Politecnico di Torino

DEPARTMENT OF ENERGY

Master's Thesis

DATA ANALYSIS: CLUSTERING OF ELECTRICITY CONSUMPTION PROFILES

ACADEMIC YEAR: 2014/2015

July 2015

Student: JOSEP OTAL PARÉS

Supervisor: SALVATORE MANCÒ

Abstract

Chapter 1 starts presenting an overview of the energy sector and the challenges it is facing; it is observed a change on the energy policies promoting the energy efficiency, encouraging an active role of the consumer, instructing them about the importance of the consumer behaviour and also protecting consumer rights. Electricity is gaining room as energy source, its share will keep increasing constantly in the following decades. In this close future, smart grids and smart meters deployment will benefit both the utility and the consumer. In this environment, new services and new business appear, focusing on the energy management field and tools, they require specialization in fields such as, computer science, software development and data science. Finally, the case of Enerbyte is presented.

Chapter 2 of this work has segmented the electricity consumers of the project of "Rubí Brilla" according to the similarities of their electrical load profiles, using the proportion of energy usage per hour (%) as a common framework. The objective behind this segmentation is to be able to provide personalised recommendations to each group in order to reduce their energy consumption and the associated costs, fostering energy efficiency measures and improving the consumer engagement.

The desired segmentation is obtained by an iterative process, based on computational clusters calculation (using "R" software) and finalized by a post-clustering analysis applying visualization and statistical techniques to detect the outliers and reallocate them to a more appropriate group. Three different clustering techniques (Hierarchical clustering, K-means clustering and Self-Organizing Maps), were tested and compared, giving similar outputs. The solution from the Hierarchical clustering is the one that better adapts to the segmentation sought, which is used as the base of the post-clustering stage to obtain the final segmentation.

The final result grouped the 121 users in 7 clusters, each group representing a distinct load profile. The first six clusters are devoted to residential consumers and the cluster 7 is devoted to business. Described and distributed as follows:

Cluster 1	Morning Peaks	(7 consumers)
Cluster 2	Late night Peaks	(17 consumers)
Cluster 3	Flat Consumers	(28 consumers)
Cluster 4	Evening Peaks	(7 consumers)
Cluster 5	Daytime Consumers	(16 consumers)
Cluster 6	Double Peaks	(18 consumers)
Cluster 7	Business	(7 consumers)

The segmentation of the electricity consumers provides knowledge and a better understanding of the consumer. In this particular case, it allows to personalise energy savings recommendations according to the consumers' specific characteristics; improving the consumer experience by being able to provide the adequate advices at the appropriate time, facts that increase the effectiveness of the energy efficiency advices' service.

Table of Contents

List of Figures

List of Tables

Glossary and Abbreviations

Abbreviation	Description
%	Percentage
€/kWh	Euros per Kilowatt hour
AMI	Advanced metering infrastructure
ANN	Artificial Neural Networks
ANSI	American National Standards Institute
API	Application Programming Interface
BPL	Broadband over Power Line
BtoBtoC	Business-to-Business-to-Customer
BtoC	Business-to-Customer
CH_4	Methane / Natural gas
CHP	Combined Heat and Power
CIS	Consumer Information System
CO_2	Carbon dioxide
CSV	*Comma-separated values, file that stores tabular data in plain-text format*
DER	Distributed energy resources
DLR	Dynamic Line Rating
DMS	Distribution Management System
EC	European Commission
ERP	Enterprise Resource Planning

EU	European Union
EV	Electric vehicle
FACTS	Flexible AC transmission systems
G2V	Grid-to-vehicle
GDP	Gross Domestic Product
GHG	Green House Gases
GIS	Geographic information
GPRS	General Packet radio Service
GPS	Global Positioning System
H	Hour
HDI	Human Development Indicator
HTS	High-temperature superconductors
HVDC	High voltage DC
ICT	Information and Communication Technology
IEA	International Energy Agency
IEC	International Electrotechnical Commission
IP	Internet protocol
kWh	Kilowatt hours
LAN	Local Area Network
MDMS	Meter data management system
MS Excel	Spreadsheet application developed by Microsoft
Mtoe	Millions of Tonnes oil equivalent
N_2O	Nitrous Oxide
NIST	National Institute of Standards Technology
OECD	Organisation for Economic Co-operation and Development
OMS	Outage Management system
OSGP	Open Smart Grid Protocol
PHEV	Plug-in Hybrid Electric Vehicles
PLC	Power line carrier
PV	Photovoltaic
Python	High-level programming language (www.python.org)
R	Programming language and software environment for statistical computing and graphics (www.r-project.org/)
SaaS	Software as a Service
SOM	Self-Organizing Map
TCP	Transmission control protocol
TLM	Transformer Load Management
TOE	Tonnes oil equivalent
UPC	Universitat Politècnica de Catalunya
UPGMA	Unweighted Pair group Method with Arithmetic mean
UPGMC	Unweighted Pair group Method using Centroids
URL	*Uniform Resource Locator.* Reference to a resource that specifies the location of the resource on a computer network and a mechanism for retrieving it.
U.S.	United States
V2G	Vehicle-to-grid
WAN	Wide Area Network
WCSS	Within-Cluster Sum of Squares
WI-FI	Wireless internet connection
WPGMA	Weighted Pair group Method with Averaging
WPGMC	Weighted Pair group Method using Centroids
ZigBee	Low-power wireless network standard similar to Wi-fi, but simpler and cheaper

Acknowledgments

I would like to thank to the company “Enerbyte Smart Energy solutions” to host me during this period, for all the facilities, help and time devoted to me while developing this thesis as well as offering me the possibility to work with them, especially to Pep Salas and Roger Segura.

Also, I want to acknowledge to KICInno Energy for providing the possibility to get enrolled in these European master programs, in this case SELECT. And a mention to the Erasmus Mundus commission for awarding me with the scholarship Erasmus Mundus Category B, which facilitated my decision to start this new adventure and experience in my life.

Thanks also to all the people involved in the SELECT program and committee, from the administration to the professors and coordinators in both KTH (Peter Hagström, Thomas Nordgreen, Nele Stoffels) and Polito (Massimo Santarelli, Salvatore Manco’, Giovanni Fracastoro).

Finally, I appreciate the support and help of my friends and my family among this time; also to all the new friends and classmates that accompanied me during this journey.

0. Thesis description

The work presented in this document is divided in two different chapters:

Chapter 1 presents a general overview of the energy sector, mainly focused on electricity. This chapter is structured from a broad and generalist perspective on the energy sector to a detailed and precise description of smart grids and the advanced metering infrastructure. Smart meter characteristics, benefits and its deployment in Europe are remarked, as the collection and treatment of the electricity data from smart meters allow new business opportunities able to provide new services that add value to both the consumer and the utility.

Chapter 2 contains the case of study which is the core of the work, which analysis and segments the residential and small business consumers by their electric consumption profile, aiming to discover knowledge in concordance to the idea to provide personalised energy efficiency recommendations. Firstly, the introduction to the case, the objectives sought and the methodology followed are described. Secondly, previous steps such as cleaning of data to remove the bad data, together with a data exploration to understand and know it better, are necessary to start the desired consumer segmentation. To carry out this segmentation different techniques are used and compared in order to find the most suitable distribution of consumers. Ending up by writing the final segmentation results, also analysing the household and householders characteristics of each group obtained to, finally, extract the final conclusions.

CHAPTER 1: ENERGY CONTEXT AND OVERVIEW

1. Energy Sector Overview

1.1. Energy and Sustainable Development

Energy has become a basic instrument for the well-being of the society, together with its increasing importance in the world; the energy sector is facing important challenges nowadays, aiming to cover all the new necessities and adapting the old standards to the new situation.

During the last decades there has been special concern on environmental issues, especially from the policy point of view. The celebration of three global environmental conferences in Stockholm (1972), Rio de Janeiro (1992) and Johannesburg (2002) prove so. A thinking evolution throughout these meetings end up with the appearance of the Sustainable Development concept, which represents a more holistic approach where economic, social and environmental dimensions need to be treated together. Figure 1 illustrates the energy role in each Sustainable Development dimension.

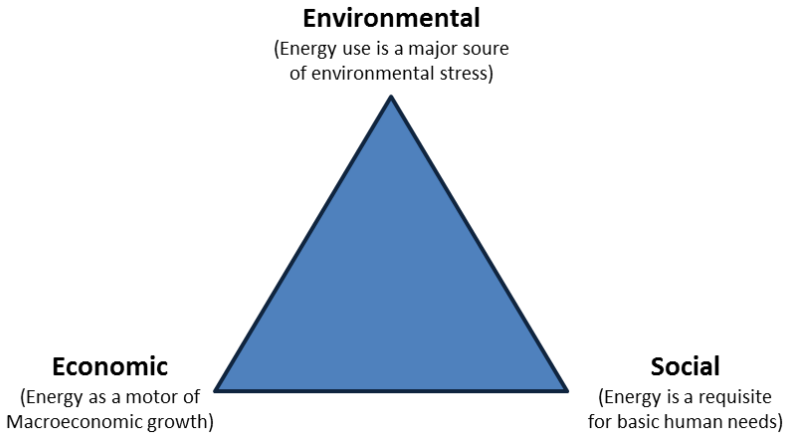


Figure 1. Energy and Sustainable Development linkages. Adapted from (Cleveland & Najam, 2008)

1.1.1. Energy & Environment

The extraction, process and use of the conventional energy sources, mainly fossil fuels such as oil, coal, natural gas; allowed, in the past, to have access to cheap and abundant energy contributing to the economic development, but also leaving behind major environmental impacts at a global, regional and local level by using these scarce and non-renewable resources.

The extraction of natural resources (mining) has significant impacts in the local ecosystems normal development; the process and use of energy implies atmospheric releases of greenhouse gases like CO_2, CH_4, N_2O among others; leading to environmental damages: climate change, acid deposition, urban smog.

A way to address this problem is switching to renewable/clean energies, for instance photovoltaic, solar thermal, wind energy, biomass, hydropower… this kind of energies are characterised by a low or nearly zero environmental impacts during extraction, conversion and use, therefore contributing to the climate change mitigation.

1.1.2. Energy & Economy

Energy is directly related to macroeconomic indicators; there is an explicit correlation between countries' energy demand and the size of their economies, measured by the countries' Gross Domestic Product (GDP) as it shows Figure 2. Principally, energy is consumed to feed the industrial processes that convert raw materials into final goods or services, creating added value.

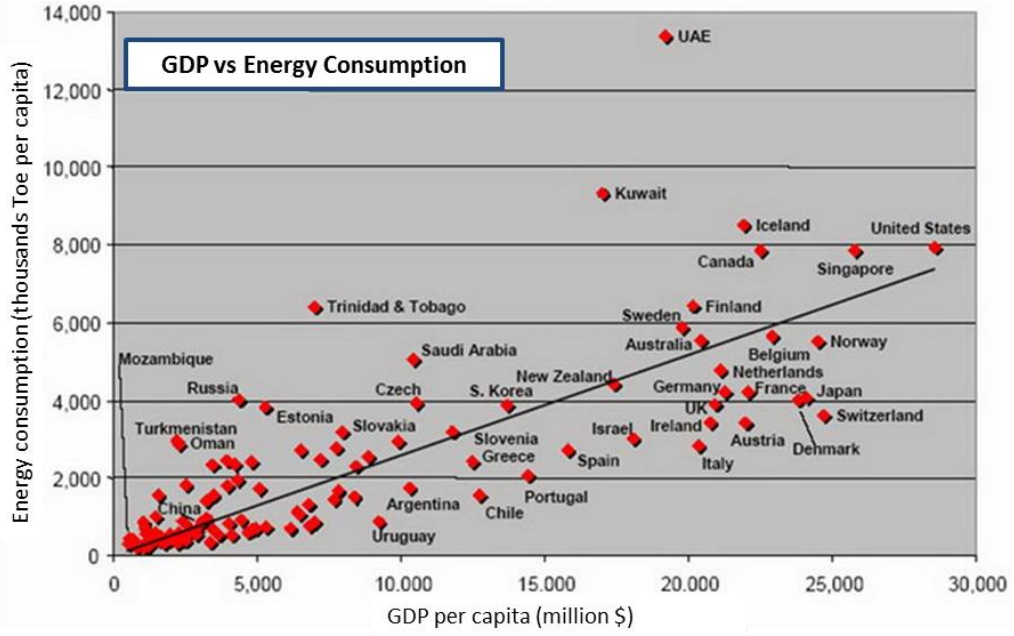


Figure 2. Correlation GDP vs Energy consumption. Energy Consumption per capita in thousands of Tons of equivalent oil (toe) and Gross Domestic Product per capita in million $. Adapted from: (Scottish Sceptic, 2013)

The previous relationship might seem obvious, although two much better indicators are:

- **Energy intensity**, relating units of energy per units of GDP [tonnes of oil equivalent-toe/$], as shown in Figure 3 where general descending of the energy intensity is seen in the various economies presented, both developed and developing countries.

- **Energy Efficiency of an economy** [$/toe], which is the energy intensity reverse. A clear classification and grouping of the top 40 economies can be found in (Corless, 2005), developed countries economies are placed in the high GDP per capita but low efficient, and most of the developing countires are also low effcient with a lower GDP per capita.

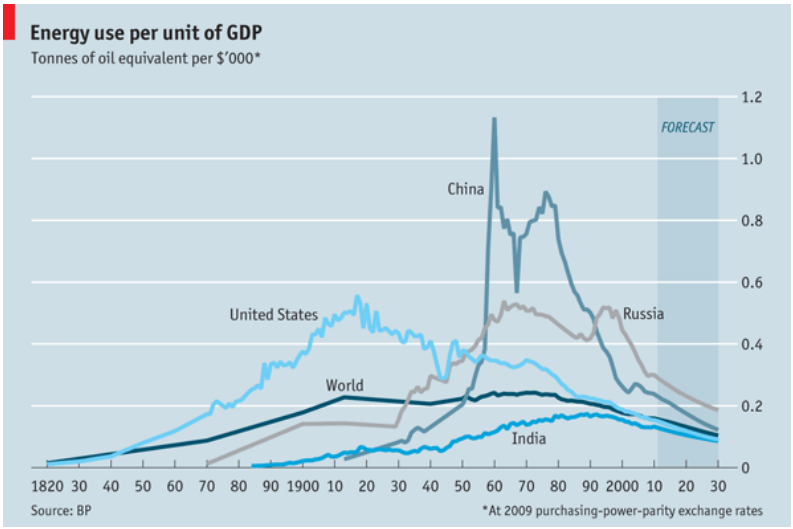


Figure 3. Energy use per unit of GDP trend. *Source :* (The Economist, 2011)

Not long ago, it has arisen the question of decoupling the energy use and economic growth, (Tverberg, 2011) by using less energy to produce the same or a higher economic outcome. The example of Germany's *Energy Transition (Energiewende) states* that it might be a room for improvement by means of energy efficiency.

1.1.3. Energy & Society

The more complex connection of energy deals with the social dimension of the Sustainable Development, when meeting the basic human needs. Energy is essential to fulfil these basic needs such as food, water, shelter, healthcare, education, unemployment…

Human Development Indicator (HDI) was created to measure and rank countries' human development in a common reference, this statistic parameter combines life expectancy, education and income indices to determine nation's social development. The Human Development Reports from the *United Nations Development Programme* recollect all countries' data referred to the HDI and many other indexes.

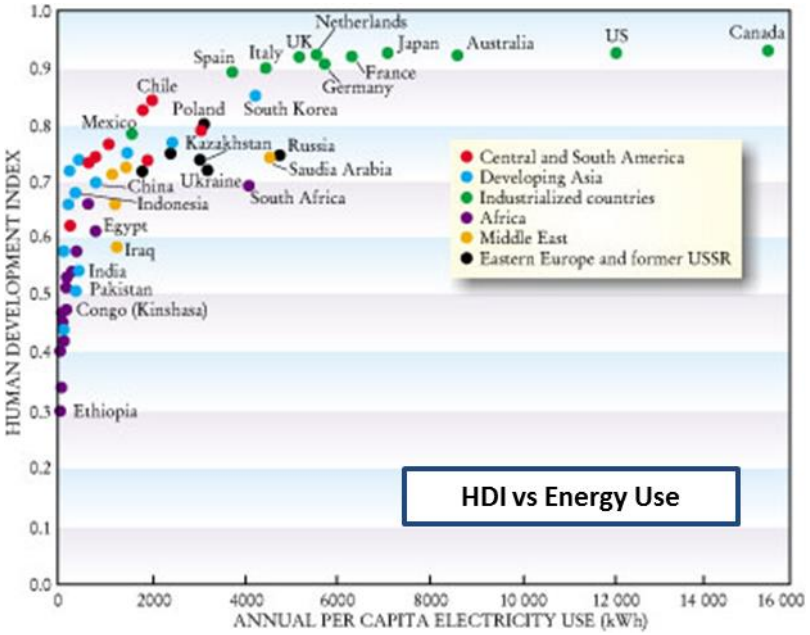


Figure 4. Energy use per country in kWh vs Human Development Index. *Adapted from:* (Pasternak, 2000)

Some European articles and reports, such as (European Commission: Energy Supply and Demand, 2014) point out the strong relationship between social indicators and energy use, apart from the fact that more population growth more energy demand, developing countries consume much less energy than developed countries from households energy requirements perspective.

A worrying fact related to the lack of access to energy services (electricity or cooking facilities), is the Energy Poverty. The International Energy Agency (IEA) corroborates that still 18% of the world population does not have access to electricity (IEA: Energy poverty, 2015). However, energy poverty is not only affecting the developing countries, the recent years pushed by the global economic crisis and the rise of energy prices, families from developed cannot face the costs of accessing to energy (Bouzarovski, 2014).

1.2. Energy Supply & Demand

The energy demand is expected to grow significantly in the coming decades, up to doubling the current demand by 2035; this increase will be led by the developing and non-OECD countries mainly China and India. In the same direction the International Energy Agency, World Energy Outlook 2012, has made the projection by 2035 presented in Figure 5.

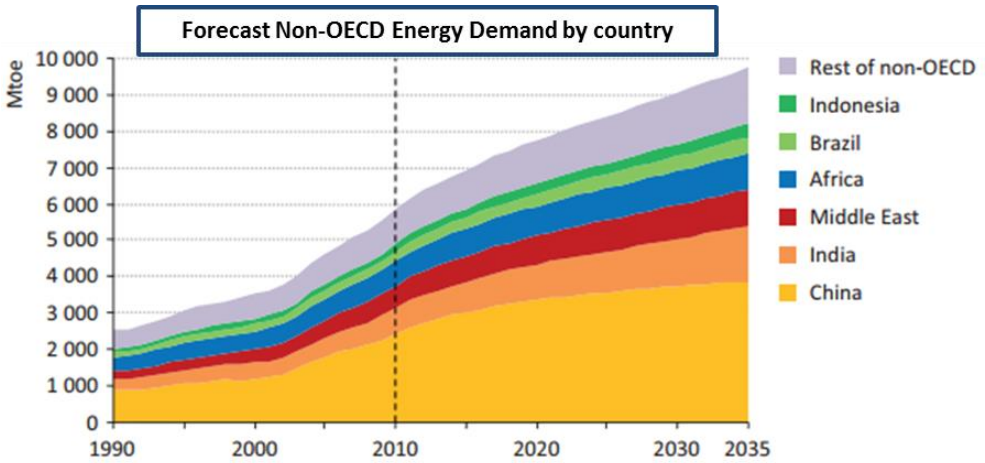


Figure 5. Forecast of Energy Demand by non-OECD countries. Source: IEA, World Energy Outlook 2012

Energy policies direct their focus to energy efficiency and to the integration of renewables starting an energy transition. A better use of the available energy and the use of clean energies will help to reduce the CO_2 emissions leading to a decarbonisation of the economy. The fully substitution of fossil fuels is yet far from reality, so the conventional primary energy sources will be still necessary to meet this demand increase, as seen in Figure 6.

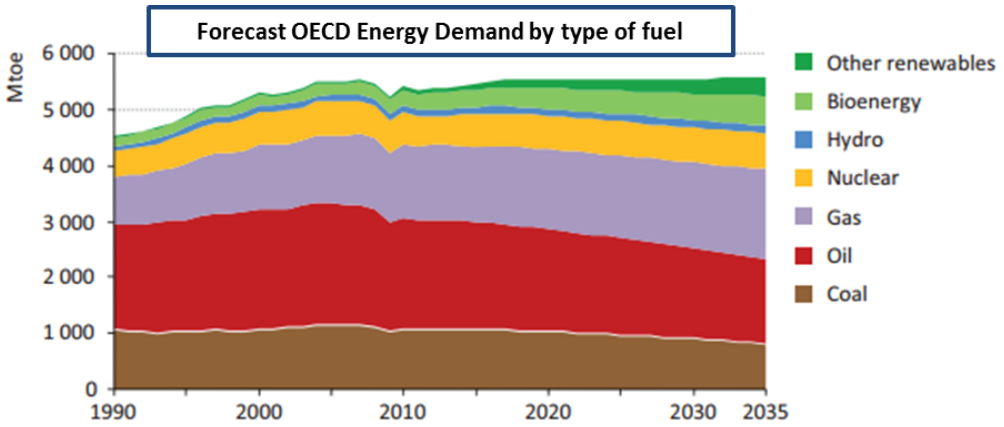


Figure 6. Forecast of Energy Demand on OECD countries by type of fuel. Source: IEA, World Energy Outlook 2012

Finally, the 2011 energy demand allocation by sector is presented in Figure 7, and it almost double the total energy consumed compared to the same figure by 1973; where the shares doesn't vary too significantly . It allows having an idea to know where and what for this energy is consumed, helping to elaborate more adequate policies to each sector.

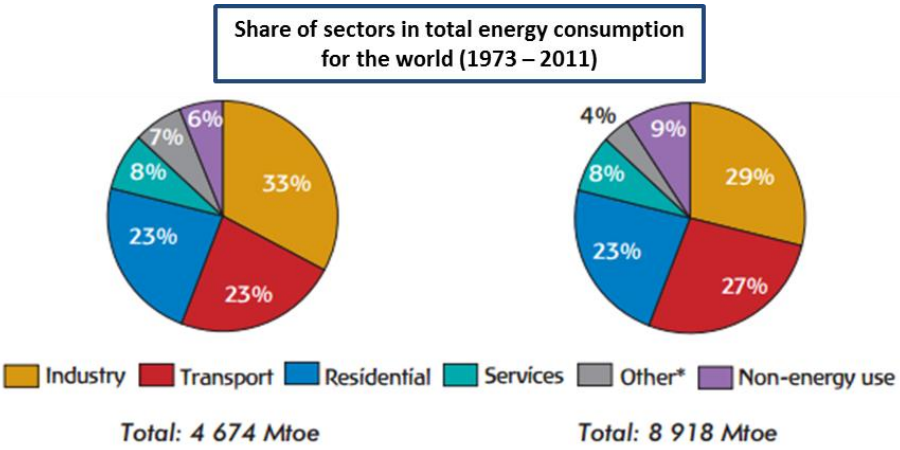


Figure 7. Comparison energy consumption by sector shares and total from 1973 (left) to 2011 (right). Source: IEA, Energy Efficiency Indicators

1.3. Energy policies and energy efficiency in Europe

Energy has always been one of the top 10 priorities in the European Union nowadays European Commission is seeking for an Energy Union in order to make energy more secure, affordable and sustainable.

1.3.1. European energy strategy

European Union countries are importing more than half of its energy for instance, around 90% of crude oil and 66% natural gas (European Commission: Imports and Secure Supplies, 2015); being the largest energy importer in the world, this fact has associated costs around 400 billion € per year.

In that sense, the fragmented market and old energy infrastructure with insufficient countries interconnections, for example 12 state members are energy isolated islands (<10% interconnection) and 6 members depend of a single supplier (European Commission: Energy Union, 2015), result to consumers vulnerability and higher prices. For instance, electricity prices in Europe are 30% higher, and gas prices can double than US ones; subtracting competitiveness to the European business and enterprises. A common market legislation and the market liberalisation can help to solve this issue.

An appropriate integration is expected to save up to 40 billion € each year and also reduce the dependence of the external suppliers. Ensuring the energy supply to all EU citizens has become a problem, moreover after the economic crisis that left more than 10% of European population without being able to pay their energy bill (energy poverty).

In the actual EU framework, strategy targets are set by 2020 aiming for (compared to 1990):

- 20% GHG reduction (40% by 2030)
- 20% increase of renewable energy in EU mix (27% by 2030)
- 20% improve of energy efficiency (27% by 2030)

Energy efficiency and use of renewable energy technologies are the most important features to achieve the targets set. Also, improving these numbers is the core of the EU energy strategy by 2030 and 2050 to mitigate the Climate Change.

1.3.2. Energy Efficiency

Energy efficiency has been gaining importance in the energy world's priorities, it is named as the "invisible fuel" (The Economist, 2015) stating that the best choice is not to waste energy. In regions like Europe with highly external energy supply dependences, an optimisation at all stages in the energy chain is a must from both environmental and economic point of view.

Rules and obligations set by the Energy Efficiency Directive (Directive 2012/27/EU, 2012) are in that direction. The target of the European Union is to reduce up to 20 % the energy consumption, achieving by 2020 an energy consumption lower than 1.474 Mtoe of primary energy or less than 1.078 Mtoe of final energy, but setting an objective adequate to each country characteristics.

This directive differentiates the energy efficiency in energy supply and energy efficiency in energy use. The measures adopted by the European Union can be summarised:

- Energy distributors and retailers must reduce annually 1,5% their energy sales
- Annual energy efficient renovation of at least 3% of buildings owned or occupied by governments
- Incentive the buildings renovation, i.e. adding insulation, double glaze windows, high efficient boilers; to improve their energy performance
- Mandatory energy performance certificates when renting or selling buildings
- Set of minimum labels or standards for a range of products, such as boilers, domestic appliances, lighting…
- Periodical energy audits for large companies and incentives for Small and Medium enterprises to undergo energy audits
- Protecting consumer rights to receive comprehensive information, access to real-time and historically energy consumption and billing data to a better consumption management
- Deployment of 200 million of electricity smart meter (72% of the total) by 2020

The energy efficiency in energy supply is principally focused on the highly-efficiency cogeneration plants (electricity + heat production), implementation of district heating and cooling and integration of renewables. Together with the energy savings attained due to suppliers obligations and the application of the white certificates (Suppliers Obligations & White Certificates, 2015)

A modernisation on energy infrastructure (transformation, transmission and distribution) leads to the final deployment of the smart grids, reducing grid losses (accounted by almost 30%) (European Environment Agency, 2012) derived from energy generation, transportation and distribution.

From the perspective of the present work, the main focus of attention is in the Energy Efficiency Directive chapter II, which is related to the energy efficiency in energy use. Aiming for a massive smart meters rollout (72% in electricity by 2020), and allowing consumers to have access to real-time and historical information of their consumption and billing; will provide inputs and knowledge to the end consumers that will enable them to have an active role to decide how and when is the best time to use energy.

In addition to the Energy Efficiency Directive, there are three other directives devoted on building's thermal and electrical demand, as buildings represent around 40% of the total energy demand in Europe (European Commission: Energy Efficiency, 2015) these directives are:

- **Energy performance of buildings directive:** include building energy certificates when selling and renting buildings; finance renovation of building elements to a low energy needs, and all new buildings must be nearly-zero energy buildings.

- **Energy labelling directive:** intending to help consumers to have more information in order to choose energy efficient products (air conditioners, television, washing machines, lights…)

- **Eco-design directive:** directed on product manufacturers, regulating the requirements of manufacturers to establish the minimum energy efficiency standards.

1.3.3. Active Consumer

Finally the paradigm has changed, for the first time policies seem to be consumer oriented ensuring them to receive a good service and able to defend their rights as energy consumer. Among the list presented in (European Commission: Consumer rights and protection, 2015), there are two points clearly related to the purpose of the present work of tracking and managing the energy use:

- Consumer has to receive and have access to accurate information to their electricity and gas consumption and billing.
- Consumer has to receive information on how much energy uses and how to use energy more efficiently. Also the benefits of using renewable energy equipment and vehicles.

Also, policies seek to give the final push to the infrastructure modernisation, not only interconnecting the countries but also by embracing innovation and new technologies to complete this new energy infrastructure.

An important transformation is occurring in the power generation model, moving from a Centralised to Decentralised energy production, illustrated in Figure 8.

The traditional Centralised power production is constituted of few and big power stations, the consumers are far from the production sites, the network is unidirectional from producer to consumer, there is no flexibility on the energy demand, so consumers have a passive paper.

In contrast, with the new Distributed production is not only delivering a service also including other features that improve the whole process. The production is also close to the consumption points, thanks to the small and medium plants built mostly by renewable energy sources (PV, wind…) or new technologies like fuel cells (BloomEnergy, 2015), the "prosumer" term appears to define those who produce and consume their own energy, the possibility to store the energy to use when and where is necessary, thanks to a network that becomes bidirectional, prioritizes the demand by using more efficiently the resources, taking advantage of ICT and the big data to shave consumption peaks and engage the customer, who will have an active role.

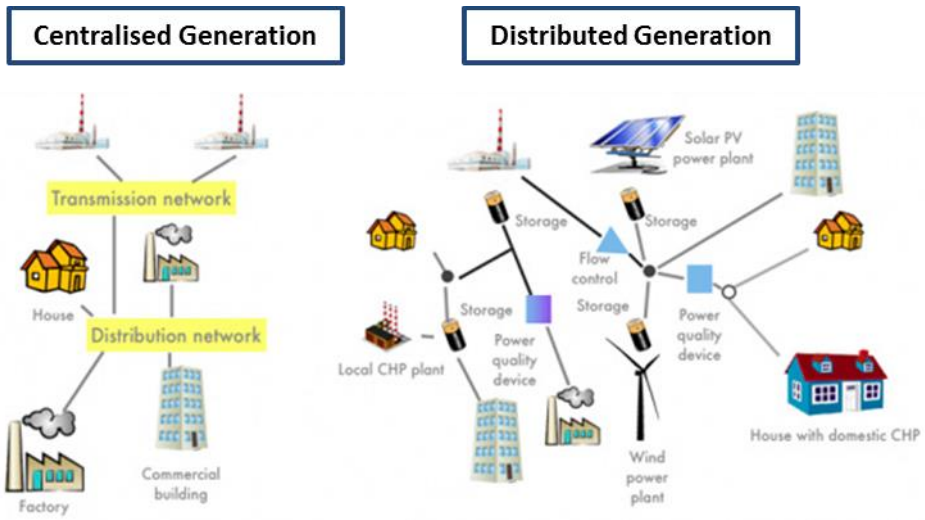


Figure 8. Conceptual model comparison between Centralized and Decentralised. Source: (European Commission: Distributed Generation, 2015)

In this new landscape, electricity is expected to be the world key energy, some studies, such as, (Armaroli & Balzani, 2011), point that society is advancing towards an electricity world: devices, cars, heat pumps… where almost everything will be powered by electricity; in this sense smart grids are necessary to update the electricity network to the meet the actual situation.

2. Smart Grids Overview

Electricity is called to be the fastest-growing component on the energy supply portfolio; the increase on electricity consumption is expected to achieve the 40% of the total energy market by 2020 (IEA: Smart grids, 2011) and reaching an important share by 2050. This growth will be essentially led by the emerging economies and developing countries, as is shown in Figure 9, while a modest increment in other developed economies will occur.

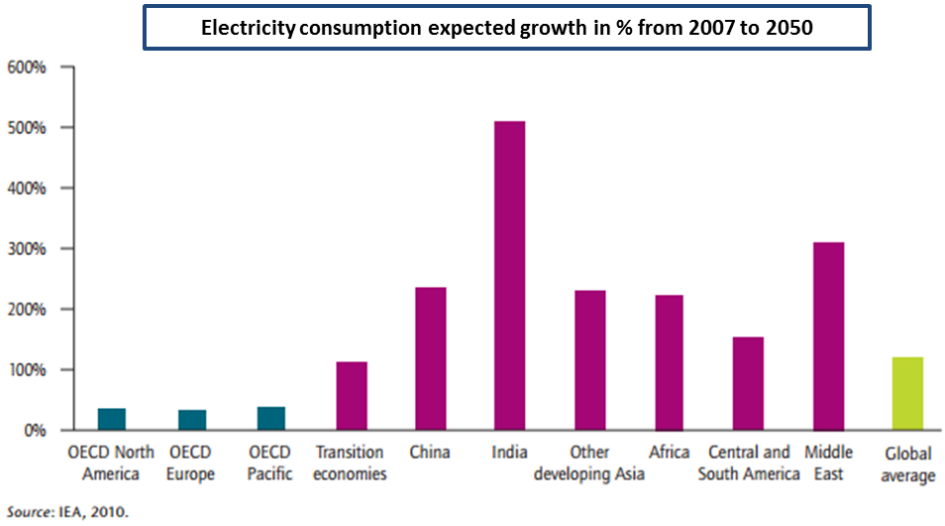


Figure 9. Electricity consumption expected % increase by regions. Source: IEA. Technology roadmap: Smart Grids.

This large growth on electricity usage is likely to arise in the:

- Residential heating and cooling due to the use of heat pumps. Either individual at home heat pumps or larger district heating and cooling heat pumps.
- The electrification of the transportation and mobility, due to vast deployment of Electric Vehicles (EV) and the Plug-in Hybrid Electric vehicles (PHEV)
- Variable electricity generation, by using PV, wind, hydro, biomass, combined heat and power (CHP) technologies, decentralising the production.

This new conjuncture will tackle the climate change to a CO_2 emissions reduction, while covering the same needs moving from rich-carbon sources to low-carbon ones. And maintain the economic and social development by using more sustainable systems; however the electrical grid will have to be ready and able to host these new features.

2.1. Smart grids

2.1.1. Definition

The concept of smart grid has been widely used these recent years, but there is not a standard definition, it can be seen as the process of making the actual electrical grid smarter in a sense of modernisation and adaptation to the new situation.

Smart grids are electricity networks that make use of the information and communication technologies (ICT) and are able to integrate other advanced technologies (distributed electricity generators, electricity storage, electric vehicles…); to maximize the efficiency, reliability and safety of the power grid, and minimize the costs and environmental impacts (U.S. Department of Energy, 2015). Allowing a flexible production and a consumer engagement, by providing information and tools to both electricity producers and consumers. As consumers will play an important role in the new electric grid landscape (Figure 10).

It that sense a difference can be made between the core of the smart grid that falls on the transmission and distribution of electricity and the peripheral ones that are the generators are consumers (Ramireddy, 2012).

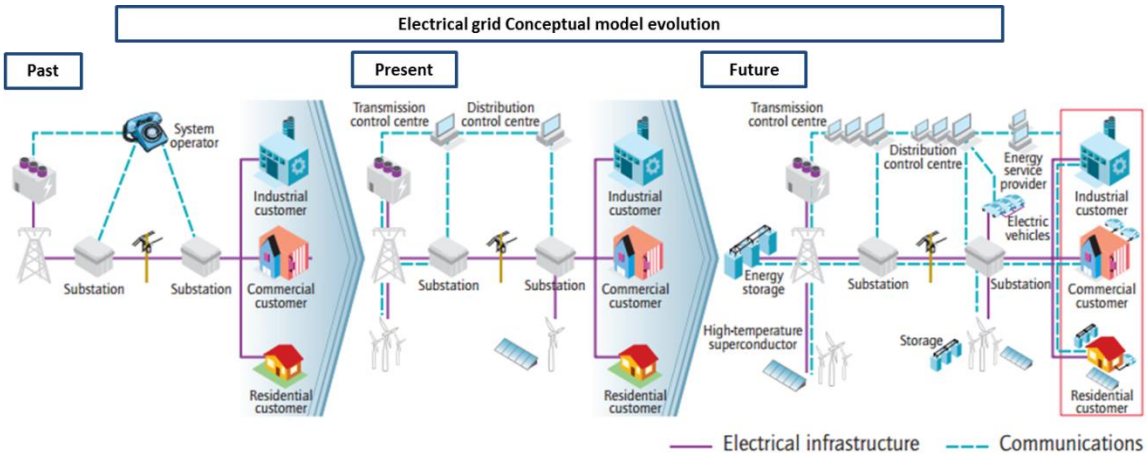


Figure 10. Smarter electricity systems, from past to future. Source: IEA. Technology roadmap: Smart Grids.

In Table 1 a summary of the differences between conventional and smart grids are presented:

Table 1. Characteristics' comparison between Conventional and Smart grids (NCCS, 2011)

Characteristic	Conventional old grid	New smart grid
Resiliency, reliability and security	Vulnerable to failures and natural disasters	Resilient to attacks and failures with rapid and even automatic restoration, self-healing
Integration generation and storage	Central generation of large power plants located in few places. Difficulty to integrate distributed energy sources.	Facility to integrate the many distributed energy sources (renewables) with "plug-and-play" as a supplement to the central stations
Consumer engagement	Consumers have a passive paper, and are under-informed	Are involved, well informed and have an active paper –demand response and possibility to be "prosumers"
Optimisation and operations efficiency	One direction power flow, far to utilise the whole capacity of the infrastructure, concurring to higher investment and maintenance costs	Two-direction communication and flow. Ability to reduce losses and use the full capacity of the system's assets.
Market evolution	Limited choices for the consumer, limited wholesale markets.	More mature, liberalization of the market lead to a growth of market and more options for the consumer

2.1.2. Smart Grid Characteristics, Functions, Benefits

In this section the main attributes of smart grid's performance and benefits are described, based the document from the US Department of Energy named Modern Grid Benefits (U.S. Department of Energy, 2015).

- **Reliability, quality and security**

The modern grid offers new means to control the transmission system, is able to monitor in real-time the grid conditions, performing self-assessments to detect, analyse, respond to, and if is necessary to, restore grid components.

Automatically diagnoses of the grid disturbances, and automated response to unexpected failures or attacks, by isolating problematic elements while preventing of so the rest of the area which is restored to normal operations. The grid's faculty to carry out these self-healing actions bring a better quality service as the interruption periods are reduced and so helping service providers to better manage the delivery infrastructure.

- **Ability to accommodate all types of generation and storage options:**

Smart grids facilitate the connection and management of the distributed generation plants of different sizes and technologies. From medium scale wind farms, to combined heat and power plant, or roof-top solar photovoltaic installations, the new grid allow to connect all the new generating plants, having the so called "plug and play" ability.

Also, the grid will be capable to include energy storage systems (batteries, hydrogen, compressed air, pumped hydro…) in order to optimise the distribution and reduce the price of the electricity. The grid would also manage the connection of the increasing number of electric vehicles (EV), which can also be a storage system.

- **Consumer engagement:**

Consumers are called to play an important role in the new electric grid, they will have information and visibility to how much they consume and at which price in a real-time basis. The increased interaction with the grid, the better options when choosing the energy supplier, the new forms of electricity pricing, such as time-of-use tariffs where energy costs more during the peak hours; are aiming to change the consumption patterns and behaviours bringing benefits by reducing the cost of the electricity and the environmental impact.

In a future, the grid itself would be able to reduce automatically consumer's consumption when power is expensive or scarce. Up to the point, where utilities can ask the consumers to switch off some appliances but being compensated for doing so.

- **Optimise assets utilisation and operating efficiency**

A better visibility of the power flow on the network facilitates distributors to understand where losses occur and where to invest, cutting the costs of maintenance and reparation as the problem or inefficiency is specifically detected at the precise time.

The intelligence of the technologies used in the smart grids optimises the performance of its assets. Increasing the operating performance when looking for the minimum cost system operation, the control devices will automatically adjust to reduce losses while operating to an optimised capacity as close as possible to the full utilisation of its assets.

- **Enable new products, services and markets**

Markets play a major role in the management of the grid, as they can control the energy, capacity, location, time and power quality. The open-access (liberalization) market will provide market flexibility due to the deregulation and new competitive services will appear to offer better choices to the customer.

2.1.3. Smart Grid conceptual model

To have a clearer idea of a smart grid in this section there is an explanation based Smart grid conceptual model from the National Institute of Standards Technology (NIST, 2013) which provides a description of the areas in which a smart grid can be divided. There are seven differentiated domains as shown in Figure 11 include the bulk production, transmission, distribution, customer, markets, operators and service providers.

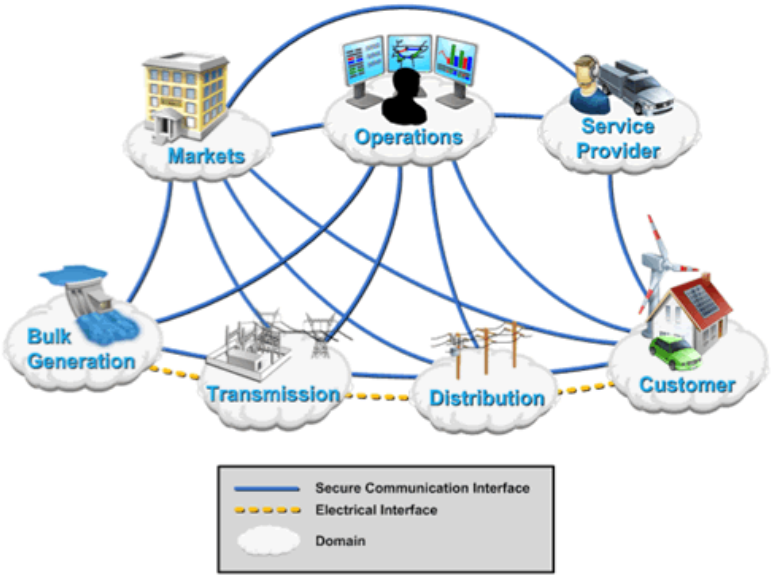


Figure 11. Conceptual illustration with all domains participating in a Smart Grid. Source: National Institute of Standards Technology (NIST)

- **Bulk generation:**

Bulk is referred to the large electricity production; this electricity can be generated from different types of energy and sizes of technologies, seen in Figure 12. Including:

- Conventional non-renewable and non-variable power plants: nuclear fission, coal and gas plants
- Renewable and non-variable: hydro, biomass, geothermal and pump storage
- Renewable and variable: solar and wind

Once the electricity is generated it is connected to the transmission domain, and the bulk generation domain communicates to the Operations, Markets and Transmission domains. The

Information and Communication Technology (ICT) infrastructure let the grid to automatically reroute the power flow when some generators fail, also allows starting and stopping electricity production, allocating distributed production and storage devices to store energy for later distribution.

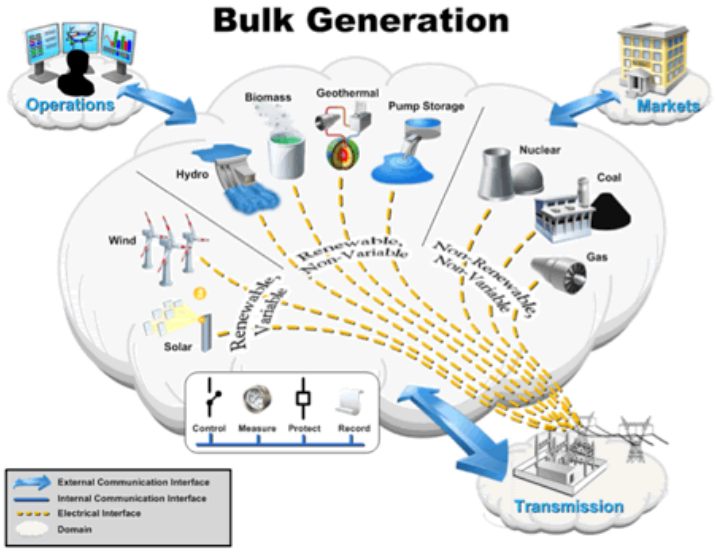


Figure 12. Deep look at the Bulk Generation domain. Source: National Institute of Standards Technology (NIST)

- **Transmission:**

Transmission is the transfer of electrical power from generation sources to distribution through multiple substations. It is electrically connected to the Generation and Distribution domains, and communicating with the Operations, and Markets domains. The transmission network operators have to balance the supply and the demand to maintain the stability of the grid. Thus may contain distributed energy sources such as storage and peaking generation units with the capacity to manage and dispatch the energy when is needed.

To do so, the Wide-area monitoring and control is necessary and it involves the generation and transmission areas, it is able to display, monitor and even optimise all power system components at real-time. This technology gives a real-time view of the power system behaviour, helps to avoid blackouts and facilitates integration of renewable energy sources. A set of equipment and systems are needed, these may include remote terminal units, substation meters, protection relays, power quality monitors, phasor measurement units.

A transmission enhancement is also needed; it is referred to the physical technologies to improve the transmission system, to name some:

- The Dynamic Line Rating (DLR) use sensors to monitor the carrying capability of a section of network, optimising the transmission assets with risking overloads.
- Flexible AC transmission systems (FACTS) maximises power transfer capability, improving transmission networks control.
- High voltage DC (HVDC) allows decreasing system losses when connecting remote wind or solar farms to large power areas.
- High-temperature superconductors (HTS): can improve the transmission performance by reducing its losses.

- **Distribution:**

Distribution domain connects electrically the Transmission domain and the Customer domain, also communicates with the Operations and Markets domains. Important changes occur in electricity distribution from historically very little communications, hierarchical and unidirectional interfaces; to the nowadays two-directional distribution grids built with much more interconnection, extensive monitoring and control devices capable to allocate the distributed generation and storage, control the demand response and the load, improving its reliability.

The Distribution grid management uses active network management technologies on the distribution and sub-stations that improve the intelligence of the grid by allowing automation distribution processes by taking real-time information to identify the failure and reduce the interruption time of restoring the network.

Capacitors banks, sectionalizers, reclosers, protection relays, distributed generators and storage devices, are devices included; capable to automatically monitor and control the voltage, compensate the reactive power. Systems used are regarding geographic information (GIS), distribution management (DMS), outage management (OMS).

- **Customer:**

Customer is called to have an important and active role in the new smart grid, so its engagement is essential for the well-functioning of the smart grids. To do so, the grid system should provide the customers with the necessary instruments to develop this paper. Customer domain is electrically connected to the Distribution domain; and communicates to several domains: Distribution, Operations, Markets, and Service Provider; see Figure 13.

Figure 13. Deep look at the Customer domain. Source: National Institute of Standards Technology (NIST)

A smart grid will enable the customer to manage not only the energy use but also the energy generation, providing control and information flow between the Customer and the other domains. There are three types of customers by size and specifications: industrial, commercial/building and residential; however all of them share a similar Advanced Metering Infrastructure, can allocate the electric vehicle charging infrastructure and dispose of similar customer side systems .

The **advanced metering infrastructure (AMI**) includes the use of different technologies, such as smart meters, communication technologies, in-home displays, gateways, severs… to a two-way flow of information providing customers and utilities with data on electricity price and consumption, including the time and amount of electricity consumed. The Meter Data Management System (MDMS), within the AMI, enables functionalities like remote load control, monitoring and control of distributed generation, in-home display of customer usage and integration with building management systems.

Electric vehicle charging infrastructure includes chargers, batteries, inverters; is expected to provide a capacity reserve and peak load shaving. With the possibility to schedule the smart charging (grid to vehicle-G2V) and discharging (vehicle to grid-V2G), it may suppose a positive effect into the billing by reducing the electricity costs and offering storage capacity for the surplus of energy when there is a low demand. Implies a direct interaction with both AMI and costumer-side systems.

Customer-side systems, refers to smart appliances, in-home displays, energy management systems and dashboards; that are used as a visual instrument to manage the electricity consumption at the residential, commercial and residential levels. Making use of this technology offers the possibility to increase the energy efficiency and reduce the peak demand, together with the consumer manual demand response or even the automated and remotely controlled appliances response to prices.

- **Markets:**

Markets are responsible to perform both pricing and balance supply and demand in the system. To do so interfaces with the generation, transmission, distribution and customer domains are crucial for this system matching, as seen in Figure 14.

Due to the fact that generation not only occurs at a large scale (Bulk Generation domain),the increasing distributed energy resources (DER) will participate actively to the grid supply and balance. The liberalization of the electricity market will help to customer to choose among different options, allowing to a fair trading and wholesaling electricity.

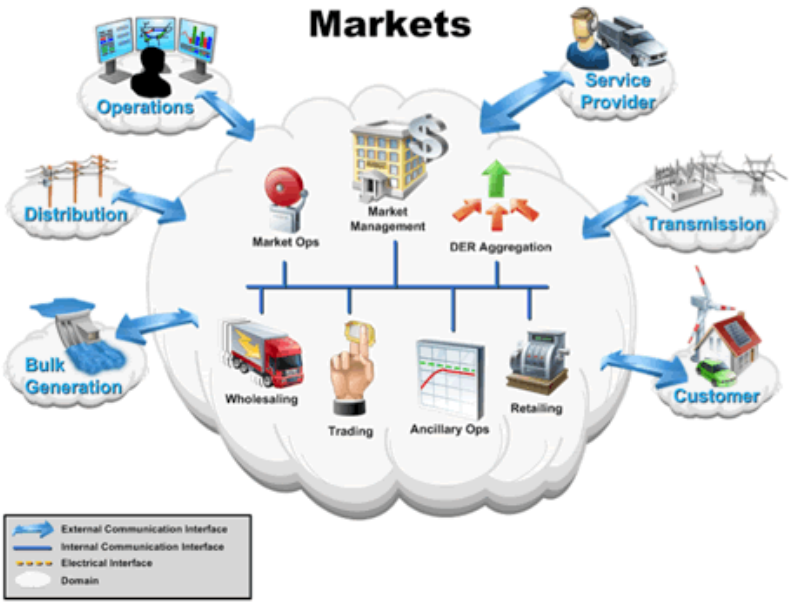


Figure 14. Deep look at the Market domain. Source: National Institute of Standards Technology (NIST)

- **Operations:**

Operators perform the ongoing management functions, typical operations may include: network operation, network operation monitoring, network control, fault management, operation feedback analysis, operational statistics and reporting, real-time network calculation, dispatcher training.

They have the role to smooth the operation of the power system, many of these tasks are responsibility of the regulated utility such as the maintenance and construction, meter reading or security management; and some of them can be outsourced to other domains.

