the 7 in Figure 5.25. As we can see they are similar with just the slight difference that in 7, the client consumes more on off-peak. Since the client already does this shift of consumption in a few months, we can advise him to do it more often, if possible. This tip can be presented with an explanation that teaches customers that off-peak consumption allows better grid management. If all customers do most of their consumption at peak hours, the EDPD network will have a reduced dimension for the consumption of all clients simultaneously in that short period. So it is much better if the clients opt for spreading their consumption through time, for example, like the washing machines that probably do not need to be turned on at a specific time like the stove to cook.

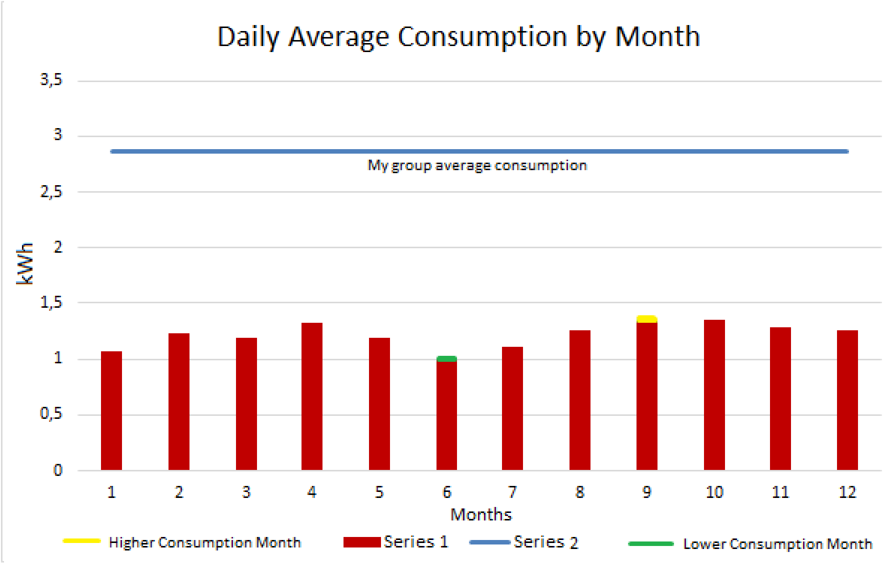


Figure 5.32: Example of what could be showed to a small regular client. Series 1 represent the consumption of the client measured through the smart meter while Series 2 represent the daily average consumption of its yearly group.

The second analysed client also has a small magnitude of consumption over the whole year; however, it is not regular as we can see from the fact it belongs to the cluster 3 of the coefficient of variation, Figure 5.11. It is an irregular client through the year despite always falling into the same cluster of magnitude. The comparative presented to this customer, Figure 5.33, makes him realize which months he consumes above the average daily consumption of his yearly group of similar consumers - the months when he probably should pay more attention to energy efficiency. Contrary to the first client, to this one, it makes sense to change the type of energy efficiency tips over the year. To this client, we can focus on the tips related to room heating since its consumption in winter months goes over its average annual group consumption. Tips to this client could be, for example, to isolate doors and windows of the home and to only heat the rooms being used. Nevertheless, this is still a client without much impact since it does not have a large consumption and always remains within the cluster of small consumers.

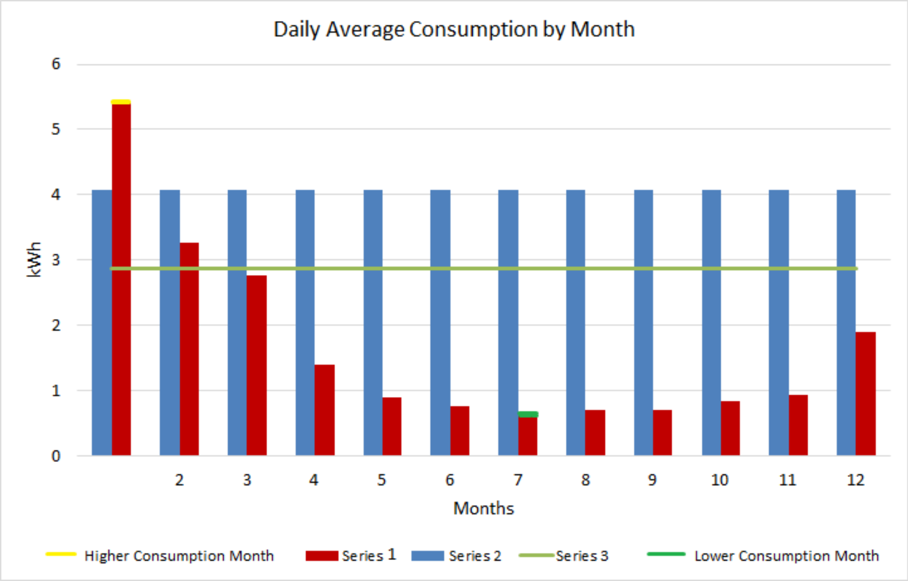


Figure 5.33: Example of what could be showed to a small irregular client. Series 1 represent the consumption of the client measured through the smart meter, Series 2 represent the daily average consumption by month of its monthly similar clients and Series 3 the daily average consumption of its yearly group.

The third client falls in the same cluster of the coefficient of variation as the second client. However, the changes in magnitude are much more significant since it shifts from the lowest cluster of magnitude to the higher. Climatization is one of the leading causes of high consumption in the colder months. In this case, the client probably does not do an efficient climate of the house, or use appliances with low efficiency and high consumption. To this client, we should reinforce the tips passed to the second client and also advise a review of appliances: "Opt, if possible, for air conditioners instead of traditional heaters like oil heaters". Moreover, in the comparative of consumption, Figure 5.34, we should alert the client for its changes of magnitude and highlighting the months that fall into the large consumers' cluster, as well as showing the daily average of its annual group and the daily averages of its monthly groups, obtained in 5.4.

5.7 Summary

To facilitate the understanding of the several exploratory analysis performed throughout the dissertation and their results, we have created Table 5.16, where it is possible to compare the results of the principal analysis made. Moreover, it works as an index to the several Sections of the dissertation allowing the reader to know immediately where each analysis is.

By analyzing the silhouette scores in the table, we can see that the first approaches that use many composite features, even when using feature reduction techniques, are worse

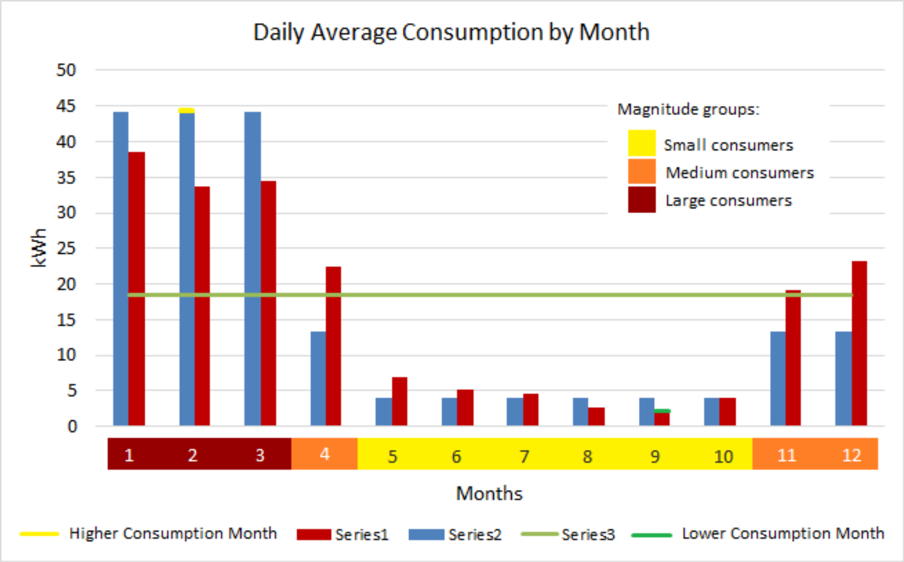


Figure 5.34: Example of what could be showed to an irregular client with different magnitudes of consumption over the year. Series 1 represent the consumption of the client measured through the smart meter, Series 2 represent the average consumption by month of its monthly similar clients (obtained by the average consumption from Figure A.14) and Series 3 the daily average consumption of its yearly group (obtained by the average consumption from Figure 5.23).

than the rest. From the moment we use separate and more specific features, the silhouette score starts to improve. This shows us that using features focused on the information we want to detect allows much more accurate results. However, all experiments were useful as they provided a comparison between the several approaches. Furthermore, from EDPD's point of view, it should be noted that each approach has different applicability in its use case.

Table 5.16: Summary of the principal experiences made through this dissertation and the respective silhouette score.

Approach	Algorithm	# Clusters	Score	Section
Composite Weeks:				5.2
Magnitude	MBC	3	0.02845	5.2.1
Time-Of-Day Tariff Periods:	MBC	3	0.08183	5.2.2
- PCA with 86 components	MBC	3	0.04434	5.2.2.1
- Autoencoders:				5.2.2.2
Sigmoid	MBC	9	-0.16920	3.2.3
Tanh	MBC	9	-0.05594	3.2.3
Regular and Irregular:				
Coefficient of Variation	MBC	4	0.5152	5.2.3
Separated Weeks:				5.3
Magnitude	MBC	3	0.5728	5.3.1
Time-Of-Day Tariff Periods	MBC	9	0.4093	5.3.2
Time-Of-Day Tariff Periods	DBSCAN	16	0.26763	5.3.3
Monthly Analysis:				5.4
Magnitude	MBC	3	0.5704	5.4.1
Time-Of-Day Tariff Periods	MBC	6	0.4489	5.4.1
Yearly Analysis:				5.4
Magnitude	MBC	9	0.3757	5.4.2
Time-Of-Day Tariff Periods	MBC	9	0.25931	5.4.2

Chapter 6

Conclusion

This chapter points out the main conclusions of the work in Section 6.1, and presents the future work in Section 6.2.

6.1 Conclusions

We analysed several clustering algorithms applied to the smart metering data in order to understand which is the best one to our application. We started without known classes and characteristics on our data set, and the purpose was to find a structure in the data and understand the natural groups present in them. Since we want the most unbiased analysis possible, we choose to use the features appropriate to each approach but without forcing any pre-conceptions about the clients. By analysing the clustering algorithms, we understand that in this case, we do not have spherical clusters easy to find by using an algorithm based on the Euclidean Distance, and that DBSCAN is not also a good choice since we do not have clusters with bizarre formats of different densities. We ended up concluding that the MBC proved to be the best approach to our data set since it tries to find the natural distribution of our data.

We analysed the clients over several perspectives, magnitude and pattern, but also over several time intervals, weekly, monthly or yearly and also analysed them based on the regularity of consumption. It made us conclude that we can indeed find patterns of consumption profiles among the clients, which allows understanding their distribution of consumption through the tariff periods of the day. Moreover, it allows splitting clients into magnitude groups much more accurate than the contracted power to understand the consumer size. Finally, the regularity was helpful to understand if the client needs a review of its group of consumption over the year and if focused tips in specific periods of the year are needed. Chi-square test was useful since it suggests that the contracted power is not the best classification of the groups of clients.

By aggregating the information obtained of all these perspectives of clustering, we created an analysis of each client the more complete possible that allows personalising the given Energy Efficiency tips. Moreover, the several clustering algorithms based on magnitudes allow the creation of a comparison between each client consumption and its cluster consumption over the year. This perspective is useful for the client to understand if he is consuming more than expected and what are the changes over the year.

One of the phases that most delayed the project was data extraction as we are using sensitive data requiring access permissions. In addition, EDPD receives a considerable amount of information from its customers daily and so realising which data could be of interest for energy efficiency analysis was also quite complex. As we mentioned earlier, the biggest challenge of this dissertation was doing a scientific study applied to a business use case. It implies the answer to the specific needs and timings of the company. However, this was also the most significant advantage as it allowed to create a project with high practical applicability in the EDPD business context.