- **Service provider:**

Service Provider includes the entities that provide services to electrical customers and utilities; performing support tasks to the business processes of power system producers, distributors and customers. The Service Provider communicates with the Markets, Operations and Customer domains. Interfacing the Markets and Customer domains, Figure 15; will enable the appearance of innovative services that will bring new business opportunities in response to the market needs and requirements. These services may be performed by the electric service provider or by third-party in support to the new services necessity.

These services range from the modernisation and enhancement of the traditional services of billing and customer account management, to the appearance of new services like the home energy management or generation. The so-called Software as a Service (SaaS) will have a major impact in this area, as it is able to provide an easy and comprehensive front-end to the consumers with the information about electricity consumption and price.

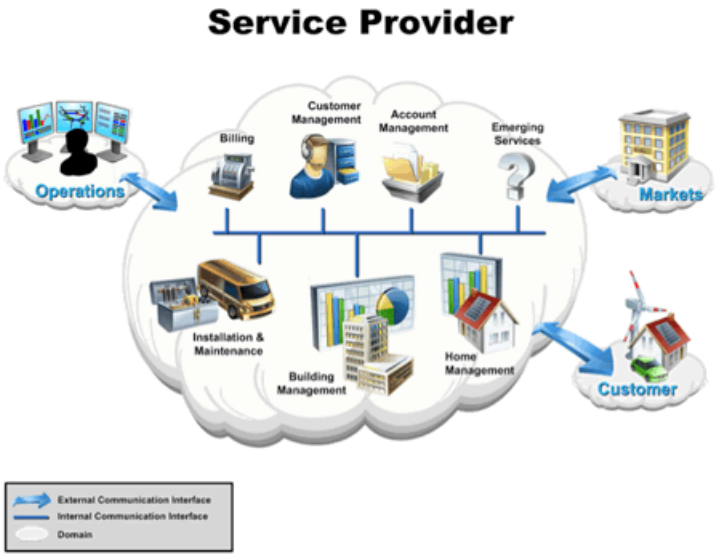


Figure 15. Deep look at the Servicer Provider domain. Source: National Institute of Standards Technology (NIST)

2.2. Smart meters & Advanced Metering Infrastructure

For the present work, a special attention is given to the smart meters and to the Advanced Metering Infrastructure AMI, as from them the consumption data is generated. This infrastructure is still under development but many regulators and governments put the focus especially to the smart meter deployment. It is the case of the European Union, inside the energy strategy for 2020 program.

Electrical markets were heavily regulated in some countries, fact that difficult the implementation of new technologies and quick changes to the conventional market scheme, due to its rigidity. This has arisen the implementation of private smart meters mainly to the industry, commercial and office building; as allows an energy management with future economic benefits. However in the residential sector the energy savings are not compensating the investment on private sub-metering equipment, and so relying on the "official" smart meter implementation to start managing their energy consumption.

2.2.1. Advanced Metering Infrastructure

What is AMI?

Advanced metering infrastructure (AMI) is a group of technologies combined together form an architecture that permits a two-way communication between the consumer and the utility (EPRI, 2007). These systems enable detail and time-based data measurement, collection and transmittal to the grid actors.

Due to the installation of smart meters and whole AMI, the so called smart metering arises as one of the keys for the Smart grids implementation. And contribute to benefit the participants of the grid:

- **Customers benefits:**

The consumers will be much better informed on how much and when they consume energy, with that information and in-home displays they are empowered to manage their consumption and costs, in order to perform energy savings actions and/or reduce their bill by having access to the price rates. The customer can receive a better and tailored service, as the utility is able to have a detailed visibility of the consumer behaviour.

- **Suppliers benefits:**

The estimated and manual lectures will be avoided; they will obtain nearly real time readings and a perceptible cost reduction to obtain the lecture. Customer service will be much better, thanks to the massive amounts of energy consumption data, new innovate offers appear such as the time-of-use tariffs, towards a consumer engagement thanks to the possibility of consumers segmentation can offer tailored services.

Collaterally, due to the AMI implementation the electrical system will be more efficient, and reliable, bringing a losses and costs reduction. Also, a rapid outrage or system faults detection and posterior quick repair of the problem, reducing interruptions and improve the service. In a more open perspective new business will arise mainly in the IT and data analysis fields, due to the large amount of data generated and needed to be treated. So, society will have the tools to overall energy savings and CO_2 emission reduction due to less energy consumption.

Anatomy of AMI:

The Advanced Metering Infrastructure is composed by several elements from the inside of the customer building until the utility. These components are: Home display units, Home Area Network, Smart meters, Wide Area Network, Data collection Head-end, Meter Data Management Systems (MDMS) and Data portal to stakeholders systems.

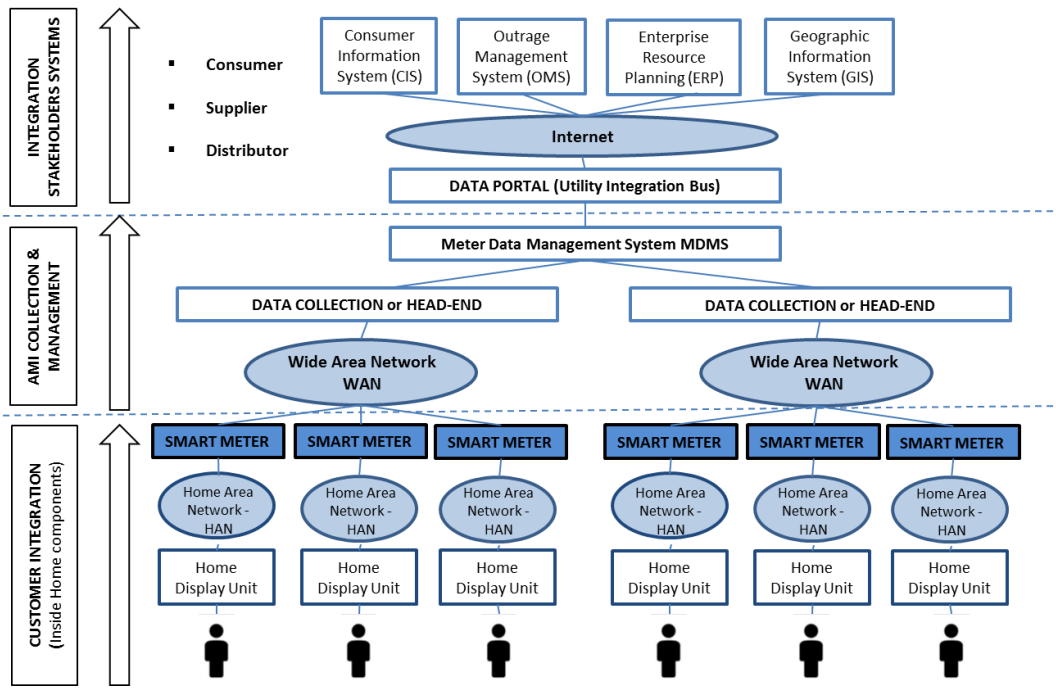


Figure 16. Advanced Metering Infrastructure diagram, from customer's home to stakeholder's services. Own figure adapted from: (Evans, 2007)

- **Home displays units:**

Are the different interfaces with the consumer, where is displayed at almost real time the information about energy use (kWh), cost (€ or $) and even CO_2 emissions. It can be different type of devices, specific equipment devices from a particular company that just has this functionality, or may be integrated in a computer through a website, an application in the smartphone or tablet. These units offer the visualisation of the data collected and emitted from the smart meter, in Figure 17 an example is shown.

Figure 17. Monitoring energy consumption in different supports. Source: (Enerbyte Smart Energy Solutions, 2014)

- **Home or Local Area Networks (HAN or LAN):**

A Local Area Network (LAN) is a computer network interconnecting computers within limited areas, houses, schools, labs… and able to exchange data between them thanks to the gateways. The Home Area Network is a specific LAN that facilitates the communication between devices and appliances within a home limits.

Figure 18 shows the smart devices (smart meters, computers, tablets), smart appliances (fridge, washing machines, washing, kitchen…), micro-generation from decentralised energy resources (DER) and others able to communicate inside the HAN. The smart devices are already in our daily life, but in a future vision of a smart home all the appliances will be connected through different communication protocols, and thus, automatically controlled.

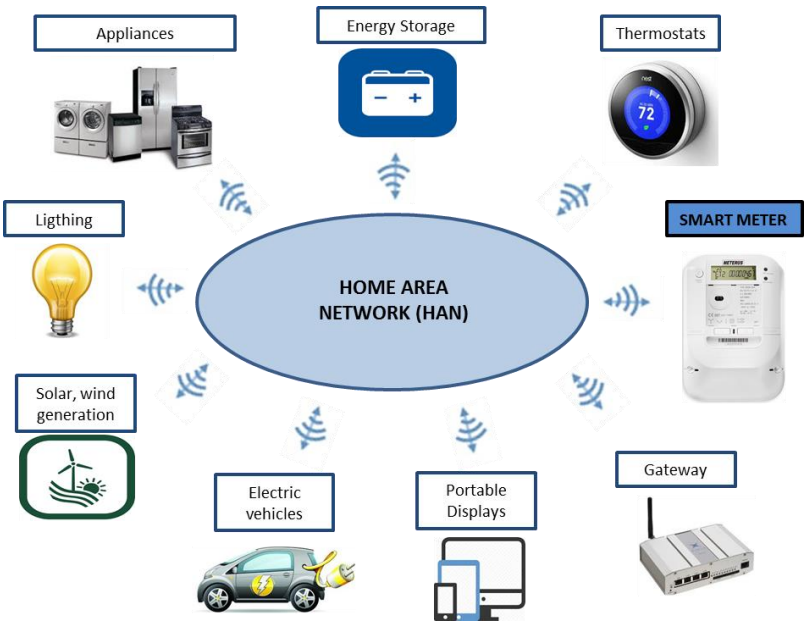


Figure 18. Conceptual model diagram of a Home Area Network (HAN) and features

- **Smart meters:**

Is the key element for the all AMI to work properly, is placed in the home or building limits and recollects all the energy consumption in intervals of time, as well as is able to provide the electricity price to the consumer. Due to its connection is able to remotely report data to both consumer and utility.

Figure 19. European Smart Meter with Open Smart Grid Protocol. Source: Meterus

- **Wide Area Networks (WAN):**

These networks are capable to cover a broader area, are the medium in which various groups of smart meters from different LANs communicate the data collected to the data concentrator or head end, as illustrated in Figure 20.

Figure 20. Wide Area Network (WAN) model, communicating and aggregating various smart meter measures from the Home and Local Area Networks

There are different ways to communicate the smart meter to other AMI components, to enumerate some communication technologies that may be appropriate:

- Power line carrier (PLC): carries data across the power lines that deliver electricity
- Broadband over Power Line (BPL): Also uses the power lines to communicate data, but requires concentrators and repeaters.
- General Packet radio Service (GPRS): uses the mobile's operators WAN to transfer data, but is not free as charges per amount of data transferred.
- Radio: uses radio waves to communicate the smart meter readings
- Wi-Fi: wireless connection to internet via an access point, it is low cost but needs high amount of power.
- ZigBee: it is a low power and low cost, wireless protocol suitable for local coverage.

- **Data collection system or head end:**

The head end is the responsible to put together all data from a large number and disparate set of smart meters. To be able to do this the head-end system should use the same protocol as the smart meter.

A protocol is system of digital rules for data exchange within or between computers. Smart meters may use various types of protocols, such as Open Smart Grid Protocol (OSGP), Transmission Control Protocol and Internet Protocol (TCP/IP), IEC, ANSI…

- **Meter data management systems (MDMS):**

MDMS is a single repository capable to store massive quantity of smart meter readings, it gathers the data. Smart meters are able to record the interval consumption, so they produce a large amount of data, for example a reading per hour or every half hour can produce around 4.000-8.000 readings per year per single meter.

- **Data portal:**

Is where the data is analysed and transformed into valuable information, by making it available to the stakeholders (customer, distributors and suppliers). MDMS database enables interaction with other information systems, see Figure 21; such as Consumer Information System (CIS) for billing systems, and the utility web site, Outage Management System (OMS), Enterprise Resource Planning (ERP), power quality management and load forecasting systems, Geographic Information System (GIS), Transformer Load Management (TLM).

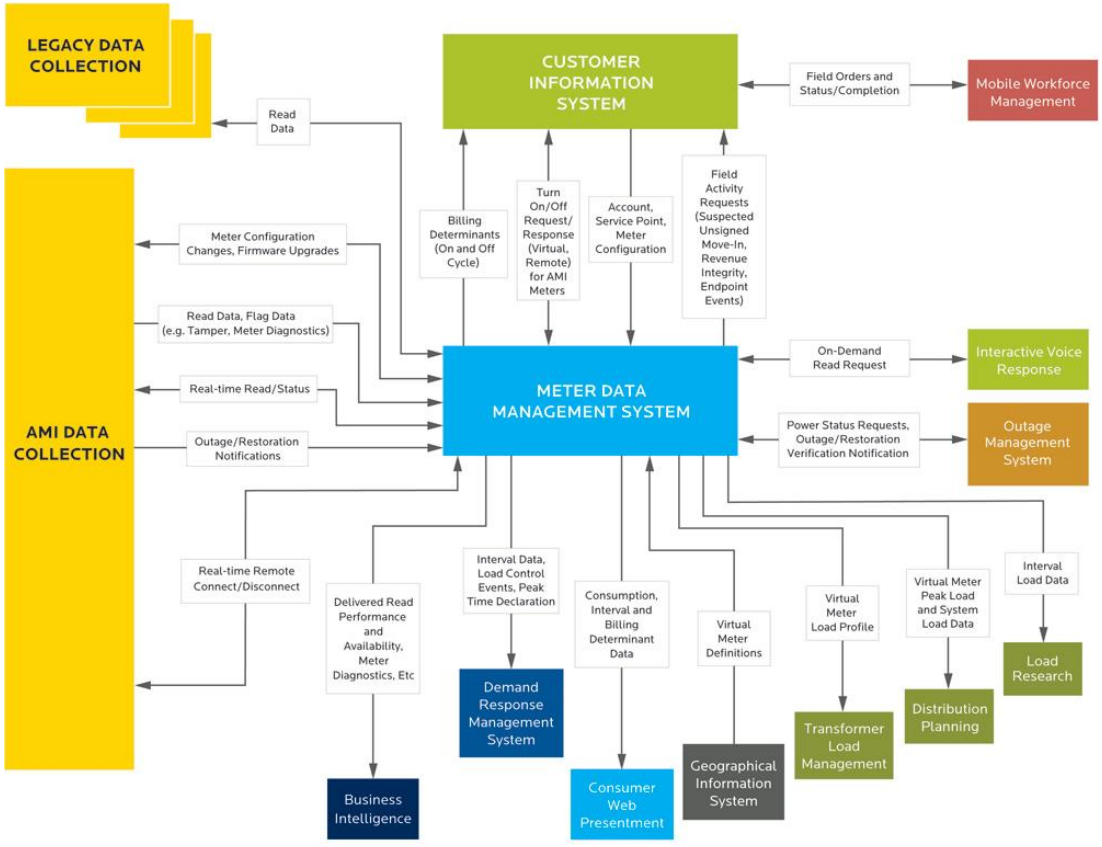


Figure 21. Meter Data Management System Inputs/Outputs and Information Systems served. Source: (Evans, 2007)

This is the whole process of data collection, communication and analysis within the AMI infrastructure; upgrading its data value by turning it into information. Then, it reaches the customer or utility and gives them the opportunity to better manage either consumption or service thanks to the support that represents to have this data processed.

2.2.2. Smart meters

Smart meter is the core of smart metering and AMI, their functionalities are far extended beyond the simple reading the energy consumer over a period of time, usually a month like the previous traditional meters do. The smart meter is capable to read the energy consumption remotely and with a much more reduced time intervals, usually configured to hourly, half hour or quarter frequency. And communicate this information to the utility and also to the consumer, due to its connectivity. However, smart meters are not exclusive of electricity, also there are water and gas smart meters available.

Many more functions are performed by the smart meter; offer a time-based price information to the customer (time-of-use tariffs), for example in peak hours when the electricity is scarce the prices are higher in order to look for a demand response to reduce its consumption during these peak hours, see Figure 22. This fact have a double benefit, gives to the customer the information to make smarter choices on when to consume the energy, and for the utility allow to prevent blackouts and offer a better customer service.

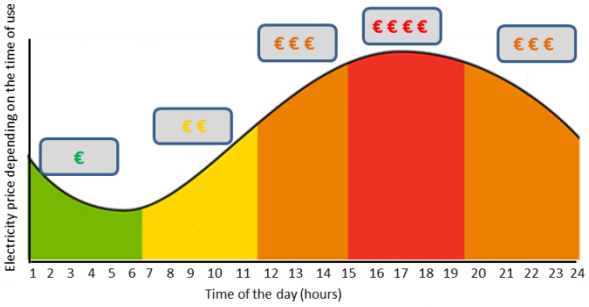


Figure 22. Time of use tariffs graphs, electricity is more expensive in the afternoon where the peak demand is

Offers the possibility to net metering, which allows to the consumers with an on-site generation plant, for instance a roof top PV installation, to consume their own energy and then export the surplus electricity to the grid reducing the energy imported from the grid, as shown in Figure 23. The term of "prosumer" or "self-consumer" arisen this last decade, since renewable energies were affordable for any household and the electricity price has continued growing. Depending on the legislation of each country net metering is able to sell this surplus or simple discount the electricity produced from the consumed from the grid.

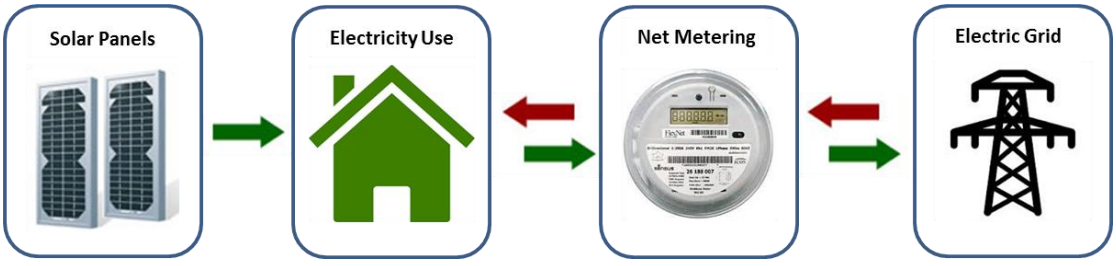


Figure 23. Scheme of net metering, two way Smart meter importing and exporting electricity with the grid

In the future in-home perspective the smart meter will have the possibility to communicate to other intelligent devices at home, the increasing penetration of smart appliances and devices with connection, mainly Wi-Fi; will bring the smart home where all these connected appliances will be remotely controlled by the household. A pilot example of a smart home can be found in the Italian project Energy@home (Energy@Home, 2010), also among quite many the NEST thermostats (NEST, 2015) are examples of devices intelligence and connection.

Smart meter rollout

The European Union has a plan to implement electricity and gas smart meters throughout the European countries, a 2014 report on the smart metering deployment (European Commission: Smart grids and meters, 2014) shows that:

- Already 200 million electricity smart meter are deployed
- Is expected by 2020, the 72% of electricity consumers will have an electricity smart meter (less than the 80% planned)
- The average cost per smart meter installed is between 200-250 €, however it is proved that electricity smart meter can save around 309 € for electricity per metering point also allowing energy savings around the 3%.

Smart Meters deployment in Europe

It seems that European smart meter implementation is growing faster than other regions such as North America, Asia pacific or Latin America.

Among European countries there are also differences between those that agreed on the large-scale rollout and those who not, for many reasons but mainly due to the fact that the cost-benefit analysis was negative. An illustration of the countries position is seen in Figure 24. To have detailed information about each country the European commission tracking reports (Country fiches for electricity smart metering , 2014) explains the countries plans regarding the smart meter rollout.

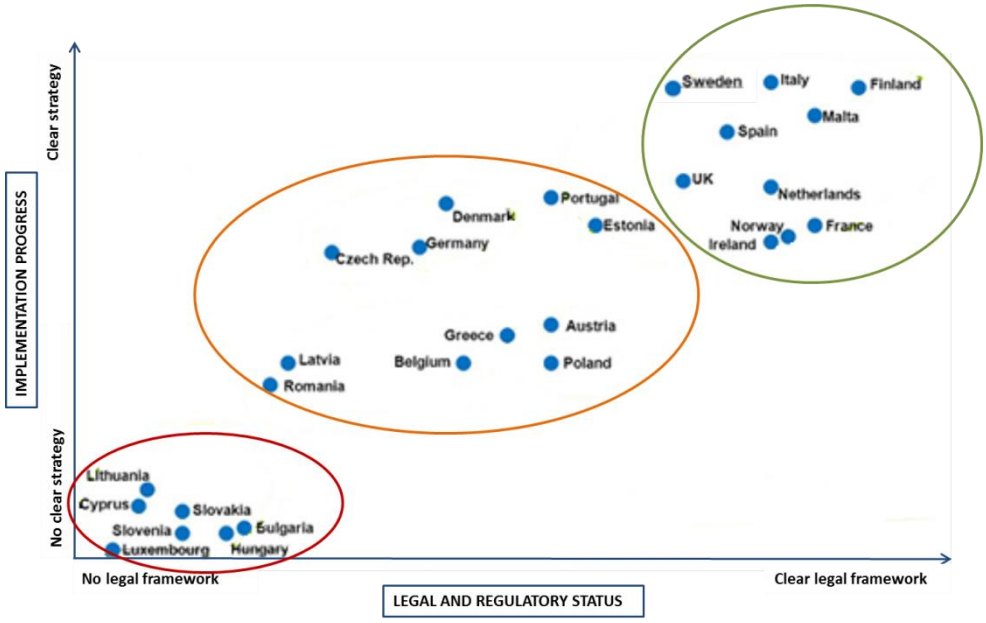


Figure 24. Strategy and Smart meter implementation in European countries. Adapted from: (Smart Regions, 2013)

Regarding the rest of the world regions, the Navigant Research study "Smart Electric Meters, Advanced Metering Infrastructure, and Meter communications: Global Market Analysis and Forecasts" (Navigant Research, 2014) predicts that by 2020 North America, Europe and Asia pacific will have installed around the 70% of smart meters, then Latin America is expected to almost reach the 50% and finally the Middle East & Africa is not expected to reach the 10%.

In North America, by 2014 only 8 states have a penetration higher than the 75%, but the majority, 28 states are moving under the 25 % penetration, he reasons for this slow implementation are mainly to the lack of clear policies, lower electricity prices in certain states and lack of utilities interest. The complete statistics can be found in a GreenTechMedia report (Lacey, 2013).

European Union regulation and recommendations

In Europe the first regulation covering the smart metering legislation is found in the Europe Electricity Directive 2009/72/EC Annex I.2., later on the European Union has set common minimum functionalities for the smart metering are included in the Recommendations 2012/148/EU. These functionalities capture the essential elements that a smart metering should have to benefit all stakeholders, and can be summarized in Table 2 (Cost-benefit analyses & state of play of smart metering deployment in the EU-27, 2014).

Table 2. The ten must-have Smart metering elements to fill with the functionalities for each stakeholder

CONSUMER	a) Provide readings directly to the consumer and /or a third party b) Update readings frequently enough to use energy savings schemes
METERING OPERATOR	c) Allow remote reading by the operator d) Provide two-way communication for the maintenance and control e) Allow frequent enough readings for networking planning
COMMERCIAL ASPECTS OF SUPPLY	f) Support advanced tariff system g) Remote On/Off control supply and /or flow or power limitation
SECURITY – DATA PROTECTION	h) Provide secure data communication i) Fraud prevention and detection
DISTRIBUTED GENERATION	j) Provide import/export and reactive metering

As stated earlier, consumer will have an active and particular importance, its functionalities are crucial. So measure the consumption and make it available to the customer ideally every 15 minutes; however in many cases if this data is available every half-hour or hourly frequency can be enough to support the advanced prices structure that is the other key functionality for the customer, in order to save costs and reduce the peak in energy demand, engaging the consumer and obtaining energy savings. But in any case providing data every 24 hours or more can be considered for an appropriate service and customer energy management.

That is why is important for the smart metering systems to allow the automatic transfer of information of both consumption data and advanced tariffs options in a standard interface.

A summary of the recommended functionalities accomplishments per county is shown in Figure 25, classifying if the actual implementations are in correlation with the recommended functionalities, partly or not.

		Func (a)	Func (b)	Func (c)	Func (d)	Func (e)	Func (f)	Func (g)	Func (h)	Func (i)	Func (j)
States In Rolling -out Smart Meters	Austria	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Denmark	YES	Partly	YES	YES	YES	YES	YES	YES	YES	YES
	Estonia	YES	Partly	YES	YES	YES	YES	YES	YES	YES	YES
	Finland	YES	Partly	YES	YES	YES	YES	YES	YES	YES	YES
	France	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Greece	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Ireland	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Italy	YES	Partly	YES	YES	YES	YES	YES	YES	YES	YES
	Luxemburg	NA	NA	YES	YES	YES	NA	NA	YES	NA	NA
	Malta	Partly	YES	YES	YES	YES	YES	Partly	YES	Partly	YES
	Netherlands	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Poland	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Romania	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Spain	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES
	Sweden	YES	Partly	YES	YES	YES	YES	Partly	YES	Partly	YES
	United Kingdom	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
States NOT Rolling -out Smart Meters	Belgium	YES	NA	YES	YES	YES	YES	YES	YES	YES	YES
	Bulgaria	YES	NO	YES	YES	YES	YES	YES	YES	YES	Partly
	Cyprus	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Czech Republic	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Germany	YES	YES	YES	YES	YES	YES	NA	YES	YES	YES
	Hungary	Partly	YES	YES	YES	YES	YES	YES	YES	YES	NO
	Latvia	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Lithuania	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Portugal	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
	Slovak Republic	YES	YES	YES	YES	Partly	YES	YES	YES	YES	Partly
	Slovenia	YES	Partly	YES	YES	YES	YES	YES	YES	YES	Partly

Figure 25. Summary of the recommended functionalities accomplished or not per country in Europe, divided by those countries involved in the smart meter rolling-out and those who not. Adapted from (Cost-benefit analyses & state of play of smart metering deployment in the EU-27, 2014)

Spain is different

From the table above, a special mention should be done to Spain which is the only country that does not allow the consumer to receive up-dated consumption data, this is not a rare issue since the Spanish electric system was heavily regulated and behaved as an oligopoly due to political interests, where the consumers have a passive paper and were not clearly informed.

The market liberalization and the fight of various collectives is pushing to change the current situation, improving the customer services by the possibility of accessing to the consumption data. But still only five distribution companies have the 90% share of the electrical grid, these companies are Endesa, Iberdrola, Unión Fenosa, E-ON, HC Energia –EDP; are not so transparent neither accessible; and probably one of the reasons for higher electricity prices in Spain (Eurostat, 2014). However, some new suppliers (Somenergia, Holaluz, Nexus…) are appearing in the market and gaining quote thanks to its transparency, commitment to the clean energy, offering competitive prices and new services. Nevertheless, the change is slow.

Regarding the billing system, the Spanish Government has approved a legislation where the retailers have to calculate the electricity bill hourly, taking advantage of the smart meter deployment, providing finally a real pricing to the consumers. It is expected that by October

1st, 2015 all the consumers with smart meters will be billed using the hourly prices. Although the smart meters were installed some years ago and the consumer is paying for having it installed.

But still, the consumer has no straightforward access to its own daily consumption; in most of the cases the consumer only knows its consumption when receiving the bills every one or two months. So, if the consumers are willing to know its daily electrical consumption they have to request explicitly to the retailer or distributor, even in this case consumers will receive it in 2 days of delay. Usually with the traditional utilities are more reluctant to changes and the new retailers are gaining market share by offering these kind of services, such as energy consumption visualization.

3. New services and business opportunities

All this evolution and regulations changes are opening the possibility to new services or business within the energy sector and more specific in the electricity market. In addition, energy represents one of the main expenses, for both residential and non-residential consumers, moreover prices has increased sharply these recent years.

These businesses could be simply divided between hardware and software:

The hardware is referred to the physical devices mainly metering and sub-metering equipment able to measure more precisely the energy consumption (electricity, gas, water); this hardware device is installed in addition to the utility meter. The companies that supply the hardware are charging an initial cost for the purchase of the product. To name some, General Electric, Schneider Electric, Itron, Elster Group, Circutor, Carlo Gavazzi, Current cost…

The software is the part of displaying the information in a monitor, all the analysis of data collection and processing to be shown in an interface. Allowing the consumer to receive and visualize the information desired. Usually, the companies that supply this software charge a monthly fee for the provision of it.

Among the larger software suppliers for energy monitoring there are: C3 Energy, Plotwatt, Fifth-play, QwikSense, E-sightenergy, Envizi, Dexma, Wattio; which are focused on industries, commercials, office buildings. However, there are also other software companies like Opower, Simple Energy, Bidgely, Smappee, Onzo, Mirubee or Enerbyte that are devoted to the residential sector; offering the service through electrical companies or public municipalities to reach more households giving a better customer service with more detailed information.

Today, these services are far more extended and offered to the non-residential clients, such as industries, commercial, offices buildings, hotels…; due to the fact that the return in terms of energy savings and cost reduction is higher for the big energy consumers. The cost of the equipment (hardware and software) does not pay-off for a residential client, and therefore few companies offer this service to residential.

These equipment and services' businesses are not related to the electricity supplier or distributor, it is a complementary service that the larger consumers need to delegate to external companies since the electric utility doesn't offer this service or because they need more detailed information by installing sub-meter to each electricity consumption device.

For the large amount of data generated and the need to be managed, companies employing data scientist and data analysts are growing faster by offering services to extract valuable information to the client business from this big amount data, by using data mining techniques or predictions applied to different fields such as banking, media, energy.

Opower

In the energy sector, Opower (Opower, 2015) has adopted a combination of computer science, behavioural science and data science to help electric utilities reduce the energy consumption and engage their residential customers. Opower analyses massive amount of data (meter data, customer insights, and operational insights) in order to translate it into a personalised experience for the households. The final aim is to help consumers to lower their energy use, the associated costs and the carbon emissions derived.

A powerful back-end based on data mining, analytics and machine learning enables this U.S. software company to retrieve valuable information, and offer innovative services to promote energy efficiency and demand response among the consumers. For instance: attach to the utility bill a personalised explicative report describing the consumer performance in comparison to their neighbours or to own last year's consumption; segmentation of the customers to group them by similar consumption behaviours, similar load patterns or similar characteristics, among many others that add value to the utility service.

Opower is the market leader offering this kind of service to utilities; in North-America, thanks to a favourable legislation and incentives, it achieved a high degree of penetration in different electric utilities; however in Europe the market is not mature yet to adopt these services.

3.1. Enerbyte

Enerbyte Smart Energy Solutions S.L. is a Software as a Service (SaaS) start-up founded in 2012 based in Sant Cugat del Vallès (Catalonia).

The need

Enerbyte has seen a need to cover, therefore a market opportunity in the actual electricity sector. Electricity retailers are facing new challenges:

- Competition is increasing due to the market liberalization
- More strict environmental legislation has been launched such as the Energy Efficiency Directive in concordance to the EU Horizon 2020
- Easy access to new technologies (solar PV), raises the decentralised energy production leading to net metering, by exchanging electricity from and to the grid.
- Unsatisfied customers, due to not understanding the energy billing process generates a distrust of the energy utility services and low service satisfaction, despite the high quality of supply

The solution

Enerbyte developed a Business-to-Business-to-Customer (BtoBtoC) business model as shown in Figure 26, instead of reaching the end-consumer one by one as the Business-to-Customer (BtoC) does. The reason to adopt the BtoBtoC model was to provide this service to small electricity consumers as residential and small business; which are the forgotten ones when talking about energy monitoring and management because they consume much less energy than corporate, office buildings or industries an investment in that sense won't have a return.

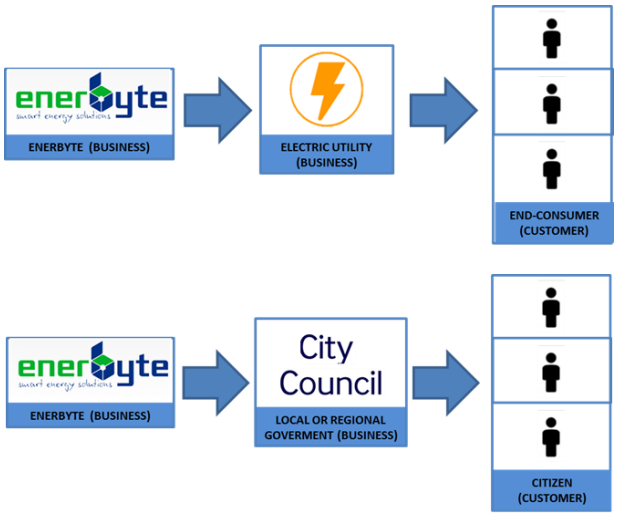


Figure 26. The B-to-B-to-C business model of Enerbyte for Electric Utilities (top) and Local and Regional Governments (down).

So, Enerbyte provides its product as a service to another business, the electricity retailer, which finally offers it to the end consumer. Although the electricity utility was the main targeted customer, recently another customer raised, the public authorities such a local or regional governments that wants to provide it to the citizen however in this case the approach is a bit different.

The objective

Enerbyte's product, named, "Personal Energy" is a multiplatform digital tool in form of app, web and reports that fosters energy efficiency between citizens and customers from electrical retailers. By the app and web service allows the end-consumer to have immediate access to the information it requests. The newsletters and the reports allow the end-consumer to have a summary about its performance related to energy consumption and costs. In Figure 27 the interfaces of each platform can be seen.

Using big data analytics, Enerbyte aims to easy create a tailored customer relationship by establishing user segmentations, with the final aim to change positively the consumer behaviour of the energy usage to obtain both energy and costs savings obtained, hence, increasing the engagement and satisfaction of the end-consumer. The final aim is to enable the end-consumer to save around a 10% of its current energy consumption when using the service.

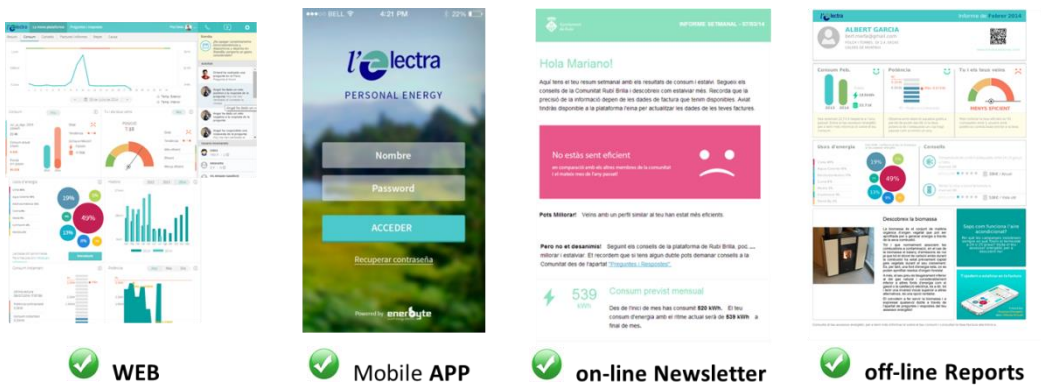


Figure 27. Personal Energy platforms and communications channels. Source: Enerbyte Smart Energy Solutions

The value proposition

The electrical utility will have an added value that will differentiate it from the competitors, as it will be able to provide a better service to the end consumer by means of monthly reports, giving them access to easy and understandable information through new communications channels engaging the consumer and increasing its satisfaction, and the features presented in Figure 28.

The end-consumer will be able to visualize its energy consumption and the related costs, compare its performance to its previous consumption and thanks to the disaggregation the consumer will know which is its major energy use.

Gamification helps the user to achieve its energy goals by obtaining points or medals for their achievements, and placing user to an ranking depending on its performance. Related to that the social community permits the interaction with other users and a common space where to shared doubts, experiences, etc…

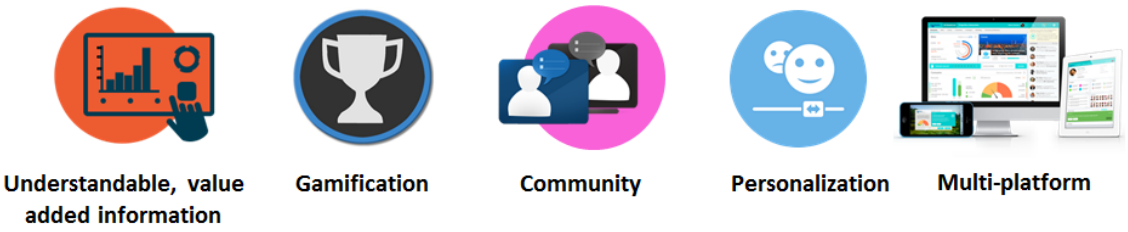


Figure 28. Features that "Personal Energy" tool offer. Source: Enerbyte Smart Energy Solutions

Personalization is where the power of data analytics appears, as it forecasts each user's consumption based on their historical data, helps to set energy saving plans. It is a tool that increases the end-consumer education, knowledge and awareness related to energy consumption, as the user receives personalised recommendations according to its consumption profile leading them to reduce its energy consumption and costs.

Projects

We can differentiate two types of projects, depending on business to which the service is provided; first, electrical utilities and second, the municipalities.

- **Electrical utilities:**

Electra Caldense (www.electracaldense.com)**:** is a small electrical utility that supplies electricity to the 8.000 inhabitants in Caldes de Montbui (Catalonia); the utility have already installed smart meters from where the measurements are taken and data is obtained. The service offered from Enerbyte is to deliver monthly reports, where the end-consumer can see its last month electricity consumption, its comparison to the same month of the previous year and the comparison to the rest of the user with the same contracted power. It has been proved an improvement of the relationship between the utility and the end-consumer, engaging the customer by improving its satisfaction.

- **Municipalities:**

Rubí Brilla (www.e-ajrubi.net/rubibrilla)**:** was the pilot project for municipalities supported and within the framework of an European project, started in 2013. Rubi is a municipality with around 70.000 inhabitants nearby Barcelona. In the case of municipalities the electrical consumption data, is not possible to be obtained from the utility smart meters, that is why it is needed to install sub-meter equipment in the houses and this lectures are used as the data to input into the platform. In the case of Rubí, the municipality has provided around 150 sub-meters to the households that have applied to the project.

Barcelona city council (ajuntament.barcelona.cat/autosuficiencia/ca/): at the beginning of the 2015 Barcelona has launched the same kind of project as the Rubí, by providing 400 sub-meter equipment for free to the citizens that have applied to the project. The sub-meter devices are able to provide accurate electrical consumption data.

Gipuskoa (www.argitu.eus): is a region in the Basque country (Spain) that also launched a project with the peculiarity that consumers from different cities are able to apply for it. The regional government provided around 350 sub-meters to the citizens that applied for the project.

CHAPTER 2: CASE OF STUDY. ELECTRICITY LOAD PROFILES CLUSTERING

4. Introduction

4.1. Preliminaries

The case study is organised in various sections starting with a description and the philosophy of the project of "Rubí Brilla" from which the electricity consumption data and house information is taken and used as input for the study; a remark is done in how consumer behaviour can contribute to foster energy efficiency; then the objectives and goals are defined.

Moving to the more technical part, the procedure of cleaning the raw data is explained to have the processed data in order to start the data analysis sought. First, an exploration and visualization of the data was carried out to extract the first outcomes from the consumers load profiles.

The core of this project focuses on the consumer's segmentation according to their electrical load profiles, various clustering methods are used and compared in order to find the most adequate consumer segmentation. Then, the house features and household characteristics are considered to find whether exists any relation between customers segments and their house properties or not.

4.2. "Rubí Brilla" project

The project of "Rubí Brilla" is a pilot and innovative project that the Rubí city council has launched by the year 2013, within the framework of a European Initiative called "Covenant of Mayors" (Covenant of Mayors, 2015).

Roughly speaking, the whole project aims to foster energy efficiency and energy-related knowledge to the small consumer, mainly residential end-consumer but also some small business; who can obtain up to 10% of savings by being active. The Rubí city council provided electricity sub-meters, for free, to the citizens who applied to the initiative, up to 150 devices, and also the access to a platform to visualise its consumption, receive recommendations to save energy and common space to share experiences and doubts among the community.

This initiative involves the participation that can be divided among Private, Public and People, detailed in Table 3:

Table 3. Summary of the participants to the Rubí Brilla project

PUBLIC		PRIVATE		PEOPLE	
Local authority	**University**	**Hardware provider**	**Software Provider**	**User**	**Local associations**
Rubí City council	UPC	Circutor and Current Cost	Enerbyte	Rubí citizen	Various

- **Public players:**

Rubí city council: was the entity that started the project and it is currently managing it. They have seen a necessity to provide tools to their citizens to increase their knowledge and awareness and foster the energy efficiency among them. The city council was the responsible to find the partners, both software and hardware providers.

Universities: an agreement with the Polytechnic University of Catalonia (UPC) was reached and students participated with the project by doing some field studies, their principal task was to perform detailed technical audits from each house adhered to the program.

- **Private partnership:**

Hardware provider: the companies that provided the sub-metering equipment to be installed in the houses, were "Circutor" and "Current Cost". A total of 150 sub-meters were supplied and installed. These devices communicate the electrical consumption data using the household's Wi-Fi and sending it to the database server where the data is stored.

Software provider: The start-up "Enerbyte smart energy solutions" is the responsible to offer the cloud-based platform where to visualize personal consumption and interact with the others users. Enerbyte also sends a newsletter to the users with their performance and some recommendation related to the energy efficiency every two weeks.

- **People:**

Users: Citizens of Rubí willing to reduce its electricity consumption, its costs and improve its energy efficiency having access to sub-meter and the platform.

Local associations: associations related to energy or environmental issues, able to prepare activities and campaigns that will have a positive impact to the project and help the citizens.

Ideally the electrical consumption data would have been obtained from the household's smart meter, but the Spanish electrical utilities are not yet offering easy access to this hourly data to the end-consumer. For this reason it was necessary the installation of the sub-meter devices are able to provide measurements each 5-15 minutes, however these devices communicate data via the consumer's Wi-Fi fact that brought many problems as it directly affected the data quality.

4.3. Consumer behaviour and data analysis

The data generated has increased exponentially this last decade, this data could contain valuable information but it needs to be explored. The electricity market is also living a data boom, for instance the electricity meter reading has moved from one read every month to host smart meters able to read electricity consumption every 15 minutes; this implies that each consumer will have around 35.000 measurements per year.

Store and manage this data is already a challenge itself, that is the reason why ICT and computer science have increased their presence in the energy sector; but also the possibility to add value to this massive amounts of data applying data mining and analytics to extract hidden information or at least be able to separate the useful one to the less interesting. Big data analytic tools are essential to add value to all this data, specifically in the energy sector could add value in terms of energy balance, energy efficiency or energy prediction; that could be profitable for both the consumer and the utility.

Consumer behaviour and demand response can contribute significantly to foster energy efficiency in the residential sector pushing a behavioural change to boost energy savings. By analysing different types of data, such as, consumers' load profiles, households' properties and householders characteristics; some patterns could be found and knowledge could be discovered in order to guide the end-consumer not only how to reduce its energy consumption but also when is the best time to do so.

Specialised software programs and programming languages are used to perform such analysis involving large sets of data; for instance: customer segmentation by clustering them by load profiles similarities, finding correlation between electrical consumption and temperature, performing consumption data exploration and visualisation or consumption prediction; among many others.

In the current project's data analysis the "R" programming language is used, "R" is a free open-source software that could be used for both experts and beginners who have the first contact with the programming languages as many on-line help material and tutorials are available, which facilitates the learning.

5. Overall Objectives

The main objective is to discover knowledge from the electricity consumption data measured in residential buildings, in concordance to the idea to provide personalised energy efficiency recommendations to the consumers. Enerbyte is aiming to stablish a methodology to automatically classify any new user into load pattern group by looking at the household and house characteristics.

This process consists in three phases described below, and graphically shown in Figure 29:

- **Phase 1:** From the data already existing, the current users are clustered/segmented into different groups according to the similarity of their electrical load profile. Similar load profiles means that the consumer behaviour and habits are almost identical.
- **Phase 2:** Once, the users are placed in one cluster or another, a further analysis is carried out in order to find the more likely properties related to the house and household of each group.
- **Phase 3:** Classify any new user to one of the existing segments by looking at the household properties and the house characteristics.

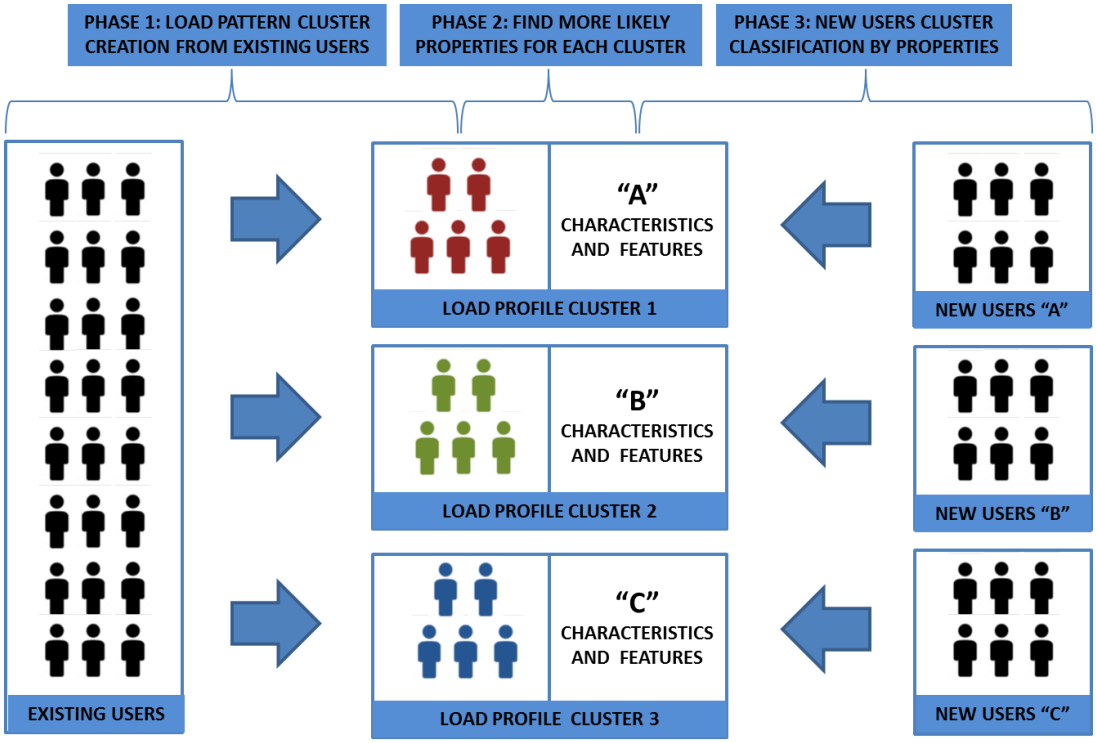


Figure 29. Diagram that describes the phases of the methodology applied in this study

The aim of this project is aligned to the Enerbyte objective, however the sample considered from the "Rubí Brilla" project is relatively small to extract rigorous conclusions and fulfil all the Enerbyte's objectives. Thus the main goal of this study is try to accomplish the Phase 1 and Phase 2 from the list above, creating the methodology and test it, to later check its validity and use it to a larger user sample.

The current project's scope is, then, to find the load profile of each residential household in order to segment the consumers into groups according to the similarity of their electrical load profiles. So, the consumers within a group will have similar consumption patterns, which will be different from the consumers' load profiles of the other groups.

Also, the household and householder's characteristics were analysed in order to find the most common attributes for the whole set of consumers and inside each group; looking for the existence of any pattern or relation when crosschecking load profiles groups and house characteristics.

Likewise, additional actions were done in terms of exploring and visualizing data in order to be incorporated into the newsletter or monthly reports that are sent to the Rubí users periodically. And in parallel to provide to the Rubí municipality the starting point for deeper data analysis that is currently willing to perform.

6. Methodology

The steps followed to conduct this work and obtain the final consumer segmentation are described below:

A first stage previous to the analysis, covers **(a) the data Collection** of the hourly electric consumption of the user and also the household and householders characteristics from the project of "Rubí Brilla", understanding how the data is presented and reshape it at convenience for the study. Once getting familiar with the dataset and having seen the problematics; a process of **(b) Cleaning the data** is carried out to remove the bad data (zeros, frozen, duplicated features). Closing this stage, a graphical representation is sought as **(c) a data exploration and visualization** to go deeper on the data used for the study, and also help to check its validity.

The second stage is devoted to the clustering analysis to obtain the desired consumer segmentation which is the core of the work; it is divided in pre-clustering, clustering and post-clustering phases.

The **(d) pre-clustering** phase consists on doing a literature research of similar studies, selecting the most adequate format of the input data to allow a fair comparison among the electrical load profile of the users, and processing the input data in that sense obtaining the proportional energy usage per hour for each user, in percentages.

In the **(e) clustering phase**; a review on the existent clustering techniques used for electric consumption data segmentation is done. Hierarchical, K-means and Self-Organising maps techniques are selected and applied individually in an iterative process, based on computational clustering calculation (using R software) in order to find the optimal number of clusters with the appropriate number of members in each of them. Finally, a visual and statistical comparison of the results of each clustering technique is performed to determine which solution is the most suitable.

The final phase, is the **(f) post-clustering**; which is devoted to obtain the final and desired consumer segmentation by the redistribution of the users. This is performed by a manual intervention done by the analyst applying visualization and statistical techniques to be able to find the outliers and reallocate them to a more appropriate group.

At the third stage, the **(g) household and householders characteristics** are analysed. Selecting those features that could add value to the analysis according to related previous papers, in this case histograms are used to find the most common characteristics inside each group of consumers.