6.2 Future Work

Our work still has some limitations, some related to the available data and some due to time constraints. These lead to the proposal of future improvements.

First of all, one of the things limiting the accuracy of the algorithm is the untrusted data. This data problems come from operational failures that should be analysed in the future. EDPD has data on operational services in the Service Orders documents. This documents contain information on smart meter issues, direct connections to the network and other operational issues that justify some data being incorrect. By considering these factors when using the data it can improve the accuracy of our algorithm. Secondly, we can gradually scale this analysis to more clients as to the clients with higher voltages that we are not considering right now. Moreover, in this dissertation, we only analyse one year of data but it is possible to apply this to more historical data and continuing analysing new data incoming. Scaling the analysis to more than one year.

This dissertation work has continuity since we have done an exploratory analysis of the Smart Metering Data with clustering algorithms, which brought a set of applications to Energy Efficiency started in the thesis but that still have work to do. The first need in the future work is analysing results to understand which tips make sense to each group of clients. We started that by analysing some specific clients of each cluster, but this is a task with still much work to do. The different approaches explored allow several conclusions and type of tips to be drawn. Moreover, the business (EDPD) will know better what the more appropriate tips based on the characteristics obtained on the clustering algorithms are. Moreover, by going more in-depth on the results with the several features and time-intervals, new ideas may come up that may make sense to present to customers on the page. One example of this is the separated weeks clustering that resulted in accurate clusters that EDPD considered useful to more in depth Energy Efficiency tip. Secondly, it will also make sense to automate all the process of the information made available to the customer. In other words, the website could automatically return focused tips and the comparative of consumption based on the client cluster.

The analysis of the clusters shows that even when we analyse more weeks, the number of clusters remain the same and the new weeks fall inside them instead of creating more clusters. This shows that the consumption have defined classes and that there is indeed visible patterns between them. Based on this, one of the main proposals for future work is to use classification algorithms based on the knowledge obtained in this dissertation. The idea is to train the classification model, based on the classes of clients obtained on the clustering phase. By learning this, it will be able to look at customer consumption and assign different consumers to the existing classes. The classification is more efficient than clustering in predicting classes when there is labelled data. The clustering, on the other hand, is useful to explore the structure of our data when we do not know which classes we have and want to understand if any natural clusters are standing out. However, it always needs the critical human sense to confirm its results, and with the increase of different data, it may become meaningless. Moreover it is computationally more complex and more susceptible to small changes and outliers. Therefore, the use of the two techniques complements each other and can lead to better prediction performance by first using clustering and then classification. The clustering algorithm should continue to be used to review existing classes, for example, once a year and see if they are still appropriate. However, the idea is that a classification algorithm is used in the day to day analysis to assign the new data coming from customers to the classes created by the clustering and approved by EDPD. When new consumption data arrives of each client, it is easier to assign it to one of the existing classes with its sharp energy efficiency tips than being always reviewing the types of groups and information given.

Finally, another proposal for future work is that by increasing the number of smart meters emitting load diagrams, more specific analyses can be created based on the time of day. These analyses will allow us to find more specific consumption characteristics, such as whether the consumer does almost all consumption at peak hours or has widespread consumption throughout the day. Cooking is the main use of energy in the house, as we can see in 2.1, and we still can not identify them due to the granularity of data. So with increasing data we could understand how to detect these type of uses in order to specialise tips of this type. This project aims to follow the technological development of smart meters by increasingly tuning the customer characteristics obtained from the consumption data and the tips given based on this information.

Bibliography

[1] INE/DGEG. *Inquérito ao Consumo de Energia no Sector Doméstico 2010*.

[2] I. E. Agency. *World Energy Outlook 2018*. 2018, p. 661. DOI: https://doi.org/https://doi.org/10.1787/weo-2018-en. URL: https://www.oecd-ilibrary.org/content/publication/weo-2018-en.

[3] A. Al-Wakeel and J. Wu. "K-means based cluster analysis of residential smart meter measurements." In: *Energy Procedia*. Elsevier, 2016. URL: https://www.sciencedirect.com/science/article/pii/S1876610216301308.

[4] D. Arden. *Applied Deep Learning - Part 3: Autoencoders - Towards Data Science*. 2017. URL: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798 (visited on 09/28/2019).

[5] C. Beard. *Smart Grids for dummies*. Chichester: Wiley, 2010.

[6] C. Beard. *Smart Metering for dummies*. Chichester: Wiley, 2010.

[7] C. Beckel, L. Sadamori, T. Staake, and S. Santini. "Revealing Household Characteristics from Smart Meter Data." In: *Energy* 78 (Dec. 2014), pp. 397–410. DOI: 10.1016/j.energy.2014.10.025.

[8] T. Cerquitelli, G. Chicco, E. D. Corso, F. Ventura, G. Montesano, M. Armiento, A. M. Gonzalez, and A. V. Santiago. "Clustering-based assessment of residential consumers from hourly-metered data." In: *2018 International Conference on Smart Energy Systems and Technologies, SEST 2018 - Proceedings* (2018). DOI: 10.1109/SEST.2018.8495863. URL: https://www.enelfoundation.org/content/dam/enel-foundation/news/Clustering-BasedAssessmentofResidentialConsumersfromHourly-MeteredData.pdf.

[9] D. Dua and C. Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[10] EDP Distribuição – Energia, S.A. *Eficiência Energética para a sua casa, Guia de boas práticas*.

[11] EDP Distribuição – Energia, S.A. "Controlador de Transformador de Distribuição (Distribution Transformer Controller – DTC) para instalação em Postos de Transformação MT/BT." 2014.

[12] EDP Distribuição – Energia, S.A. *Boas práticas para poupar energia.* 2018. URL: https://www.edpdistribuicao.pt/pt-pt/sustentabilidade/ser-eficiente/boas-praticas.

[13] F. Fahiman, S. M. Erfani, S. Rajasegarar, M. Palaniswami, and C. Leckie. "Improving load forecasting based on deep learning and K-shape clustering." In: *2017 International Joint Conference on Neural Networks (IJCNN).* 2017, pp. 4134–4141. DOI: 10.1109/IJCNN.2017.7966378.

[14] V. Figueiredo, F. Rodrigues, Z. Vale, and J. Gouveia. "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques." In: *Power Systems, IEEE Transactions on* 20 (June 2005), pp. 596 –602. DOI: 10.1109/TPWRS.2005.846234.

[15] K. Gajowniczek and T. Ząbkowski. "Data Mining Techniques for Detecting Household Characteristics Based on Smart Meter Data." In: *Energies* 2015 (July 2015), pp. 7407–7427. DOI: 10.3390/en8077407.

[16] N. Hargreaves, S. M. Pantea, A. Carter, and G. Taylor. "Foundations of a Metamodel Repository for Use With the IEC Common Information Model." In: *Power Systems, IEEE Transactions on* 28 (Nov. 2013). DOI: 10.1109/TPWRS.2013.2270302.

[17] N. A. Heckert, J. J. Filliben, C. M. Croarkin, B Hembree, W. F. Guthrie, P Tobias, and J Prinz. *Handbook 151: NIST/SEMATECH e-Handbook of Statistical Methods.* 2002. URL: https://www.itl.nist.gov/div898/handbook/prc/section2/prc242.htm.

[18] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks." In: *Science* 313.5786 (2006), pp. 504–507. ISSN: 0036-8075. DOI: 10.1126/science.1127647. eprint: https://science.sciencemag.org/content/313/5786/504.full.pdf. URL: https://science.sciencemag.org/content/313/5786/504.

[19] F. Iglesias and W. Kastner. "Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns." In: *Energies* 6.2 (2013), pp. 579–597. ISSN: 19961073. DOI: 10.3390/en6020579.

[20] K. H. Johansson. *Smart Infrastructures for a Sustainable City: Stockholm Case Studies | 2014 Future Cities Conference.* URL: https://pt.slideshare.net/futurecitiesproject/2014-future-cities-conference-joel-silveirinha-the-internet-of-everything-30676458/6.

[21] L. Krippahl. "Aprendizagem Automática (Machine Learning) - Lecture Notes." 2018.

[22] J. Kwac, C.-W. Tan, N. Sintov, J. Flora, and R. Rajagopal. "Utility customer segmentation based on smart meter data: Empirical study." In: Oct. 2013, pp. 720–725. DOI: 10.1109/SmartGridComm.2013.6688044.

[23] A. Lavin and D. Klabjan. "Clustering time-series energy data from smart meters." In: *Energy Efficiency* 8.4 (2015), pp. 681–689. ISSN: 15706478. DOI: 10.1007/s12053-014-9316-0. arXiv: 1603.07602.

[24] E. Lutins. *DBSCAN: What is it? When to Use it? How to use it. - Medium.* 2017. URL: https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818 (visited on 02/09/2019).

[25] G. Mauro. *I reverse-engineered a $500M Artificial Intelligence company in one week. Here's the full story.* 2017. URL: https://medium.com/startup-grind/i-reverse-engineered-a-500m-artificial-intelligence-company-in-one-week-heres-the-full-story-d067cef99e1c (visited on 02/07/2019).

[26] F. McLoughlin, A. Duffy, and M. Conlon. "A clustering approach to domestic electricity load profile characterisation using smart metering data." In: *Applied Energy* 141 (Mar. 2015). DOI: 10.1016/j.apenergy.2014.12.039.

[27] MissingLink.ai. *7 Types of Activation Functions in Neural Networks: How to Choose?* URL: https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/.

[28] T. Räsänen and M. Kolehmainen. "Feature-Based Clustering for Electricity Use Time Series Data." In: Springer, Berlin, Heidelberg, 2009, pp. 401–412. DOI: 10.1007/978-3-642-04921-7_41. URL: http://link.springer.com/10.1007/978-3-642-04921-7{_}41.

[29] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models." In: *The R Journal* 8.1 (2016), pp. 205–233. URL: https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf.

[30] S. Sharma. *Activation Functions in Neural Networks - Towards Data Science.* 2017. URL: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6.

[31] *Tipologia de Áreas Urbanas 2014.* Tech. rep. Sistema de Metainformação, Instituto Nacional de Estatística.

[32] A. Tureczek, P. Nielsen, and H. Madsen. "Electricity consumption clustering using smart meter data." English. In: *Energies* 11.4 (Apr. 2018). ISSN: 1996-1073. DOI: 10.3390/en11040859.

[33] Universidade EDP - Campus Online. "Introdução à Transformação Digital." 2019.

[34] Wang, Wei and Huang, Yan and Wang, Yizhou and Wang, Liang. *Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction.* 2014. URL: https://www.cv-foundation.org//openaccess/content_cvpr_workshops_2014/W15/papers/Wang_Generalized_Autoencoder_A_2014_CVPR_paper.pdf.