7. Procedure of data collection and data cleaning

7.1. Objective of cleaning data

To perform any data analysis the data quality is essential also the preparation of this data into the right format is the key to start the analysis sought. Hence, it is important to know beforehand the purpose of the analysis by having one or other purpose the procedure to go into the data processing may change completely. A good data preparation will also report better and more accurate results.

The steps to be followed can be summarised as shown in Figure 30:

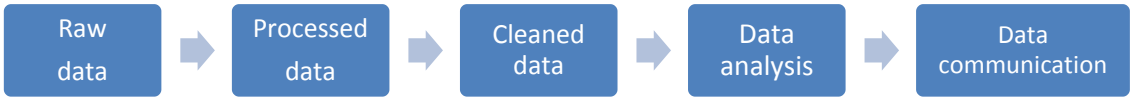


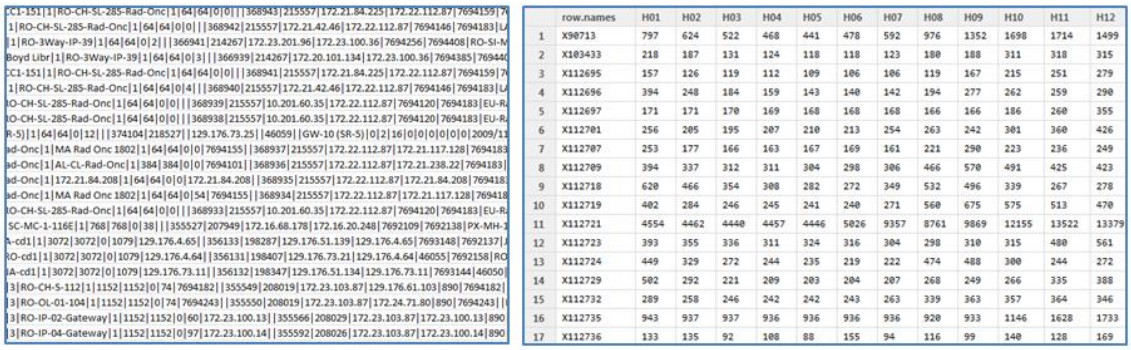
Figure 30. States of the data when performing a data analysis. Source: Coursera "Data Scientist toolbox" course

Once the data motivation is clear, so once having the specific idea to what we want to extract from the data, we can start looking at how the raw data is presented. It is not likely to receive the data with the precise format we would like to have in order to start the analysis, usually raw data is hard to visualise and therefore analyse.

Processing the raw data is needed in order to transform the data into the appropriate format, this data ready for the analysis can be called tidy data or processed data. To achieve it computational data processing is used, by writing and running a piece of code called processing script that will allow getting the data ready.

When the data is processed, the next step before starting the data analysis is the cleaning the data. What cleaning data does is to prevent and correct the errors or the incompleteness of the dataset, for instance with duplications, missing data, NAs…

In Figure 31, an example between raw and processed data can be seen; the image on the right is what we would like to see our data in order to run the computation data processing:



	row.names	H01	H02	H03	H04	H05	H06	H07	H08	H09	H10	H11	H12
1	X90713	797	624	522	468	441	478	592	976	1352	1698	1714	1499
2	X103433	218	187	131	124	118	118	123	180	188	311	318	315
3	X112695	157	126	119	112	109	106	106	119	167	215	251	279
4	X112696	394	248	184	159	143	140	142	194	277	262	259	290
5	X112697	171	171	170	169	168	168	168	166	166	186	260	355
6	X112701	256	205	195	207	210	213	254	263	242	301	360	426
7	X112707	253	177	166	163	167	169	161	221	290	223	236	249
8	X112709	394	337	312	311	304	298	306	466	570	491	425	423
9	X112718	620	466	354	308	282	272	349	532	496	339	267	278
10	X112719	402	284	246	245	241	240	271	560	675	575	513	470
11	X112721	4554	4462	4440	4457	4446	5026	9357	8761	9869	12155	13522	13379
12	X112723	393	355	336	311	324	316	304	298	310	315	480	561
13	X112724	449	329	272	244	235	219	222	474	488	300	244	272
14	X112729	502	292	221	209	203	204	207	268	249	266	335	388
15	X112732	289	258	246	242	242	243	263	339	363	357	364	346
16	X112735	943	937	937	936	936	936	936	920	933	1146	1628	1733
17	X112736	133	135	92	108	88	155	94	116	99	140	128	169

Figure 31. The left side image shows a typical raw dataset in a ".csv" format; and the right side shows the processed data set in a data frame format, easier to visualise.

A dataset is composed of rows and columns, where, usually, each row represents an observation and each column a different variable. This data can be either qualitative (categorical) such as gender, country, city; or quantitative (numeric) such as electricity consumption, height, price, distance. Again and as said before depending on the type of data the approach used to the analysis may vary, because is no treated the same way numeric and categorical data.

In the next sections, the following types of data are treated to be used as input in the analysis, in our particular case:

- ✓ Electricity consumption data
- ✓ Households and householders information

7.2. Electricity consumption data

The data regarding the electricity consumption is the main one used in the later analysis; it is numerical data as its records the measurements of electrical consumption for every household participating into the project described in section 4.2.

The following Figure 32 summarizes graphically how the electrical consumption data used for the analysis is obtained:

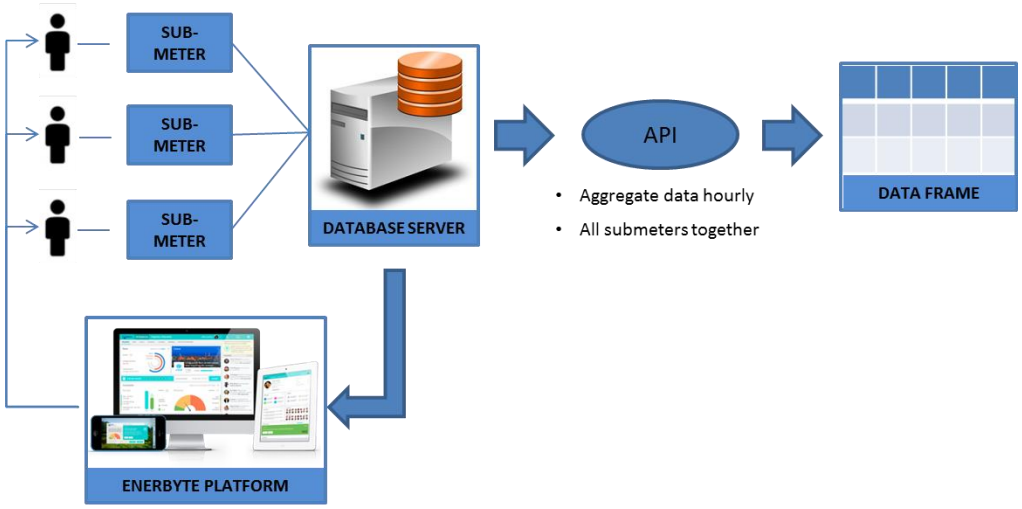


Figure 32. Diagram to explain how the data used for the analysis is obtained

The electrical consumption measurements are obtained from the sub-metering equipment installed in the residential buildings (and some small business); these devices communicate the data through the Local Area Network (LAN), using the household's Wi-Fi and sending it to the database server of the device manufacturer where all data is stored.

Communication via Wi-Fi brought many problems that directly affected the data quality, as if the Wi-Fi is switched off, no communication can be stablished interrupting the sending and reception of data. Then, the data for this time period is represented by zeros as it was no consumption; thus altering the dataset values which need to be eliminated in order to perform an accurate data analysis.

From the database server the Software as a Service (SaaS) company, in this case Enerbyte, access to the data stored and proceed to upload it into the web-app platform for the visualization of each user.

Finally, the data used for the current analysis was obtained through an Application Programming Interface (API) configured by Enerbyte in order to:

- First, aggregate all the data hourly. The sub metering equipment is able to measure the consumption in high-resolution frequencies, each 5-15 minutes according to the device. For the aim of this study described in section 5, the hourly aggregation is enough and a good starting point if a further deeper study is sought in the future.

- Second, all the households consumption data are in the same file, fact that facilitates the analysis procedure as reading one file is enough to analyse the data.

Hence, by using the *"url"* that permits the access to the database server a *".csv"* file is downloaded, this file is the input data regarding electrical consumption that will be used for the data analysis. This data is considered time series numerical as it records the electrical consumption and time of it.

The *".csv"* files have the characteristic that are easy to read by software prepared to conduct a data analysis, for instance *"MS Excel", "R", "Python"* among others. *"R"* is the software used to carry out the current analysis.

7.2.1. Dataset description

The dataset contains for each different household the hourly consumption data per day since the sub-metering equipment was installed beginning of 2014 until the March 2015 data. The dataset contains a total of 47.000 observations (rows) and 26 columns that are:

- "idmeter": sub-meter identification number
- "date": day, month and year of the measurement
- "00:00 -23:00": a column per hour measurement from 00:00 to 23:00.

The size of this dataset is relatively manageable due to mainly three facts.

- The number of sub-meter installed are a small amount accounting for 157 sub-meters
- The measurements period are in most of the cases around year length
- The data was aggregated hourly, fact that reduced the final number of observations and therefore the dataset's size

	row.names	idmeter	date	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00
1	61	90713	24/01/2014	0	0	0	0	0	0	0	0	0	0	0	0	0	382
2	62	90713	25/01/2014	287	592	559	488	913	287	436	638	270	1893	1138	2380	1746	301
3	63	90713	26/01/2014	418	744	814	353	274	285	268	281	623	2193	1151	2078	2625	2408
4	64	90713	27/01/2014	889	876	949	532	681	1102	499	1556	1907	2219	2262	549	592	1322
5	65	90713	28/01/2014	705	1119	898	570	533	513	1289	1221	2164	2232	2490	1965	2410	776
6	66	90713	29/01/2014	1755	406	538	482	289	637	1386	1380	2283	2521	2502	2433	535	693
7	67	90713	30/01/2014	1681	846	687	297	487	503	1172	1442	1852	2508	3112	2341	593	828
8	68	90713	31/01/2014	263	339	263	322	374	658	772	752	2213	902	2320	1610	1599	533
9	69	90713	01/02/2014	576	1026	664	589	322	298	589	354	330	1738	2332	1318	329	1407
10	70	90713	02/02/2014	372	693	590	775	665	633	570	297	1106	2106	1276	2870	1929	1492
11	71	90713	03/02/2014	962	519	257	265	575	897	672	2118	2300	2398	1930	264	1302	730
12	72	90713	04/02/2014	2454	2228	927	1042	740	700	883	1349	2195	2820	2360	2355	501	1434
13	73	90713	05/02/2014	1660	742	302	615	242	854	1146	1754	2132	2565	1029	559	537	697
14	74	90713	06/02/2014	1153	569	835	332	509	673	1231	1879	2232	2812	2440	2062	1792	706
15	75	90713	07/02/2014	986	281	279	262	290	645	1025	1552	2200	2617	2450	267	610	623
16	76	90713	08/02/2014	375	384	531	279	571	253	253	544	1112	2223	2100	2409	1727	733
17	77	90713	09/02/2014	576	824	266	266	855	263	313	552	510	1556	2217	2441	1998	935
18	78	90713	10/02/2014	698	623	890	565	584	1072	306	1566	2162	2655	2413	1392	1614	738

Figure 33. View of the first rows of the dataset in the "R software". Values' units in Watts [W] per each hour

7.2.2. Cleaning data

Although this data is processed and presented in a tidy format, it needs to be cleaned. Cleaning the data means detect and delete all the possible elements that could affect to the well-being of the analysis; in this particular case:

- **First**, we need to select the sub-meters that are currently being used in the project; as the column under the label "idmeter" contains both current and old sub-meters identifiers. Hence, this step will consist on selecting only the valid sub-meters that are currently working and have consistent data.

- **Second,** from the selected sub-meters in the first step we should detect and delete those periods of data that contains "0" as a measure. The "0" values are quite common due to data communication system used is the Wi-Fi or router of the house, so when it is switched off no data is send and stored.

- **Third**, detect and select those periods that present the so called "frozen" values. The "frozen" values are the repetition of the same value for consecutive observations; this is due to a hardware problem during the data communication which froze the measurement and repeats it over a time.

It is essential to do these three steps to have a valid and verified output of the analysis and to obtain more accurate results; otherwise the as the input data is somehow "corrupted" it may affect the analysis.

After the **first step**, the sub-meters dataset was reduced from the initial 157 sub-meters to **123 meters,** which are shown in Table 4, and accounts for a total of **35.757 observations**. The discarded sub-meters are due to facts that don't add value to the analysis, such as:

- No data was stored
- All the data stored were "zeros"
- Only few observations (less than a month) were stored

Table 4. Enumeration of the sub-meters selected for the study

idmeters selected											
1	90713	22	112761	43	112820	64	112886	85	112940	106	113008
2	103433	23	112763	44	112822	65	112887	86	112942	107	113009
3	112695	24	112764	45	112824	66	112890	87	112950	108	113015
4	112696	25	112767	46	112826	67	112897	88	112951	109	113016
5	112701	26	112770	47	112833	68	112900	89	112952	110	113018
6	112707	27	112773	48	112838	69	112901	90	112953	111	113020
7	112709	28	112777	49	112839	70	112904	91	112960	112	113023
8	112718	29	112782	50	112840	71	112910	92	112966	113	113030
9	112719	30	112783	51	112849	72	112912	93	112970	114	113031
10	112723	31	112784	52	112854	73	112914	94	112977	115	113035
11	112724	32	112786	53	112858	74	112917	95	112978	116	113036
12	112729	33	112788	54	112862	75	112919	96	112979	117	113037
13	112732	34	112789	55	112863	76	112920	97	112980	118	113041
14	112737	35	112799	56	112868	77	112922	98	112981	119	113043
15	112738	36	112802	57	112873	78	112928	99	112992	120	113044
16	112739	37	112810	58	112876	79	112931	100	112993	121	113045
17	112742	38	112811	59	112877	80	112932	101	112994	122	1049953413
18	112745	39	112814	60	112878	81	112934	102	112995	123	1928113559
19	112746	40	112815	61	112879	82	112935	103	112997		
20	112749	41	112817	62	112880	83	112936	104	112999		
21	112754	42	112818	63	112882	84	112939	105	113000		

The **second step**, is to delete the all the rows that contains at least one "0" observation. The criteria to delete those observations is to remove any row that contained at least one "0", due to the fact that a zero observation can alter and distort the daily profile and thus the validity of the study. An example of deleted rows of zeros is shown in Figure 34. After this step the number of rows was reduced down to 31.169.

	row.names	idmeter	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00
1	773	112695	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	774	112695	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	775	112695	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	776	112695	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	777	112695	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	778	112695	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	779	112695	149	152	156	186	149	142	142	148	231	160	139	458	826	247	192	238	201
8	780	112695	240	166	190	125	160	140	187	178	169	159	286	237	166	115	140	165	187
9	781	112695	228	215	144	187	131	161	179	146	229	197	160	569	594	140	148	163	219
10	782	112695	236	146	189	126	200	189	172	145	371	623	641	809	229	200	183	214	206
11	783	112695	402	255	231	226	151	138	144	299	309	132	127	217	155	127	136	156	148

Figure 34. View of a dataset were "0" rows are present and need to be removed. Rows 1 to 6.

The **third step** is to delete the considered "frozen" rows. The criterion is to remove any row that contained at least three consecutive "frozen" values, so the final dataset will only contain the rows with the same value repeated in two consecutive hours.

The decision adopted was taken after analysing Table 5 where different frozen values were set to cut the dataset and showing the rows that will remain after eliminating the "frozen" rows.

Also, a randomness tolerance is considered as in some cases it is possible that in reality two consecutive hours have the same measure, as it can be seen that from the difference and the amount of deleted rows from setting 1 o 2 number of consecutive repeated values.

Table 5. Various frozen values set to cut the dataset and the implied rows removed. Highlighting the 2 or less rows as it was the chosen one

Number consecutive repeated values	Final dataset number of rows	Deleted rows
1	27.718	3.541
2 or less	**28.863**	**2.306**
3 or less	29.280	1.889
4 or less	29.550	1.619
5 or less	29.800	1.369

An example of deleted rows of "frozen" values is shown Figure 35.

	row.names	idmeter	date	00:00	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00
1	19819	112769	24/04/2014	255	218	134	88	120	131	95	109	152	161	106	133	106	100	130
2	19820	112769	25/04/2014	296	248	128	83	128	127	80	132	162	121	168	150	135	170	158
3	19821	112769	26/04/2014	297	362	163	94	127	121	89	133	118	98	238	90	128	125	87
4	19822	112769	27/04/2014	142	381	305	217	135	98	103	133	134	99	154	128	85	123	128
5	19823	112769	28/04/2014	230	241	77	134	112	108	137	86	226	290	129	128	203	199	133
6	19824	112769	29/04/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
7	19825	112769	30/04/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
8	19826	112769	01/05/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
9	19827	112769	02/05/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
10	19828	112769	03/05/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
11	19829	112769	04/05/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
12	19830	112769	05/05/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
13	19831	112769	06/05/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
14	19832	112769	07/05/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195
15	19833	112769	08/05/2014	195	195	195	195	195	195	195	195	195	195	195	195	195	195	195

Figure 35. View of a dataset were "frozen" rows are present and need to be removed. Rows 6 to 15. The frozen value is the 195 in this particular case.

So, the final number of observations will be **28.863 rows.** The piece of code used in "R" to compute this data cleaning used can be found in the ***Appendix A: Cleaning data***.

Now, the dataset is clean and ready to be analysed. The final dataset is composed by 28.863 rows, 26 columns (including "idmeter","date", "00:00-23:00") for 121 different "idmeters" (2 "idmeters" were eliminated due to the cleaning data process); and it will be the one used as input to start the analysis. Each row represents the daily consumption divided by hours per each day ("date") for each user ("idmeter" column).

7.3. Household and householder's data

Additional information of households and householders are also studied and crossed with the electrical consumption data, to achieve the goal sought. These households' features and householders' characteristics are obtained from a quick online questionnaire in the platform ("info-house") and from the extensive and face-to-face technical audits.

A short and easy to fill questionnaire allocated in the software platform. Once having access to the platform, the user is able to enter this personal information regarding its house characteristics, devices and others features, seen in Figure 36. Every user's answers are recollected and aggregated to the database server of Enerbyte; and a table is generated where all this data is stored and almost ready to be treated.

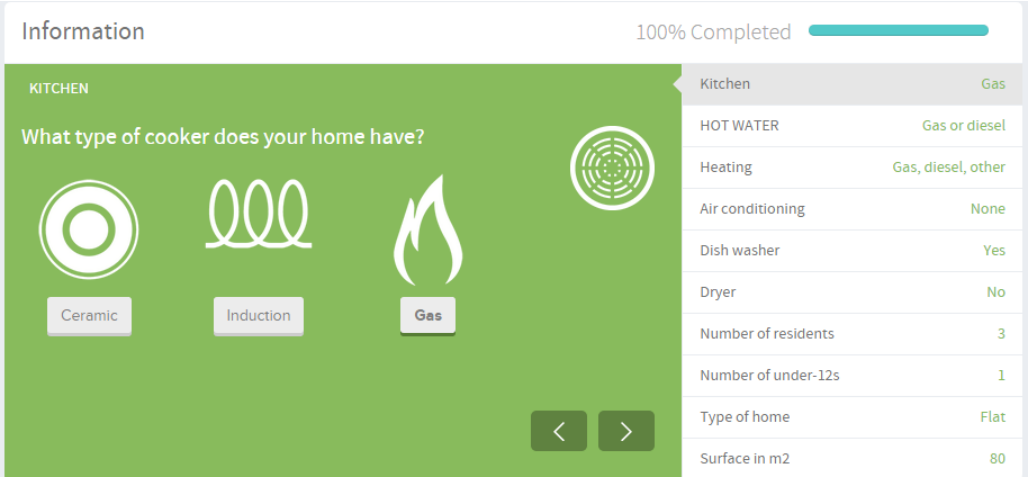


Figure 36. Image of the online questionnaire in Enerbyte's platform. Source: Enerbyte

The complete and extensive technical audits were performed by the UPC students as part of the collaboration with the project of "Rubí Brilla". This audit account for more than 250 questions to each of the consumers, most of the questions are referred to quite detailed information that are not considered in the current study, only those which have valuable interest for the analysis will be extracted and used; for instance the age ranges of the residents, house's year of construction, house typology, property or rental etc…The audit results are present in a Excel sheet format, fact that facilitate its visualization and future extraction.

7.3.1. Datasets description

Info-house

The initial dataset from the information obtained from the platform questionnaire is a table composed by and 3 columns ("idmeter", "feature description", "value") and 1.449 rows, but a single idmeter has many rows. So, the first step would be to shape the dataset into the ideal format which would be to have a row per each idmeter, and each column to be a different feature having as many columns as different features; see Figure 37.

	idmeter	variable	value
1	112932	CA_CALELE_Tipus_Electric	4
2	112932	EL_RENTV_Unitats	1
3	113036	CU_TIPUS_Electric	1
4	113036	CU_TIPUS_Induccio	0
5	113036	CU_TIPUS_Gas	0
6	113036	CU_TIPUS_Vitroceramica	0
7	113036	ACS_EQ_Electric	2
8	113036	CA_CALELE_Electric	2

	idmeter	AA_TIPUS_Electric	AA_US_TempConsigna	ACS_EQ_Electric
1	90713	1.0	22.0	1.0
2	103433	2.0	NA	1.0
3	112695	NA	NA	NA
4	112696	NA	NA	NA
5	112701	2.0	24.0	2.0
6	112709	NA	NA	NA
7	112718	NA	NA	NA
8	112719	1.0	23.0	1.0

Figure 37. Views of the dataset before (left) and after (right) the reshaping.

The “info-house” dataset contains a total of:

- 125 different “idmeters”
- 14 different variables (Table 6)

This means that not all the features are available for each of the “idmeter”, as well not all the 14 features will be used for this analysis only the ones that will be selected in section 10. At this stage the aim is to present the dataset in the best format to facilitate the later analysis.

Table 6. The household and householder’s variables obtained from the “info-house” questionnaire

Info-house variables	
1	Air conditioning (yes or no)
2	Air conditioning set temperature (ºC)
3	Electric boiler domestic hot water (yes or no)
4	Heating system electric (yes or no)
5	Heating system set temperature (ºC)
6	Electric Kitchen (yes or no)
7	Gas Kitchen (yes or no)
8	Dryer (yes or no)
9	Dishwasher (yes or no)
10	Number of adults occupants (> 12 years)
11	Number of children occupants (< 12 years)
12	Power contracted (kW)
13	Area (m2)
14	Type of home (flat or house)

Technical audit

Once organised the “info-house” dataset, the desired data from the technical audits should be incorporated and merged to the “info-house” dataset, in order to have all the data available in the same shape and in a single table.

This audit was done to 97 consumers, lower than the 125 consumers (“idmeters”) from the info-casa dataset, this means that it won’t possible to have a complete information for some of the consumers and the correlation sought it won’t be as accurate as desired; due to the dataset’s incompleteness.

The “technical audit” dataset contains a total of:

- 97 different “idmeters”
- 10 different variables (Table 7)

Table 7. The household and householder's variables obtained from the "technical audit"

Technical audit variables selection	
1	Contract type (property or rent)
2	Occupants number less than 3 years old
3	Occupants number between 3 and 9 years old
4	Occupants number between 10 and 17 years old
5	Occupants number between 18 and 24 years old
6	Occupants number between 25 and 65 years old
7	Occupants number older than 65 years old
8	House typology
9	Year of construction
10	Area (m2)

7.3.2. Merging data

The final dataset that will be object of the analysis is composed of **125 rows** each one corresponding to a different idmeter, and **20 columns** each of them hosting a variable regarding the house information. As stated before, the dataset is not complete as some variable's data are not available, this fact compromises the quality of the data to be analysed it could distort the expected results and it may not allow to perform the desired analysis.

As noticed in previous section, some of the variables are the repeated in both "info-house" and "technical audit" (i.e. Area and house typology), others are already discarded (i.e. temperature set for heating and air conditioning). The complete list of selected variables can be seen in Table 8, where from 1 to 5 is referred to general house properties, from 6 to 13 the occupant ages ranges and finally from 14 to 20 the in-home devices and features are described.

Table 8. All household and householder's variables that could be considered for the analysis

All variables	
1	Contract type (property or rent)
2	Type of home (flat or house)
3	Area (m^2)
4	Year of construction
5	Power contracted (kW)
6	Number of adults occupants (> 12 years)
7	Number of children occupants (< 12 years)
8	Number of occupants with less than 3 years old
9	Number of occupants between 3 and 9 years old
10	Number of occupants between 10 and 17 years old
11	Number of occupants between 18 and 24 years old
12	Number of occupants between 25 and 65 years old
13	Number of occupants older than 65 years old

14	Air conditioning (yes or no)
15	Electric boiler domestic hot water (yes or no)
16	Heating system electric (yes or no)
17	Electric Kitchen (yes or no)
18	Gas Kitchen (yes or no)
19	Dryer (yes or no)
20	Dishwasher (yes or no)

These variables need a further analysis, for instance the variables regarding the occupants' age ranges need to be defined according to the related literature, as now these ranges division are mixed one to another. So, the final and chosen variables that would be object of study will be selected and justified in section 10; so far the dataset is ready to be analysed.

8. Data exploration and visualization

8.1. Introduction

Before being able to do the customer segmentation, knowing the data that will be analysed is essential to get a sense of what is in the dataset, which particularities and problematic points could be found to deal with. Getting familiar with the electricity usage data and also with the "R software" commands that will be used when treating this data.

When exploring the data, three aspects should be taken into account:

- **First,** regards the dataset format as in most of the cases it needs to be reshaped into the best format to have the desired graphical output.
- **Second,** the dates' treatment is a challenge when exploring the data, as time series data needs further and different approaches. Especially, if division between weekdays and weekends, or monthly distinction is needed.
- **Third,** focus on the specificity of the electricity use data, studying which is the best path to extract the information or conclusions sought.

At this stage the data used is hourly data. For a further and deeper study more accurate data would be needed, for instance in 5 or 15 minutes resolution, together with the house's features, being able to discover the causes of the profile shape and the peaks.

8.2. Objective

The aim is to organise and visualise all the data using different representations according to the specific goal of the visualization. It is interesting to provide a tool that allows the visual representation for each individual customer and satisfies all the purposes sought. This facilitates the comparison among consumers, permits to access easily to the individual electric consumption main characteristics that, for instance, facilitate a quick audit on its consumption and detect possible irregularities or problems.

Among these purposes are:

- Differentiate the weekdays and weekends load profile, are compare them to the load profile output accounting all days not differentiating weekdays and weekends.
- Differentiate the load profile for each day of the week (Monday-Sunday), to know if among the weekdays the load profiles are similar or not (Friday may differ from Monday-Thursday)
- Representation in absolute values, percentage of consumption per hour, accumulated in a period of time
- Study if there is a consumption difference among months and seasons

- Find the characteristic load profile per each month of the year
- Be able to specify exact dates and plot the consumption at that period
- Using different kinds of visualization graphs shapes (bars, points, lines, boxplots…) and colours to facilitate the information visualization
- Facilitate the comparisons between users by using interactive graphs

8.3. Load profiles Visualizations

In this section the electrical consumption is analysed in different levels of detail. The procedure followed starts from analysing the whole gross data and gradually moving towards a more detailed data analysis by segmenting and grouping the data for a more accurate study.

Two kinds of values were used to plot the electrical consumption; the absolute units of power consumption [kWh] and the proportional hourly energy usage in percentages (%) were used. When analysing each consumer individually the important values are the absolute units in terms of kWh; but to fairly compare various consumers' profiles is necessary to create a common normalized framework where to contrast the curves' shape; this can be done with the hourly percentages of energy consumption (Opower, 2015).

As said before, depending on the analysis' goal one parameter or another would be used. So, although the possible approaches to the consumption data an endless, below different representation ways are shown that would be useful to know and better understand the consumer. This section uses various consumers for the different representations, in order to see variety of load profiles and also test the validity of the illustrations.

8.3.1. All data available

The load pattern for each customer is obtained by considering all data available, without making distinction between weekdays and weekends, for the period selected. In the case presented in Figure 38, the load profile in terms of hour proportion of energy usage of the user 112696 is shown; this user has the highest in the evening and another peak of consumption at lunch time. In Figure 39, the load profile in terms of power of the user 112854 is displayed, and it can be seen another evening peak.

These illustrations are useful to get a sense of the user's load profile in both absolute and percentage terms. The load profile represented was obtained by calculation the mean for each hour of the day (00:00 to 23:00h) input all the days available in the dataset.

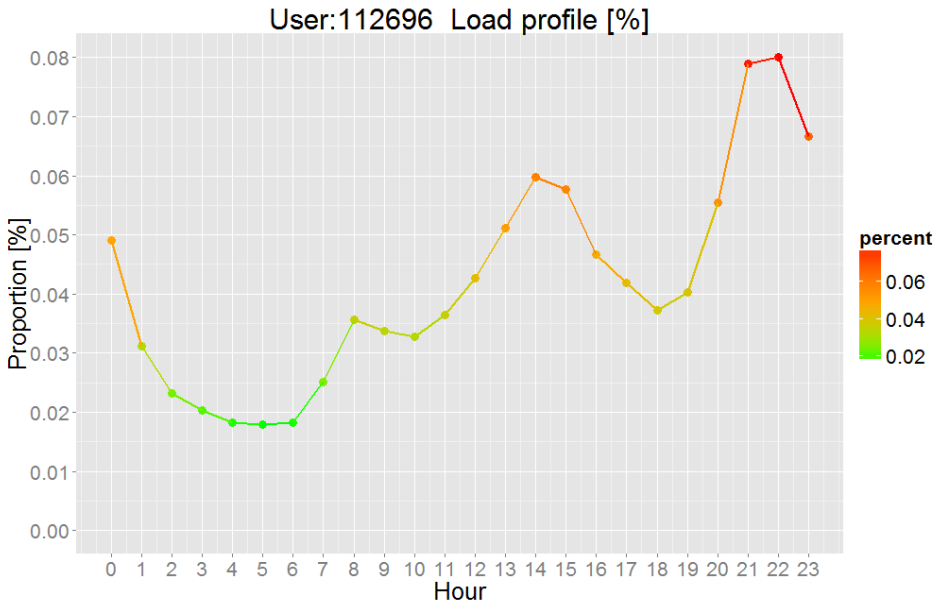


Figure 38. User 112696 load curve expressed in hourly proportion (%), represented in points and lines

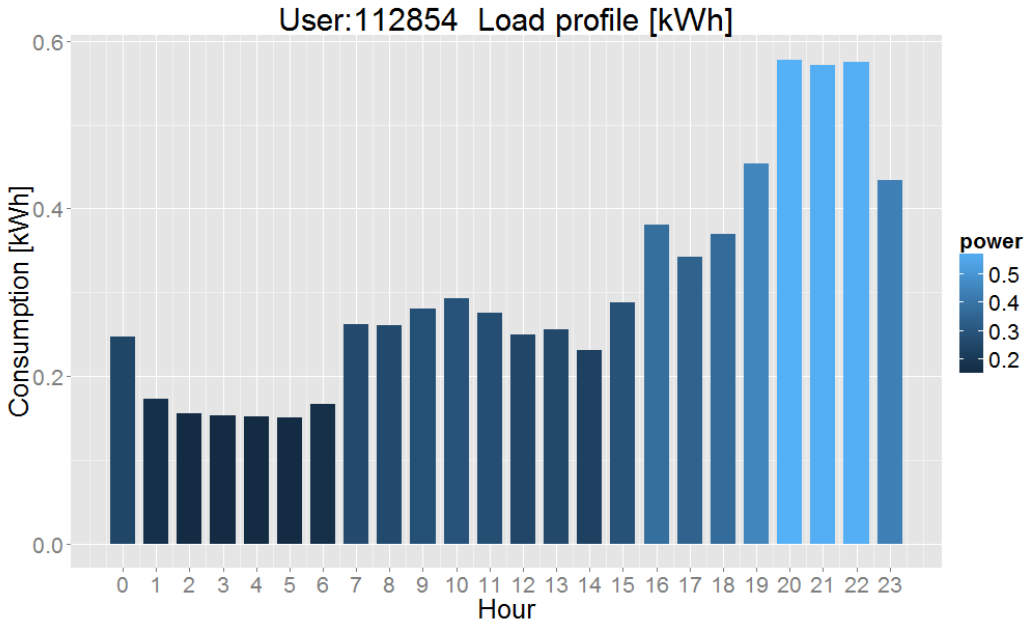


Figure 39. User 112854 load curve expressed in hourly absolute consumption units (kWh), in graph bars

Among the multiple possible representations, the circular one illustrated in Figure 40 is another way to represent the load profile giving more emphasis on the hours that present the higher and lower consumption, the example is from the user 112723.

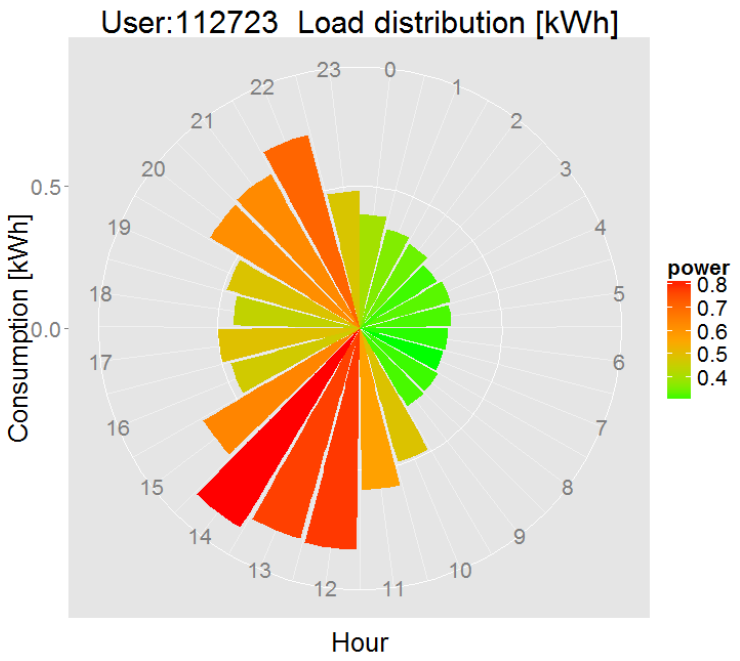


Figure 40. User 112723 load curve expressed in hourly absolute consumption units (kWh), represented in bars in a circular plot

8.3.2. Weekdays and weekends separately

The analysis can be concentrated on the separation of the weekdays and weekends, and find the load patterns for both of them. Doing so, it would be visible how similar or different are both curves and how similar or different are those with the load pattern obtained taking into account all the data together without differentiating weekdays and weekends, as presented in section 8.3.1.

The user 112992, was used in this example; as it can be appreciated in Figure 41, Figure 42, Figure 43, it exists a significant difference between the weekdays load profile and the weekend load profile.

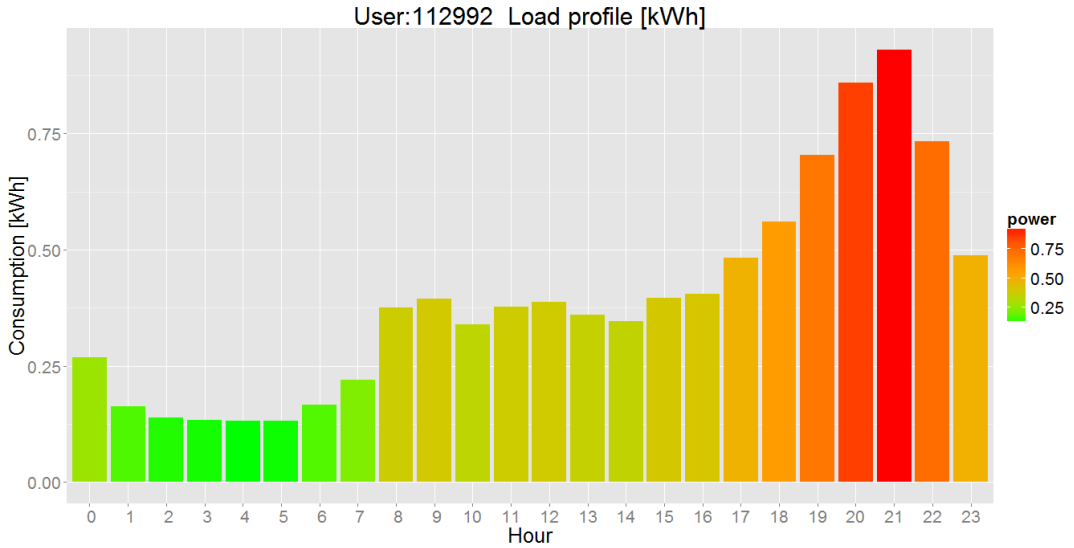


Figure 41. User 112992 load curve expressed in hourly absolute consumption units (kWh), in graph bars

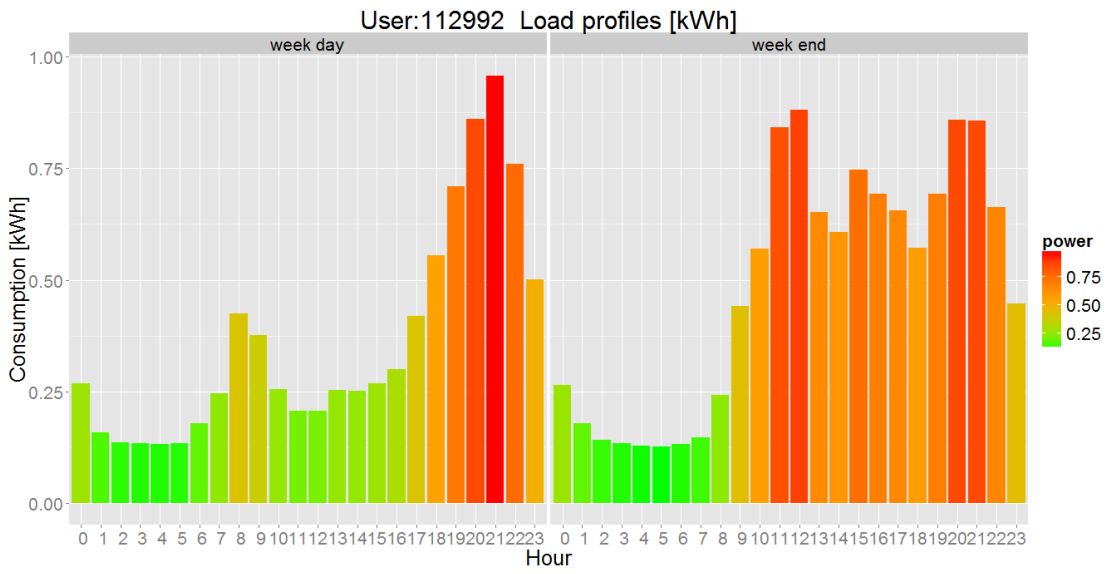


Figure 42. User 112992 Weekdays (left) and Weekends (right) load curves expressed in hourly absolute consumption units (kWh), in graph bars

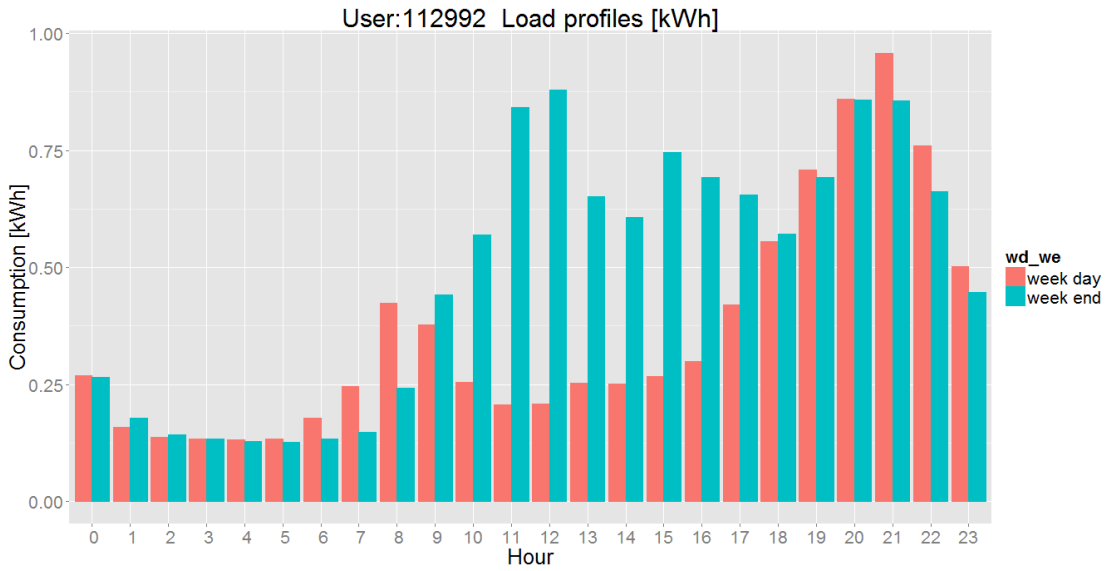


Figure 43. User 112992 Weekdays (red) and Weekends (blue) load curves expressed in hourly absolute consumption units (kWh) in graph bars in the same plot

The previous figures allow the comparison and identification of the differences between the profile of working and not working days, for a better understanding of the consumption behaviour in these two periods. With that information, the advices and measures to take can be defined more accurately. Both load profiles are obtained by calculating the mean for each hour of the day (00:00 to 23:00h) for the weekdays and the weekends available in the dataset, but also a specific period of time can be selected.

The same representation can be done with hourly percentages of energy usage (%), the output would be the same and it can be used to compare two or more users load profiles, and identify if their consumption patterns are similar and eventually give the same kind of advices.

Finally, a more detailed look at the data is done by dividing the load patterns per each day of the week, as it shows Figure 44 for the same user 112992. This disaggregation per days of the week permits to see that from Monday to Thursday the load profiles are almost identical, on Friday the evening peak is a bit longer in hours; and as it was seen before the weekend days are completely different to the rest of the days due to the fact that the households are at home during the weekends.

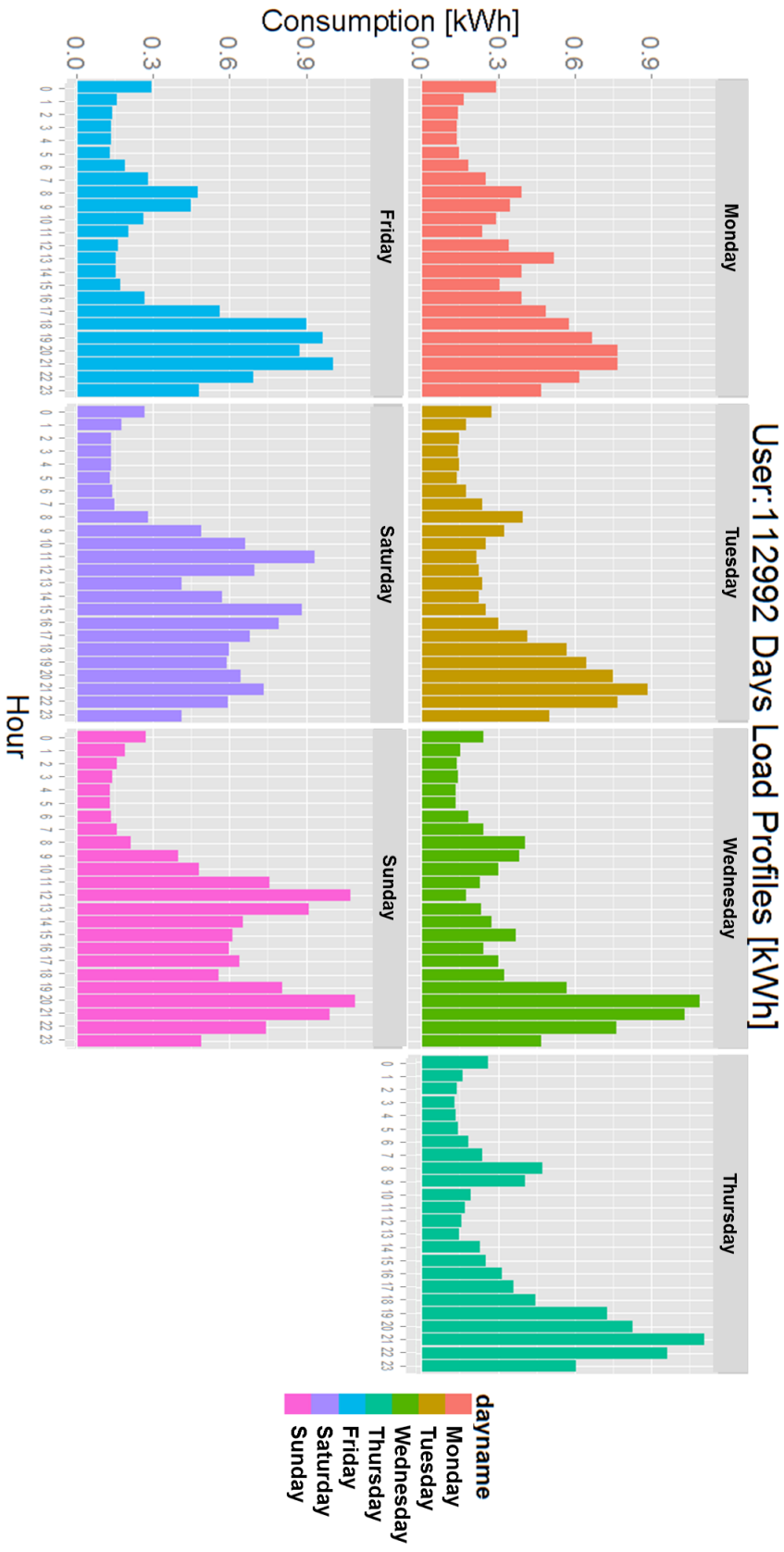


Figure 44. User 112992 Days of the week load curves expressed in hourly absolute consumption units (kWh) and represented in bars

8.3.3. Months and seasons

Looking at the load profile split by month is the aim of this section, where also the weekdays and weekends load profiles per each month are drawn; it can be used to detect some behaviour changes in specific months or differences between summer and winter months. Below, Figure 45 shows the monthly load profiles for the user 112754.

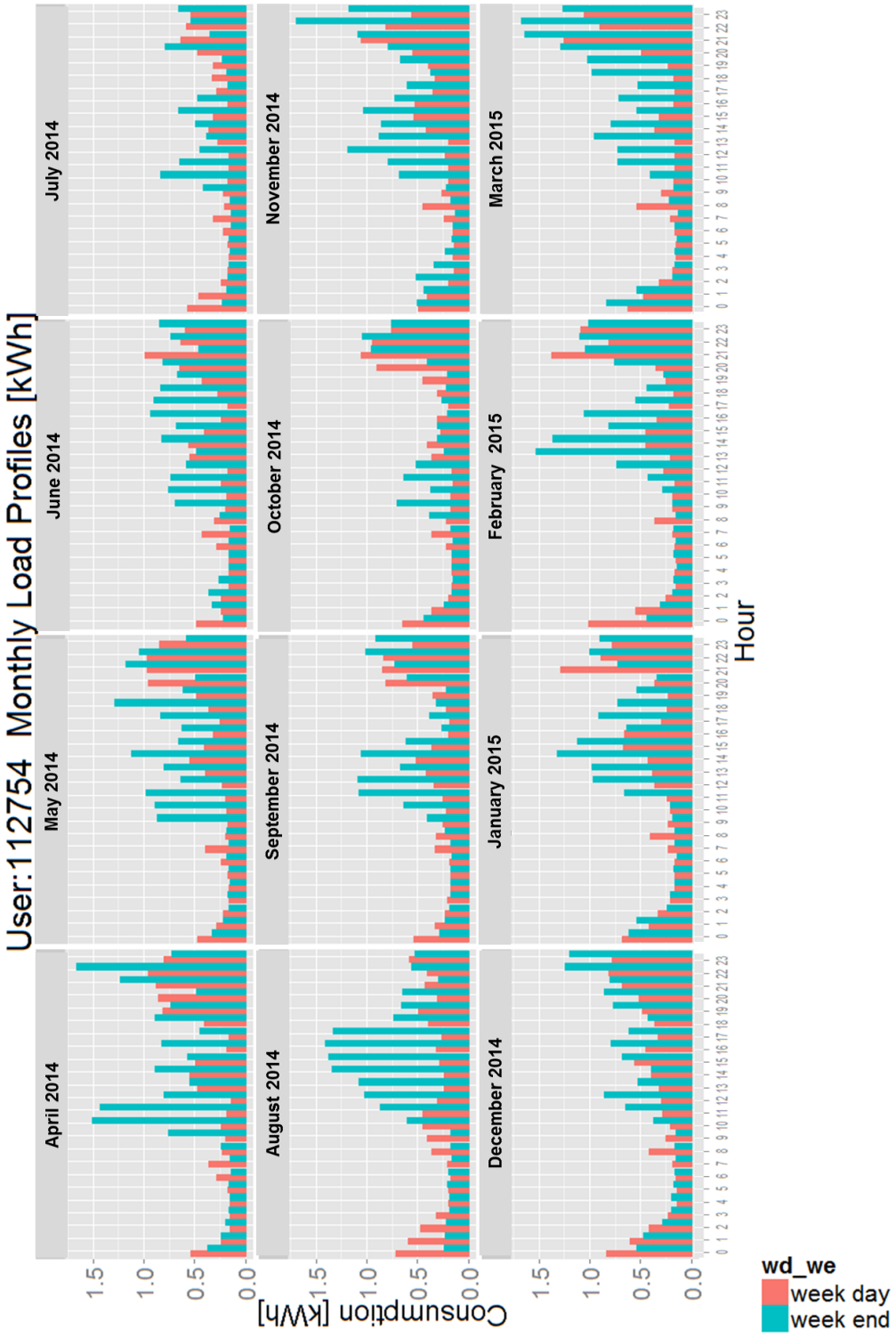


Figure 45. User 112754 monthly load profiles curves, separated by weekdays (red) and weekends (blue) expressed in hourly absolute consumption units (kWh) and represented in bars

8.3.4. Accumulated daily and monthly

The data is daily and monthly aggregated. This aggregation is facing a problem due to the fact that days with bad data have been eliminated, and so leaving empty data spaces. Empty data days may distort the results when comparing daily or monthly accumulated consumption, as probably the observations won't have the same length which eventually leads to an unfair comparison.

Figure 46 and Figure 47 show the daily and monthly aggregated data consumption for the user 90713. In this case the period of the aggregation can be selected, in the figures one year is drawn from April 2014 to May 2015; it can be observed in Figure 46 the empty space where data was eliminated due to was considered bad data.

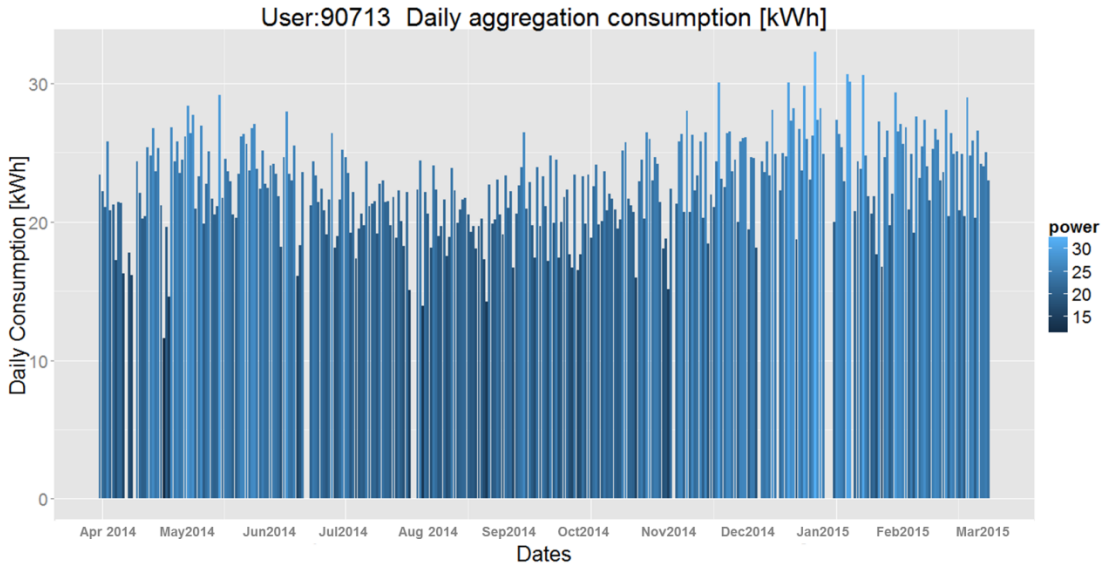


Figure 46. User 90713 daily aggregation consumption for one year period expressed absolute units (kWh)

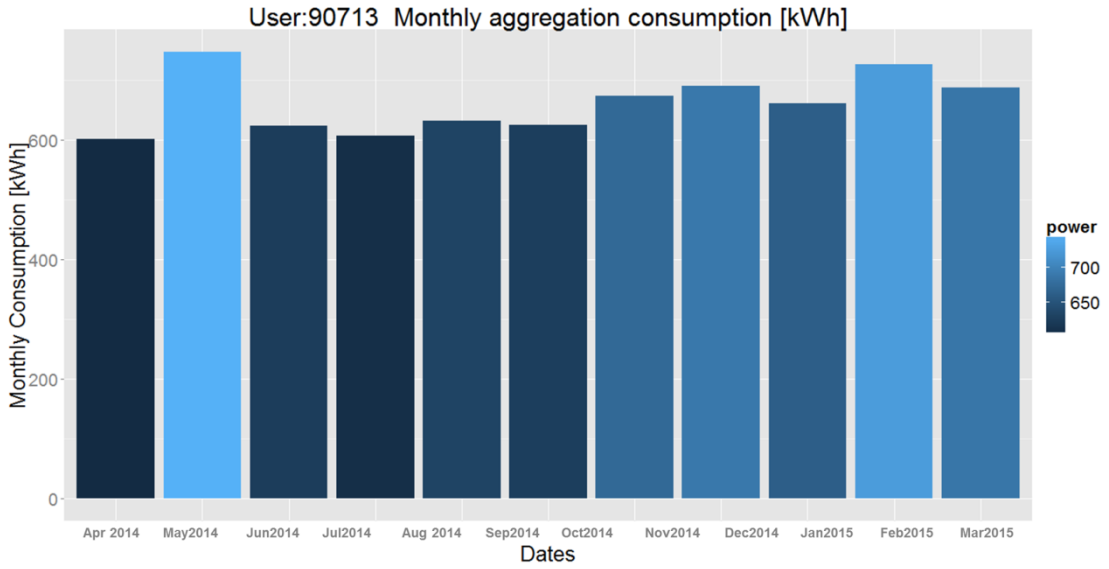


Figure 47. User 90713 monthly aggregation consumption for one year period expressed absolute units (kWh)

8.3.5. Periods and comparisons

The period representation illustrates the data in two axis: "x" axis are the dates (month and year) and "y" axis is the consumption (kWh). It can be observed in this case the consumption profile for a full week, Figure 48, and two weeks representation where empty spaces appear due to the data cleaning process eliminated the data from those days, Figure 49, where weekdays are plotted in red and the weekend is plotted in blue.

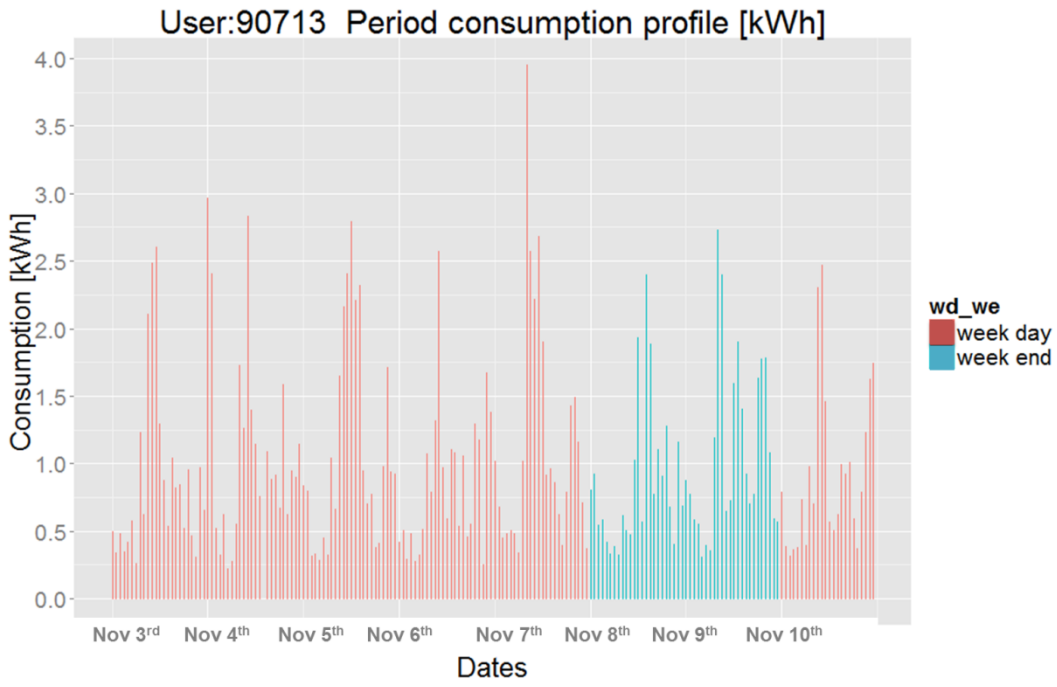


Figure 48. User 90713 consumption profile for a week (Nov 3rd – 10th) period expressed absolute units (kWh)

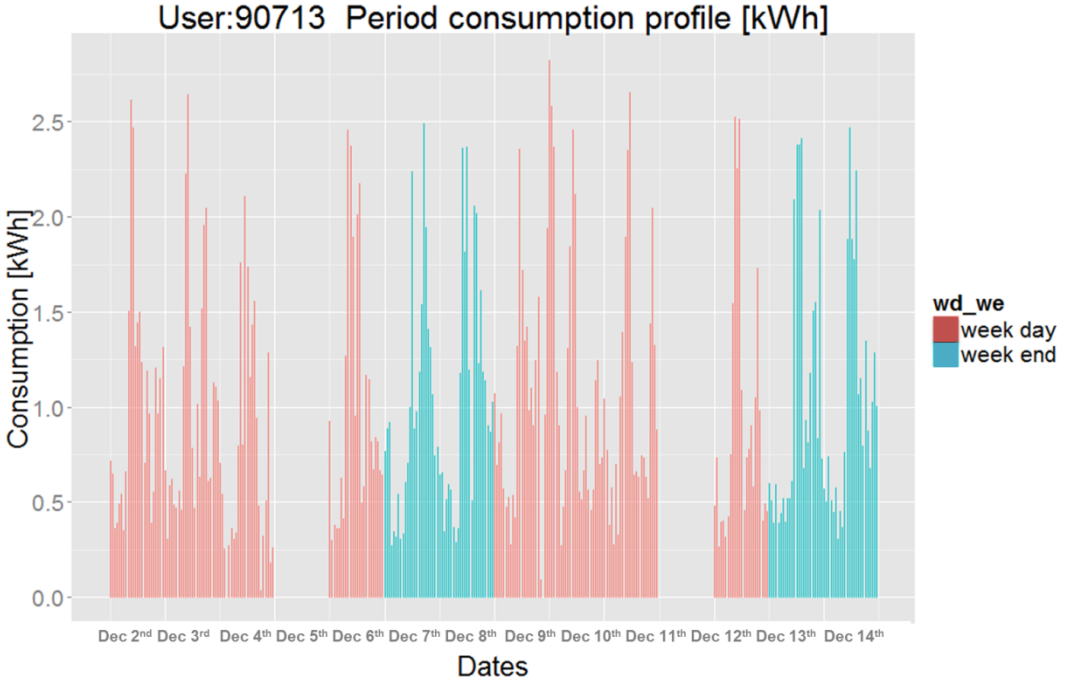


Figure 49. User 90713 consumption profile for two weeks (Dec 2nd – 14th) period expressed absolute units (kWh)

8.4. Final comments

The pieces of code used with the functions are described in the ***Appendix B: Load profiles visualization***. To finalize the exploration and visualization of this data, some remarks are done on various aspects:

- All the above representations are devoted to get a first sense of which is the load profile for each consumer without going much into detail as the input is hourly electrical consumption data. Those graphs can be used, for instance, to inform periodically monthly the users of its monthly profile divided in weekdays and weekends by attaching it to the monthly reports (Figure 50) or newsletters.

Figure 50. Example of a report where the load profile curve could be incorporated

- Another application is to provide to the Rubí municipality the starting point for deeper data analysis that is currently willing to perform. This deeper analysis will involve both, electrical consumption data in higher resolution, 5-15 min, and the study of the detailed house audits and household characteristics already available. Doing so, the consumption peaks can be associated to specific features of the house; for example knowing if the hot water heating system is electric or the house has an electric kitchen.

- The idea is to show different approaches to electrical consumption visualizations, which may be useful and easy to understand for the final consumer in order to provide a good and closer customer service. Further work on statistical parameters from the data studied can be done, for example using histograms to see the most common consumption value at each hour; this information may not have value for the final user but it does for the enterprise, utility or municipality that provide the service.

- Regarding the separation of load profiles between weekdays and weekends, it is seen that the load profiles differ for weekdays and weekends. Table 9, shoes that there are no significant differences among the consumption in weekday and in a weekend day. For the sake of the current study, this figure gives a hint for later discussions whether both weekdays and weekends' profiles should be considered or only the weekdays profiles; when clustering the consumers load patterns.

Table 9. Total consumption aggregated for all Rubí project citizens, and share for weekdays and weekends

Type of days	Total aggregated consumption [kWh]	Percentage respect to the total	Aggregated Consumption per day [kWh/day]
Weekdays (5 days)	323.085	70%	64.617
Weekends (2 days)	136.465	30%	68.232
Both weekdays and weekends (7 days)	459.550	100%	65.650

- To improve the visualization quality, in order to not to have empty spaces for those dates with missing data and also allow a fairly comparison when aggregating data; a rule should be created to associate a consumption value to those days. So that, the periods visualisation won't be empty as the most common or the mean consumption values will be plotted, and when it comes to data aggregation, for instance monthly, the total number of days with consumption will be the same, hence this normalization will permit a fairly comparison between months , as the Figure 47 shows.

9. Consumer segmentation by load profiles

9.1. Introduction

Understanding the customer and being able to provide a personalised customer experience is the key for any successful business, the electrical utilities are not an exception. Clustering electrical consumers by its load profile allows the utility to personalise its actions and services as not all customers have the same specifics needs and characteristics.

Data analytics provide the knowledge and techniques to do so, in this case clustering electricity consumption patterns will help the utility or a third party to distinguish the customers by its consumption behaviour; although each consumer is different it exists some consumption patterns were the consumer can be place in.

In that direction, a study done by Opower (Shilts & Fischer, 2014), has analysed more than 800.000 electrical consumption curves of different residential consumers in the US with the aim to discover the "Load Profile Curve Archetypes", which are referred to the most common load profile groups where a consumer can be identified with.

In this particular case, Opower has used weather-normalized hourly electrical consumption for a typical weekday, and by applying a clustering technique (vector quantization) they identified five archetypes for a later grouping of customers, as shows Figure 51.

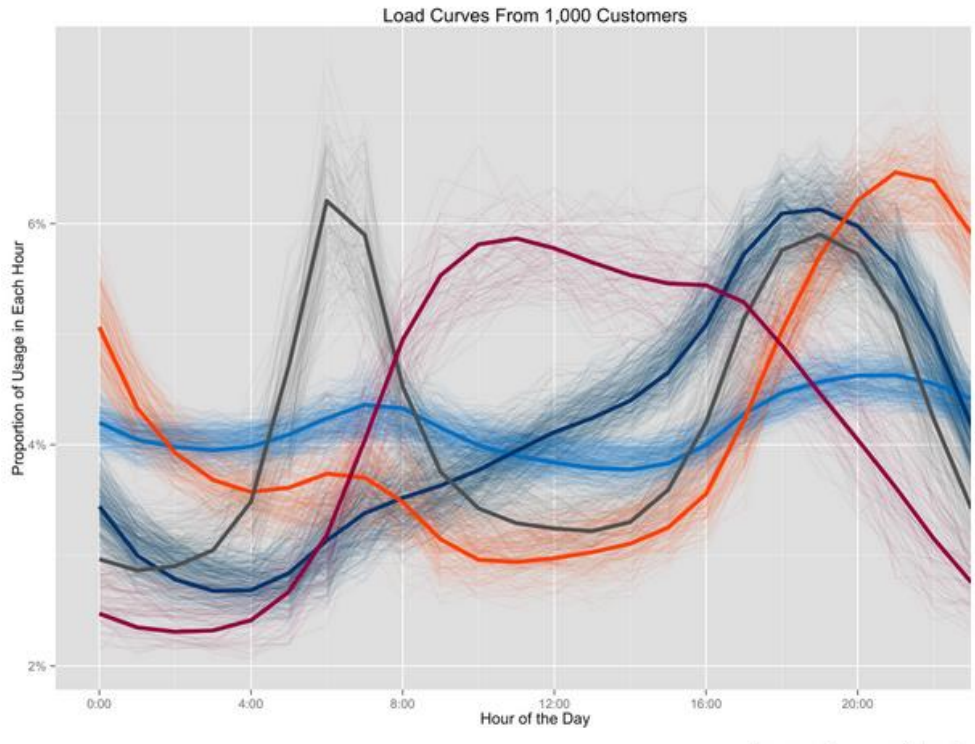


Figure 51. The five "Load curve archetypes" from the Opower study. Source: Opower (2014)

These five patterns can be described as follows:

- Flat consumers (clear blue), they have more or less the same proportional consumption each hour, no significant peaks.
- Evening peak (dark blue), consumers that have an elevate consumption proportion in the evening, a peak can be seen between 19:00-22:00h, lower proportion during night time.
- Night peak (orange), consumers that have an elevate consumption proportion in the late evening, a peak can be seen after 23:00h.
- Double peak (grey), consumers that have two significant peaks one in the morning and another in the evening, and low consumption proportion during the daytime
- Day consumers (red), they have an elevate proportion consumption during the daytime (noon and afternoon).

The benefits sought by performing such study are oriented to the utility in order to improve the customer experience by personalizing their actions depending on the customer's "Load Archetype". For instance, more personalised communications can be launched delivering the right message at the right time to the right customer in form of notifications or alerts; the utility deliver more targeted peak reduction programs or it can prepare offers that are relevant to the customer according to its characteristics like suggesting a change in the tariff.

9.2. Clustering objective

The objective of this part of the project is to find the optimal segmentation of "Rubí Brilla" consumers by their consumption load profile in order to cluster them into groups with the final aim to be able to offer personalised recommendations and advices according to the group they are placed.

The idea behind the current analysis follows the approach of the Opower's study mentioned before (Shilts & Fischer, 2014), although the difference is that the objective of the Opower was to define the common Load Profile Archetypes but not placing the customers to a specific group as they analysed "anonymous" load profiles.

Our aim is to assign every consumer into a segment and also find the most adequate number of clusters to accomplish the objective. To do so, various clustering techniques were used and various iterations were done to define the optimal number of groups. Previously, the representative load patterns of each of the 121 residential households and small business are sought; they are illustrated in Figure 52.

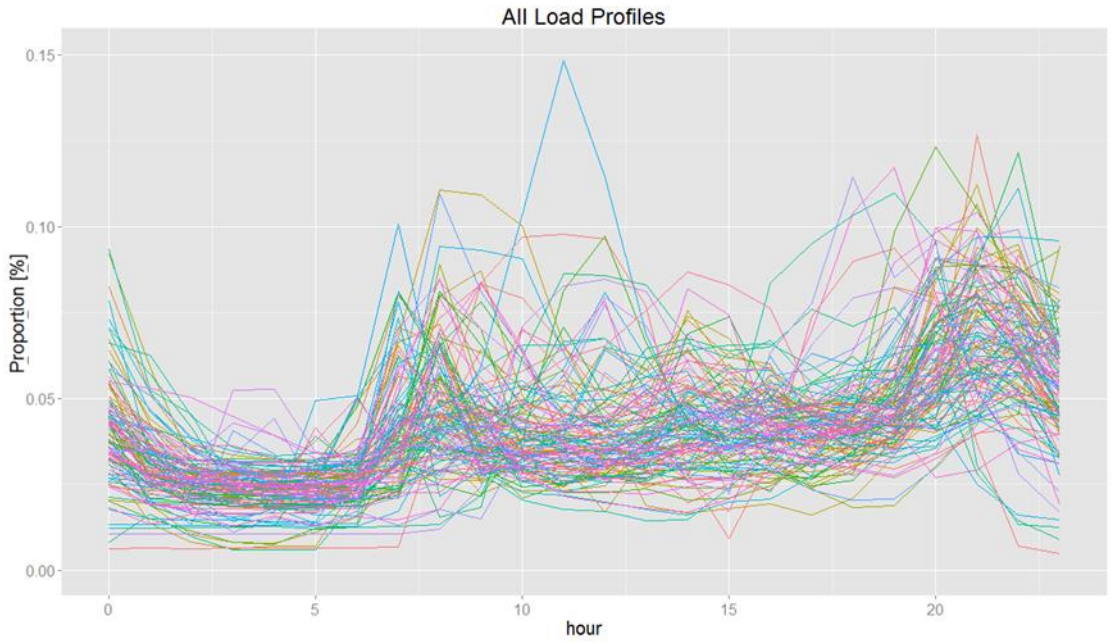


Figure 52. Visualization of all 121 loads profiles from the project of "Rubí Brilla"

9.3. Clustering overview

9.3.1. Cluster analysis theory

The aim of the cluster analysis is to group a set of observations based on one or various properties that make them similar to each other, hence the objects inside the same group will be more similar to any other object placed in other groups. The properties to group or cluster the observations are set by the analyst according to the aim of the grouping; this means that the same set of observations can be clustered in many different ways; depending on what is considered similar, it is a task of the analyst to define the appropriate similarity/distance metric.

In addition to that, there are so many clustering algorithms and techniques to group observations, each one may also lead to a different grouping result. The selection of the appropriate technique is essential to perform the desired clustering analysis; also the typology of the input data to be clustered is the main factor that will make choose one technique or another, taking into account that not every algorithm is valid for a specific problem.

Finally, again the analyst should be able to create an easy visualization of the results and also interpret and communicate the finding to the easy understanding of the audience, Figure 53 summaries the cluster analysis steps.

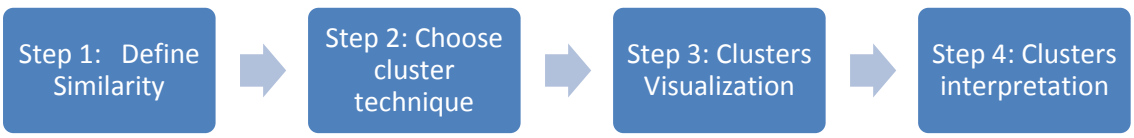


Figure 53. Steps to perform any cluster analysis

Cluster analysis is a task used in exploratory data mining and statistical analysis to discover knowledge, find patterns or retrieve information among many others. The most extended fields where clustering is used are business and marketing, biology and bioinformatics, computer science and the digital world (Coursera, 2014).

9.3.2. Clustering and classification

At this point it is adequate to differentiate between clustering and classification, see Table 10, in order not to misunderstand concepts and define clearly what represents each one:

Table 10. Basic characteristics to differentiate Clustering and Classification

CLUSTERING	CLASSIFICATION
The similarity characteristics of the data are not known in advance	Exists a training dataset that was used to previously to subset this data in groups
The dataset is divided into clusters or groups. Objects inside the same group have similar properties to each other; and differ from instances of other groups	New data is classified based on the training dataset. Algorithms find the group to which each new object belongs to due to its common characteristics
It is an Unsupervised learning, due to the absence of a training dataset able to provide prior knowledge	It is a Supervised learning task due to the existence of a training dataset, that has labelled or grouped the data previously
The objective is to label or group the observations, according to their similarity	The objective is each new data object into an existing group

It has to be noticed that in this study the data to be clustered is residential and small business electricity consumption measures, which is considered a time series data as consists of successive measurements made over a time interval. This is to say that per each hour there is a measure of the power consumption.

As the aim is to group the electrical consumption profiles, this implies that the clustering has to be done in a pattern or set of points belonging to the same object, so the procedure and the output visualization will be different to the common two-dimensional single points clustering, as shown in Figure 54. In section 9.4.1 is discussed which are the most adequate properties and parameters to be used for the load profile clustering.

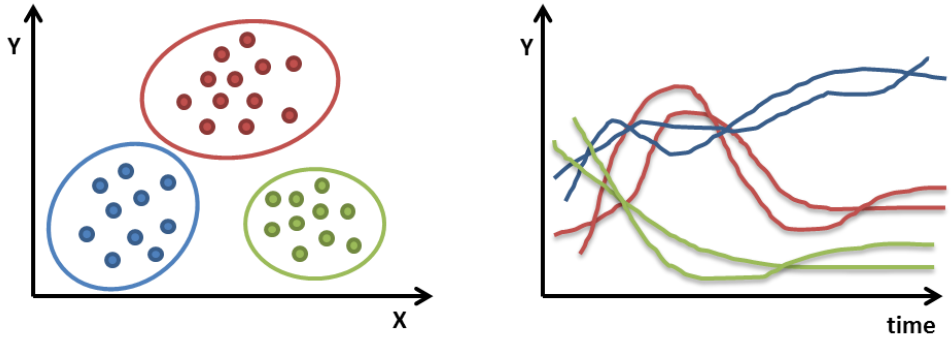


Figure 54. Schematic plots to differentiate single points' clustering (left) and pattern or time series cluster (right)

9.3.3. Times series clustering

This section is devoted to review the existing clustering techniques, understand the algorithms, distances parameters and linkages criteria that these methods involve, in order to be able to determine the technique or techniques that are used in this project.

Many different clustering techniques have been already used in order to group the electrical loads into similar patterns, in that sense an exhaustive literature review is done in (Chicco, 2012) and it is summarised in the following Table 11.

Table 11. Summary of the clustering techniques and papers, adapted from (Chicco, 2012)

Method	References
Adaptive Vector Quantization (AVQ)	(Tsekouras, et al., 2007)
Entropy-based (Renyi)	(Chicco & Sumaili Akilimali , 2010)
Follow-the-leader (FDL)	(Chicco , et al., 2003) (Chicco , et al., 2004) (Chicco , et al., 2005), (Yu , et al., 2005), (Chicco, et al., 2006)
Fuzzy and ARIMA	(Nazarko, et al., 2005) (2005)
Fuzzy k-means (FKM)	(Chicco , et al., 2005), (Chicco, et al., 2006), (Tsekouras, et al., 2007)
Hierarchical clustering (HC)	(Chicco , et al., 2005), (Chicco, et al., 2006), (Tsekouras, et al., 2007)
K-means (KM)	(Marques, et al., 2004), (Chicco , et al., 2005), (Chicco, et al., 2006), (Tsekouras, et al., 2007)
Multivariate statistics (MANOVA)	(Verdu , et al., 2006)
Probabilistic Neural Network (PNN)	(Gerbec, et al., 2004) , (Gerbec , et al., 2005)
Self-Organizing Map (SOM)	(Marques, et al., 2004), (Verdu , et al., 2006), (Chicco, et al., 2006), (Valero , et al., 2007), (Räsänen, et al., 2010)
Support vector clustering (SVC)	(Chicco & Ilie, 2009)

The table above presented many different methods that have distinct clustering approaches, it can be found: from agglomerative techniques like the Hierarchical clustering where there is no need to previously set a number of clusters; to partitioning methods such as k-means where the number of clusters is necessary to be previously set. The fuzzy methods which states that each instance has a certain degree of belonging to a group, even they can belong to more than one group; instead of assigning the instances to a specific group. The unsupervised learning-based Self-Organizing Maps; the supervised ones like the neural networks and some statistical based like the multivariate statistics. Other more recent techniques are the Entropy-based or the Support Vector Clustering.

Hence, it is observed that diverse techniques can be used to achieve the electrical load pattern grouping, each technique has its own particular approach to reach the same final goal. In that sense, for the sake of the current analysis the considered methods are:

- ✓ Hierarchical clustering
- ✓ K-means clustering
- ✓ Self-Organizing Maps (SOM)

The algorithms above represent three approaches that differ one to another, the aim is to test them three, iterate, define how similar are the resulting groups of each technique and which one offers the best output according to the objective of the study. The SOM is an artificial neural network that allows reducing high dimensional data to a 2-dimensional space relying on unsupervised learning to group the vectors into regions of a map. The Hierarchical is an agglomerative bottom-up approach that allows to graphically represent its results in a dendrogram; fact that permits to visualize the appropriate distance and linkage criteria when clustering. Finally, the k-means was chosen, it is the most common clustering method of vector quantization and it is needed to previously define the number of clusters. They are deeply explained in sections 9.5, 9.6 and 9.7.

9.3.4. Distances and linkages

In order to decide which observations belong to a cluster or another, a measure of similarity or dissimilarity between the observations needs to be established. This is achieved by the use an appropriate distance and linkage criteria. Distance or metric is the distance between a pair of observations and linkage determines the distance between sets of observations. The most common types of distances and linkages, and also the ones that "R software" supports are described below.

Distances

To explain the distance vectors "x" and "y" are used as example, in a dataset the x and y will be the rows, each row representing a different instance; and the from 1 to n represent the features of the vector.

$$x = (x_1, x_2, \dots, x_n) \qquad y = (y_1, y_2, \dots, y_n)$$

- **Euclidean distance:** it is the distance of the straight line that between two points, calculated as follows.

$$d(x,y) = d(y,x) = \sqrt{\sum_{1}^{n} (y_i - x_i)^2}$$

(Eq. 1)

- **Squared Euclidean distance:** is the squared of the Euclidean distance, it is normally used to empathise the difference between objects by placing progressively greater weight on farther objects.

$$d^2(x,y) = d^2(y,x) = \sum_{1}^{n} (y_i - x_i)^2$$

(Eq. 2)

- **Manhattan distance:** is the distance calculated as the sum of the absolute differences of Cartesian coordinates, also known as rectilinear distance or taxicab distance.

$$d(x,y) = \|x - y\|_1 = \sum_{1}^{n} |x_i - y_i|$$

(Eq. 3)

- **Canberra distance:** it is the weighted version of the Manhattan distance. Used for comparing ranked lists or for intrusion detection.

$$d(x,y) = \sum_{1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

(Eq. 4)

- **Maximum distance:** also known as Chebyshev distance, where the distance between two vectors is the greatest of their differences along any coordinate dimension

$$\|x - y\|_\infty = \max_i |x_i - y_i|$$

(Eq. 5)

Other types of distances are "Binary distance",”Minkovski distance” among many other, but are not considered in this study. In Figure 55 the difference between Euclidean and Manhattan distances can be visualised.

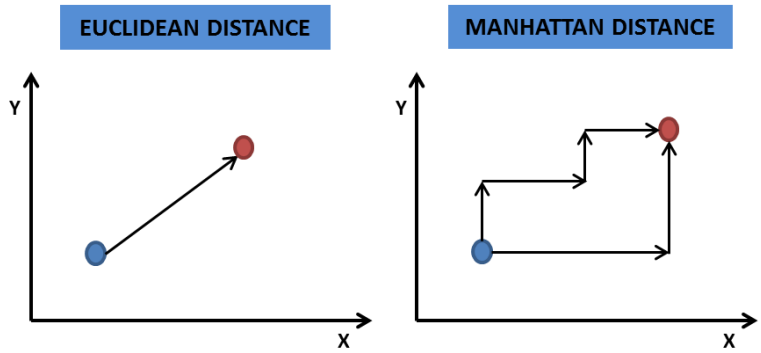


Figure 55. Illustration of the Euclidean and Manhattan distances

Linkages

To explain the linkages cluster “A”, “B”, “C”, “D” and “K” is another random cluster, are used as example.

- **Single-linkage or nearest neighbour:** the distance between two clusters corresponds to the minimum distance between any two members in two clusters. Aim to find similar clusters. Adopts a “friends of friends” cluster strategy.

$$\boldsymbol{D(A,B) = \min_{a \in A,\ b \in B} d(a,b)}$$

(Eq. 6)

- **Complete linkage or furthest neighbour:** is the opposite approach of single linkage, assumes that the distance between two clusters corresponds to the maximum distance between any two members in two clusters. Aims to find similar clusters.

$$\boldsymbol{D(A,B) = \max_{a \in A,\ b \in B} d(a,b)}$$

(Eq. 7)

- **Average linkage or UPGMA (Unweighted Pair group Method with Arithmetic mean):** the distance between two clusters is defined as the average distance between all pairs of the two clusters' members. Unweighted means that all pairwise distances contribute equally, the mean distance between all pair of elements each belonging to one cluster.

$$D(A,B) = \frac{d(a,k)|a| + d(b,k)|b|}{|a| + |b|}$$

(Eq. 8)

- **Centroid linkage or UPGMC (Unweighted Pair group Method using Centroids)**: the geometric centre (centroid) of each cluster is computed first, and the distance between the two clusters equals the distance between the two centroids. Unweighted means that all pairwise distances contribute equally.

$$D(A,B) = \|\tilde{x}_A - \tilde{x}_B\| \quad where \quad \tilde{x}_A = \frac{1}{n_A}\sum_{1}^{n_A} a_A$$

(Eq. 9)

- **WPGMA (weighted Pair group Method with Averaging):** the difference to the UPGMA is that the distances do not contribute equally. When two clusters A and B are joined together, the new distance to a cluster C is the mean between distances A-C and B-C.

$$D(A,B) = \frac{d(a,k) + d(b,k)}{2}$$

(Eq. 10)

- **WPGMC (weighted Pair group Method using Centroids):** the difference to the UPGMC is that the distances do not contribute equally, as the centroids are weighted. When two clusters are joined together, the new centroid is the midpoint between the joined centroids.

$$D(A,B) = \|\tilde{x}_A - \tilde{x}_B\| \quad where \quad \tilde{x}_A = \frac{1}{2}\ (\tilde{x}_C + \tilde{x}_D)$$

(Eq. 11)

- **Ward linkage or Ward's minimum variance criterion:** where the criterion for choosing the pair of cluster to merge at each step is based on the optimal value of an objective function that minimizes the total with-in cluster variance. At each step, the pair of clusters with the minimum between-cluster distances are merged, leading to a minimum increase in total within-cluster variance at each step, this increment is a weighted squared distance between cluster centres.

$$D(A,B) = d^2(A,B) = \frac{|A| * |B|}{|A| + |B|}\|A - B\|^2$$

(Eq. 12)

A schematic representation of the some of these linkages approaches are illustrated in Figure 56.

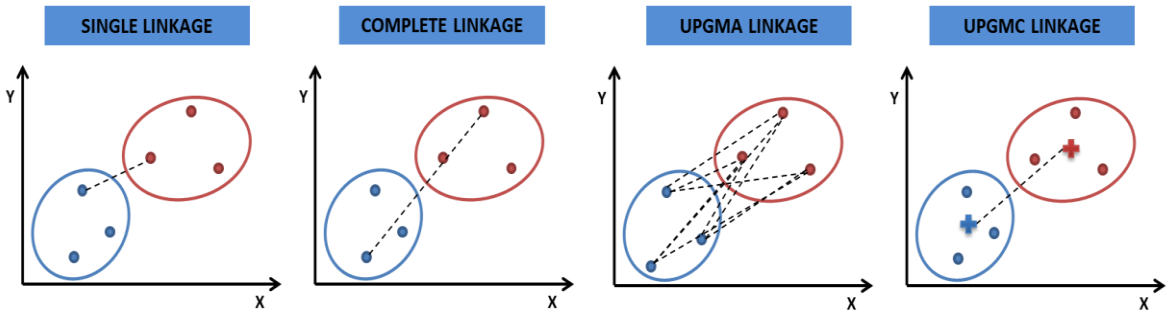


Figure 56. Schematic view of the single, complete, UPGMA and UPGMC linkages

9.4. Pre-clustering phase: Input data processing

9.4.1. Input data and features discussion

The input data used to carry out the clustering analysis vary depending on the specific aim of the load profile's segmentation, and it is up to the analyst to decide which are the most convenient input data units; the use of absolute values in kWh, the use percentages (%) or dimensionless data. Also, whether consider all the dataset, including both weekdays and weekends' consumption or just consider the weekdays' consumption. And finally, decide if additional features should be taken into account, or it is enough to consider the 24 hourly consumption measures.

In literature the input data is used in many different shapes and including various additional features, in the below list there are a few examples extracted from the vast literature and studies performed regarding this topic;

- In (Beckel, et al., 2012) the absolute mean values of electrical consumption, in kWh, were used determining the mean daily consumption, the maximum and minimum consumption, the mean power per periods (morning, night, evening, noon). Then it also uses other figures in order to make the comparison more accurate, calculating the ratios between the minimum, the mean and the maximum, also the ratios between the periods; among other statistical properties such as the variance.

- In (Ardakanian, et al., 2014) where, giving personalised recommendations is also the objective of studying the consumption profiles are using the absolute mean values, in kWh, separating weekdays and weekend days, and also aiming to find a profile per each day of the week separately.

- In (Chicco, 2012), it compares the load patterns in terms of their curve shape by using a representative load pattern obtained by dividing a daily load pattern by the reference power which is the peak value of a daily pattern. Hence, relative values in form of dimensionless ratios are used here to perform the cluster analysis.

- Among all the research done, the one that adapts better to the aim of this study are the **percentages of energy usage per hour**, also used by "Opower" and explained in (Shilts & Fischer, 2014)when performing the customer segmentation for electric utilities. The load profiles are found by the proportion of energy usage each hour respect to the sum of the total daily energy use, in percentages (%).

9.4.2. Input data and features calculation

After cleaning the data in section 7.2.2, the dataset is composed by 28.863 rows corresponding to the observation of the 121 different customers differentiated by the number of the "idmeter"; and 26 columns that represents the variables (including "idmeter","date", "00:00-23:00").

The procedure to obtain the **Percentage of energy usage per hour** is explained in the ***Appendix B: Load profiles visualization.*** The process is divided in three steps:

The first step, is to select the 24 columns that contains the consumption data from "00:00-23:00", and create a new column that contains the sum of the values of each row. So, the data frame is formed by 1 column and 28.863 rows.

The second step, is to get the hourly proportion of the energy usage per each row, this is obtained by dividing the input dataset by the each row summation dataset obtained from the first step. Each row is identified by the correspondent "idmeter", so the dataset is formed by 25 columns ("idmeter" and "00:00 to 23:00 measures") and 28.863 rows.

Finally, the **third step** consists on obtaining the mean hourly percentage values per each "idmeter", so per each customer. It is obtained by taking separately all the measures of every "idmeter" and then the mean of each hour (column) is calculated, and finally collected into the same data frame.

Hence, eventually the dataset is composed by 121 rows and 25 columns. As said before, each row correspond to a different customer, the first column stated the identification number of the customer ("idmeter") and the following columns represents the percentages of energy usage per each hour from 00:00 to 23:00. Some additional features have been calculated in order to be considered when clustering the load patterns:

- Percentage figures: finding the minimum, the maximum and the median hourly percentages
- Percentages by periods: divided by morning (6:00-11:00h), noon(12:00-17:00h), evening(18:00-23:00h) and the night period (00:00-05:00h)
- Ratios between the percentage figures, and also ratios between the periods calculated
- Some statistical parameters such as the variance and the standard deviation.

9.4.3. Input data and features selection

In order to find the most adequate input data and features that eventually will create the desired clusters according to the aim of the study. The following cases of input data and features were examined:

- Considering only the power consumption during the weekdays to find similarities among the observations. Not taking into account the weekends; as the aim is to find the load profiles and place them into the representatives load patterns, the weekdays are the ones that provide the load profile more accurately as during the weekend the consumption habits and activities are different.
    - Firstly, only a vector of 24 variables (00:00-23:00h) is used to only 24 h; and
    - Secondly, in addition to the 24 hour's vector the figures of the median, minimum, maximum, and period figures (night, morning, noon, evening) are also contemplated.

- All data is considered, both weekdays and weekends.
    - Firstly, only a vector of 24 variables (00:00-23:00h) is used to only 24 h; and
    - Secondly, in addition to the 24 hour's vector the figures of the median, minimum, maximum, and period figures (night, morning, noon, evening) are also contemplated.

The decision taken is to consider **only the weekdays** (Saturdays and Sundays are excluded) as they share corresponds to the 70% of the total consumption in a typical week, and also it was weekends' load curve are completely different to the typical weekdays, fact that could distort the result sought.

Discarding the weekends, has also a backward which is that around 30% of the data won't be used; and in a relative small dataset as the one of this project reduces the total number of rows from the 28.863 to the 20.576 rows.

Regarding the features examined, the **24 hour's vector (00:00-23:00h)** is considered for this first approach to the calculation of the similarities between load profiles, as it is the one that best adapts to the cluster methods that will be used. In later steps, the figures of the median, minimum, maximum, and periods (night, morning, noon, evening) could be incorporated.

Remark that the input data used is not normalised by the temperature, so it does not differentiate between seasons. The possible outlier values were not eliminated, although they may influence to the final result. A further point of study would be to set some limits into the values in order to eliminate these outliers.

9.5. Hierarchical clustering

9.5.1. Hierarchical Theory

Hierarchical clustering method builds a hierarchy of clusters by the help of a tree diagram named dendrogram. Two strategies can be followed when using hierarchical clustering; the agglomerative (bottom-up) approach or the divisive (top-down) approach.

In the current study, the agglomerative (bottom-up) approach is followed; in which each observation starts being a cluster, then the method aims to find the closest observations and put them together according to its similarity forming clusters; the same procedure is followed successively until all the observations are part of the same cluster.

At the end of the process the dendrogram illustrates how close the observations are to each other; however to reach this, a metric distance and a linkage approach from the ones described in section 9.3.4 need to be defined.

Finally, the last step is to cut this dendrogram to form the clusters which are the aim of the study. A schematic dendrogram can be seen in Figure 57. Choosing the numbers of clusters is also not a simple task as the analyst has to decide how many clusters can better solve the considered situation.

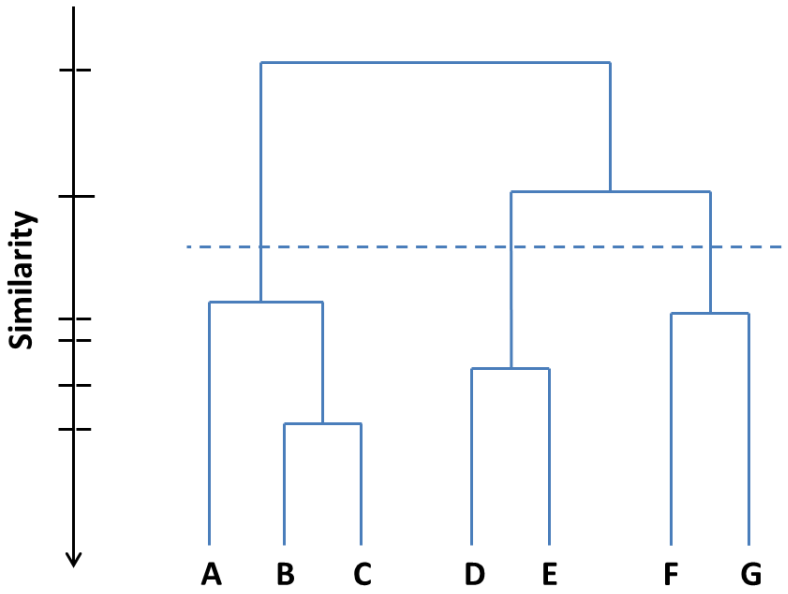


Figure 57. Dendrogram example with the observations (A to G) at the bottom, merged by similarity. The dashed horizontal line cuts the tree in 3 clusters.

9.5.2. Cluster formation

As stated in section 9.4.3, for this approach the decision taken was to cluster the customers by using the weekdays' data, and the vector with the 24 variables corresponding to the hourly percentage of energy usage from 00:00 to 23:00h was considered.

Dendrogram comparisons of distances and linkages

To perform a Hierarchical clustering is needed to define the type of metric (distance) between observations and the linkage between clusters. Various iterations combining the possible distances and linkages were done, in order to determine which combination better adapts to the goal. Some of the combinations considered dendrograms can be found on the ***Appendix C: Dendrograms comparison***.

One particularity of the hierarchical clustering is that it allows a visual representation in dendrograms. The dendrograms presented below are the three that better fulfilled the objectives sought, Figure 58 presents a Euclidean distance with Ward linkage, Figure 59 presents Manhattan distance with Ward linkage and Figure 60 presents Canberra distance with Ward linkage.

In the vertical axis, on the left side, the values state the similarity distances calculated, and in the horizontal axis all the different “idmeters” numbers are placed. A horizontal line was drawn in each of the dendrograms, that could be, at a first glance, an adequate number of clusters division.

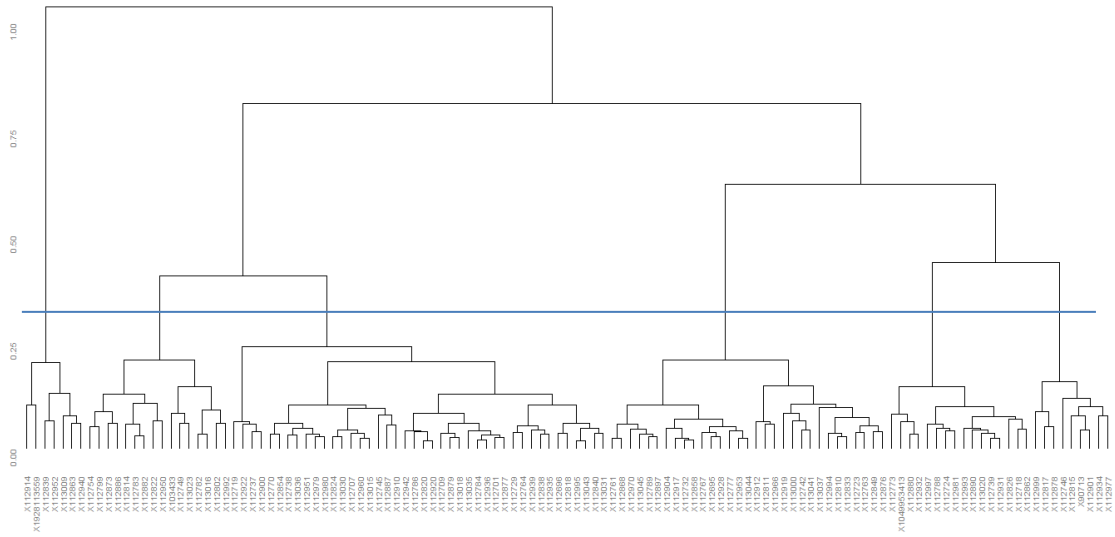


Figure 58. Dendrogram obtained with Euclidean distance and Ward linkage. Cutting line to obtain 6 clusters

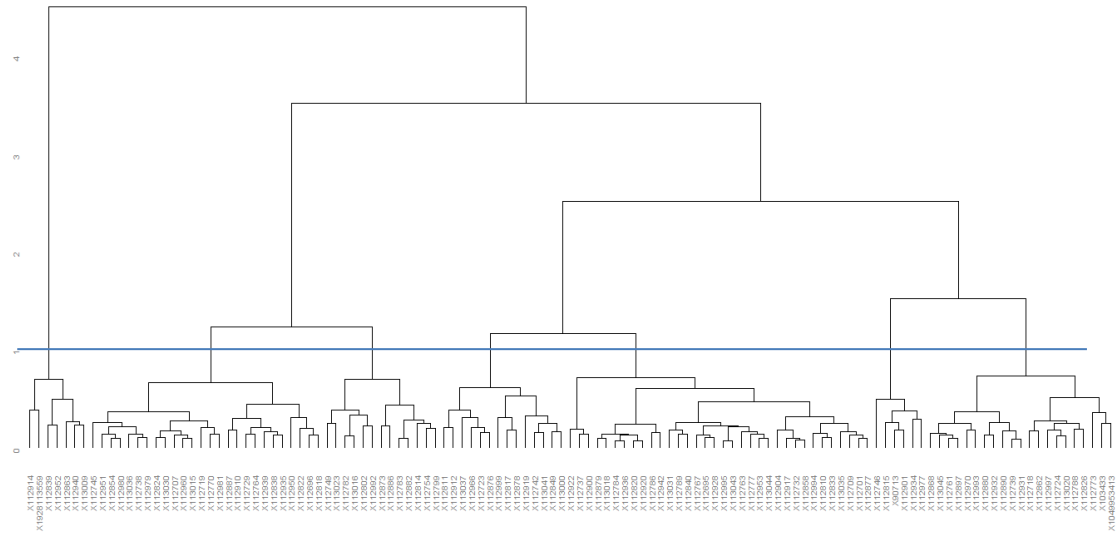


Figure 59. Dendrogram obtained with Manhattan distance and Ward linkage. Cutting line to obtain 7 clusters

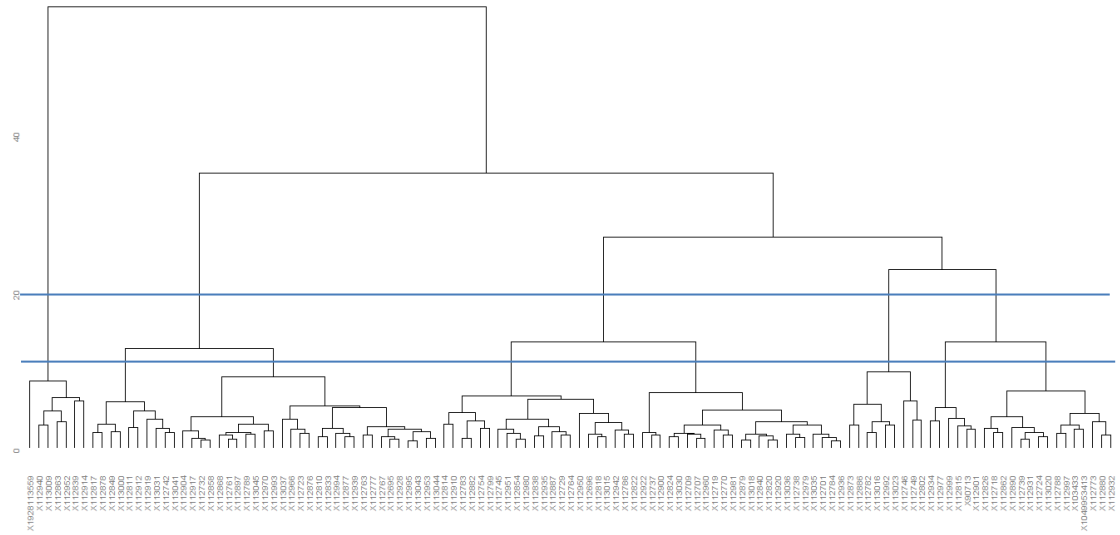


Figure 60. Dendrogram obtained with Canberra distance and Ward linkage. Cutting line to obtain 5 or 8 clusters

After analysing the previous figures, it was decided to cut the dendrograms to obtain a number of groups according to the aim of the project, so small and large number of groups were discarded and the range of clusters was set between 5 and 8; expecting to obtain differentiate load profiles curves on each of them.

- Cluster the Euclidean distance with Ward linkage dendrogram (Figure 58) in 6 groups and in 7 groups.
- Cluster the Manhattan distance with Ward linkage dendrogram (Figure 59) in 7 groups.
- Cluster the Canberra distance with Ward linkage dendrogram (Figure 60) in 5 and 8 groups.