Appendix A

Auxiliary Figures

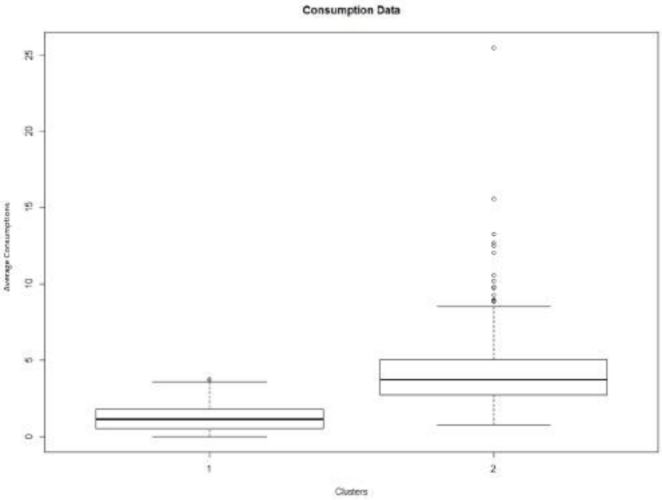
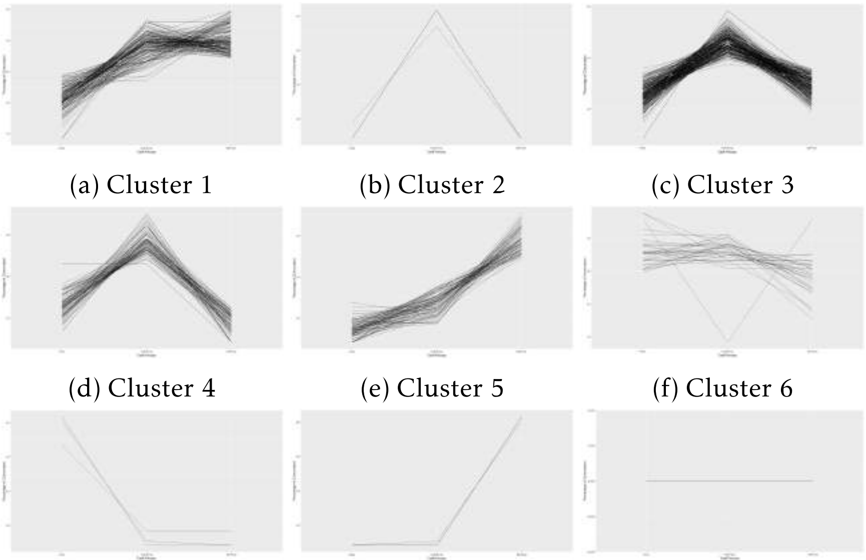


Figure A.1: Box plot of the average consumption of the clients of each cluster obtained by running the MBC over the absolute sum of consumption of consumption by tariff period of one week (3rd to 9th of March).



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

(g) Cluster 7 (h) Cluster 8 (i) Cluster 9

Figure A.2: Hairball plot of the nine clusters obtained by running the MBC over the percentage of consumption by tariff period of one week (3rd to 9th of March). The y axis is the percentage of consumption and the x the three tariff periods Peak, Half-Peak and Off-Peak.

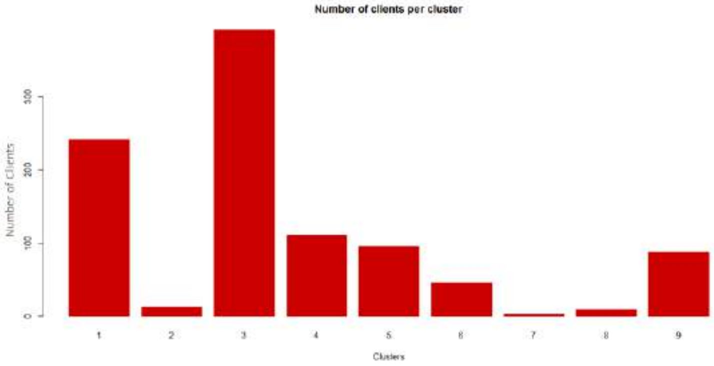
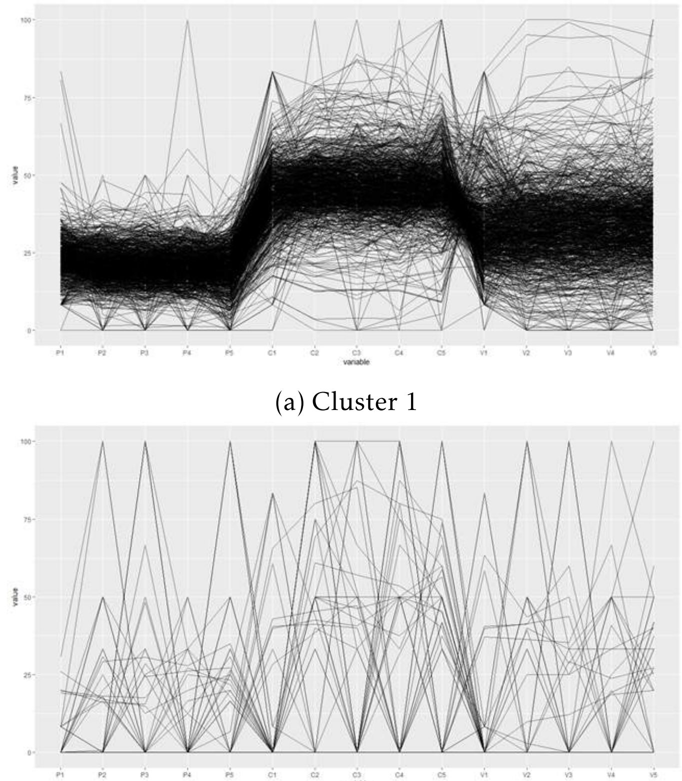


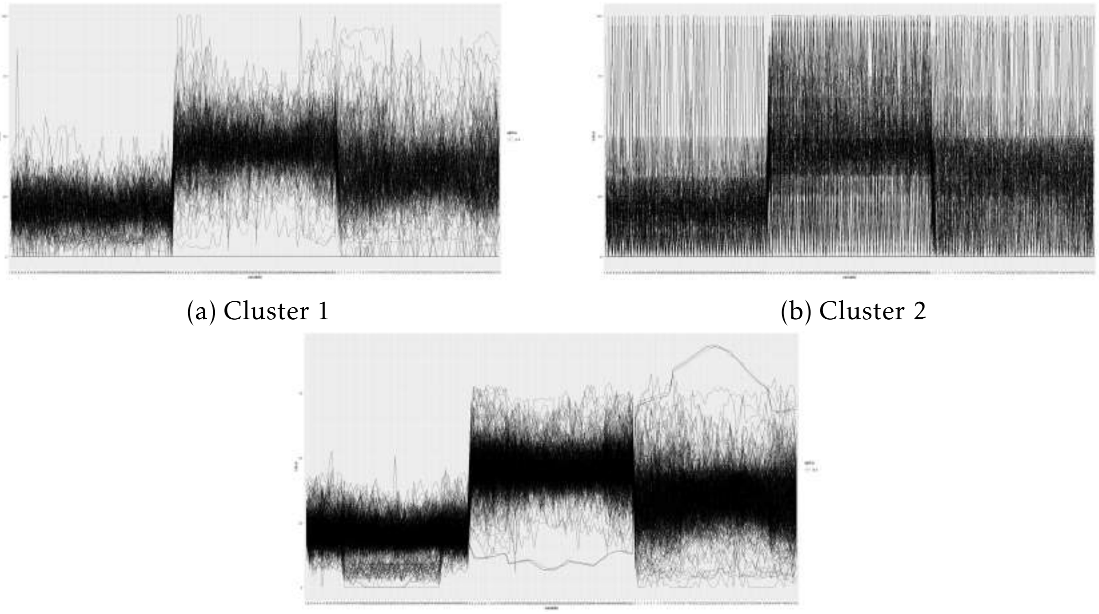
Figure A.3: Size of the clusters obtained by running the MBC over the percentage of consumption by tariff period of one week (3rd to 9th of March).