Discussion

As there are many options when clustering, the work of the analyst is to decide which one is better for the interests sought. Below there is the explanation and reasons for each one of them, for a better understanding and further details, the ***Appendix D: Hierarchical clustering formation*** should be looked when reading this part:

- **Euclidean distance & Ward linkage**

Clustering in 6 groups permits a well-definition and differentiation among clusters, when plotting all the load curves per cluster these are compacted and some but few outliers can be seen. Later, it was considered the option to obtain 7 clusters to define better the in-groups load profiles, but still clusters 3 and 4 presented a large number of members and the new cluster created was only 4 members.

- **Manhattan distance & Ward linkage**

The data was clustered in 7 groups, it provided well-defined clusters when plotting all the load curves very few outliers can be seen in each cluster. So, the difference among each cluster's representative load profile is clear. However, a remark should be done regarding cluster 5, it seems that it grouped a conglomerate of different and dispersed load profiles that could not fit in any other cluster, this issue will be further analysed.

- **Canberra distance & Ward linkage**

The problem of clustering in only 5 groups implied that inside the same cluster many different load profiles were assigned, thus no definition of the cluster can be clearly done. This leads to stablish a higher number of clusters, 8, when plotting the load profile per each cluster they are less dispersed. Referring to cluster 6, it grouped a conglomerate of dispersed load profiles, when plotting its representative load profile it doesn't correspond to the cluster 6 members' load profiles, finally say that cluster 4 and 5 are almost identical and they could be treated as one.

In all of the cases, similar representative load profile appear and can be differentiated; the most common were: (1) high morning peak, (2) two peaks during the morning and the afternoon as it was a business or store, (3) high evening peak with a moderate morning one, (4) high morning and high evening peaks, (5) night peak around 00:00 and (6) flat curve with no peaks.

9.5.3. Chosen solution

The hierarchical solution adopted was to cluster the data in 7 groups, by using the **Manhattan distance & Ward linkage**, the numbers of members in each cluster can be seen in Table 12.

The decision was taken looking at both the load profile visualization and statistical parameters; it should be said that the cluster solution using Euclidean distance provided similar results to the ones finally selected using the Manhattan distance. The number of members assigned to each cluster, identified in Table 13, range from the minim 6 (cluster 1) to 34 (cluster3); so a distributed users allocation is also achieved; avoiding clusters with 1 to 5 members which could mislead the analysis.

Table 12. Number of members per each cluster Manhattan Distance and Ward Linkage

Manhattan Distance and Ward Linkage							
Cluster number	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
Number of consumers	6	21	34	26	14	13	7

Table 13. All idmeters in each cluster Manhattan Distance and Ward Linkage

Cluster 1	Cluster 2		Cluster 3		Cluster 4		Cluster 5	Cluster 6	Cluster 7
X90713	X103433	X112890	X112695	X112879	X112696	X112887	X112723	X112749	X112839
X112746	X112718	X112897	X112701	X112900	X112707	X112910	X112742	X112754	X112863
X112815	X112724	X112931	X112709	X112904	X112719	X112935	X112811	X112782	X112914
X112901	X112739	X112932	X112732	X112917	X112729	X112939	X112817	X112783	X112940
X112934	X112761	X112970	X112737	X112920	X112738	X112950	X112849	X112799	X112952
X112977	X112773	X112993	X112763	X112922	X112745	X112951	X112876	X112802	X113009
	X112788	X112997	X112767	X112928	X112764	X112960	X112878	X112814	X1928113559
	X112826	X113020	X112777	X112936	X112770	X112979	X112912	X112873	
	X112862	X113045	X112784	X112942	X112818	X112980	X112919	X112882	
	X112868	X1049953413	X112786	X112953	X112822	X112981	X112966	X112886	
	X112880		X112789	X112994	X112824	X113015	X112999	X112992	
			X112810	X112995	X112838	X113030	X113000	X113016	
			X112820	X113018	X112854	X113036	X113037	X113023	
			X112833	X113031			X113041		
			X112840	X113035					
			X112858	X113043					
			X112877	X113044					

Observing Figure 61, Figure 62, Figure 63, each of the 7 clusters obtained presents a clearly defined representative load profile, easy to identify and differentiated from the other clusters` profiles. Also, statistical parameters were computed in order to have a numerical-based decision. The distance between each of the cluster member's load profile and each cluster mean (representative load profile in Figure 63) is calculated and aggregated in order to find the total distance to the mean. The results will show, numerically, which cluster has more dispersed members and will identify the outliers, namely residuals members of the cluster.

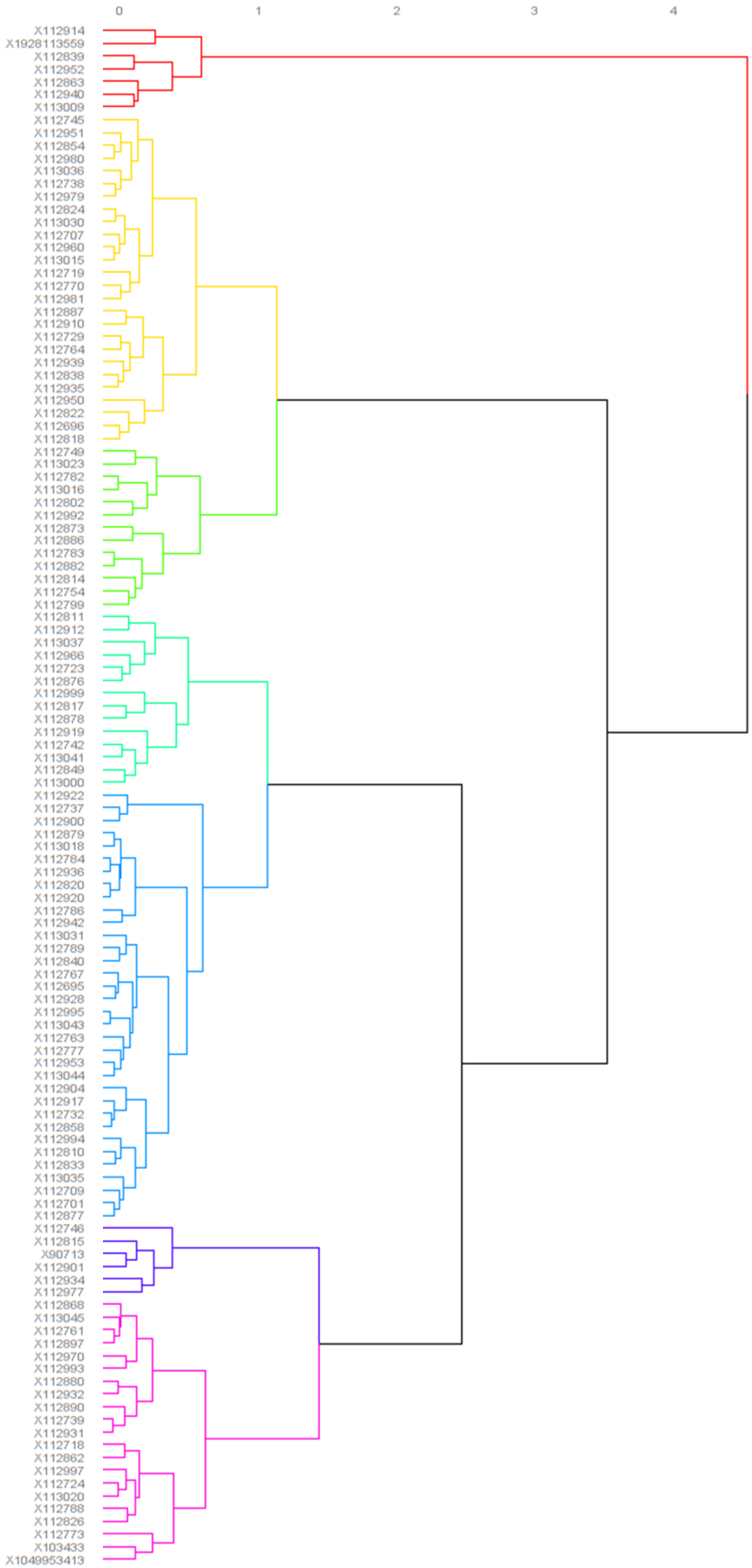


Figure 61. The 7 clusters coloured in the Manhattan Distance and Ward Linkage's dendrogram to easy visualization, on the left (y axis) the similarity and on the x-axis the "idmeters" numbers.

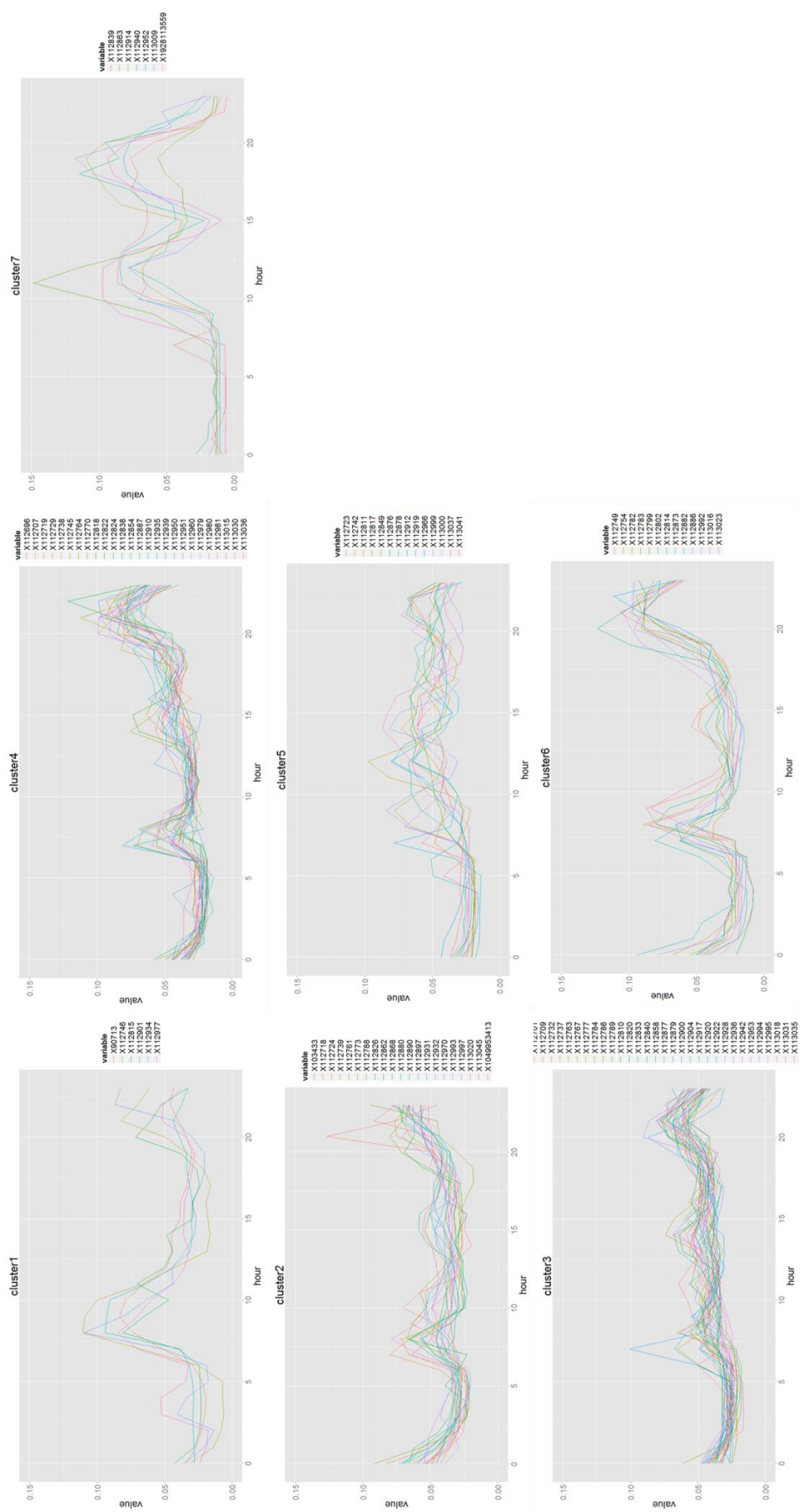


Figure 62. Plotting all the idmeters of each of the 7 cluster, to see the variability and check the clustering results from the Manhattan Distance and Ward Linkage clustering

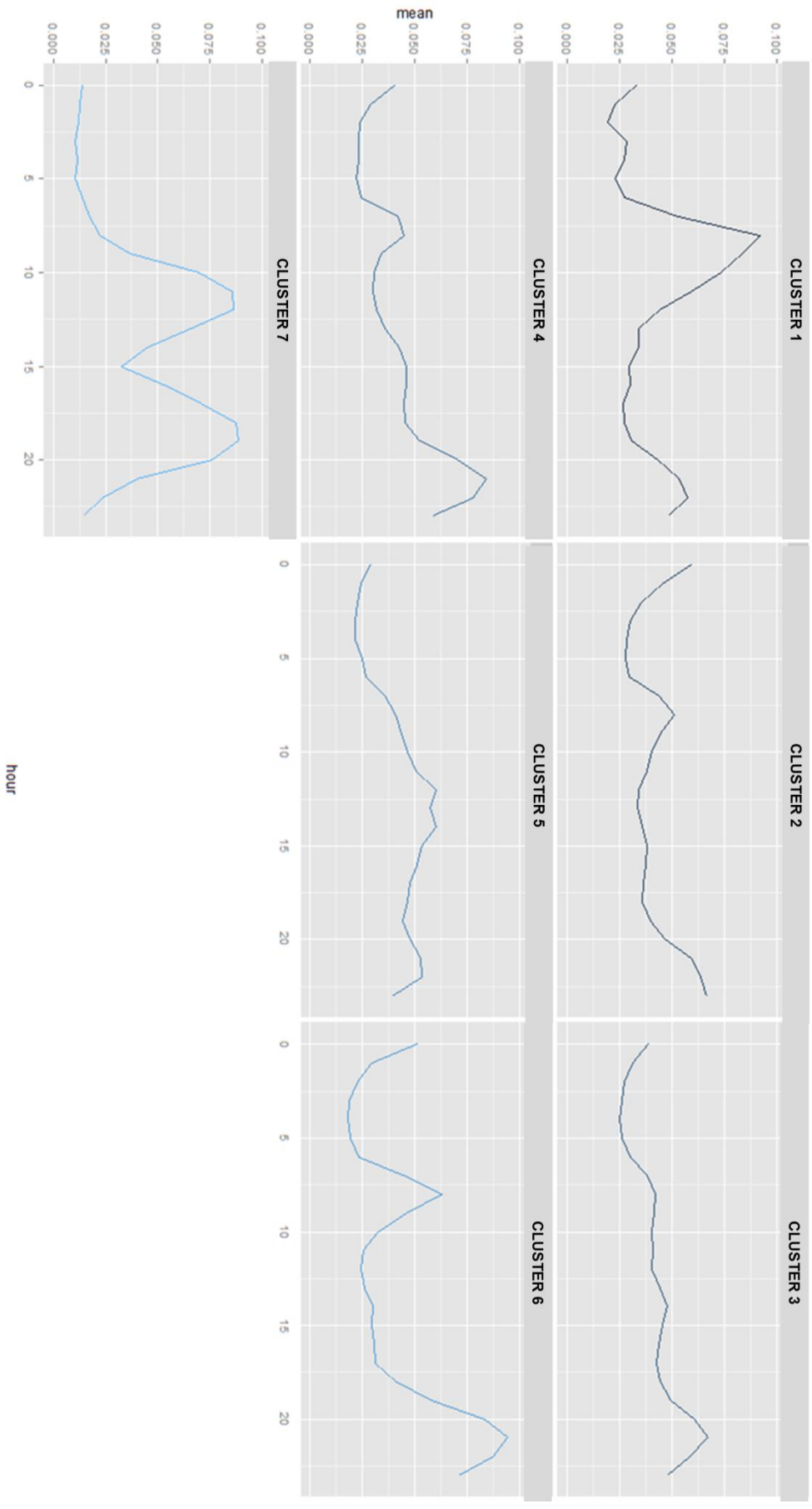


Figure 63. Mean load profile curve of each of the 7 clusters obtained from Manhattan Distance and Ward Linkage

Figure 63 shows the means of the load profiles curves per each cluster, the resulting profiles correspond, at some measure, to the archetypes of load patterns that the Opower's study mentioned. Below a description of the seven clusters mean load profiles is done:

- **Cluster 1:** Presents a high morning peak, almost 10% of the daily energy use is consumed in that peak that is around 8:00h. Then, there is no activity as the proportion of energy use in the afternoon decreases down to 3%, similar at the night percentage, and the evening presents a peak of around 6% at the dinner time. Many reasons can cause this morning peak, the house features can help to determine the causes; but it could be the morning shower, the coffee machine, microwave…

- **Cluster 2:** Presents a not sharp morning peak and a quite regular consumption during day; the peculiarity of this cluster is the there is a night peak at 00:00. The inhabitants of the house are staying longer at night, usually the young people or young couples are associated to that type of load profile.

- **Cluster 3:** Presents a quite flat consumption during the whole day, around 5%. With small evening peak, around 21:00h, that reaches the 7%.

- **Cluster 4:** Presents a pronounced evening peak reaching around the 7,5%, and also a small morning peak that reach almost 5% which is the same proportion as the afternoon consumption.

- **Cluster 5:** This cluster is a bit ambiguous; it looks like it groups all the load that not clearly belong to any of the other clusters and are place in this cluster 5. However, it seems that the load profile mean indicates that some activity is perform during the day in the house, it could mean that people stays home during the day (pensioners, unemployed people,…). This cluster needs a further study to extract some conclusions.

- **Cluster 6:** Presents two sharp peaks, one in the morning and other in the evening, that reach almost 7% and 10% respectively. Then the energy usage during the rest of the day and during the night is so low, constant at 2.5%, so no activity during these periods; it means that the households are not home during the day; could correspond to working people with children.

- **Cluster 7:** It presents two peaks prolonged in the morning (9:00) and in the afternoon (19:00), following the schedule of a small business or local, also the night consumption proportion is so low. So, it seems it is not a residential consumer.

As stated before, many solutions are possible, and it is not likely that the solution sought will be automatically given by the computational method. So, in order to accurate the final solution, the analyst has to study and modify the assignation, if necessary. For instance, check if cluster's outliers could be placed to another cluster where it can fit better; it is likely that cluster 5 members would need to be reallocated.

Residual calculations

Calculating the distance of each load profile respect to its cluster mean load profile, will allow the checking the cluster assignation and determining the cluster's outliers, which are the furthest load profile's to the mean and they might be reallocated in another cluster.

In Table 14, the two members of each cluster that have the highest total distance to its mean load profile are shown, these members could be considered as the cluster's outliers as they are the ones that differ more from the cluster mean (representative load profile). The distance of each member of the cluster to its mean is in the ***Appendix E: Hierarchical residual distances calculations.***

Table 14. First two "idmeter's" members more distanced from the cluster load profile mean

Cluster 1	Dist to Mean 1	Cluster 2	Dist to Mean 2	Cluster 3	Dist to Mean 3	Cluster 4	Dist to Mean 4
X112746	0.3116457	X112773	0.31491590	X112922	0.27586741	X112950	0.27174340
X112815	0.2131447	X103433	0.27869635	X112942	0.21441096	X112719	0.21464751

Cluster 5	Dist to Mean 5	Cluster 6	Dist to Mean 6	Cluster 7	Dist to Mean 7
X112999	0.3104575	X112749	0.2928833	X112914	0.3530615
X113037	0.2591932	X112802	0.2770582	X1928113559	0.2815018

Part of the code written to perform this Hierarchical clustering can be found in ***the Appendix F: Hierarchical functions in "R".***

9.6. K-means clustering

9.6.1. K-means Theory

K-means clustering is the most common partitioning method of vector quantization used for cluster analysis in data mining. It aims to partition the dataset's observations into a number of k clusters defined a priori; where each of the observations belongs to the cluster with the nearest mean or centroid.

In order to achieve the best division of observations into groups, k-means algorithm aims to minimize the total distance between the cluster's members and its corresponding mean (representative of the group); this total distance, known as, within-cluster sum of squares (WCSS) is defined as:

$$arg\ min \sum_{1}^{k} \sum_{1}^{n} \|x_i - \bar{\mu}\|^2$$

(Eq. 13)

Where the distance between an observation (x) an the cluster mean is calculated by the term $\|x_i - \bar{\mu}\|^2$, so the WCSS is the summation of all the set of observations (n) in each (k) cluster, and the iteration process of the algorithm aims to minimizes this distance.

The k-means algorithm it is also referred as "Lloyd's algorithm"; and uses an iterative refinement approach to reach the minimum WCSS. Conceptually the k-means algorithm is described below and a schematic example is represented in Figure 64 in order to easily understand the algorithm process (Wikibooks, 2015) .

Step1: Define the initial groups' means or centroids, as starting point for the algorithm. The common initialization methods are (1) Forgy which randomly chooses k observations from the dataset and use them as initial means or centroids; and (2) Random Partition which first, assigns a cluster to each observation and it computes the initial mean to be the centroid.

Step2: Assign each observation to the cluster that has closest centroid or mean. To algorithm calculates the distance in order to find the most similar cluster. The Euclidean distance is used as a metric and variance is the measure for the cluster scatter.

Step 3: Recalculate the values of the centroids, as the average of all data points in the cluster. The values of the centroids are updated.

Step 4: Repeat steps 2 and 3 iteratively until the observations are no reassigned, so they don't change groups

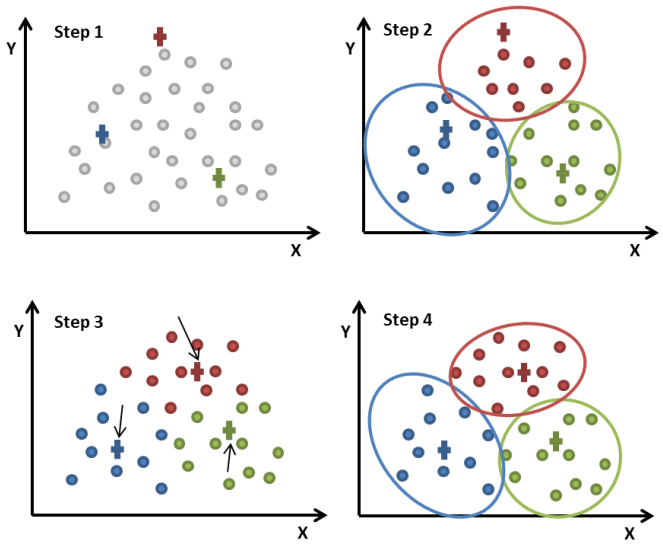


Figure 64. K-means steps procedure demonstration for clusters k=3

In the current project the k-means clustering is transferred to the load profiles segmentation, so instead of points the grouping will be done for larger dimensional vectors, represented by lines.

K-means algorithm is a quite fast and efficient cluster technique, however:

- As the number of k cluster should be fixed as an input parameter, an appropriate choice of the number of cluster may lead to not desired results. Although, it's possible to run diagnostic checks to determine an appropriate number of cluster; or make of the dendrogram obtained from the hierarchical clustering as a reference.

- It does not guarantee to find the global minimum distance, as a heuristic algorithm the result depends on the initial clusters, so the result will represent the local minimum for those starting conditions. So, multiple algorithm runs should be needed to be able to determine the closest solution to the global minimum.

The k-means algorithm is rather easy to implement in large datasets at it has been used for market segmentation, geostatistics, computer vision etc.. However those limitations for some other applications derived to the appearance of some variations, such as the k-medians, k-medoids, Expectation-maximization (EM), fuzzy k-means, among many other partitioning methods; that are not studied in this project.

9.6.2. Cluster formation

As stated in section 9.4.3, for this approach the decision taken was to cluster the customers by using the weekdays' data, and the vector with the 24 variables corresponding to the hourly percentage of energy usage from 00:00 to 23:00h was considered.

K-means algorithms comparisons

To perform the K-means clustering various arguments are needed to be defined and required for the "R" software; the following expression shows necessary inputs to run the k-means clustering:

```
kmeans (x, centers, iter.max, nstart, algorithm)
```

Where "x" is the dataset, "centers" are the number of clusters, "iter.max" are the number of maximum iterations to run, "nstart" is the number of random sets to start the clustering and finally the "algorithm" which could be (1) Hartigan-Wong, (2) Lloyd, (3) Forgy or (4)MacQueen. These algorithms are described below and detailed in (Wikibooks, 2015).

- **Lloyd:** given a set of initial centre vectors (k), Lloyd's algorithm consists on repeatedly (1) moving every initial centre vector to the cluster centroid and (2) updates it by recomputing the distance from each point to its nearest centre; until convergence.
- **Forgy:** it chooses randomly k observations from the dataset and use them as initial centre vectors (c_0). And then, repeatedly (1) assigns each observation to the closest initial centre vector, and (2) replace each of the initial centre vectors (c_0) with the mean of the observations assigned in it, creating the centre (c_1) for the next iteration.
- **MacQueen:** this algorithm works by repeatedly moving all cluster centres to the mean of their respective Voronoi planes. A Voronoi plane is the subset regions that each cluster takes when partitioning a plane.
- **Hartigan and Wong:** K-means allocates each observation to one of K clusters to minimize the within-cluster sum of squares.

Various tests were run combining the different arguments regarding the number of clusters, the starting vectors and the algorithms to find the solution that better fulfilled the objectives sought when clustering the load profiles. From the knowledge obtained when running the Hierarchical clustering (previous section), the number of clusters considered was set 7 for the four different k-means algorithms.

It should be said that a different solution can be obtained each time the function is invoked, if the *set.seed()* function is used guarantees that the results are reproducible, so inside each argument combination many solutions are reviewed. The clustering solutions that were more appropriate are presented:

- Clustering 7 groups using the Lloyd's algorithm. (seed=4)
- Clustering 7 groups using the Forgy's algorithm (seed=9)
- Clustering 7 groups using the MacQueen's algorithm (seed=13)
- Clustering 7 groups using the Hartigan-Wong's algorithm (seed=20)

Discussion

In this section the clusters that better adapts to the goals of clustering the users in different load profiles. Below the four solutions are discussed, their detailed statistical and graphical data can be found in ***Appendix G: K-Means clustering comparison.***

On one hand, the solutions selected from **Lloyd's algorithm, Forgy's algorithm, MacQueen's algorithm;** presented almost identical characteristics.

- The distribution of members per cluster is quite similar where two clusters has a large number of members, around 30-40 members and three small ones with less than 10 members.
- The representatives load profiles, which are the means of all load profiles in the cluster, are well-defined and differentiated among clusters using each of the three algorithms. Those are: (1) high morning peak, (2) two peaks during the morning and the afternoon as it was a business or store, (3) high evening peak with a moderate morning one, (4) high morning and high evening peaks, (5) night peak around 00:00 and (6) flat curve with no peaks.
- The statistical parameters which define the total within clusters distance and between clusters distance are 0,266 and 0,330.

On the other hand, the **Hartigan-Wong's algorithm**, was discarded due to the fact that cluster 5 has only 2 members, and also presented an almost identical representative load profile as cluster 5 which only has 6 members. This implies that, for instance, cluster 2 has 44 members which hosts various types of load profiles that could be split in various groups. Although the statistical parameters of distance between and within clusters are ranged in values similar to the other three algorithms.

9.6.3. Chosen solution

The K-means solution adopted was to cluster the data in 7 groups by using the Forgy's algorithm, but as said before the Lloyd's and MacQueen's algorithms could also be chosen. Due to the fact at this stage is not possible to stablish big differences among them.

The number of members assigned to each cluster, identified in Table 15, range from the minim 7 (cluster 4) to 37 (cluster 5); so a distributed users allocation is also achieved; avoiding clusters with 1 to 5 members which could mislead the analysis. Table 16 shows the within distance clusters where, as expected, the clusters with large number of members (cluster 5 and 7) present a higher within cluster distance. In Table 17 all idmeters per cluster are listed.

Table 15. Number of members per each cluster defined by Forgy's algorithm

Forgy algorithm cluster's members							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Number of consumers	9	17	8	7	37	8	35

Table 16. Within cluster sum of squared distanced per each cluster defined by Forgy's algorithm

Forgy algorithm Within cluster sum of squares distance							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Within cluster distance	0.01736	0.03583	0.02881	0.03677	0.06278	0.02135	0.06401

Table 17. All idmeters in each cluster defined by Forgy's algorithm

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5		Cluster 6	Cluster 7	
X112709	X112718	X90713	X112839	X112695	X112897	X112749	X103433	X112854
X112719	X112724	X112746	X112863	X112701	X112904	X112782	X112696	X112879
X112737	X112739	X112815	X112914	X112723	X112912	X112799	X112707	X112882
X112887	X112773	X112817	X112940	X112732	X112917	X112802	X112729	X112920
X112900	X112788	X112901	X112952	X112742	X112928	X112886	X112738	X112935
X112910	X112826	X112934	X113009	X112761	X112953	X112992	X112745	X112936
X112919	X112862	X112977	X1928113559	X112763	X112966	X113016	X112754	X112939
X112922	X112868	X112999		X112767	X112970	X113023	X112764	X112942
X112981	X112873			X112777	X112994		X112770	X112950
	X112880			X112789	X112995		X112783	X112951
	X112890			X112810	X113000		X112784	X112960
	X112931			X112811	X113018		X112786	X112979
	X112932			X112833	X113031		X112814	X112980
	X112993			X112840	X113037		X112818	X113015
	X112997			X112849	X113041		X112820	X113030
	X113020			X112858	X113043		X112822	X113035
	X1049953413			X112876	X113044		X112824	X113036
				X112877	X113045		X112838	
				X112878				

Observing Figure 65, each of the 7 clusters obtained presents a clearly defined load profile aggregation, easy to identify and differentiated from the other clusters` profiles. The distance between each of the cluster member's load profile and each cluster mean (representative load profile in Figure 66) is calculated and aggregated in order to find the total distance to the mean. The results will show, numerically, which cluster has more dispersed members and will identify the outliers, namely residuals members of the cluster.

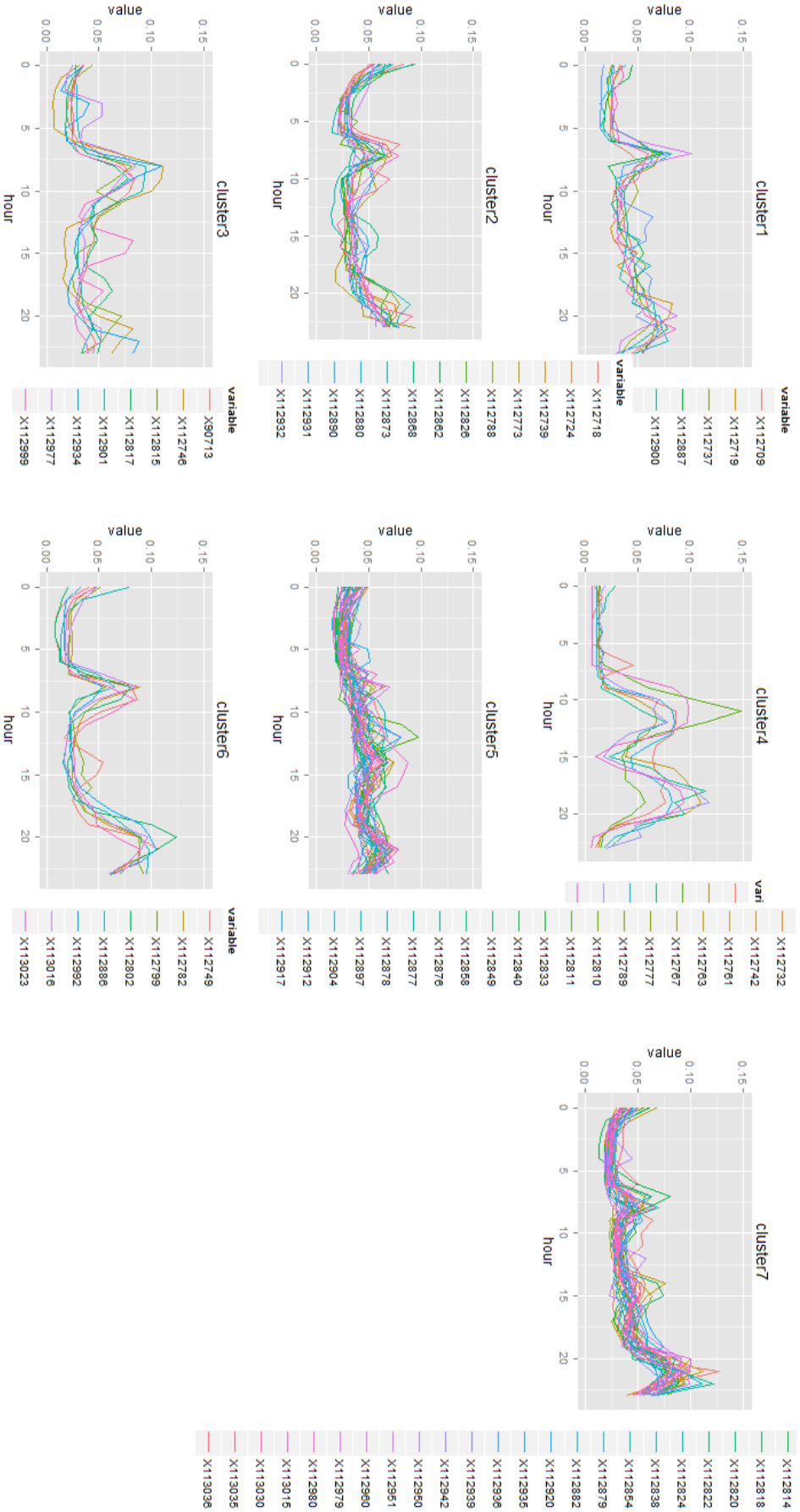


Figure 65. Plotting all the idmeters of each of the 7 cluster, to see the variability and check the clustering results from the k-means Forgy's clustering

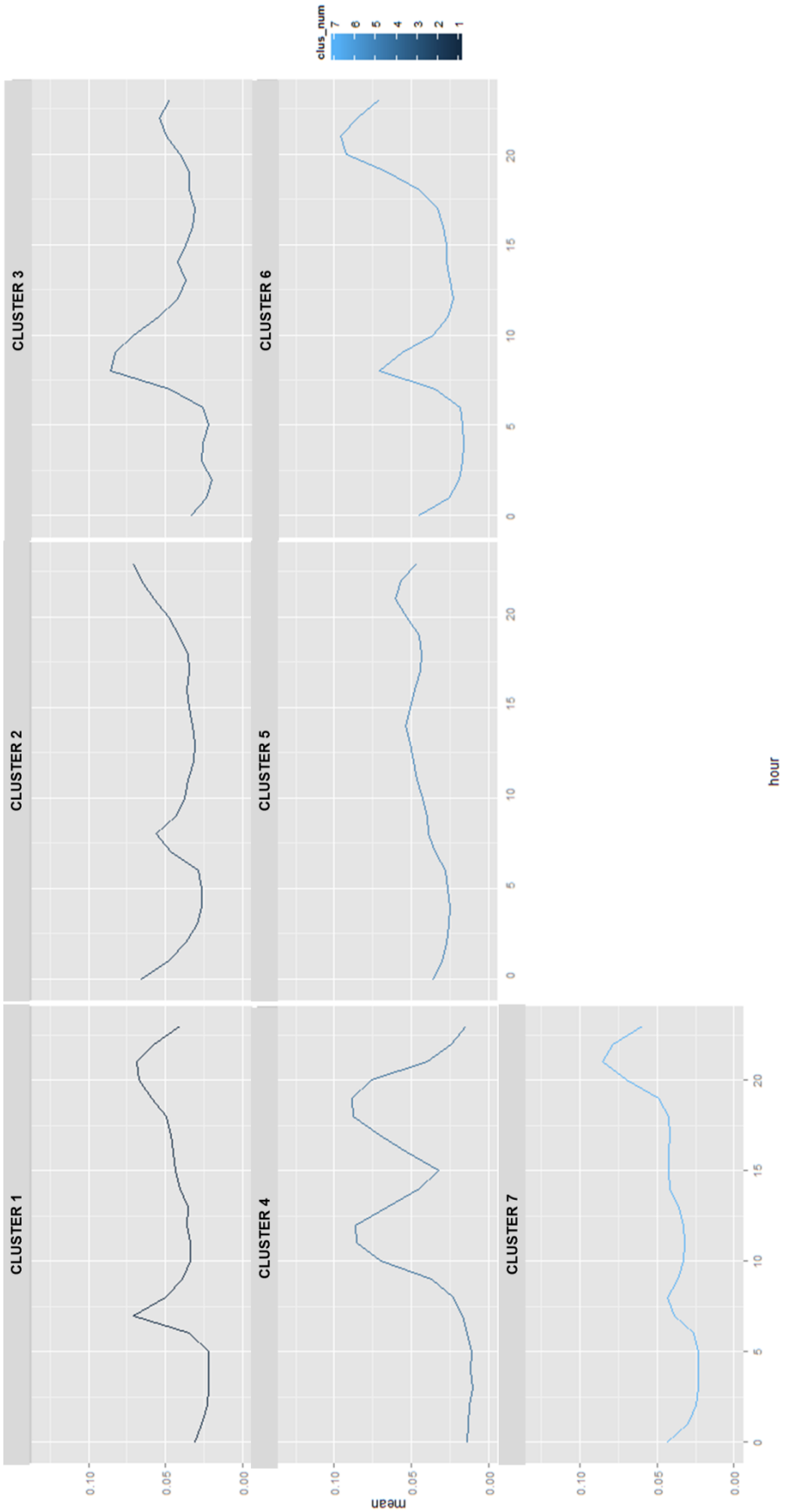


Figure 66. Mean load profile curve of each of the 7 clusters from k-means clustering using Forgy's algorithm

Figure 66 illustrates the means of the load profiles' curves per each cluster, the resulting profiles could be considered as the archetypes of load patterns. Below a description of the seven clusters' mean load profiles are done, it should be notice that they are quite similar to the ones obtained from the Hierarchical clustering in section 9.5.3.

- **Cluster 1:** there are two peaks, morning and evening both of them represents around 7% of the energy usage. It is similar to the cluster 6, although this cluster one presents a higher proportional values between peaks that ranges from 3-5%, it could mean that some activity is done at home.

- **Cluster 2:** Presents a not sharp morning peak and a low consumption percentage during day of around 3% similar to the night hours proportion; this means that there is no activity at home during the day. The peculiarity of this cluster is the there is a night peak at 00:00. The inhabitants of the house are staying longer at night, usually the young people or young couples are associated to that type of load profile.

- **Cluster 3:** Presents a high morning peak, around 9% of the daily energy use is consumed in that peak that is around 8:00h. Then, there is no activity as the proportion of energy use in the afternoon decreases down to 3%, similar at the night percentage, and there is a small evening peak of around 5% at the dinner time.

- **Cluster 4:** It presents two peaks prolonged in the morning (9:00) and in the afternoon (19:00), following the schedule of a small business or local, also the night consumption proportion is so low. So, it seems it is not a residential consumer.

- **Cluster 5:** This cluster presents a flat profile, with the higher values during the day at lunch and dinner time. It could mean that people stays home during the day (pensioners, unemployed people,…).

- **Cluster 6:** Presents two sharp peaks, one in the morning and other in the evening, that reach almost 7% and 10% respectively. Then the energy usage during the rest of the day and during the night is so low, constant at 2.5%, so no activity during these periods; it means that the households are not home during the day; could correspond to working people with children.

- **Cluster 7:** Presents a significant evening peak around 21:00h of almost 9%. There is also a small morning peak, almost not noticeable which does not reach the 5% of energy usage. During the days it also presents a continuous consumption of around 4-5% until the evening peak.

Similarly to the hierarchical cluster, the k-means clustering solution provided by the computational method, it is not likely to be the solution sought but it does an important step ahead to be closer to the final desirable solution. For instance, identify the cluster's outliers and try to place them into another similar cluster, or try to deep study those cluster with larger number of members and detect if any could be reassigned.

Residual calculations

Finding the distance of each load profile respect to its cluster mean load profile allows to determine the cluster's outliers, which are the furthest load profile's to the mean and they might be reallocated in another cluster.

In Table 18, the two members of each cluster that have the highest total distance to its mean load profile are shown, these members among others could be considered as the cluster's outliers as they are the ones that differ more from the cluster mean (representative load profile). The distances of each member of the cluster to its mean are in the ***Appendix H: K-means residual distances calculations***.

Table 18. First two "idmeter's" members more distanced from the cluster load profile mean

Cluster 1	Dist to Mean 1	Cluster 2	Dist to Mean 2	Cluster 3	Dist to Mean 3	Cluster 4	Dist to Mean 4
X112919	0.2459039	X112873	0.30980746	X112746	0.3677728	X112914	0.3530615
X112910	0.1832258	X112773	0.30472691	X112999	0.2703058	X1928113559	0.2815018

Cluster 5	Dist to Mean 5	Cluster 6	Dist to Mean 6	Cluster 7	Dist to Mean 7
X113037	0.33669788	X112749	0.2557940	X112950	0.26696678
X112912	0.25079317	X112886	0.2206823	X103433	0.25520192

Part of the code written to perform this K-means clustering can be found in the ***Appendix I: K-means functions in "R".***

9.7. Self-Organizing Maps (SOM)

9.7.1. Self-Organizing Maps Theory

The Self-Organizing Map (SOM) is a clustering and data visualization technique that projects high dimensional data to a 2-dimensional space in a hexagonal grid. SOM is an artificial neural network that relies on unsupervised learning to group the input vectors into a map composed by nodes or neurons.

A weight vector with the same dimension as the input data vectors is associated with each node and positioned into the map space. A vector from the dataset is placed onto the map by finding the closest weight vector to it, so the closest node. Then, clustering techniques can be applied to group neighbouring nodes into clusters.

Like other artificial neural networks, SOM has two steps; the first is the training in which builds the map using input examples and the second, the mapping in which automatically classifies a new input vector. During the training process SOM uses each data vector to update the closest centroid and also the nearby centroids, this ordered set of centroids means that the centroids that are close to each other are more similar that further centroids; so SOM enforces neighbourhood relations that will eventually end up in forming clusters. And facilitate the interpretation and visualization of clustering results.

A schematic example of a Self-Organizing Map is illustrated in Figure 67. The summarised steps of the algorithm are (Wikibooks:SOM, 2015):

Step1: Define the initial
Step2: Repeatedly, select an input vector, determine the closest centroid (node) to the vector and updated this centroid and the nearby centroids.
Step 3: until the centroids don't change or the limit is exceeded
Step 4: Assign each object to its closest centroid and return the centroids and clusters

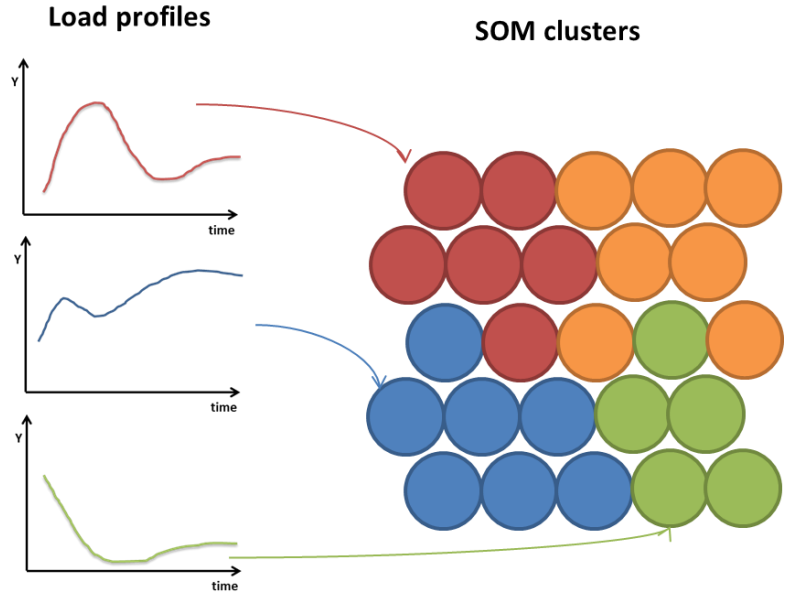


Figure 67. SOM example scheme, placing load profiles to the SOM's nodes, each colour represents a cluster

9.7.2. Map and clusters formation

As in the previous sections Cluster formation 9.5.2, 9.6.2, the consumption data used is again the same composed by the weekdays', and the vector with the 24 variables corresponding to the hourly percentage of energy usage from 00:00 to 23:00h was considered.

When clustering using the SOM, two steps are needed; first is to create the Map that will reduce the data to a 2-dimensional space; and second apply the most convenient clustering technique, for instance hierarchical or k-means clustering.; to achieve the final clustering.

Map formation

It was decided to reduce the initial dataset of 121 consumers, into a 4x4 nodes map; as explained in the SOM theory the algorithm will allocate the most similar consumer's load profiles in the same or neighbours nodes. The decision to reduce the data it's up to the analyst, it could also be any other nodes number such as 3x3, 5x5. It should be said that SOM is more appropriate and commonly used for higher dimensional datasets, accounting for thousands of observations; however it also work for the current project.

The first outcome when running the code in R and plotting the SOM is presented in Figure 68 (left image) and also the number of members grouped in each node Figure 68 (right image). Table 19 presents the relation between each "idmeter" and the node to which is placed.

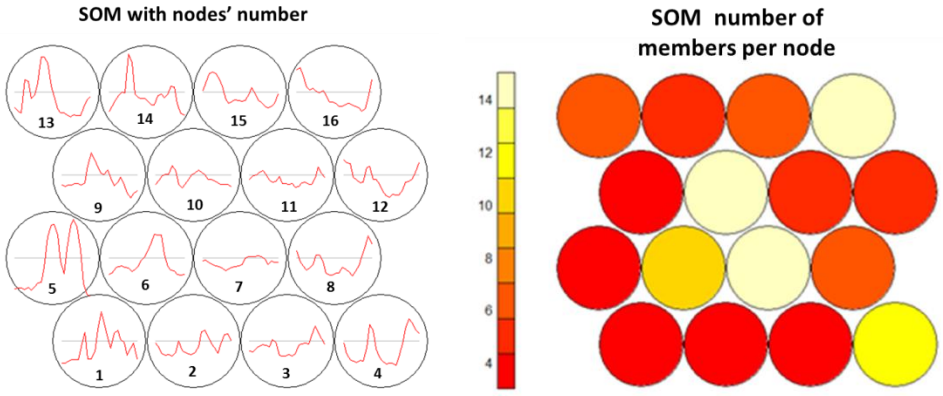


Figure 68. Left side image shows the SOM with the nodes identification. The right side image shows the number of members per node in a colour scale

Table 19. Identification of each idmeter number into the node where is placed

Idmeter	Node	Idmeter	Node	Idmeter	Node	Idmeter	Node	Idmeter	Node
X90713	12	X112770	6	X112849	14	X112920	7	X112995	16
X103433	3	X112773	4	X112854	6	X112922	11	X112997	4
X112695	16	X112777	16	X112858	16	X112928	16	X112999	16
X112696	10	X112782	1	X112862	4	X112931	4	X113000	16
X112701	7	X112783	2	X112863	13	X112932	4	X113009	13
X112707	7	X112784	7	X112868	8	X112934	12	X113015	10
X112709	7	X112786	10	X112873	1	X112935	10	X113016	1
X112718	4	X112788	3	X112876	14	X112936	7	X113018	7
X112719	6	X112789	8	X112877	10	X112939	10	X113020	4
X112723	16	X112799	6	X112878	9	X112940	13	X113023	5
X112724	4	X112802	5	X112879	7	X112942	3	X113030	6
X112729	10	X112810	15	X112880	4	X112950	7	X113031	10
X112732	16	X112811	14	X112882	2	X112951	6	X113035	11
X112737	11	X112814	2	X112886	1	X112952	13	X113036	7
X112738	6	X112815	12	X112887	10	X112953	16	X113037	14
X112739	4	X112817	9	X112890	4	X112960	7	X113041	15
X112742	15	X112818	10	X112897	8	X112966	16	X113043	15
X112745	15	X112820	7	X112900	11	X112970	8	X113044	16
X112746	5	X112822	10	X112901	12	X112977	12	X113045	16
X112749	5	X112824	6	X112904	16	X112979	6	X1049953413	3
X112754	2	X112826	4	X112910	10	X112980	6	X1928113559	13
X112761	8	X112833	15	X112912	11	X112981	7		
X112763	14	X112838	10	X112914	13	X112992	6		
X112764	10	X112839	13	X112917	7	X112993	8		
X112767	16	X112840	7	X112919	9	X112994	15		

9.7.3. Cluster formation: Chosen solution

To obtain the clusters in the map a clustering technique should be applied to the map by grouping the nodes most similar nodes. For instance k-means or hierarchical clustering could be applied; in this case the Hierarchical clustering is applied again using the Manhattan distance and the Ward linkage in order to find 7 clusters, the map obtained is the one in Figure 69. Table 20 presents the “idmeter” numbers in each cluster and Table 21 shows the list of all idmeters per cluster.

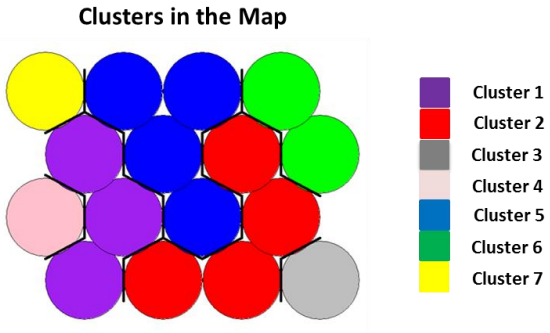


Figure 69. The 4x4 SOM with the 7 clusters in differentiated by colours

Table 20. Number of members per cluster

Forgy algorithm cluster's members							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Number of consumers	18	19	12	4	41	20	7

Table 21. Identification of each idmeter number into the correspondent cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5		Cluster 6	Cluster 7
X112782	X112754	X112718	X112746	X112701	X112838	X90713	X112839
X112873	X112783	X112724	X112749	X112707	X112877	X112815	X112863
X112886	X112814	X112739	X112802	X112709	X112887	X112901	X112914
X113016	X112882	X112773	X113023	X112784	X112910	X112934	X112940
X112719	X103433	X112826		X112820	X112935	X112977	X112952
X112738	X112788	X112862		X112840	X112939	X112695	X113009
X112770	X112942	X112880		X112879	X113015	X112723	X1928113559
X112799	X1049953413	X112890		X112917	X113031	X112732	
X112824	X112761	X112931		X112920	X112763	X112767	
X112854	X112789	X112932		X112936	X112811	X112777	
X112951	X112868	X112997		X112950	X112849	X112858	
X112979	X112897	X113020		X112960	X112876	X112904	
X112980	X112970			X112981	X113037	X112928	
X112992	X112993			X113018	X112742	X112953	
X113030	X112737			X113036	X112745	X112966	
X112817	X112900			X112696	X112810	X112995	
X112878	X112912			X112729	X112833	X112999	
X112919	X112922			X112764	X112994	X113000	
	X113035			X112786	X113041	X113044	
				X112818	X113043	X113045	
				X112822			

In Figure 70, each of the 7 clusters obtained presents a clearly defined load profile, easy to identify and differentiated from the other clusters` profiles. The 7 representative load profiles (Figure 71) are almost the same as the previous clustering techniques, it should be noticed that cluster numbers differ from one technique to another but the representative load profiles are the same. In this case:

- Cluster 1 represents the Day-time consumers
- Cluster 2 represents the Evening peaks
- Cluster 3 represents the Double peaks
- Cluster 4 represents the Business
- Cluster 5 represents the Flat consumers
- Cluster 6 represents the Late night peaks
- Cluster 7 represents the Morning peaks

Comments

SOM is useful to reduce high dimensional data, in this case the sample of 121 consumers was relatively small to fully take advantage of the SOM algorithm, although it worked well and the results obtained are acceptable; other techniques such as the hierarchical and k-means clustering have adapted better to the project goal.

A backward of SOM is that it is difficult to justify and keep track of the grouping nodes as the process is automatic, so the analyst cannot control the grouping of observations into the nodes, only when applying the hierarchical clustering technique can visualize the nodes grouping into clusters but cannot act into the nodes formation.

SOM potential could be applied if, for instance, the aim was to find the archetypes of the 22.000 daily load profile curves. This high-dimensional data could be reduced into a map and then clustering at convenience; using SOM for this case will allow a quick computation and allocation of the load profiles.

Part of the code written to perform this SOM clustering can be found in the ***Appendix J: SOM clustering.***

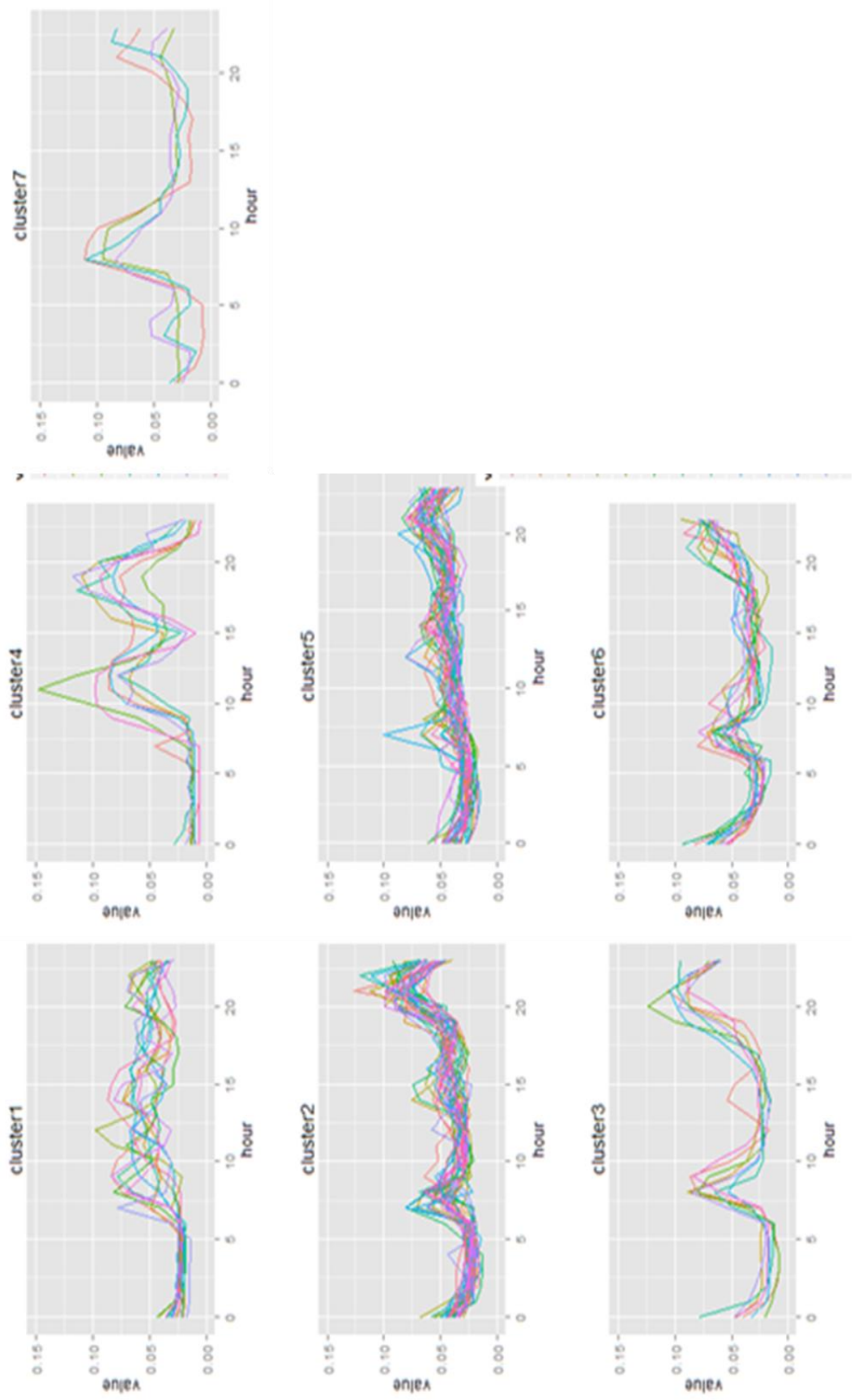


Figure 70. Plotting all the idmeters of each of the 7 cluster, to see the variability and check the clustering results from the SOM method

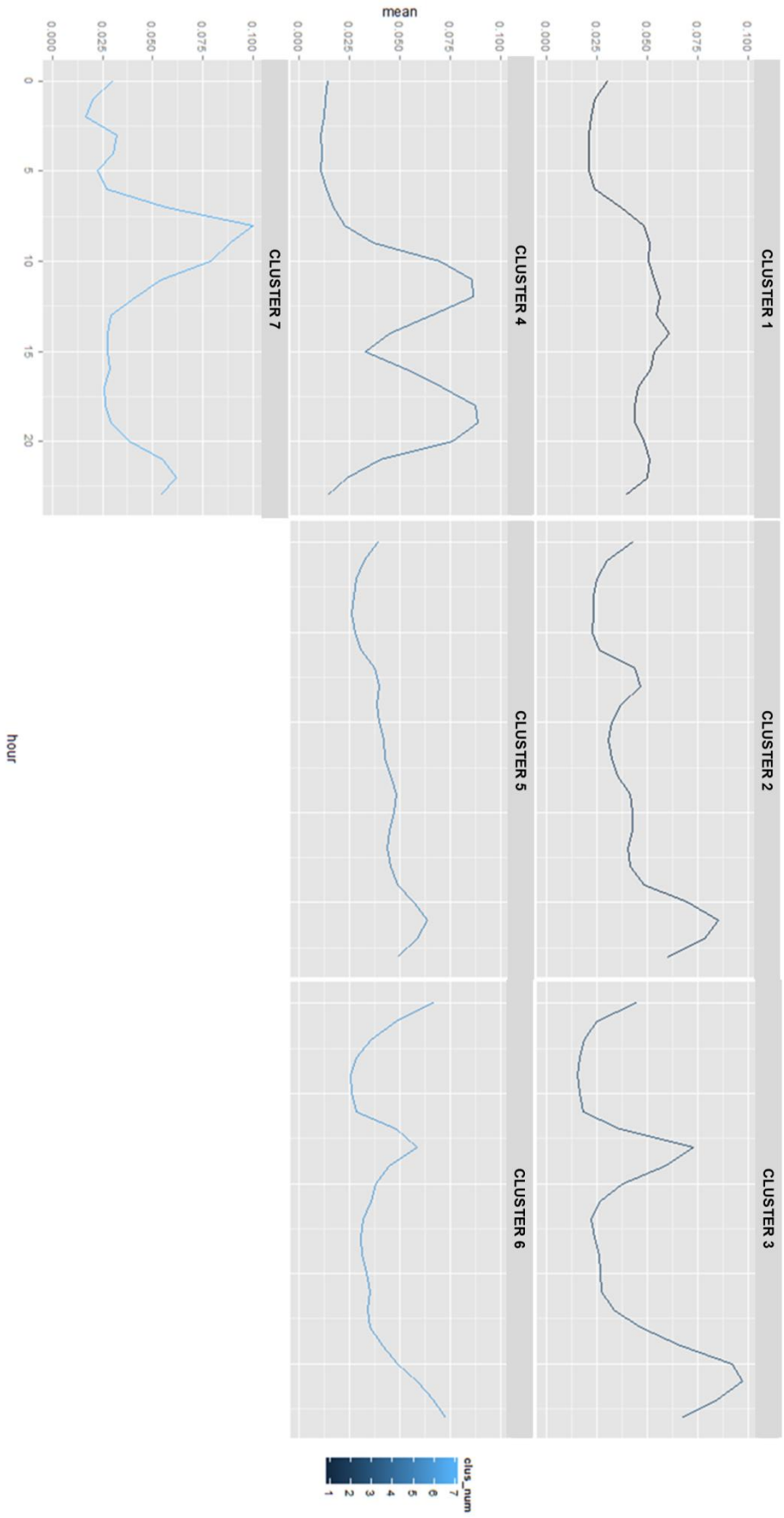


Figure 71. Mean load profile curve of each of the 7 clusters obtained from SOM map and hierarchical clustering

9.8. Post-clustering: Results and conclusions

9.8.1. Final Clustering Solution

After having obtained the three clustering solutions from the three methods (SOM, Hierarchical, K-means), none of them adapts exactly and perfectly to the consumer segmentation pursued; as it may exist some cluster's outliers or some undefined clusters which would need a further exploration that could likely lead to a members' re-assignation.

An analysis of the positive and negative points of each method that were found when carrying out this project are presented in Table 22:

Table 22. Pros and cons found for each clustering method used

Method	SOM	Hierarchical	k-means
Pros	High dimensional data to 2-dimensional into a map	Dendrogram visualization, easy to keep track and make clustering decisions	Statistical data associated to the clustering
	Simple algorithm, easy to explain	No need to specify number of clusters in advance	Fast computational method
Cons	Need to apply a clustering technique after obtaining the SOM	Dendrogram difficult to visualize for large datasets	Local minimum due to random starting points (no global minimum)
	Difficult to keep track of the of observation distribution into nodes	Slower computational method	Need to specify number of clusters in advance

The base solution that is used to find the final clusters is the one obtained by using the **Hierarchical Clustering** seen on section 9.5.3. It is considered to be the closest one to the final cluster formation, and the one that would need apriori less modifications to reach the results sought; as it already presents well-defined and differentiated clusters which are in concordance to the objectives sought.

From the previous computational clustering results, it was seen that the best segmentation is to define seven clusters that represent the following load profiles illustrated in Table 23:

Table 23. Description of the 7 representative load profiles obtained from the analysis

Load profile definition	Criteria
1- Morning peak	A morning peak higher than 7,5%, low energy usage during the day and evening peak lower 7,5%
2- Late night peak	A night peak is reached at around midnight at 23:00h (later than the evening peak). Also exists a morning peak
3- Flat consumers	An almost flat profile appears, with an evening peak upper to 5% as the higher point

4-	**Evening peaks**	A evening peak higher than 7,5% and small morning peak lower than 5%
5-	**Daytime consumers**	It presents the highest period of consumption during the day
6-	**Double peak**	Present both morning and evening peaks, and a very low consumption during the daytime
7-	**Business**	Not residential. The load profile peaks follows the schedule of a small business or local

The work done in this stage consists basically on analysing deeper each cluster determined by the computational methods in order to identify the possible redistribution of cluster's members.

The already well-defined clusters, such as cluster 1 and 7 won't differ from the base clustering members, moreover their members are relatively small in number and they not present any outlier that could be placed in other cluster, hence they are not further analysed.

However, a special attention is put on the clusters with higher number of members, such as cluster 2, cluster 3 and cluster 4. As, in some of them (cluster 3 and 4) when plotting all the members' load profiles is still not clear to visualise and detect if inside each cluster there are some sub-groups or outliers that could fit better in other cluster. On the other hand, in cluster 2 the outliers are easy to detect and re-assign to another cluster with similar characteristics. Cluster 5 is a case apart, as it is the most disperse cluster when plotting their members, however, its representative load profile shows a different load patterns from the other clusters.

The hierarchical clustering is applied now for each of the clusters (cluster 2, 3, 4, 5 and 6) separately, using again the Manhattan distance and the Ward linkage; the idea is to divide each group in several, in order to determine which members inside each cluster are considered well allocated and which ones could fit better in another cluster. In the following diagram (Figure 72) the clusters' members redistribution is presented and some of the most relevant changes are described below:

- The base cluster 2 (late night peak), five of the members are reallocated to the flat consumers group.

- The base cluster 3 (flat consumers), it was the cluster with larger members and the one that suffers a bigger change. Performing the division of this cluster it is detected that some load profiles present evening peaks according to the ranges adopted, and o they are placed in the evening peaks' cluster.

- In the base cluster 4 (evening peak), some members are reallocated to the double peak segment; as when performing the cluster sub-division these members present also a high morning peak.

- In the base cluster 5 (daytime consumers), just one member was reallocated to the morning peak segment.

- The base cluster 6 (double peaks), two members were moved to late night peak segment.

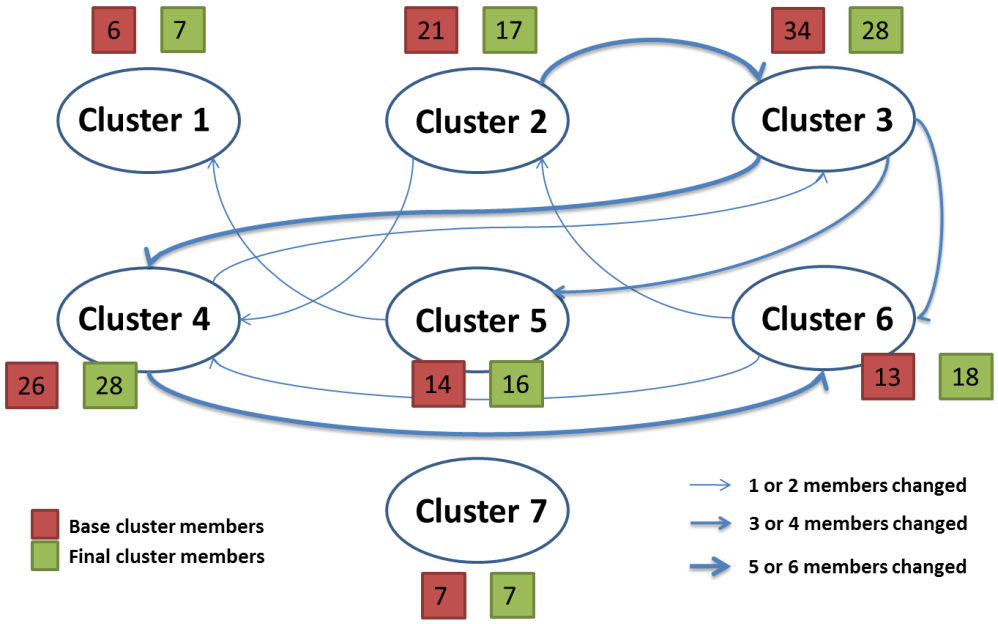


Figure 72. Clusters' members redistribution to define the final solution

The results of each cluster's division and the identification of the members that has been reallocated are detailed in
***Appendix K: Base clusters* sub-division.**

In the end, to reach the segmentation desired, a combination of computational methods that allows the automatization of the procedure together with the analyst manual review are needed. The final segmentation members are presented in Table 24 and Table 25.

Table 24. Final number of consumers distributed per cluster

Final clustering solution							
Cluster number	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
Number of consumers	7	17	28	28	16	18	7
Share of total consumers	6%	14%	23%	23%	13%	15%	6%

Table 25. Final distribution and idmeters identification per cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
X90713	X112718	X112695 X112897	X103433 X112939	X112723	X112719	X112839
X112746	X112724	X112701 X112904	X112696 X112942	X112742	X112737	X112863
X112815	X112739	X112709 X112917	X112707 X112950	X112810	X112749	X112914
X112817	X112773	X112732 X112920	X112729 X112951	X112811	X112754	X112940
X112901	X112788	X112761 X112928	X112738 X112960	X112833	X112770	X112952
X112934	X112826	X112763 X112935	X112745 X112979	X112849	X112782	X113009
X112977	X112862	X112767 X112936	X112764 X112980	X112876	X112783	X1928113559
	X112868	X112777 X112953	X112786 X112992	X112878	X112799	
	X112873	X112784 X112970	X112818 X112995	X112912	X112802	
	X112880	X112789 X112993	X112820 X113015	X112919	X112814	
	X112886	X112840 X113018	X112822 X113030	X112966	X112882	
	X112890	X112858 X113043	X112824 X113031	X112994	X112887	
	X112931	X112877 X113044	X112838 X113035	X112999	X112900	
	X112932	X112879 X113045	X112854 X113036	X113000	X112910	
	X112997			X113037	X112922	
	X113020			X113041	X112981	
	X1049953413				X113016	
					X113023	

In Figure 73, the final 7 clusters obtained after the redistribution of the otuliers is presented, each cluster has a clearly defined load profile, easy to identify and differentiated from the other clusters` profiles. The 7 final representative load profiles (Figure 74) are similar to the ones obtained from the computational methods, so that the redistribution of the above mentioned members does not vary the cluster's representative load profiles.

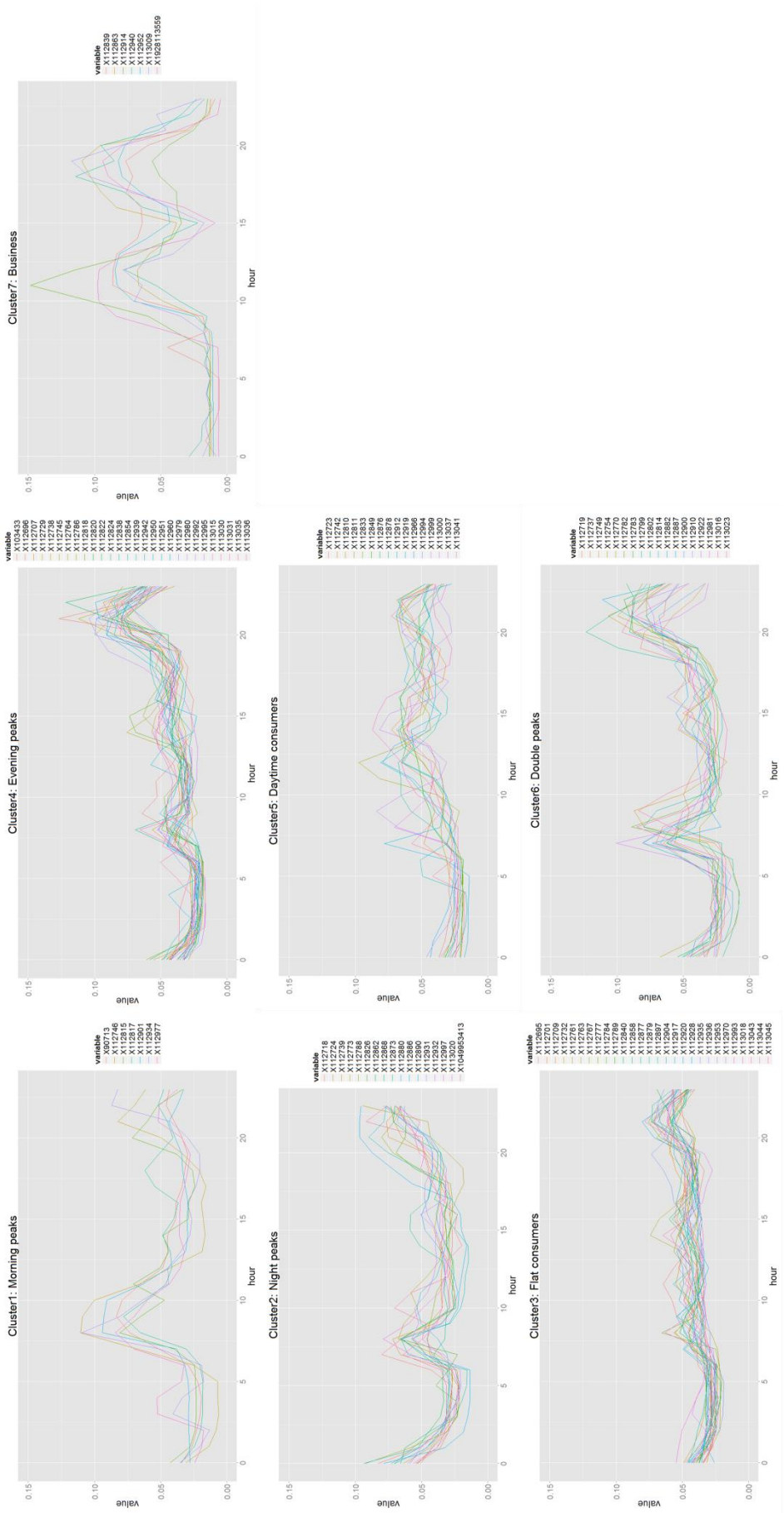


Figure 73. Plotting all the idmeters of each of the final 7 clusters

Figure 74. Mean load profile curve of each of the 7 clusters obtained after the redistribution

Cluster description & advices

After having segmented the consumers by their characteristic load profile, it is possible to provide some personalised energy-saving recommendations for each segment's consumers.

- **Electricity tariffs change:**

The costs related to the electricity consumption have a significant importance, so their analysis are useful in order to study if a possible change in tariff will pay-off and so, be less costly to the consumer. Also, doing this exercise it would useful to aware people on which tariff the have contracted, and the one that would be better according to their load profile.

In Spain, there are three different types of electricity tariffs (Figure 75) for residential customers:

- Normal tariff, which presents a similar price between 0,12 -0,13 €/kWh
- Discrimination tariff (day-night), which presents a lower price (00:00h to 13:00h) around 0,06-0,07€/kWh and higher price (14:00-23:00h) around 0,15€/kWh.
- Electric vehicle tariff, similar to the discrimination, but the price at night hours (01:00h-07:00h) is a bit lower; as during this time is when the electric vehicles should be charged.

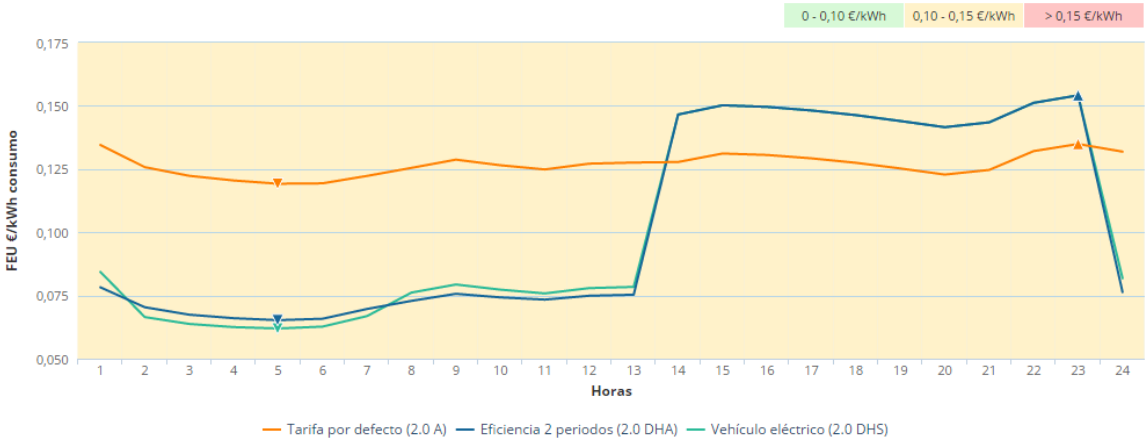


Figure 75. Example of June 1st, 2015 hourly electricity prices per each tariff. Source: (Red Eléctrica de España, 2015)

A simple calculation is done in order to determine which could be the better tariff for each cluster; in reality this calculation should be done individually for each consumer. In this example, a comparison between normal tariff and discrimination tariff is done taking the cluster's representative loads profile and multiplied for the price of electricity at each hour.

The result was that for every cluster the recommended tariff would be the discrimination tariff. It is likely that a significant number of consumers don't know which tariff and the contracted power they have at home; moreover

- **Base consumption :**

The base consumption refers to the minimum consumption that the house constantly has, this is related to the night hours and the hours where there is no activity at home, for instance when the households are working.

Minimizing this base consumption is key to reduce the consumption when there is no activity and no need to use electricity. Therefore, this base consumption period should present a linear and regular profile, and as lower the better.

Focusing on the representative load profiles from Figure 74; clusters 1, 2 and 6 (morning peak, evening peak and double peaks) present this base consumption between the peaks; in these three cases the base consumption represents a relatively small share as the presence of the peaks represent a high proportion of energy usage. But still is energy consumed that is not "used".

The base consumption in Cluster 3 (flat consumers) is quite important as there are no peaks, the hourly night consumption is above the 2,5%, which in the end represent a significant share for a period such as the night where no activity and consumption should be needed.

Some recommendations in the sense to reduce the base electricity consumption could be:

- Devices that have 24h consumption such as the refrigerator and the freezer, is recommended to have an energy efficiency label A^{+} , A^{++}, A^{+++}. As stated in (Fonseca, et al., 2009) the refrigerator consumes around 28% of the total electricity use

- Avoid the so-called stand-by consumption by switching off completely the devices. The stand-by mode allows the devices not to be completely switched off, and leaves a LED switched on that is believe to consume around 10% of the consumption with that device switched on. For instance, switch off the TV, wi-fi, computer… at night or when leaving home.

- A consumer's behaviour change is needed in order to make people understand the importance of switching off the lights, the air conditioning, disconnecting the chargers; when is no needed.

- **Peaks reduction:**

The peaks and the power contracted should be treated together, as higher power contracted, higher cost. So being able to reduce the peaks will lead a reduction of the peak power therefore to the power contracted and the costs associated.

Also, nowadays the time-of-use tariffs allows to set different prices each hour, the utilities in order to shave the peaks set higher prices to the peak time, usually the evening. So, again a behavioural change is needed in order to reduce the peaks, the simultaneity coefficient of devices used at the same time should be reduced. For instance, use the washing machine at night, out of peak hours.

However, a further investigation on the households' features and householders' characteristics is needed in order to be able to determine the possible causes and the origin of the load patterns and whether exists or not any correlation with the households' features and the load patterns; with so it would be possible to provide more accurate and detailed energy-savings recommendations.

For example, the load profile it won't reflect the same whether the kitchen is electric or gas; also whether the hot water boiler is electrical or gas; due to the fact that it they are electric the peaks are likely to be visible. However, these and other characteristics are further discussed in the next section.

10. Households and Householders features analysis

10.1. Introduction

As part of the Phase 2 described in section 5, after having clustered the consumers by their load profile, the next step is to find the more likely characteristics and features of the households and householders inside each cluster. In order to be able to associate some specific features to a specific cluster's load profile.

Some studies have been done in analysing households and householders' features; for instance, in (Beckel, et al., 2012) the answers to extensive and detailed questionnaires of 3.488 private households were recollected by the Irish Commission for Energy Regulation (CER); where 18 household properties were considered and differentiated as follows:

- Properties related to the occupants of the household: number of adults, children, occupants at home during the day, employment and social class
- Properties related to the dwelling: type of dwelling, relationship to property, year of construction, floor area, number of bedrooms.
- Properties related to the home appliances: Type of cooking, domestic hot water heating, space heating, number of appliances, among others.

After analysing these features, they could detect the group of households with large energy savings potentials, mainly due to the high number of electrical appliances. And concluding that the characteristics that were more relevant to the define a consumer with high savings potential are: type of occupants employment, number of adults/children, type of space heating, type of domestic hot water, heating, total number of home appliances and dwelling construction year.

However, for the current study case it is not available such amount of detailed data as in (Beckel, et al., 2012); the features' data is much more limited in terms of samples and properties. The data referred to the household and householders is described in section 7.3, this data is not fully complete and also needs to be further treated in order to eliminate bad data and duplicated features. Once refined, the data is able to be analysed, in this case a graphical analysis using histograms is considered to represent the results and be the base to extract conclusions.

10.2. Objective

The ideal goal pursued is to discover if exists any pattern or relation when crossing the electrical load profiles groups with the additional information related to households and householders features. So, find the responsible features that could be the cause that define a specific load profile or another. In order to achieve this objective a significant number of samples is needed to consider the output results as a valid indicator and to be representative to extrapolate and use them for other purposes; for instance to determine the load profile of new users by knowing their characteristics.

Nonetheless, the actual features' dataset presents some limitations (small dataset, missing some relevant features, etc…) also the members in each cluster is so small, for instance the larger cluster has only 28 members. These facts, led to change the initial analysis' purpose to a one more adequate according to the dataset available, moving to a less ambitious analysis limited to find the most common household's and householder's features within each cluster. Hence, the output sought is to have a representative table stating the most common features for each load profile cluster.

10.3. Procedure

10.3.1. Dataset refining

The joined dataset from section 7.3.2 which merged the data from the info-house and the technical audit accounted for a total of 20 features, but some of them were duplicated, others were incomplete or not relevant. So, a restructuration of the table was necessary to continue the study, the actions performed are listed below:

- Add a column with the cluster number "cluster", to identify each idmeter.
- There were two "Area" columns and two "Typology" columns; those features were duplicated when merging the datasets, so it was necessary to check these columns and reduce to unique column for "Area" and another for "Typology"
- The "Age ranges" were also simplified to "Children: Less than 12 years", "Adults: from 13 to 65 years" and "Pensioners: older than 65". The old dataset had the ages divided in different ranges.
    - On the technical audits the age ranges were: "Less than 3 years", "Between 3 and 9 years", "Between 10 to 17", "Between 18 to 24", "Between 25 to 65","More than 65".
    - The info-house only differentiated between "Less 12 years" and "Adults: more than 12 years"
- The maintained features' columns were: "Year", "Contracted Power", "Air conditioning", "Domestic hot water", "Space heating", "Kitchen type", "Drier", "Dishwasher"

10.3.2. Features analysis

The analysed features dataset is composed by 125 "idmeters" and the 14 properties described below. These properties can be divided depending on their nature as:

- House properties: Typology, contract, Area, Year, Contracted power (5 features)
- Sociological: Age ranges by pensioners, adults, children (3 features)
- Home appliances: Air conditioning, Domestic Hot Water, Space heating, Kitchen type, Drier, Dishwasher (6 features)

To have a sense of the quality of the data, Figure 76 shows the histograms for each feature taken into account for the whole dataset without differentiating the clusters. It should be noticed that the number of "Not available (NA)" data is quite significant mainly in the home appliances properties; fact that, added to the small dataset, reduces the reliability nd could distort the results.

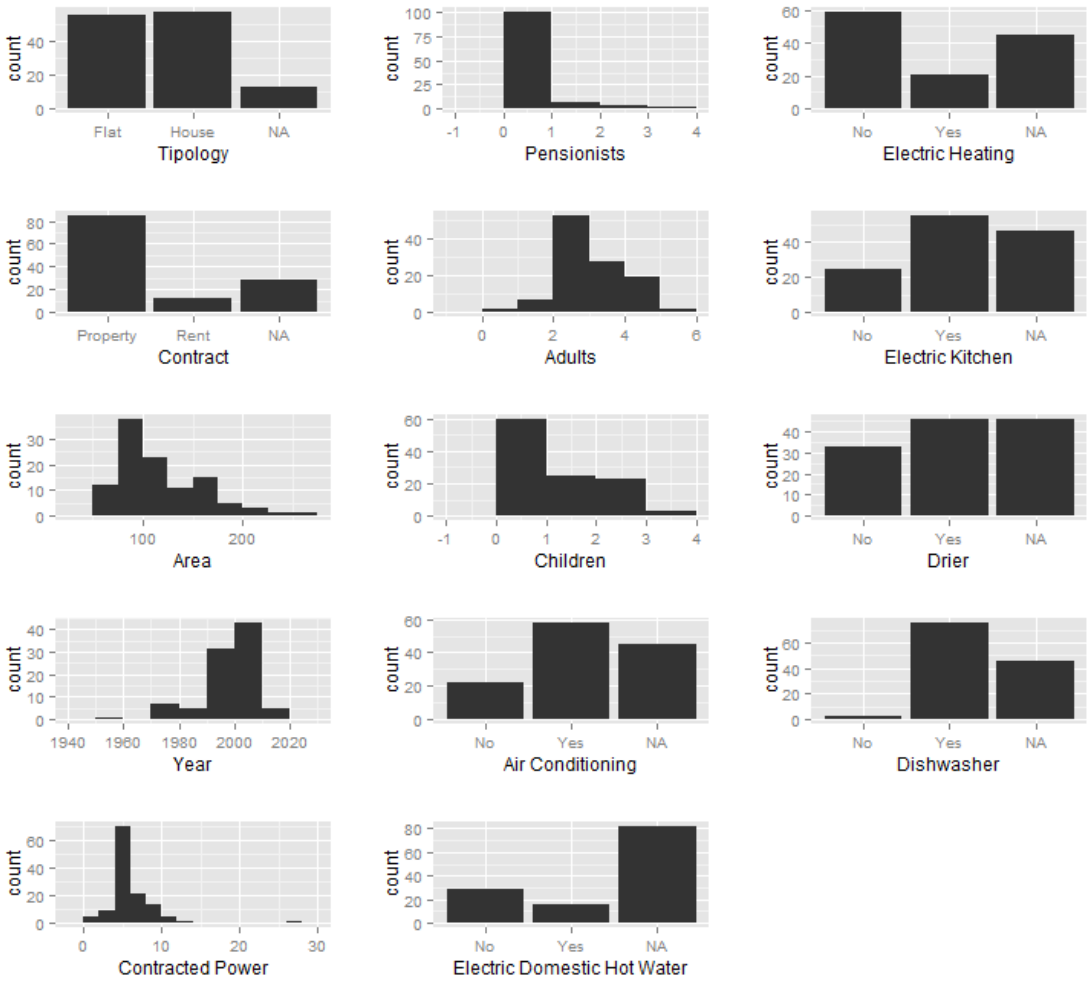


Figure 76. Histograms for each features considered, taking into account the whole dataset

The same exercise was done for each cluster, the histograms figures can be seen in the ***Appendix L: House features analysis***; and the outcomes as explained in the next section.

10.4. Results and Conclusions

Table 26 summarizes and states the most common property for each cluster.

Table 26. Most common property figure for each cluster

CLUSTER	1	2	3	4	5	6	7
House properties							
Typology	House	Flat	House	House	Flat	Flat	NA
Contract	Property	Property	Property	Property	Property	Property	NA
Area	150 m^2	<100 m^2	90-110 m^2	90-100 m^2	90-100 m^2	<100 m^2	NA
Year	2000-2010	1995-2000	2000-2010	1990-2000	2000-2010	2000-2010	NA
Contracted power	5 - 10 kW	<5 kW	4-5 kW	4-5 kW	4-5 kW	>4,4 kW	NA
Sociological (ages ranges)							
Pensioners	0	0	0	0	0	0	NA
Adults	2 and 4	2	2-3	2-3	4	2	NA
Children	0	0	1*	1-2*	0	1-2*	NA
Home appliances							
Air conditioning	No	Yes	Yes	Yes	Yes	Yes	NA
Domestic Hot water	NA	NA	NA	NA	NA	NA	NA
Heating system	Gas	NA	Gas	Gas	Gas	Gas	NA
Kitchen	Electric	Electric	Electric	Electric	Electric	Electric	NA
Drier	Yes	NA	Yes	NA	No	Yes	NA
Dishwasher	Yes	Yes	Yes	Yes	Yes	Yes	NA

* "0" is the most common but the counts of 1-2 children is higher than in the other clusters

A brief description of each cluster features and possible correlation suppositions are presented below:

- **Cluster 1 (morning peaks):** big property houses, probably occupied by a couple of 2 adults with little children or a couple with a pair of grown kids considered adults in the analysis. Not using electricity for heating and cooling as no air conditioning neither electric heating are installed. The power contracted is the highest one among all the clusters, which could be related to the house area.
- **Cluster 2 (late-night peak):** relative small flats, low power contracted below 5kW, it may be young couples with no children as no pensioners are in this cluster. The occupancy of young couples could explain the peak at night.
- **Cluster 3 (flats consumers):** relative small and modern houses (after year 2000), 2-3 adults, with 1 child. No electric heating, but there are some electric appliances electric (kitchen, drier, dishwasher, air conditioning…)
- **Cluster 4 (evening peaks):** almost the same number of houses and flats in property, with an area lower than 100m^2, built between 1990-2010. It could be families with little children as the number of 1-2 children in this cluster is higher than the other clusters.

- **Cluster 5 (day-timers):** almost only adults are presents in this larger number than in other clusters (2-4 adults), no children, no pensioners, less home electric appliances, different schedules
- **Cluster 6 (double peaks):** property flat, new buildings after year 2000, it may be occupied by couples with small children. No electric heating, but with air conditioning and electric appliances as the kitchen, drier and dishwasher. The time of waking-up and bring the children to the bed could cause the peaks.
- **Cluster 7 (business):** There is no data available as the initial aim the project was devoted to residential. The technical audits and the on-line info-house are not yet adapted to business.

It was not possible to detect any pattern or clear evidence that could relate some of the features analysed to be the cause of a load profile pattern or another. Issues, such as; the small dimension of the sample studied, the number of missing values ("NA") in it, and the fact the features available were not the most suitable (adding i.e. type of employment, people home during the day…); made not possible to reach the desired output when analysing the features in each load profile cluster.

Most of the home appliances features are common to each cluster, with air conditioning, a non-electric heating system, electric kitchen, drier, dishwasher, and not available data for the domestic hot water system. Also it does not exist any significant difference among the house properties and the sociological features in the different clusters, in order to be able to affirm that one or some specific features are the cause of the load profile pattern.

Remark that, the scope of analysing the features for each cluster would have been more adequate if a larger sample with better data quality and more specific features were available. So, the current process could not add any value to the project; but it is a procedure that could be followed when a larger dataset would be available.

11. Final Conclusions

Energy efficiency

Energy management tools are not extended within the residential sector, since the energy consumed is so small compared to business and industries, the savings obtained don't pay-off the investment. However, there is room for providing such tools to residential consumers, for instance by creating a community as Enerbyte is doing with the collaboration of public administrations and electricity utilities.

The project of "Rubí Brilla" is an example of a public administration pioneer initiative to engage the consumers and foster the energy efficiency among them, aiming to provide energy knowledge, understanding and guidance to the user in order to reduce its consumption on average up to 10%. Nonetheless, it was seen that providing the technical consumption data in kWh to the householders has limited influence and effect, so is necessary to go beyond and translate the technical information into call-to-action measures, guiding the user to smart energy choices (like a GPS), have a better response from the householder. Also the need of sub-metering devices to obtain the consumption data limits the scalability, as the cost associated is significant. To be able to reproduce this kind of project in a large-scale the access to the data from the utilities smart meters is necessary.

So, the combination of the smart meter deployment and the big data analytics are called to play an important on the energy sector. The smart meters generate large amounts of raw data that need to be managed and, once analysed can be converted into useful information that benefits both the utility and the consumer, as aim to improve the customer engagement and the quality of the service. Creating new business opportunities, mainly related to data science and data analysis in response to the market needs.

Data quality

Data quality is the key to perform a proper analysis and extract reliable conclusions from it. In this study, the dataset was not the optimal neither quantitatively nor qualitatively. A much larger dataset, for instance counting for thousands of users, would have been more adequate to the purpose of the study than the "Rubí Brilla" small sample studied of only 121 users.

In addition to that, the electrical consumption data from the sub-metering equipment presented some issues that difficult the analysis at some point and demanded a pre-clustering stage for treating and cleaning the data. As the communication between the sub-metering device and the server is done through the Wi-Fi of the consumer, when it is switched-off the consumption data is lost and counted as "0" like if it had been no consumption.

So, when the possibility to access to the hourly smart meter consumption data becomes a reality, the current study purposes will find a suitable framework since the numbers of users could be much higher and the data quality should be almost perfect as are the measures from the electricity utility to bill their customers.

Customer segmentation results

The main objective of obtaining consumer segmentation by the similarity of their load profiles was satisfactorily achieved; however complementary objective regarding the features analysis of each cluster couldn't add value to the project. The common framework used to compare and group fairly the different load profiles, represented by calculating the proportion of energy usage per hour in percentage worked well according to the purpose of clustering the consumers by their load profiles.

The 7 groups of consumers (Morning peak, Late night peak, Flat consumers, Evening peaks, Daytime consumers, Double peak, Business) was appropriate in terms of number of groups and size. Limiting the number of clusters to 7 was adequate, as each group represents a different load profile shape, moreover if one of the groups is devoted to the business and the six remaining to the residential users.

Less than 7 groups would have mixed some different load profiles to the same cluster; and, higher than 7 groups would have been almost no difference between some groups' loads profiles. Also the size of each group was appropriate as the smallest cluster has 7 members and the largest 28. Avoiding clusters of 1 or 2 members fact that would have distorted the results.

To obtain the final clusters it was necessary an iterative process, based on computational clusters calculation (using software) and finalized manually by the analyst applying visualization and statistical techniques to be able to find the outliers and reallocate them to a more appropriate group. The clustering techniques used (Hierarchical, K-means, SOM) have given similar outputs, but it was the Hierarchical technique the one that better adapted to outcome sought, moreover it makes use of the dendrogram, which facilitates the clustering visualization and the partitioning possibilities.

Methodologies comparison

The methodology used in this project is simplified compared to the procedures of the referenced papers, as larger datasets and more complex parameters are used. (Ardakanian, et al., 2014), (McDonald, et al., 2014), (Kavousian, et al., 2015) normalize the electricity consumption considering the weather and seasonal effects, although for the present study there was no temperature data available. This normalization is advisable as it mitigates the electricity consumption correlated to the temperature; however this is more appropriate for aggregated daily consumption. But for instance in (Beckel, et al., 2012), the electricity consumption is not normalized with the temperature.

Most of these methodologies use the absolute values in kWh, as they were more focused on identify the users with higher energy savings potential. But for the segmentation purpose, the decision to use percentages of energy usage fulfils the expectations sought. Providing consumers segments similar to the load profile archetypes from the Opower study (Shilts & Fischer, 2014) basis of the current study (i.e. Evening peaks, double peaks, late night peaks, flat consumers).

There are some similarities among the methodologies, mainly regarding the pre-clustering phase of cleaning the data, and also regarding the clustering phase as the algorithms and techniques used are the most common ones (Hierarchical, K-means, SOM), although some are more complex as in (Kwac, et al., 2014).

Further steps

Use and test this methodology to a larger and complete dataset, for example using the utility smart meter data. More advanced clustering techniques could be used in order to improve some stages of the procedure, together with a deeper literature review regarding smart meter data treatment and segmentation.

The actual analysis can incorporate additional figures or processes leading to even more accurate results; such as, the use of weather normalised electricity consumption data, use more features than only the 24h vector to cluster, eliminate outliers when finding the mean load profile of a user, perform the same study using the weekends profiles, also, apply other and more advanced clustering techniques.

However, it was seen that hourly data was enough for the purpose of clustering load profiles, but if what is sought is to determine the cause of the curve shape and the peaks hourly data won't be enough. So that, it will make necessary to use 5 minutes electricity consumption data and detailed appliances and features data.

Personalised Service

Consumer segmentation allows the utility to understand individual and groups of consumers offering tailored services, as not all the consumers are equal, so each of them needs different personalised energy-savings recommendations at a different time of the day. By providing that, this will improve the effectiveness of energy efficiency programs and also it will have a positive effect on the energy reduction recommendations. As well as, improving the effectiveness of the demand response programs aiming to shift the time of energy usage to avoid peaks or reduce the peaks duration.

Personally

At a personal level, it was really challenging to perform such project in a relative short time period, furthermore it was my first experience using programming languages in this case "R" and also regarding the usage of clustering techniques and the correspondent algorithms. So, a lot of time was invested in getting familiar with them, in addition as all the process and knowledge acquired was self-learning, I felt a bit lost at some stages and without a methodology to follow when running such analysis.

I was bit disappointed of the limitations and difficulties the data available provoked, but I'm proud of the methodology, the process, the work done, the experience I gained, the results, and the possibility to apply it to a larger dataset. And I'm sure that the project quality would have improved if I had with more time and support.

12. Bibliography

Ardakanian, O. et al., 2014. *Workshop Proceedings of the EDBT/ICDT 2014 Joint Conference on CEUR-WS.org.*

Armaroli, N. & Balzani, V., 2011. Towards an electrcity-powered world. *Energy Environmental Science,* pp. 4, 3193-3222.

Beckel, C., Sadamori, L. & Santini, S., 2012. Towards automatic classification of private households using electricity consumption data. *Embedded Sensing Systems for Energy-Efficiency in Buildings: Proceedings of the Fourth ACM Workshop, (BuildSys '12),* pp. pp.169-176.

BloomEnergy, 2015. *Fuel Cell: Distributed Generation.* [Online]
Available at: http://www.bloomenergy.com/fuel-cell/distributed-generation/

Bouzarovski, S., 2014. Energy poverty in the European Union: landscapes of vulnerability. *WIREs Energy Environ,* p. 3: 276–289. doi: 10.1002/wene.89.

Chicco , G. & Ilie, I., 2009. Support vector clustering of electrical load pattern data. *IEEE Trans. Power Syst,* 24(3), pp. 1619-28.

Chicco , G. et al., 2003. Customer characterisation options for improving the tariff offer. *IEEE Trans. Power Syst,* 18(1), pp. 381-7.

Chicco , G. et al., 2005. Emergent electricity customer classification. *IEE Proc Gener Transm Distrib,* 152(2), pp. 164-72.

Chicco , G. et al., 2004. Load pattern-based classification of electricity customers. *IEEE Trans. Power Syst,* 19(2), pp. 1232-9.

Chicco, G., 2012. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy,* June, Volume 42, p. pp. 68–80.

Chicco, G., Napoli , R. & Piglione, F., 2006. Comparisons among clustering techniques for electricity customer classification. *IEEE Trans. Power Syst,* 21(2), pp. 933-40.

Chicco, G. & Sumaili Akilimali , J., 2010. Renyi entropy-based classification of daily electrical load patterns.. *IET Generation Transm Distribution,* Volume 4, pp. 736-45.

Cleveland, C. J. & Najam, A., 2008. *Energy and sustainable development at global environmental summits.* [Online]
Available at: http://www.eoearth.org/view/article/152457/

Corless, P., 2005. Analysis of top 40 largest national economies (GDP per capita vs Energy Efficiency). 30 September.

Cost-benefit analyses & state of play of smart metering deployment in the EU-27, 2014. *Benchmarking smart metering deployment in the EU-27 with a focus on electricity.* [Online]
Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014SC0189&from=EN

Country fiches for electricity smart metering , 2014. *Benchmarking smart metering deployment in the EU-27 with a focus on electricity.* [Online]
Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014SC0188&from=EN

Coursera, 2014. *Data Science Course.* [Online]
Available at: https://www.coursera.org/specialization/jhudatascience/1

Covenant of Mayors, 2015. *Commited to local sustainable energy.* [Online]
Available at: http://www.covenantofmayors.eu/

Directive 2012/27/EU, 2012. *Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency.* [Online]
Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1399375464230&uri=CELEX:32012L0027

Enerbyte Smart Energy Solutions, 2014. [Online]
Available at: http://www.enerbyte.com/

Energy@Home, 2010. [Online]
Available at: http://www.energy-home.it/

EPRI, 2007. *Advanced Metering Infrastructure (AMI).* [Online]
Available at: https://www.ferc.gov/EventCalendar/Files/20070423091846-EPRI%20-%20Advanced%20Metering.pdf

European Commission: Consumer rights and protection, 2015. *Energy: Markets and consumers.* [Online]
Available at: http://ec.europa.eu/energy/en/topics/markets-and-consumers/consumer-rights-and-protection

European Commission: Distributed Generation, 2015. *Energy Research.* [Online]
Available at: http://ec.europa.eu/research/energy/print.cfm?file=/comm/research/energy/nn/nn_rt/nn_rt_dg/article_1158_en.htm

European Commission: Energy Efficiency, 2015. *Buildings.* [Online]
Available at: http://ec.europa.eu/energy/en/topics/energy-efficiency/buildings

European Commission: Energy Supply and Demand, 2014. [Online]
Available at: http://ec.europa.eu/research/energy/print.cfm?file=/comm/research/energy/gp/gp_imp/article_1082_en.htm

European Commission: Energy Union, 2015. *Energy Union: secure, sustainable, competitive, affordable energy for every European.* [Online]
Available at: http://europa.eu/rapid/press-release_IP-15-4497_en.htm#_ftn1

European Commission: Imports and Secure Supplies, 2015. *Energy.* [Online]
Available at: http://ec.europa.eu/energy/en/topics/imports-and-secure-supplies

European Commission: Smart grids and meters, 2014. *Markets and consumers.* [Online]
Available at: https://ec.europa.eu/energy/en/topics/markets-and-consumers/smart-grids-and-meters

European Environment Agency, 2012. *Energy losses and energy availability for end users in 2008 (% of primary energy consumption).* [Online]
Available at: http://www.eea.europa.eu/data-and-maps/figures/energy-losses-and-energy-availability-1

Eurostat, 2014. *Electricity prices for household consumers, first half 2013.* [Online]
Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/File:Electricity_prices_for_household_consumers,_first_half_2013_(1)_(EUR_per_kWh)_YB14.png

Evans, J., 2007. *Article: Asset Management for AMI.* [Online]
Available at: http://www.electricenergyonline.com/show_article.php?mag=45&article=337

Fonseca, P. et al., 2009. Characterization of the household electricity consumption in the EU, potential energy savings and specific policy recommendations. *ECEEE 2009 SUMMER STUDY.*

Gerbec , D., Gasperic , S., Simon , I. & Gubina, F., 2005. Allocation of the load profiles to consumers using probabilistic neural networks. *IEEE Trans. Power Syst,* 20(2), pp. 548-55.