(a) Cluster 1

(b) Cluster 2

Figure A.4: Pattern of consumption by tariff period obtained by running the MBC over the percentage of consumption by tariff period over five weeks. Plot ordered by tariff period. In the x axis P, V and C represent respectively the three tariff periods Peak, Off-Peak and Half-Peak of the different weeks. The y axis represent the consumption value for each period.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

Figure A.5: Hairball plot of the clusters obtained by running the MBC over the matrix of the 86 principal components of the percentage of consumption by tariff period over the fifty two weeks of 2018. In the x axis it is represented the fifty two weeks of each tariff period in the order peak, half-peak and off-peak.

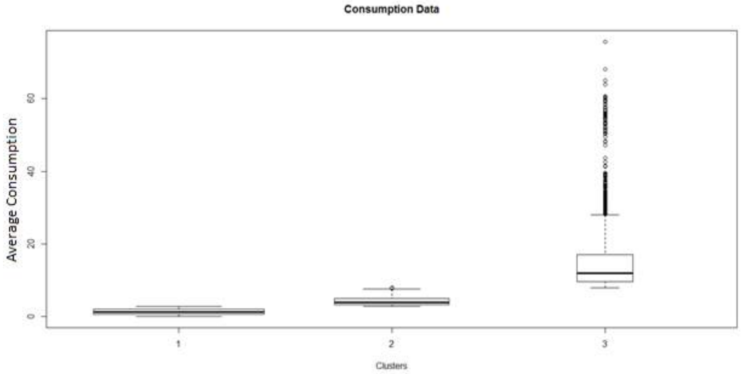


Figure A.6: Box plot of the average consumption of the clients of each cluster obtained by running the MBC over the absolute sum of consumption per week per client.

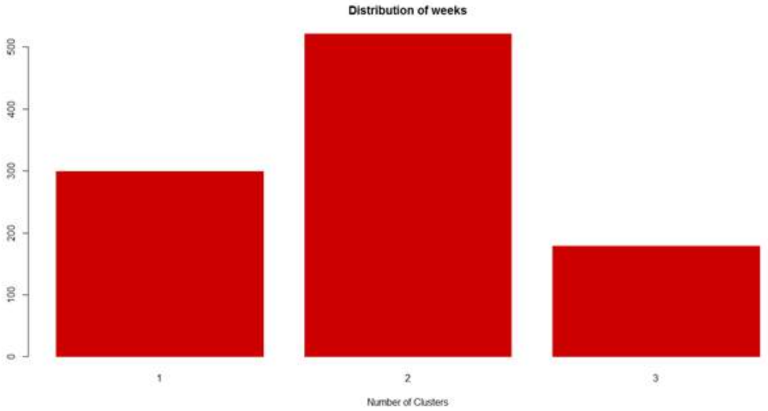


Figure A.7: Distribution of clients weeks per cluster, i.e. distribution of the number of clusters each client belongs to. Clusters obtained by running the MBC over the absolute sum of consumption per week per client. The y axis is the number of clients that have their weeks in only one, two or three clusters (x axis).

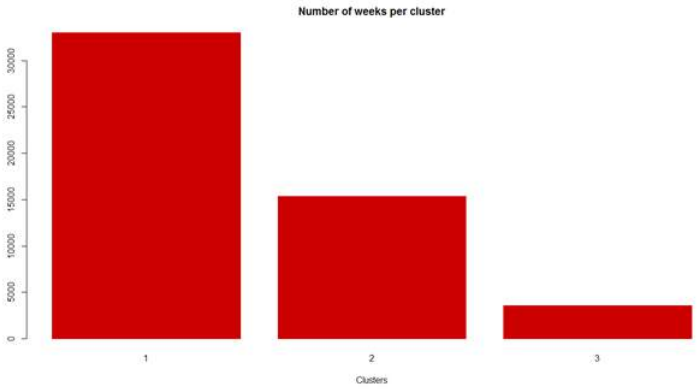
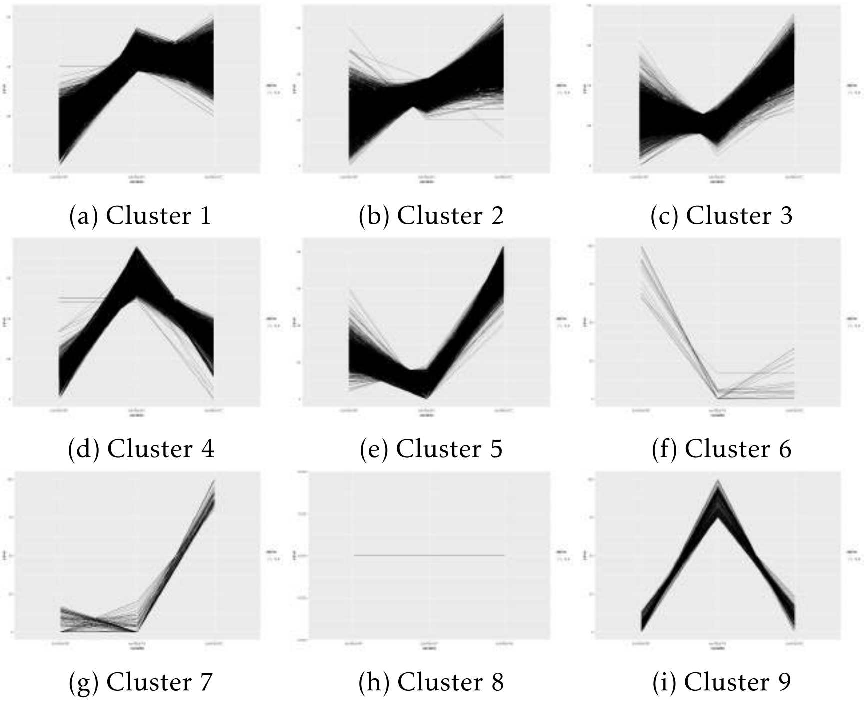


Figure A.8: Number of *weeks of clients* (y axis) in each cluster, i.e. size of clusters obtained by running the MBC over the absolute sum of consumption per week per client.