Gerbec, D., Gasperic, S., Simon, I. & Gubina, F., 2004. Determining the load profiles of consumers based on fuzzy logic and probability neural networks. *IEE ProcGener Transm Distrib,* Volume 151, pp. 395-400.

IEA: Energy poverty, 2015. *Topics: Energy Poverty.* [Online]
Available at: http://www.iea.org/topics/energypoverty/

IEA: Smart grids, 2011. *Technology Roadmap: Smart Grids.* [Online]
Available at: https://www.iea.org/publications/freepublications/publication/smartgrids_roadmap.pdf
[Accessed 9 April 2015].

Kavousian, A., Rajagopal, R. & Fischer, M., 2015. Ranking appliance energy efficiency in households: Utilizing smartmeter data and energy efficiency frontiers to estimate and identify thedeterminants of appliance energy efficiency in residential buildings. *Energy and Buildings,* Volume 99, pp. 220-230.

Kwac, J., Flora, J. & Rajagopal, R., 2014. Household energy consumption segmentation using hourly data. *IEEE,* 5(1), pp. 420-430.

Lacey, S., 2013. *The US Smart Meter Market Is Far From Saturated.* [Online]
Available at: http://www.greentechmedia.com/articles/read/smart-meter-penetration

Marques, D. et al., 2004. A comparative analysis of neural and fuzzy cluster techniques applied to the characterization of electric load in substations.. *Proc. IEEE/PES Transmission and Distribution Conference and Exposition,* November.pp. 908-13.

McDonald, B., Pudney, P. & Rong, J., 2014. Pattern recognition and segmentation of smart meter data. *ANZIAM,* Issue 54, pp. 105-150.

Navigant Research, 2014. *Smart Meters.* [Online]
Available at: http://www.navigantresearch.com/research/smart-meters

Nazarko, J., Jurczuk , A. & Zalewski, W., 2005. ARIMA models in load modelling with clustering approach. *Proc. IEEE power Tech, St. Petersburg, Russia.*

NCCS, 2011. *Smart Grid Technology Primer: a Summary.* [Online]
Available at: https://www.nccs.gov.sg/sites/nccs/files/Smart%20Grid%20Primer.pdf

NEST, 2015. [Online]
Available at: https://nest.com/

NIST, 2013. *NIST Smart Grid Conceptual Model.* [Online]
Available at: http://www.sgiclearinghouse.org/ConceptualModel

Opower, 2015. [Online]
Available at: http://www.opower.com/

Pasternak, A., 2000. Global energy Futures and Human Development: A framework for analysis. *Lawrence Livermore National Laboratory,* October.

Ramireddy, V., 2012. *An overview of smart power grid.* [Online]
Available at: http://electrical-engineering-portal.com/an-overview-of-smart-power-grid

Räsänen, T. et al., 2010. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data.. *Appl Energy,* 87(1), pp. 3538-45.

Red Eléctrica de España, 2015. *Red Eléctrica de España (REE).* [Online]
Available at: http://www.esios.ree.es/web-publica/pvpc/
[Accessed 1 June 2015].

Scottish Sceptic, 2013. *Enerconics: The Relationship between Energy and GDP.* [Online]
Available at: http://scottishsceptic.co.uk/2013/10/18/enerconics-the-relationship-between-energy-and-gdp/

Shilts, E. & Fischer, B., 2014. *Opower labs: We plotted 812,000 energy usage curves on top of each other..* [Online]
Available at: http://blog.opower.com/2014/10/load-curve-archetypes/]

Smart Regions, 2013. *Smart Regions.* [Online]
Available at: www.smartregions.net

Suppliers Obligations & White Certificates, 2015. *Institute for Energy and Transport (IET).* [Online]
Available at: http://iet.jrc.ec.europa.eu/energyefficiency/white-certificates

The Economist, 2011. *Articles: Energy intensity is converging across the world.* [Online]
Available at: http://www.economist.com/blogs/dailychart/2011/01/energy_use

The Economist, 2015. *Energy Efficiency: Invisible Fuel.* [Online]
Available at: http://www.economist.com/news/special-report/21639016-biggest-innovation-energy-go-without-invisible-fuel

Tsekouras, G., Hatziargyriou, N. & Dialynas , E., 2007. Two-stage pattern recognition of load curves for classification of electricity. Volume 22 (3), pp. 1120-8.

Tverberg, G., 2011. *Article: Is it really possible to decouple GDP Growth from energy Growth?.* [Online]
Available at: http://ourfiniteworld.com/2011/11/15/is-it-really-possible-to-decouple-gdp-growth-from-energy-growth/

U.S. Department of Energy, 2015. *Office of Electricity Delivery and Energy Reliability.* [Online]
Available at: https://www.smartgrid.gov/the_smart_grid/

Valero , S. et al., 2007. Methods for customer and demand response policies selection in new electricity markets. *IET Generation, Transm Distribution,* 1(1), pp. 104-10.

Verdu , S. et al., 2006. Classification, filtering, and identification of electrical customer load patterns through the use of self organizing maps. *EEE Trans. Power Syst,* Volume 21, pp. 1672-82.

Wikibooks:SOM, 2015. *Wikibooks: Data Mining Algorithms In R/Clustering/Self-Organizing Maps (SOM).* [Online]
Available at: http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Self-Organizing_Maps_%28SOM%29
[Accessed 9th April 2015].

Wikibooks, 2015. *Wikibooks: Data Mining Algorithms In R/Clustering/K-Means.* [Online]
Available at: http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/K-Means
[Accessed 9th April 2015].

Yu , I., Lee , J., Ko , J. & Kim, S., 2005. A method for classification of electricity demands using load profile data. *Proc. Fourth Annual ACIS Intern. Conf Comput Inf Sci,* pp. 164-8.

13. Appendices

Appendix A: Cleaning data

To remove "0" values the "apply" function is used to go to every row and detect the "0" observations per row, and eliminate those rows that have at least on "0". To remove the rows that contain frozen values, the procedure followed for every row was to subtract to each value its previous one, this implies that the result of subtracting two repeated values will be "0", thus detecting the number of "0" values per row will therefore allow us to detect the "frozen" values.

Piece of the code used in "R" to remove the rows that contain at least one "0" and "frozen" values

```
…

##  Remove rows with at least one "0" value
rubi_cut<-rubi[3:26]
      rubi_zero <- rubi[apply(rubi_cut==0, 1, sum)<=0, ]


## Remove rows with at least 2 consecutives same values
x<-rubi_zero[3:26]
   y<-x[-1]
     diff <- y-x[1:length(x)-1]
        rubi_net<-rubi_zero[apply(diff==0,1,sum)<=2,]
…
```

Appendix B: Load profiles visualization

To obtain the proportion energy usage per hour of each consumer, the steps are done:

- First find the total value of each row by adding the 24 hours values, using the "rowSums" function
- Second, divided each hour value by the total obtained
- Third, find the mean of all the values for each consumer, using "colMeans" for each consumer.

Piece of the code used to calculate the Proportion of energy usage per hour for every customer observation

```
…

##1. Get the sum of each row
  rubi_sum<-rubi_net[3:26]
       row_sum<-as.matrix(rowSums(rubi_sum))
          names(row_sum)<-"sum"

##2. Division to get the percentages per hour
  division<-as.data.frame(rubi_sum/row_sum)
      division$idmeter<-NULL
         division_id<-cbind(rubi_net$idmeter,division)
           names(division_id)<-c("idmeter",c(0:23))
```

```
                division_id<-as.data.frame(division_id)

##3. Column means
    cast_99<-as.data.frame(lapply(split(division_id, division_id$idmeter),colMeans))
      cast100<-as.data.frame(t(cast_99))
          cast100$idmeter<-NULL

     hour_percent<-cast100

…
```

The dates' column is converted into a "xts" object, which allows to assign a number 1 to 7 to the dates where 1 is Monday and 7 is Sunday. Later on, the weekdays are selected by selecting "1:5", and the weekend days by selecting dates that have di

Piece of the code used to separate the Weekdays and Weekends in the data set

```
…

## Separate Weekday and Weekends
library(xts)
rubi2<-as.xts(rubi_net,rubi_net$date)

## Weekdays
weekdays<-rubi2[.indexwday(rubi2)  %in%  1:5]  #labels=c("Monday","Tuesday","Wednesday",
"Thursday", "Friday")
w_days<-as.data.frame(dates=index(weekdays), coredata(weekdays))

## Weekends
weekendss<-rubi2[!.indexwday(rubi2) %in% 1:5]        #labels=c("Saturday","Sunday")
w_ends<-as.data.frame(dates=index(weekends), coredata(weekends))

…
```

To plot the load profile the "R" library "ggplot2" is used, where the function "ggplot" allows to plot the load profiles in different formats (points, lines, bars…) and add colours, labels etc…

Piece of the code used to plot the load profiles curves in R

```
…

 library(ggplot2)
ggplot(rubi_net_perc,aes(hour, percent,group=1))+
geom_bar(width=0.9,aes(fill=percent),stat = "identity")+
 labs(title="LOADPROFILE",x="Hour",y="Proportion[%]")+
scale_fill_gradientn(colours=c("green","orange","red"))+
scale_x_continuous(breaks=c(seq(0,23,by=1)))+scale_y_continuous(breaks=c(seq(0,0.15,by=0.01)))
```

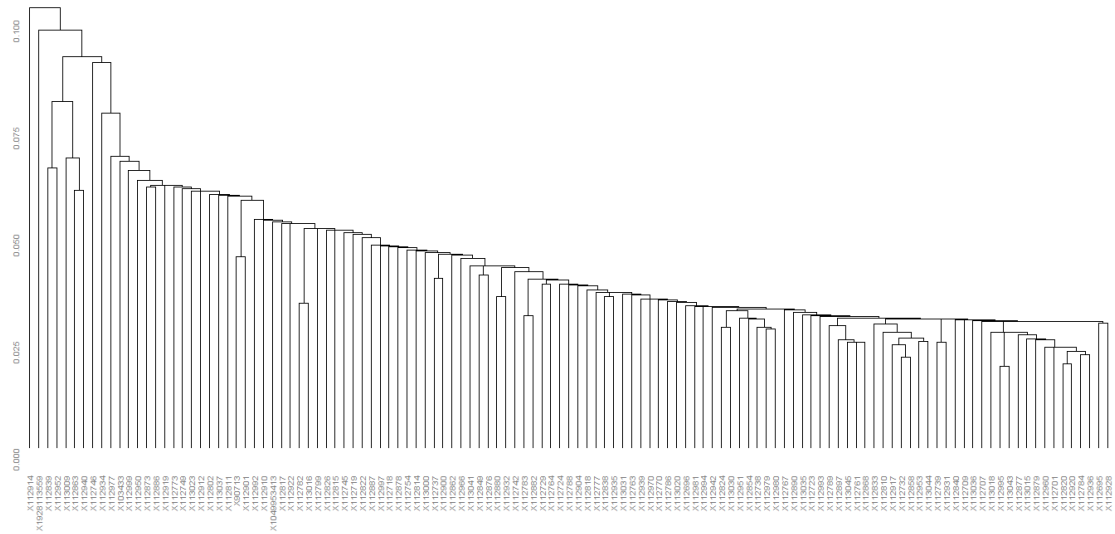
The function "read_dt_w" allows to read the ".csv" file with the data of the idmeter selected, in the example number 90713. Then the starting and ending dates are selected, and finally the "aggregate" function permits to sum all data values.

Piece of the code to select the idmeter, date and aggregation data in R

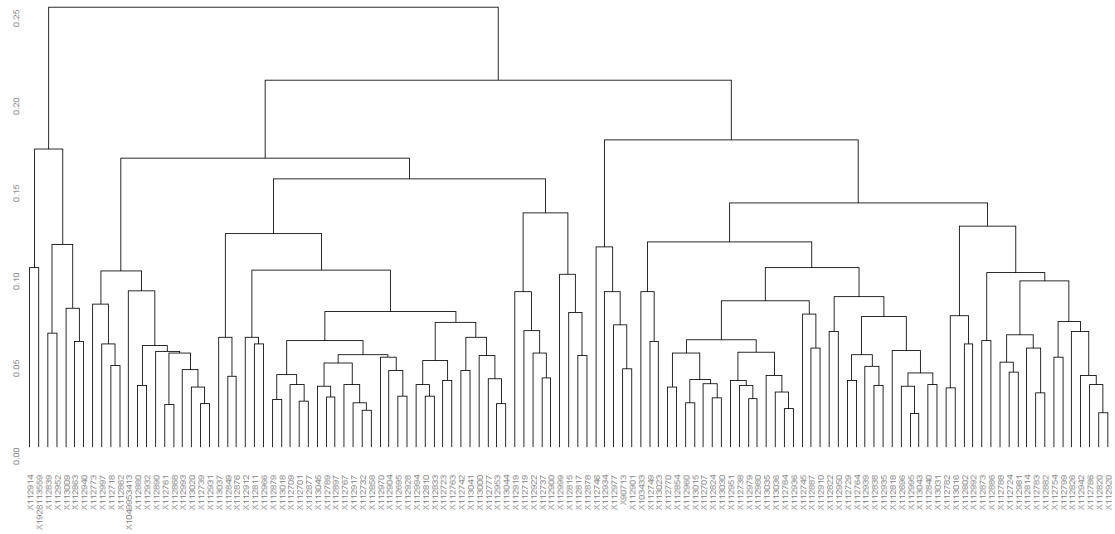
```
…
## select the idmeter number
rubi_net<-read_dt_w(90713)

##select the dates
mon1<-rubi_net['2014-04::2015-02']

## data aggregation using sum function
added<-aggregate( cbind( power ) ~ day + month + year , data = mon1 , FUN = sum )
…
```

Appendix C: Dendrograms comparison

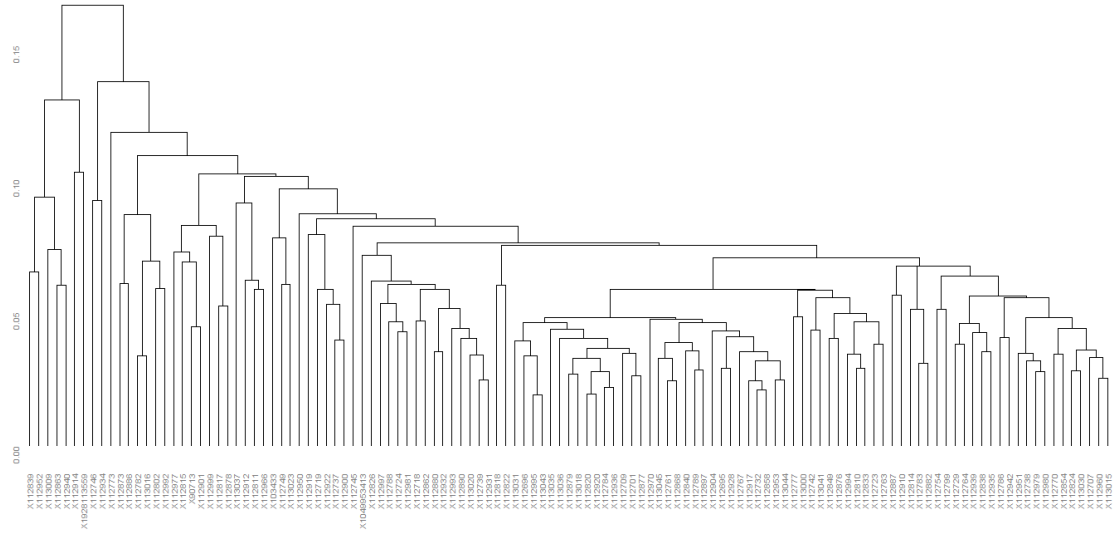
C.1. Euclidean distance and single linkage dendrogram



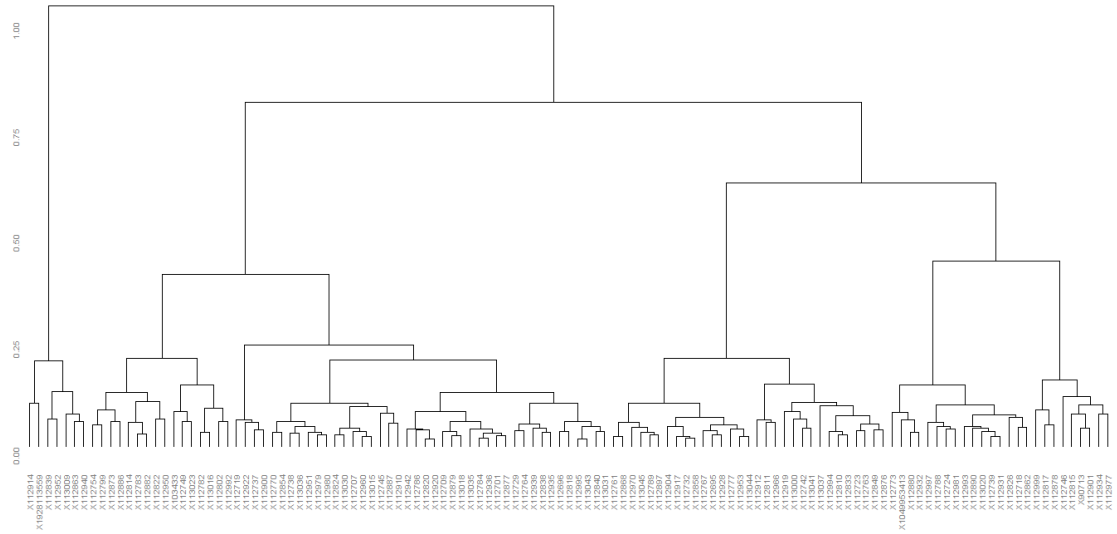
C.2. Euclidean distance and complete linkage dendrogram



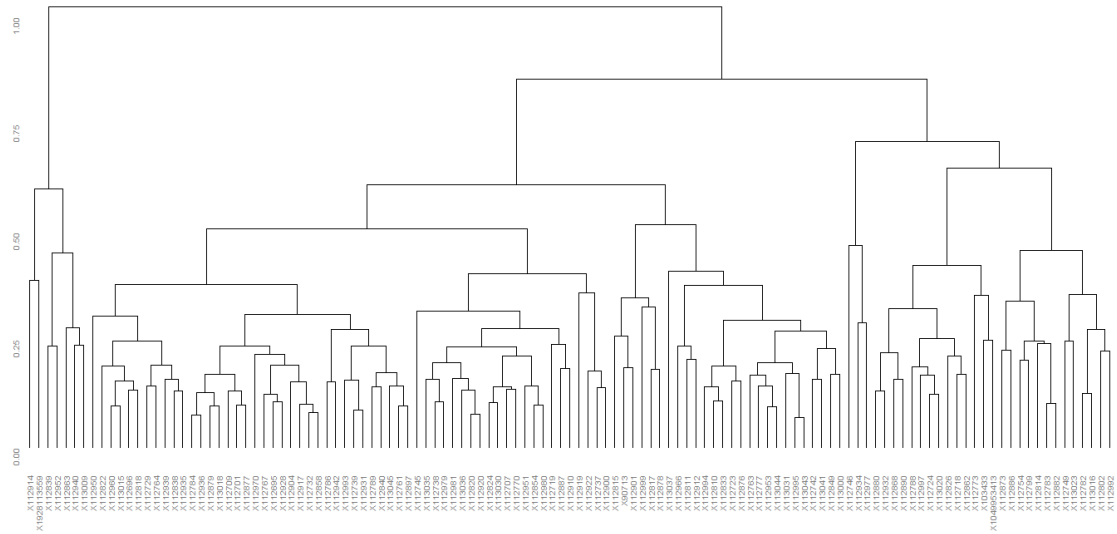
C.3. Euclidean distance and average (UPGMA) linkage dendrogram



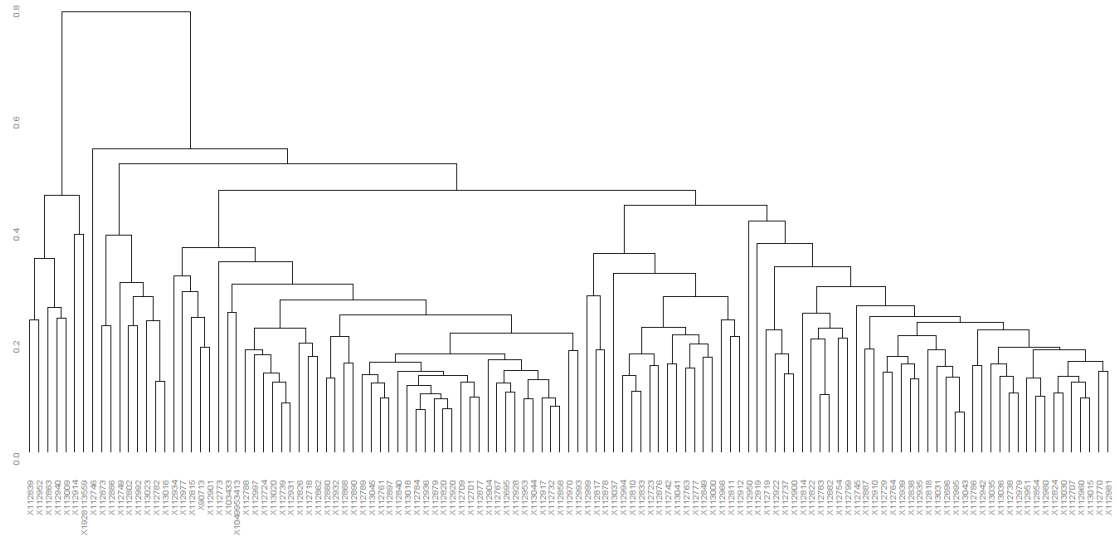
C.4. Euclidean distance and Ward linkage dendrogram



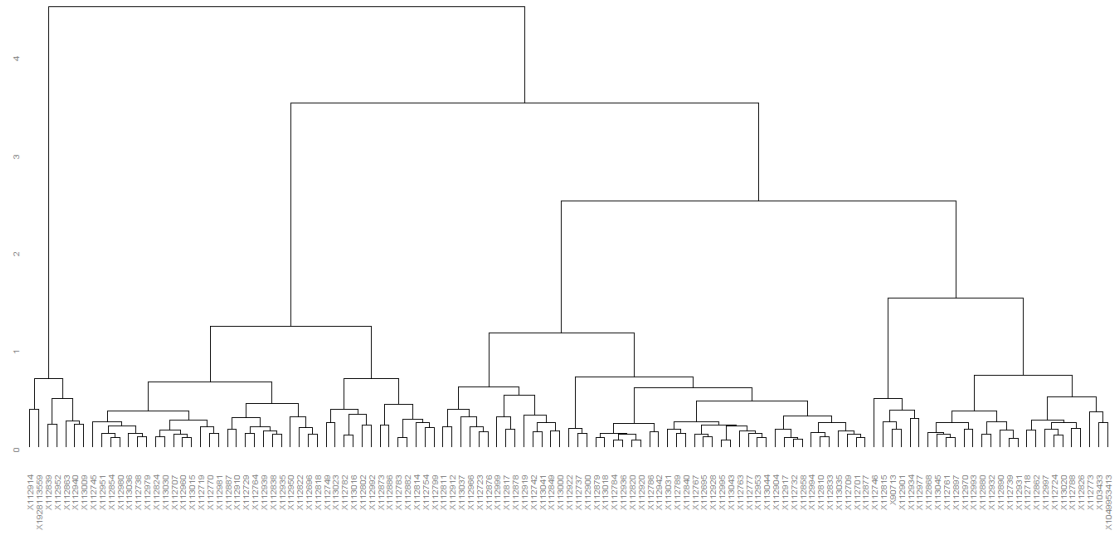
C.5. Manhattan distance and Complete linkage dendrogram



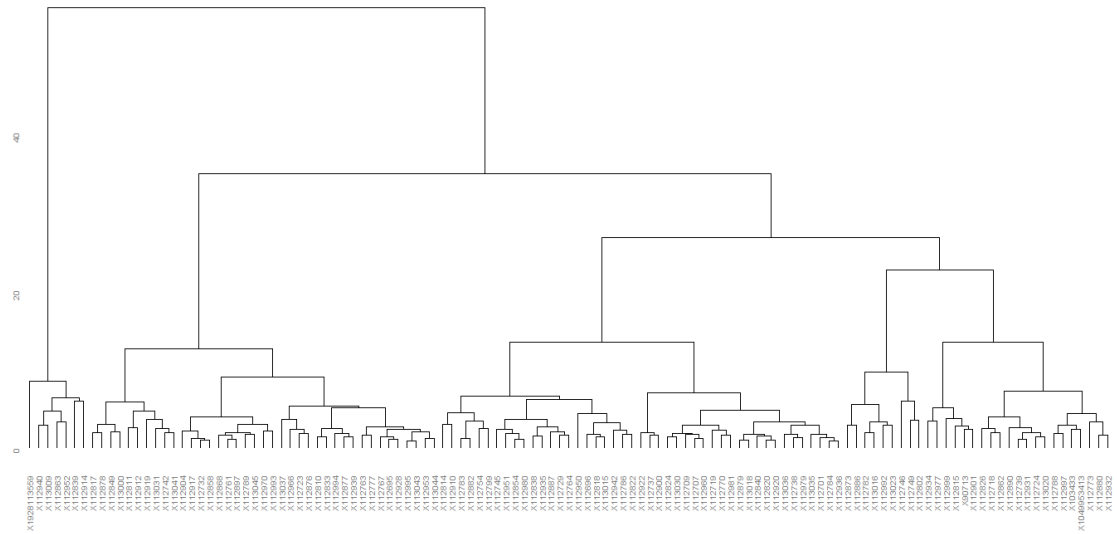
C.6. Manhattan distance and McQuitty linkage dendrogram



C.7. Manhattan distance and Ward linkage dendrogram



C.8. Canberra distance and Ward linkage dendrogram



Appendix D: Hierarchical clustering formation

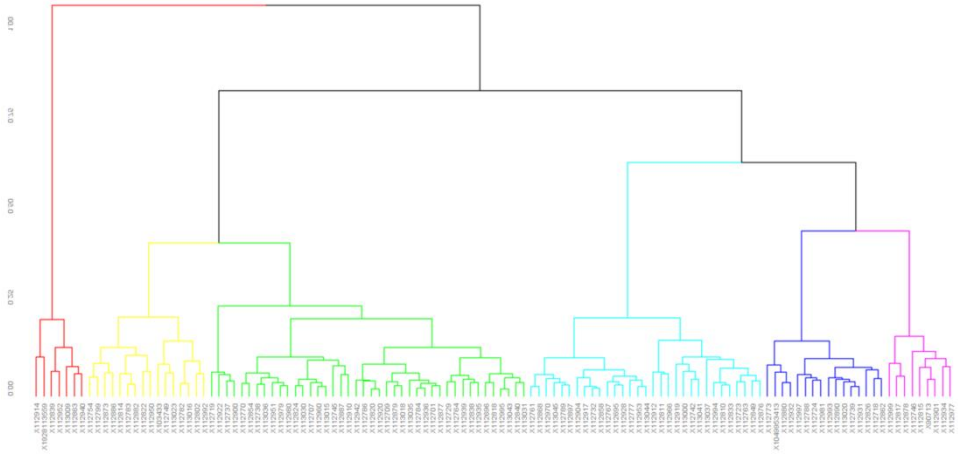
D.1. Euclidean distance and Ward linkage dendrogram

D.1.1. 6 Clusters

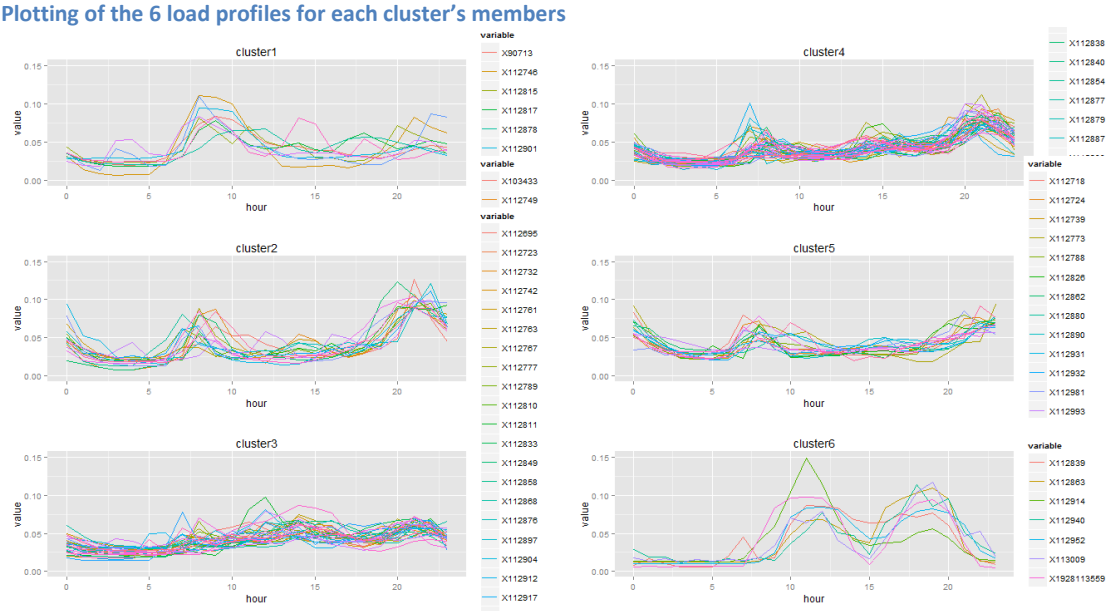
Number of members per each of the 6 clusters

Canberra Distance and Ward Linkage						
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Number of consumers	9	16	31	42	16	7

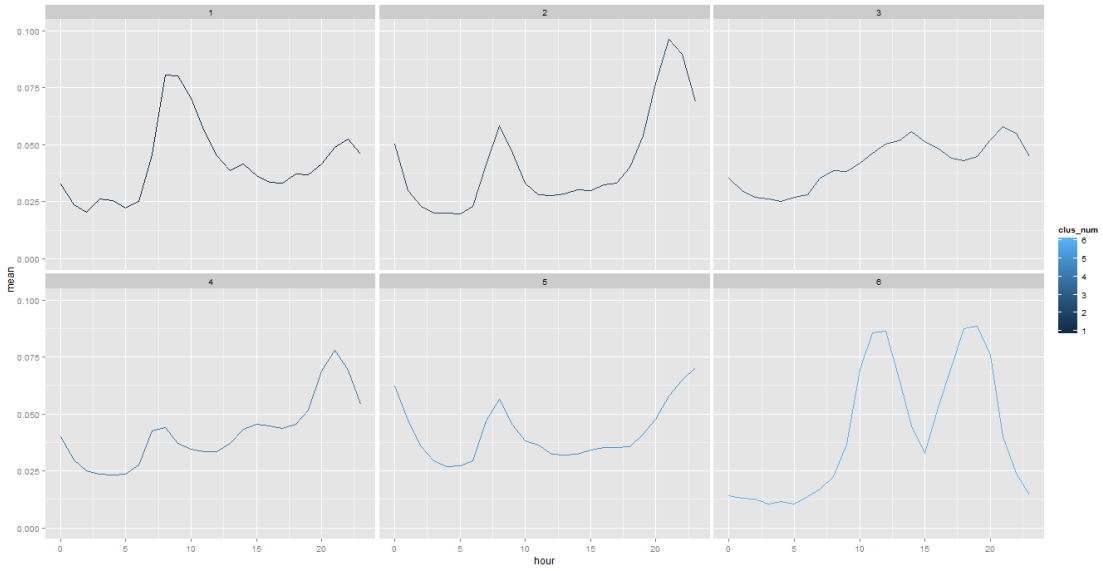
6 Clusters colored in the dendrogram



Plotting of the 6 load profiles for each cluster's members



Mean 6 load profile curves for each cluster obtained

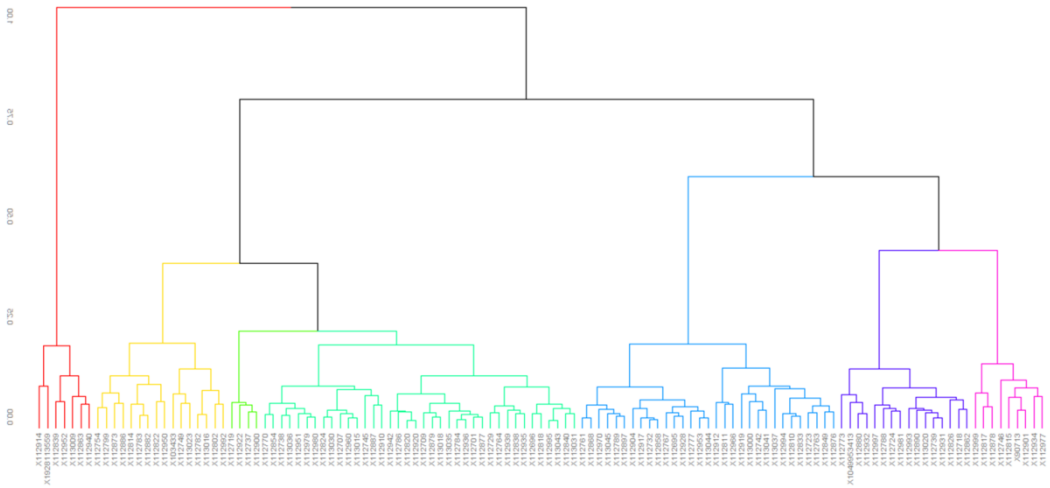


D.1.2. 7 Clusters

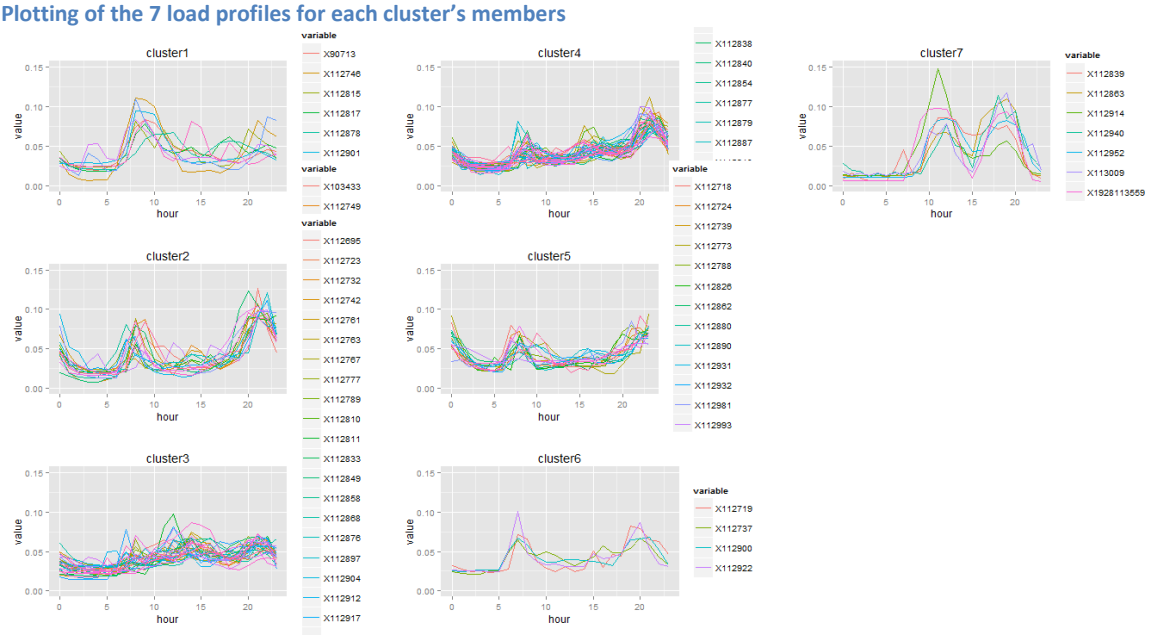
Number of members per each of the 7 clusters

Canberra Distance and Ward Linkage							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Number of consumers	9	16	31	38	16	4	7

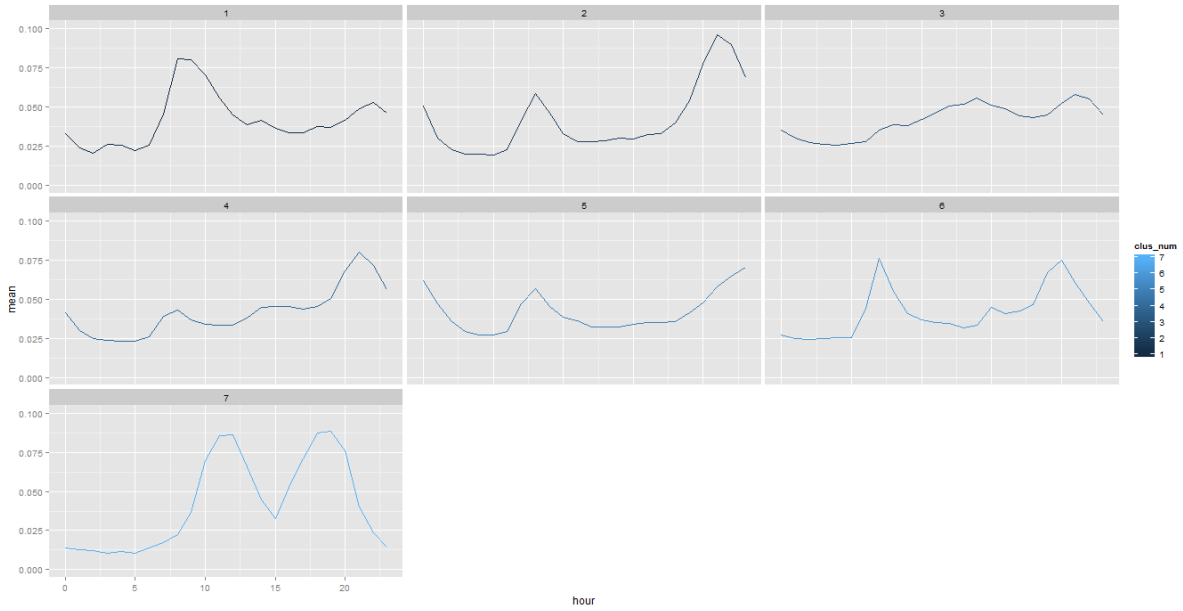
7 Clusters colored in the dendrogram



Plotting of the 7 load profiles for each cluster's members



Mean load profile curve for each cluster obtained

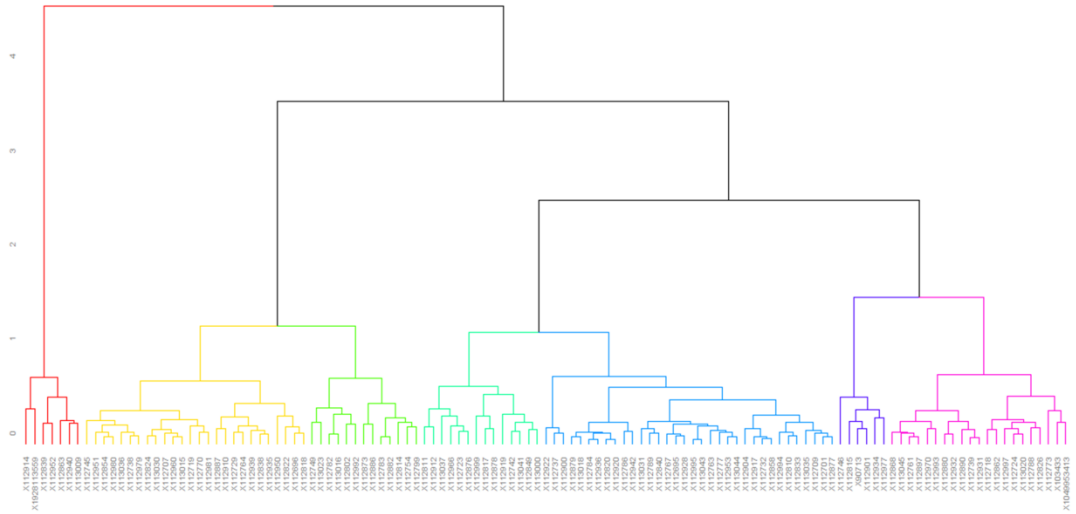


D.2. Manhattan distance and Ward linkage

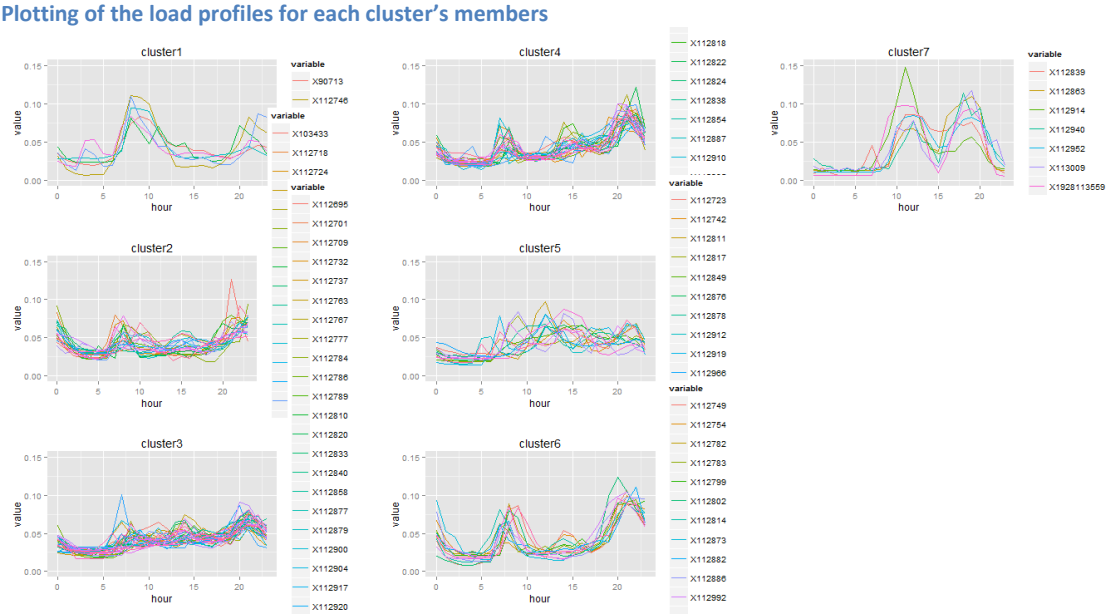
Number of members per each of the 7clusters

Manhattan Distance and Ward Linkage							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Number of consumers	6	21	34	26	14	13	7

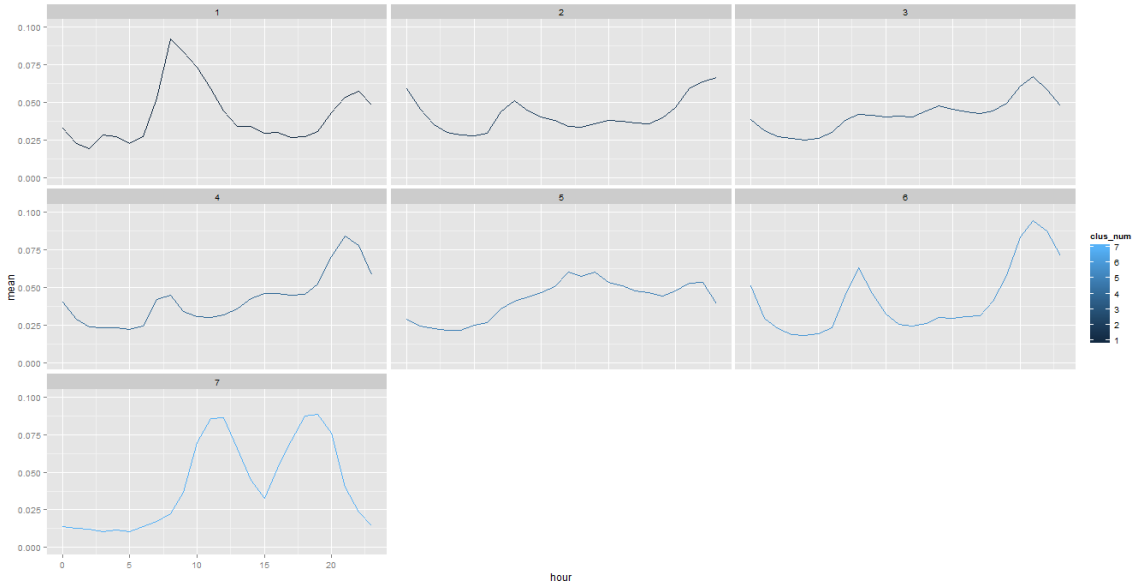
7 Clusters colored in the dendrogram



Plotting of the load profiles for each cluster's members



Mean of the 7 load profile curve for each cluster obtained



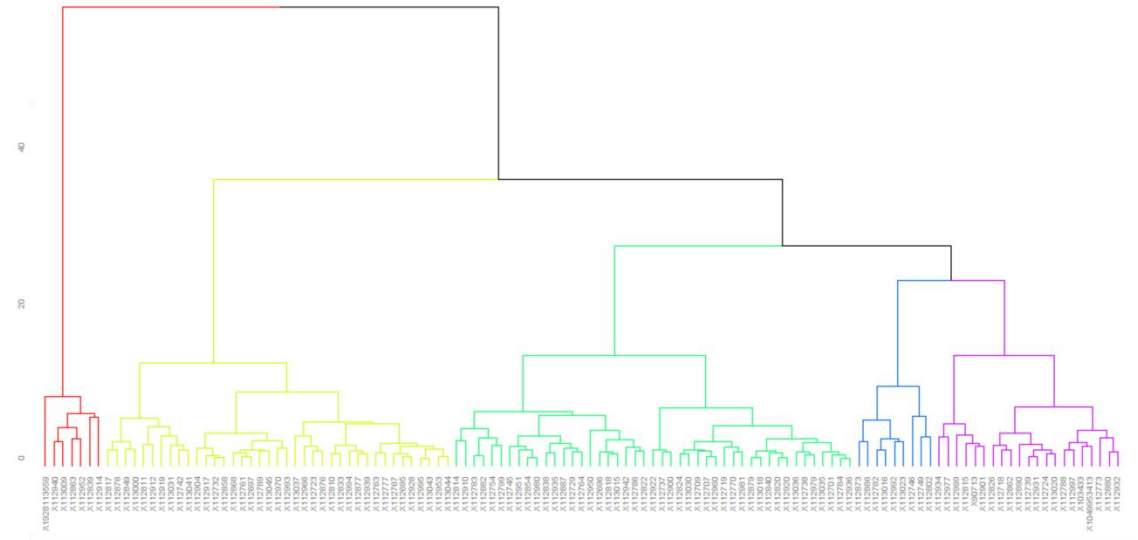
D.3. Canberra distance and Ward linkage

D.3.1. 5 clusters

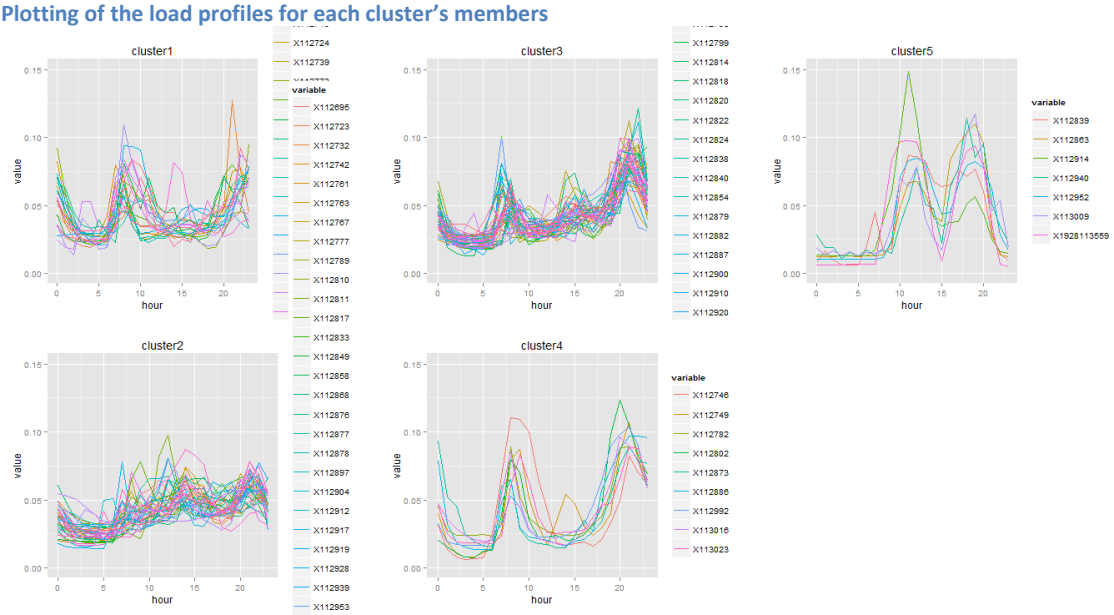
Number of members per each of the 5 clusters

Canberra Distance and Ward Linkage					
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of consumers	21	39	45	9	7

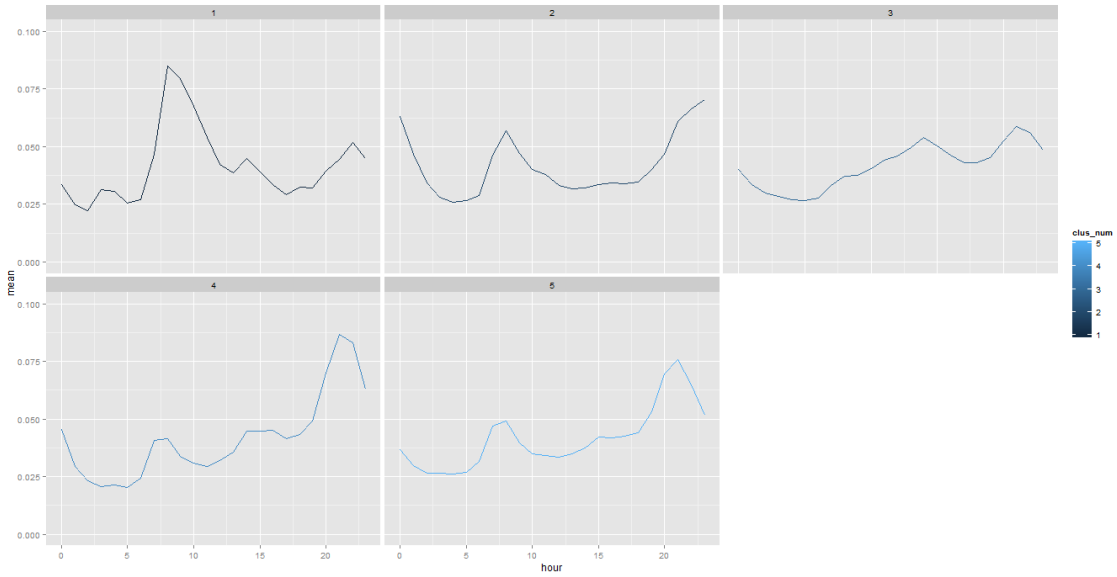
5 Clusters colored in the dendrogram



Plotting of the load profiles for each cluster's members



Mean load profile curve for each cluster obtained

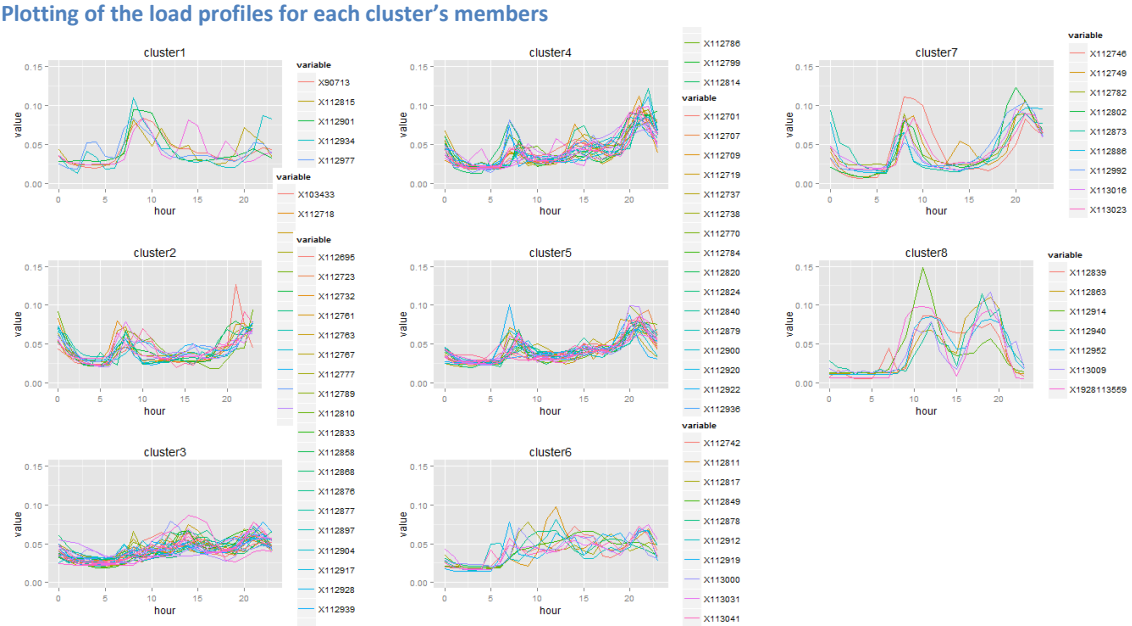


D.3.2. 8 clusters

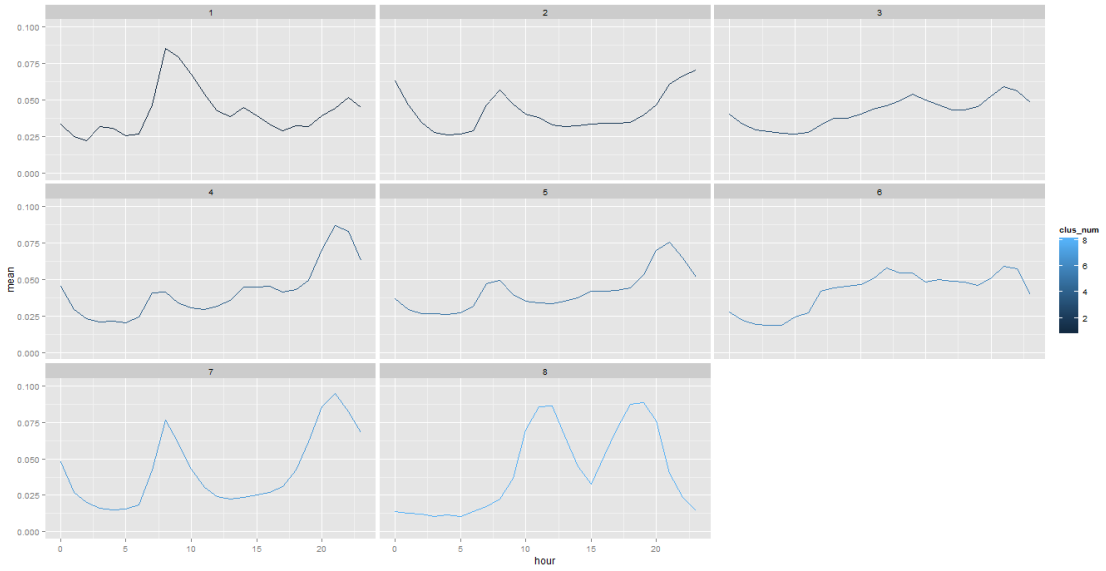
Number of members per each of the 8 clusters

Canberra Distance and Ward Linkage								
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Number of consumers	6	15	29	22	23	10	9	7

Plotting of the load profiles for each cluster's members



Mean load profile curve for each cluster obtained



Appendix E: Hierarchical residual distances calculations

E.1. Distance to the mean

The aggregated distance of each member to the cluster's mean is shown in the table below, it can be seen that the Euclidean aggregated distance is less than the Euclidean aggregated distance; due to the fact that the Euclidean cluster 6 is composed of only 4 members which lead to a small distances aggregation.