(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

(g) Cluster 7 (h) Cluster 8 (i) Cluster 9

Figure A.9: Hairball plot of the nine clusters obtained by running the MBC over the percentage of consumption by tariff period per week per client. The x axis is the percentage of consumption and the y the tariff period.

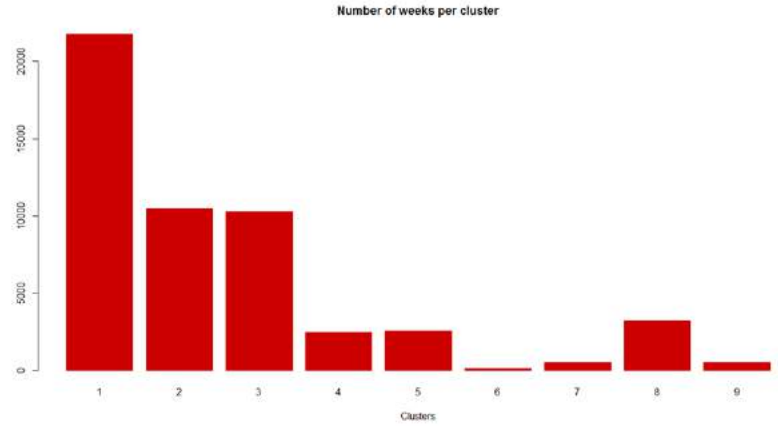


Figure A.10: Number of weeks (y axis) in each cluster, i.e. size of clusters obtained by running the MBC over the percentage of consumption by tariff period per week per client.

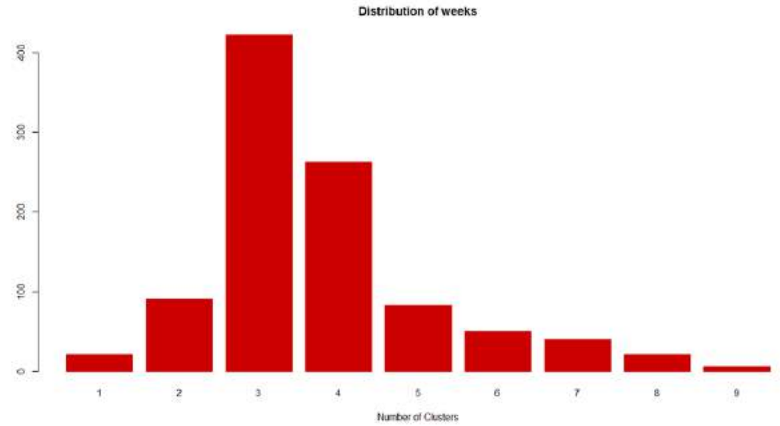


Figure A.11: Distribution of clients weeks per cluster, i.e. distribution of the number of clusters each client belongs to. Clusters obtained by running the MBC over the percentage of consumption by tariff period per week per client. The *y* axis is the number of clients and the *x* the number of clusters.

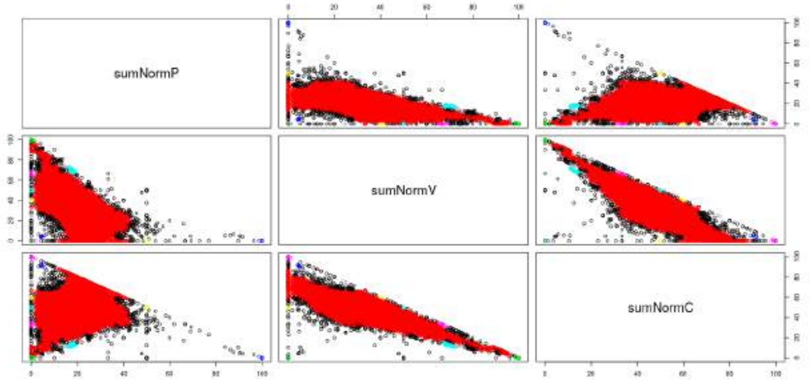
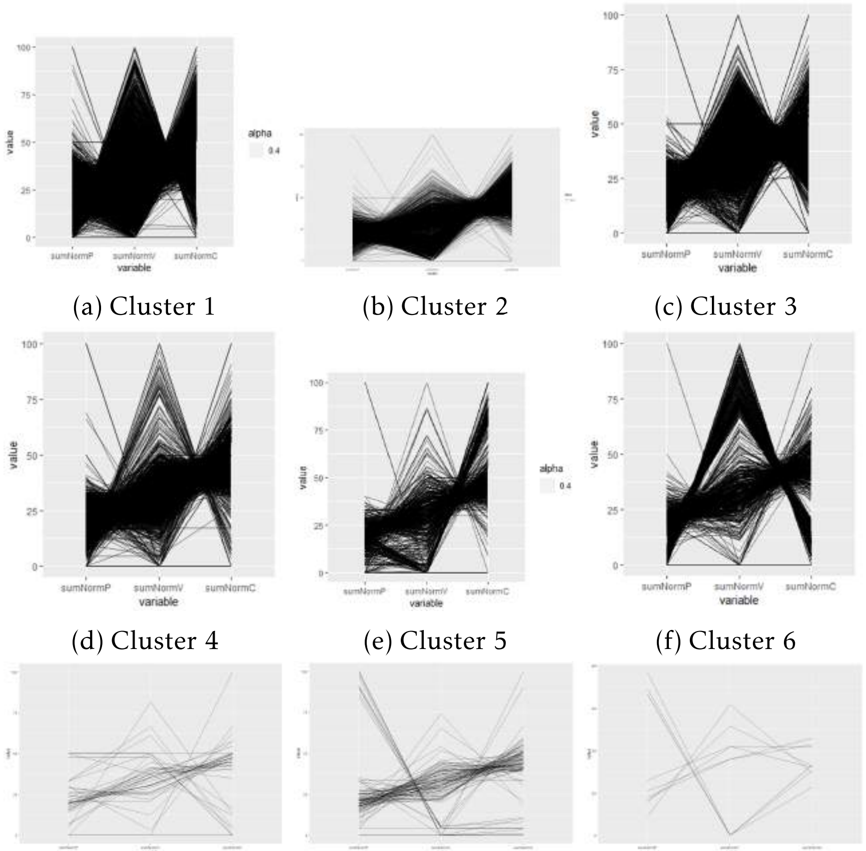


Figure A.12: Clusters obtained by running the DBSCAN, with *epsilon* of 2 and *minPts* of 15, over the percentage of consumption by tariff period per week per client.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

(g) Cluster 7 (h) Cluster 8 (i) Cluster 9

Figure A.13: Hairball plot of the nine clusters obtained by running the Hierarchical Clustering Algorithm over the percentage of consumption by tariff period per week per client.

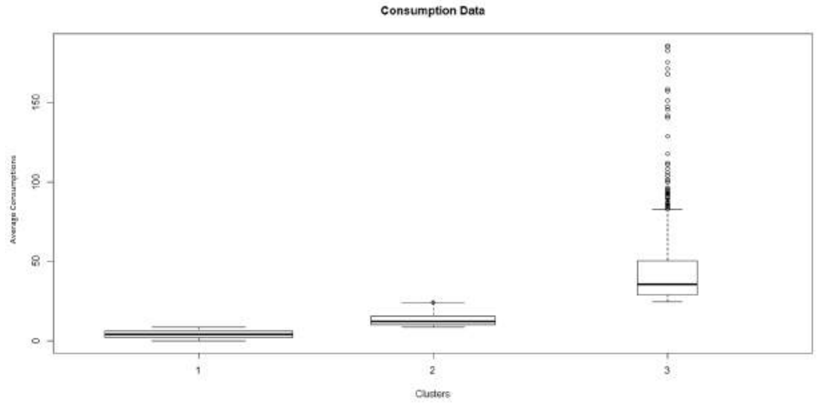


Figure A.14: Box plot of the average consumption by month of the clients of each cluster obtained by running the MBC over the absolute sum of consumption of each month.

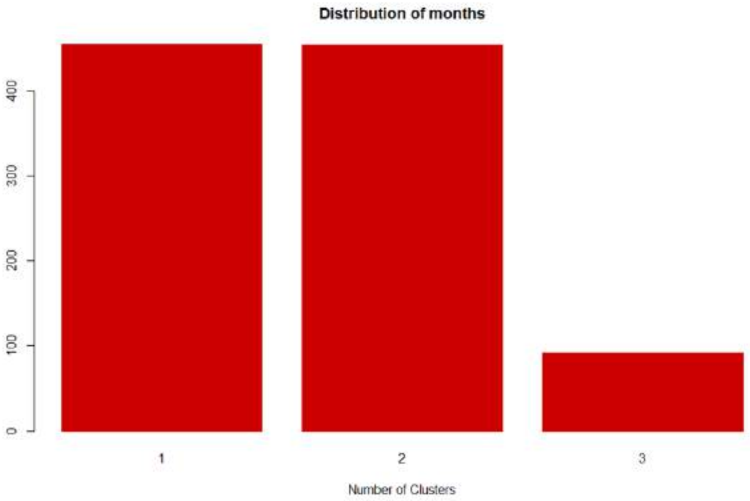
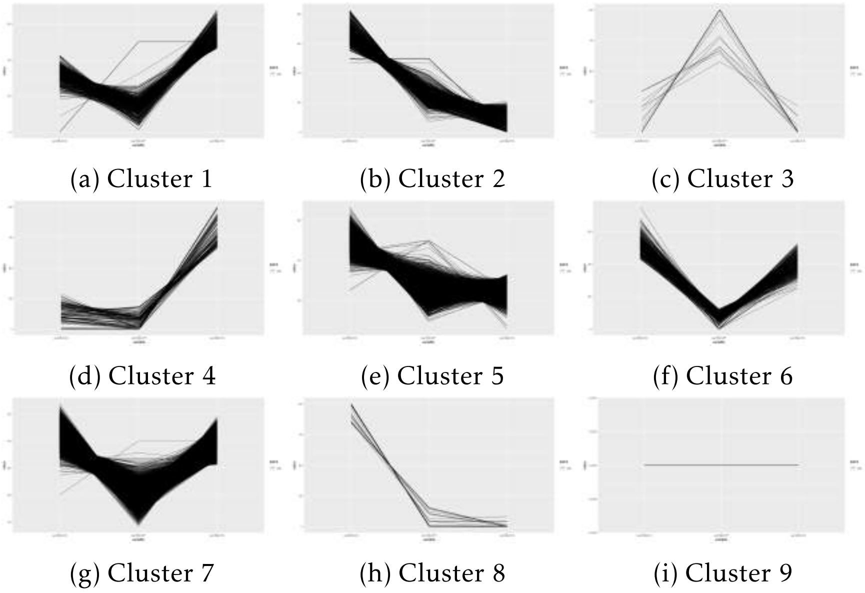


Figure A.15: Number of *months of clients* (y axis) in each cluster, i.e. size of clusters obtained by running the MBC over the absolute sum of consumption by week for each client.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

(g) Cluster 7 (h) Cluster 8 (i) Cluster 9

Figure A.16: Hairball plot of the nine clusters obtained by running the MBC over the percentage of consumption by tariff period of each month of 2018.