Comparison among Euclidean and Manhattan aggregated clusters distances

EUCLIDEAN AND WARD		MANHATTAN AND WARD	
Cluster	**distance**	**Cluster**	**distance**
Cluster 1	0,2255076	**Cluster 1**	0,2086356
Cluster 2	0,2153522	**Cluster 2**	0,1689538
Cluster 3	0,162276	**Cluster 3**	0,1358658
Cluster 4	0,1378701	**Cluster 4**	0,1490822
Cluster 5	0,1577359	**Cluster 5**	0,2034691
Cluster 6	0,1198553	**Cluster 6**	0,2029609
Cluster 7	0,252591	**Cluster 7**	0,252591
AGGREGATEED DISTANCE	1,2711881	AGGREGATEED DISTANCE	1,3215584

E.2. Residuals

The following table shows each cluster member's distance respect to the cluster mean, highlighting the two with higher distances.

Each cluster member's distance to the cluster mean, highlighting the two with higher distances

Cluster 1	dist to Mean 1	Cluster 2	dist to Mean 2	Cluster 3	dist to Mean 3	Cluster 4	dist to Mean 4
X112746	0.3116457	**X112773**	0.31491590	X112922	0.27586741	X112950	0.27174340
X112815	0.2131447	**X103433**	0.27869635	X112942	0.21441096	X112719	0.21464751
X112977	0.2095507	**X112718**	0.20984158	X112737	0.20518847	X112745	0.19827058
X112934	0.2066820	**X1049953413**	0.20724579	X112763	0.19569810	X112818	0.19087939
X90713	0.1566011	**X112826**	0.20162737	X112900	0.18892009	X112910	0.17966412
X112901	0.1541896	**X113045**	0.20129802	X112786	0.17958958	X112822	0.17846375
		X112862	0.18365219	X112833	0.17398367	X112696	0.17825987
		X112970	0.18298311	X112904	0.15706262	X112981	0.17148808
		X112761	0.18106981	X112994	0.15607604	X112935	0.17103526
		X112890	0.17130899	X112777	0.15403618	X112887	0.16034537
		X112788	0.16652927	X112695	0.15278095	X112729	0.15946907
		X112868	0.16356952	X113031	0.14999552	X112939	0.15263649
		X112724	0.15576553	X113035	0.14354081	X112979	0.14445456
		X112997	0.15135792	X112879	0.13516887	X112770	0.13828084
		X112932	0.15011446	X112789	0.13234977	X112951	0.13625597
		X112993	0.14306492	X112920	0.12867692	X112764	0.13583522
		X112880	0.12799036	X113043	0.12480618	X112738	0.13446362
		X112897	0.11783588	X113018	0.12264866	X113036	0.13308756
		X113020	0.10089499	X112928	0.11929259	X112854	0.12157758
		X112931	0.08969590	X112810	0.11825736	X112707	0.11189466
		X112739	0.04857223	X112820	0.11443098	X113015	0.10667367
				X112709	0.11383248	X112980	0.10441407
				X112877	0.11035765	X113030	0.10100852
				X112917	0.10919750	X112838	0.09857586
				X112995	0.10226179	X112824	0.09431463
				X112701	0.10142231	X112960	0.08839804
				X112840	0.10109488		
				X112953	0.09772379		
				X112767	0.09607565		
				X113044	0.09156161		
				X112732	0.09104306		
				X112784	0.08819956		
				X112858	0.08771756		
				X112936	0.08616610		

Cluster 5	dist to Mean 5	Cluster 6	dist to Mean 6	Cluster 7	dist to Mean 7
X112999	0.3104575	X112749	0.2928833	X112914	0.3530615
X113037	0.2591932	X112802	0.2770582	X1928113559	0.2815018
X112919	0.2558509	X112873	0.2467687	X112940	0.2686330
X112817	0.2353625	X112814	0.2270393	X112839	0.2544544
X112912	0.2334895	X112886	0.2139956	X112863	0.2445030
X112811	0.2210057	X112754	0.2096413	X113009	0.2189858
X112966	0.2054016	X112992	0.1956796	X112952	0.1469973
X113041	0.1798340	X113023	0.1891108		
X112878	0.1771708	X112882	0.1747765		
X112742	0.1719250	X112783	0.1689626		
X112876	0.1646843	X113016	0.1584391		
X112723	0.1586258	X112799	0.1472939		
X113000	0.1575425	X112782	0.1368432		
X112849	0.1180248				

Appendix F: Hierarchical functions in "R"

The function "dist()" allows to calculate a matrix with all the distances between all observations, in this case using the Manhattan distance. Then, the function "hclust()" allows to select which linkage method use, in this case Ward linkage. The "ggdendrogram" function plots the matrix into the hierarchical tree. Finally, the "cutree()" function is used to divide the dendrogram in the numbers of clusters indicated, in this case k=7 clusters.

```
…

## Distances: distance is choosen in dist()
p<-dist(hour_percent,method="manhattan")

## Hierarchical clustering method hclust()
      p1<-hclust(p, method="ward.D")

## Dendrogram
library(ggplot2)
    library(ggdendro)
        ggdendrogram(p1, rotate = F, size = 3, labels=T)

## Cluster division by cutree()
      clus <- as.data.frame(cutree((p1), k=7))  ## k indicates number of clusters

…
```

Appendix G: K-Means clustering comparison

G.1. Lloyd algorithm 7 clusters (seed 4)

Number of members per each of the 7 clusters from k-means Lloyd algorithm

Lloyd algorithm cluster's members							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Number of consumers	35	7	41	6	8	8	16

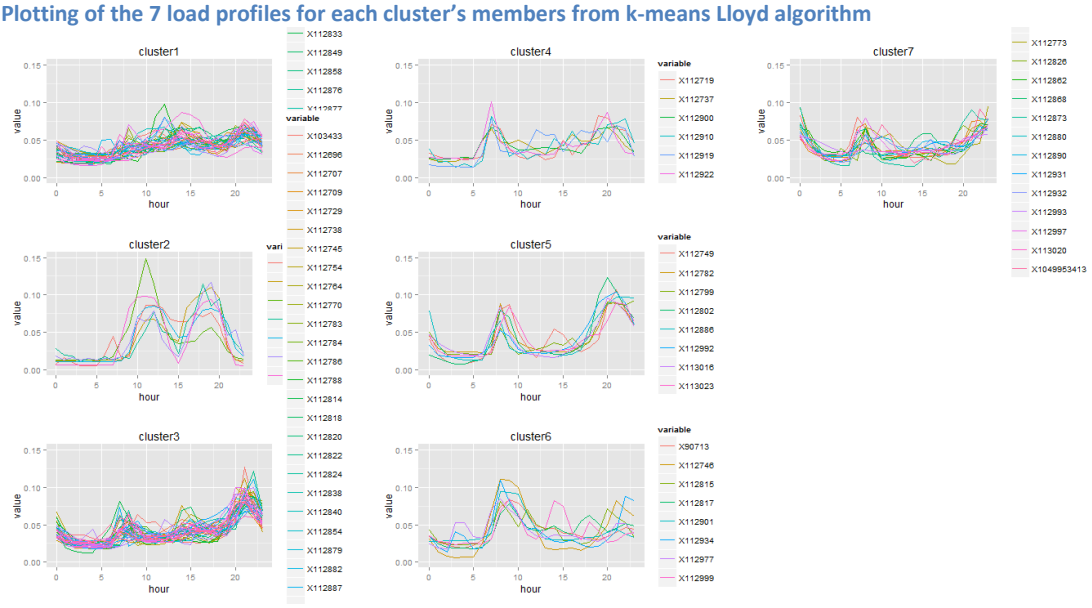
Within clusters sum of squares distance per each of the 7 clusters from k-means Lloyd algorithm

Lloyd algorithm Within cluster sum of squares distance							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Within cluster distance	0.0594	0.0367	0.0751	0.01137	0.02135	0.0288	0.03386

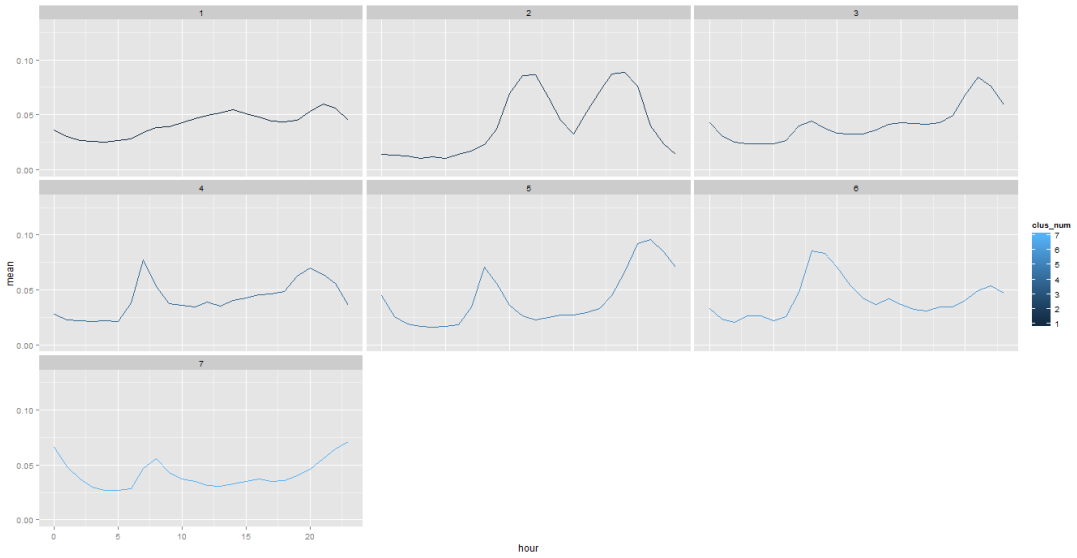
The total within clusters distance is 0.2668178

The distance between clusters is 0.3295435

Plotting of the 7 load profiles for each cluster's members from k-means Lloyd algorithm



Mean 7 load profile curves for each cluster obtained from k-means Lloyd algorithm



G.2. Forgy algorithm 7 clusters (seed 9)

Number of members per each of the 7 clusters from k-means Forgy algorithm

Forgy algorithm cluster's members							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Number of consumers	9	17	8	7	37	8	35

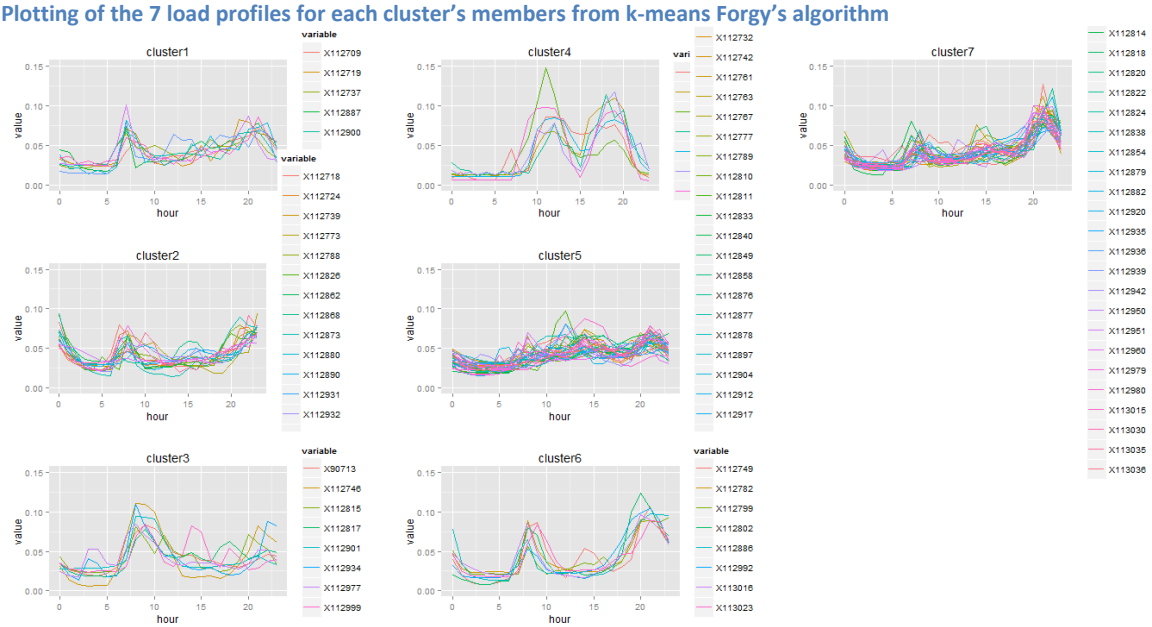
Within clusters sum of squares distance per each of the 7 clusters from k-means Forgy algorithm

Forgy algorithm Within cluster sum of squares distance							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Within cluster distance	0.01736	0.03583	0.02881	0.03677	0.06278	0.02135	0.06401

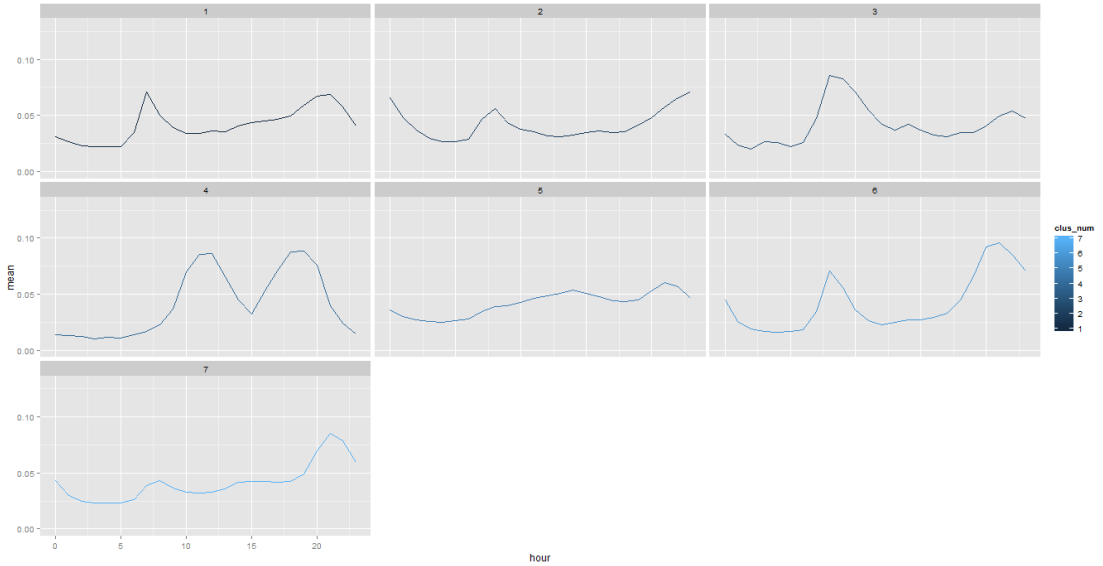
The total within clusters distance is 0.2669388

The distance between clusters is 0.3294226

Plotting of the 7 load profiles for each cluster's members from k-means Forgy's algorithm



Mean 7 load profile curves for each cluster obtained from k-means Forgy's algorithm



G.3. MacQueen algorithm 7 clusters (seed 13)

Number of members per each of the 7 clusters from k-means MacQueen's algorithm

MacQueen algorithm cluster's members							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Number of consumers	19	6	37	9	35	8	7

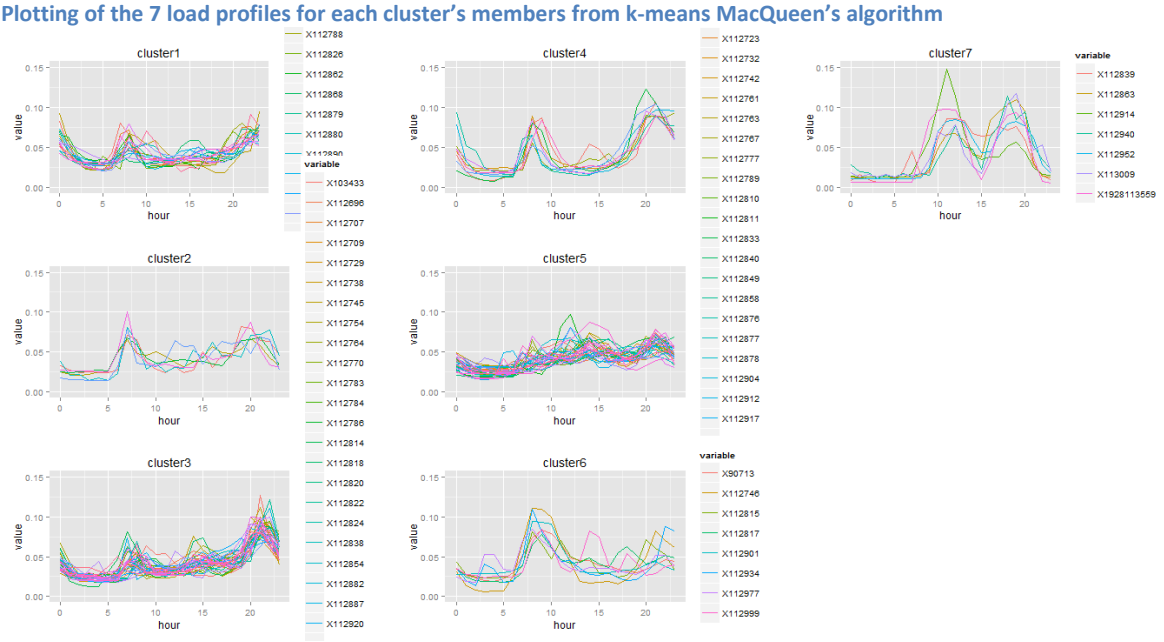
Within clusters sum of squares distance per each of the 7 clusters from k-means MacQueens's algorithm

MacQueen algorithm Within cluster sum of squares distance							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Within cluster distance	0.03419	0.01137	0.06867	0.02726	0.05962	0.02881	0.03677

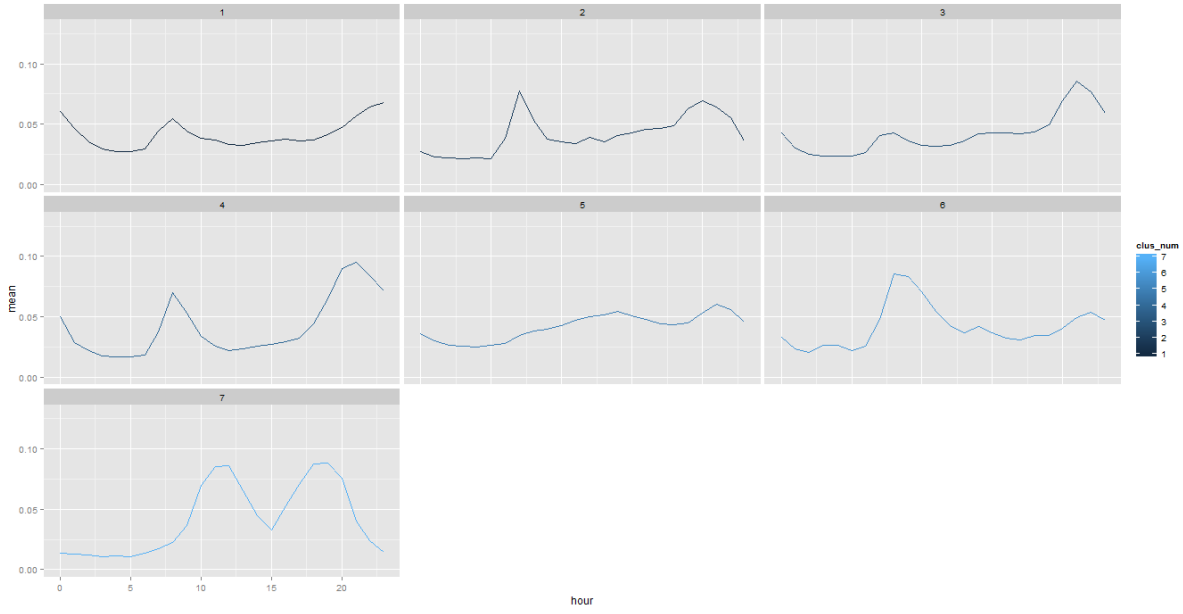
The total within clusters distance is 0.2667249

The distance between clusters is 0.3296364

Plotting of the 7 load profiles for each cluster's members from k-means MacQueen's algorithm



Mean 7 load profile curves for each cluster obtained from k-means MacQueen's algorithm



G.4. Hartigan-Wong algorithm 7 clusters (seed 20)

Number of members per each of the 7 clusters from k-means Hartigan-Wong's algorithm

Hartigan-Wong algorithm cluster's members							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Number of consumers	9	44	35	2	5	8	18

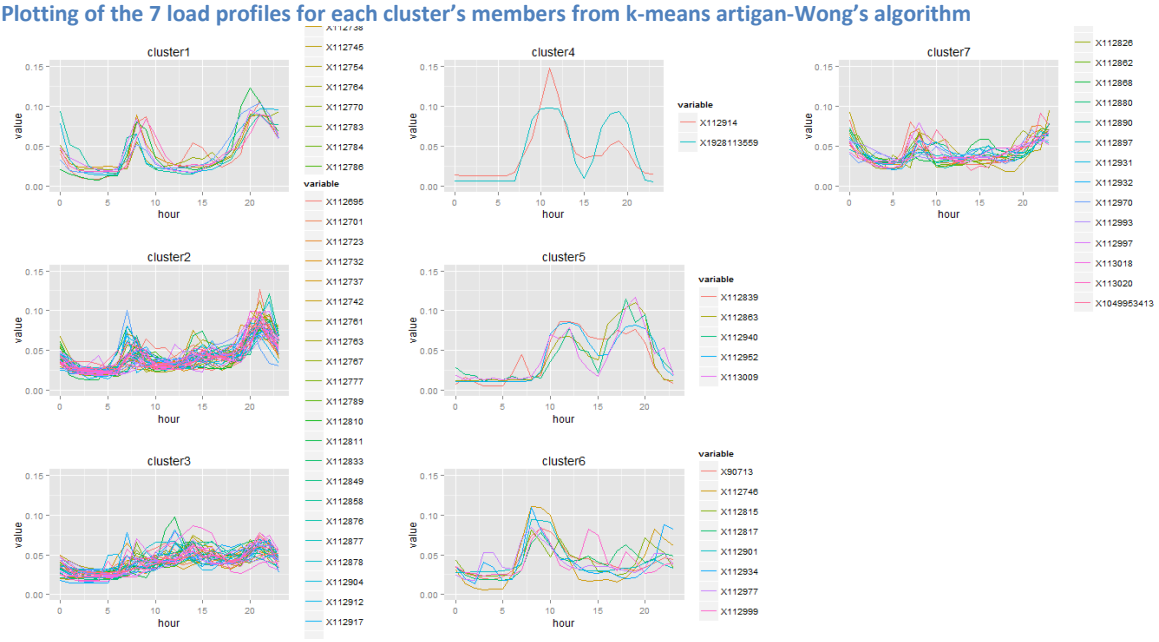
Within clusters sum of squares distance per each of the 7 clusters from k-means Hartigan-Wong's algorithm

Hartigan-Wong algorithm Within cluster sum of squares distance							
Cluster number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Within cluster distance	0.02726	0.09264	0.06395	0.00529	0.01446	0.02881	0.03324

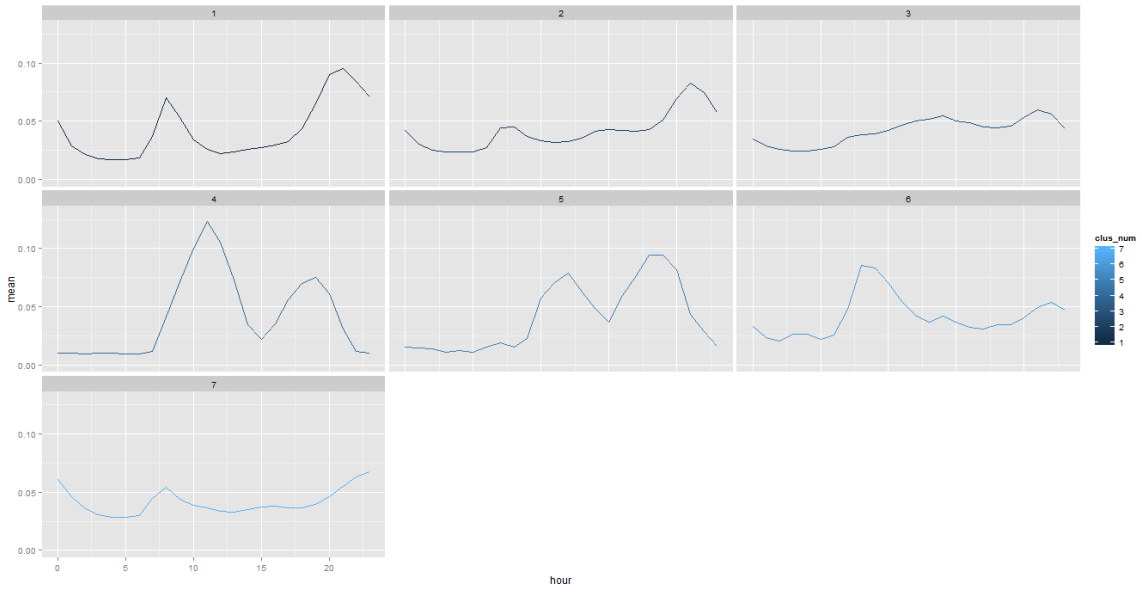
The total within clusters distance is 0.2656916

The distance between clusters is 0.3306698

Plotting of the 7 load profiles for each cluster's members from k-means artigan-Wong's algorithm



Mean 7 load profile curves for each cluster obtained from k-means Hartigan-Wong's algorithm



Appendix H: K-means residual distances calculations

H.1. Distance to the mean

The aggregated distance of each member to the cluster's mean is shown in the table below.

K-means clustering, using Forgy's algorithm distances to the mean of each cluster

K-means, Forgy's algorithm								
Clusters	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	AGGREGATEED DISTANCE
dist	0,1633829	0,1676934	0,2203558	0,252591	0,1527241	0,1841199	0,1506885	1,2915556

H.2. Residuals

Each cluster member's distance to the cluster mean, highlighting the two with higher distances

Cluster 1	dist to Mean 1	Cluster 2	dist to Mean 2	Cluster 3	dist to Mean 3	Cluster 4	dist to Mean 4
X112919	0.2459039	X112873	0.30980746	X112746	0.3677728	X112914	0.3530615
X112910	0.1832258	X112773	0.30472691	X112999	0.2703058	X1928113559	0.2815018
X112922	0.1779059	X1049953413	0.20351252	X112934	0.2430865	X112940	0.2686330
X112887	0.1771052	X112868	0.19096153	X112815	0.2056351	X112839	0.2544544
X112981	0.1762038	X112718	0.17471141	X112977	0.2021639	X112863	0.2445030
X112719	0.1574524	X112932	0.17349957	X112817	0.1851979	X113009	0.2189858
X112737	0.1390845	X112826	0.17049452	X112901	0.1632162	X112952	0.1469973
X112900	0.1226932	X112993	0.15890963	X90713	0.1254682		
X112709	0.0908713	X112788	0.15173132				
		X112862	0.15029183				
		X112880	0.14596942				
		X112724	0.14552932				
		X112890	0.14360521				
		X112997	0.14056858				
		X113020	0.11101639				
		X112931	0.10472562				
		X112739	0.07072581				

Cluster 5	dist to Mean 5	Cluster 5	dist to Mean 5	Cluster 6	dist to Mean 6	Cluster 7	dist to Mean 7	Cluster 7	dist to Mean 7
X113037	0.3366978	X112904	0.1369114	X112749	0.2557940	X112950	0.26696678	X113035	0.14660485
X112912	0.2507931	X112877	0.1344389	X112886	0.2206823	X103433	0.25520192	X112786	0.14583931
X112811	0.2366526	X112917	0.1337624	X112802	0.1998793	X112814	0.23755889	X112854	0.14564670
X112878	0.2349515	X112840	0.1297287	X112799	0.1838040	X112745	0.20916515	X112980	0.14121432
X112970	0.1970680	X112701	0.1292283	X113023	0.1694592	X112754	0.19230955	X112879	0.13174101
X112966	0.1931201	X112763	0.1283313	X112992	0.1651639	X112818	0.18842320	X112738	0.11879178
X112849	0.1921468	X112723	0.1273581	X113016	0.1545605	X112935	0.17919008	X112838	0.11816856
X113041	0.1875601	X112767	0.1259386	X112782	0.1236163	X112783	0.17577819	X113036	0.11583111
X112742	0.1829092	X113043	0.1173588			X112729	0.16944931	X112936	0.11332178
X113000	0.1826833	X113045	0.1149415			X112942	0.16633637	X113015	0.10912113
X112876	0.1816516	X112777	0.1130053			X112882	0.16566772	X112824	0.10871013
X112994	0.1747328	X112732	0.1092429			X112696	0.16389307	X112820	0.10787117
X112833	0.1623321	X112810	0.1091133			X112822	0.15854572	X113030	0.10757291
X112761	0.1620684	X112995	0.1049080			X112951	0.15807666	X112784	0.10224810
X113018	0.1576107	X112928	0.0970840			X112770	0.15543135	X112920	0.09841826
X113031	0.1482497	X112858	0.0899713			X112764	0.15400553	X112707	0.09383317
X112897	0.1439637	X112953	0.0895190			X112979	0.15064921	X112960	0.07339608
X112695	0.1418353	X113044	0.0557397			X112939	0.14911863		
X112789	0.1371809								

Appendix I: K-means functions in "R"

The function "set.seed()" allows to save the results obtained by the simulation run. As the results obtained don't assure the global minimum, so each simulation results may differ from one to another as random initial points are used for the iterations; that's is the reason why is useful to use the "set.seed()" function.

Then, the function "kmeans()" allows to make the partition of the dataset in the example case the dataset is "hour_percent", and willing to partition it in 7 clusters, iteration 100 times before finding the minimum, starting by 100 different vectors as initial points and using the "Forgy's" algorithm.

```
…
set.seed(9)
fit <- kmeans(hour_percent, 7,iter.max=100,nstart=100, algorithm="Forgy")
…
```

Appendix J: SOM clustering

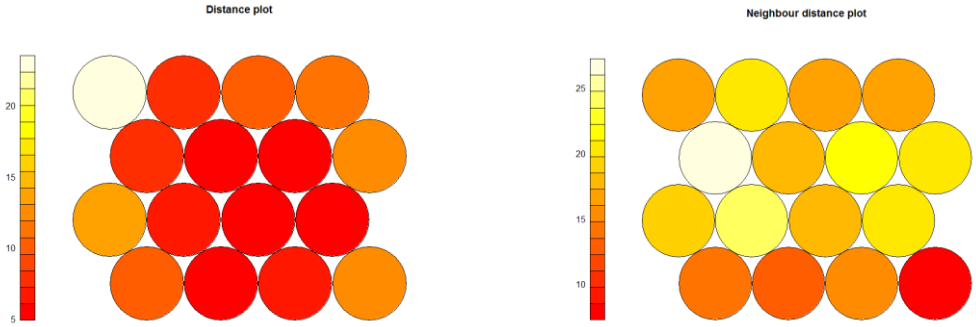
The "library (som)" and the "library (kohonen)" are the packages in "R" where the "som" function is found. In the "som()" function it needs to be input the dimensions of the map, in this case 4x4 with a hexagonal disposition and circular nodes.

```
…
som_model<-som(cast4, grid=somgrid(4,4,"hexagonal"),
               rlen=1000,
               alpha=c(0.05,0.001),
               keep.data = TRUE,
               n.hood="circular")
 plot(som_model)
…
```

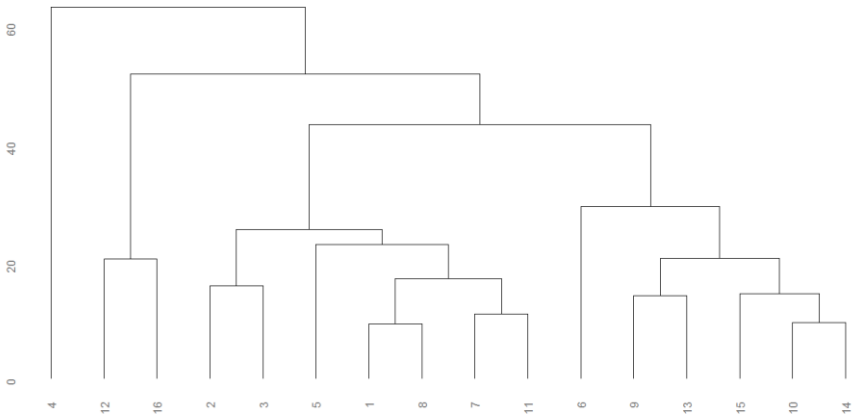
4x4 SOM map Manhattan distance and Ward linkage

Additional SOM plots from the 4x4 solution chosen in section 8.7.

Left side image shows the SOM distances inside each node. The right side image shows the distances to the neighbours nodes.



Dendrogram of the 16 nodes from the SOM using hierarchical clustering using Manhattan distance and Ward linkage

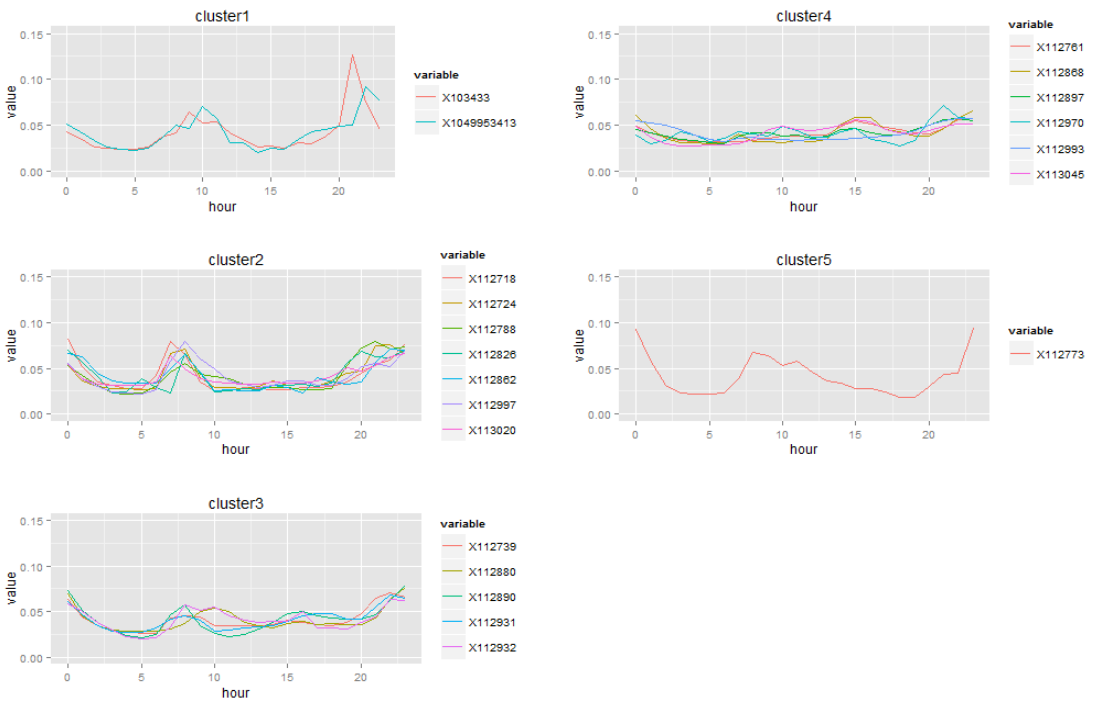


Appendix K: Base clusters sub-division

K.1. Base Cluster 2 sub-division

The base cluster 2 has 21 members, and the final cluster 2 ends up with 17 members.

- It loses 5 members to the final cluster 3 (flat consumers)
    - X112761, X112897, X112970, X112993, X113045
- It loses another member which is considered to fit better in cluster4 (evening peak).
    - X103433

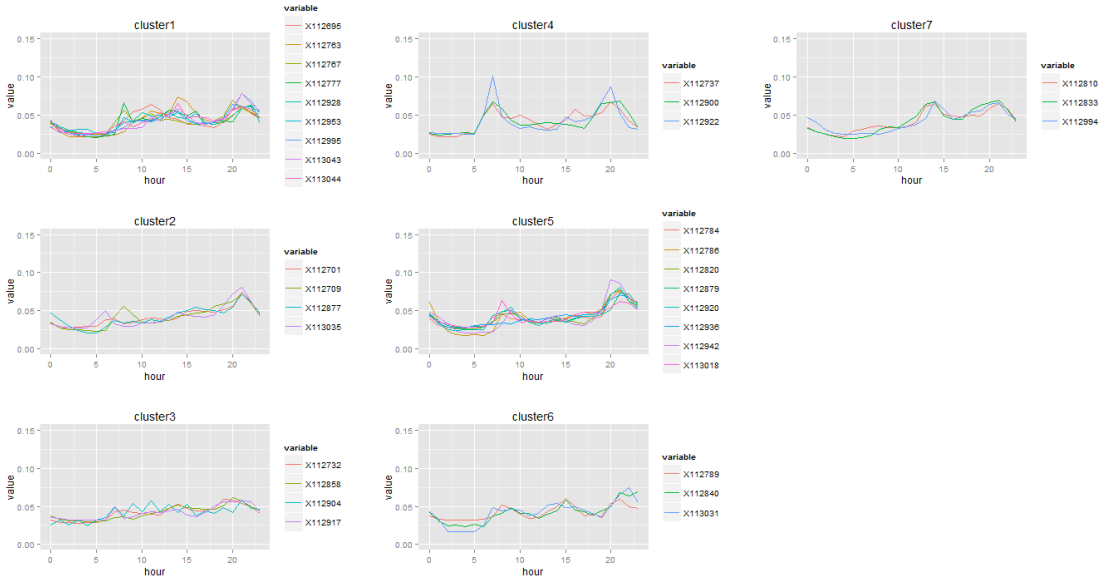


K.2. Base Cluster 3 sub-division

The base cluster 3 has 34 members, and the final cluster 3 ends up with 28 members.

- It loses 6 members to the final cluster 4 (evening peaks)
  - X112786, X112820, X112942, X112995, X113031, X113035
- It loses 3 members which fit better in cluster 5 (daytimers).
  - X112810, X112833, X112994
- It loses 3 more members which fit better in cluster 6 (double peaks).
  - X112737, X112900, X112922

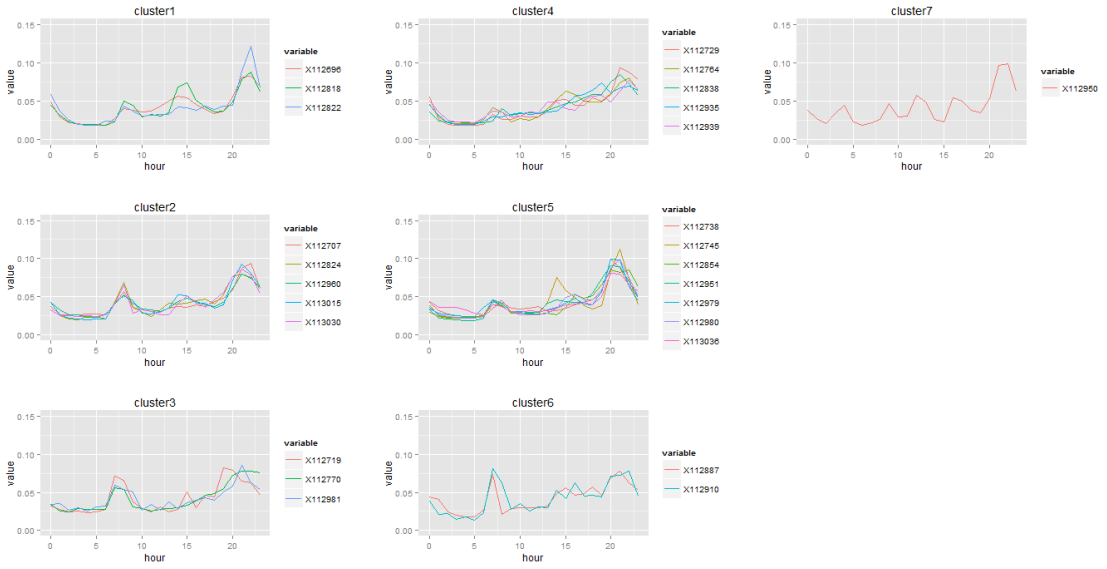
Then, it also receives some members from other clusters to reach the final 28 members.



K.3. Base Cluster 4 sub-division

The base cluster 4 has 26 members, and the final cluster 4 ends up with 28 members.

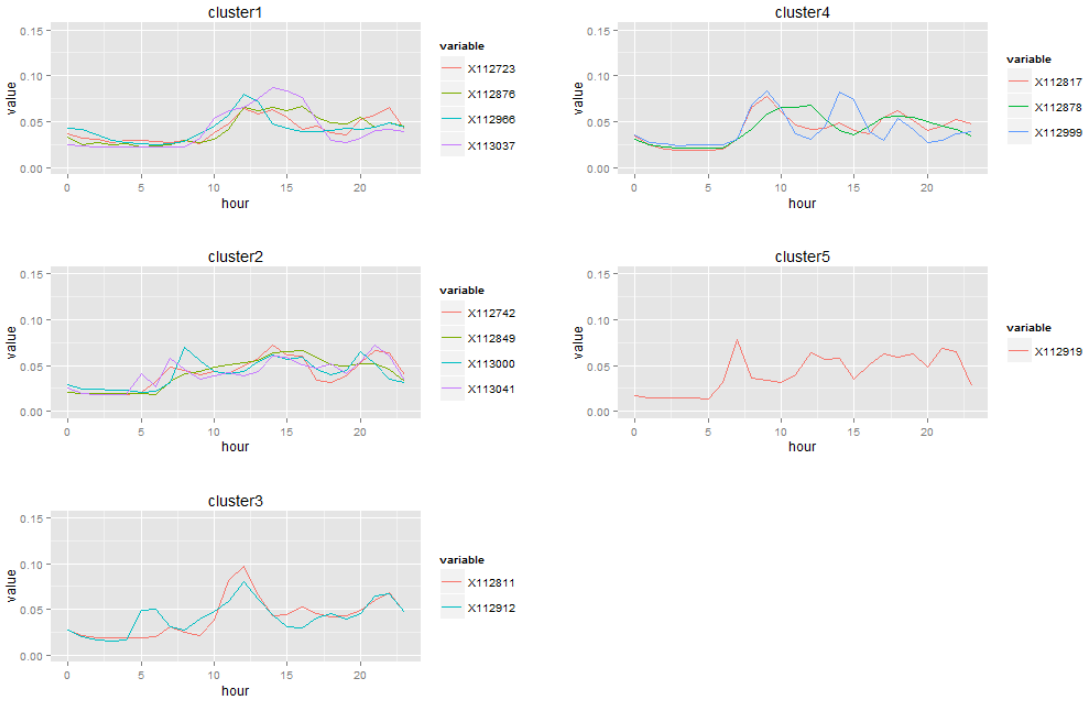
- It loses 1 member which fit better in cluster 3 (flat consumers).
  - X112935
- It loses 5 members to the final cluster 6 (double peaks)
  - X112719, X112770, X112887, X112910, X113981



K.4. Base Cluster 5 sub-division

The base cluster 5 has 14 members, and the final cluster 5 ends up with 16 members.

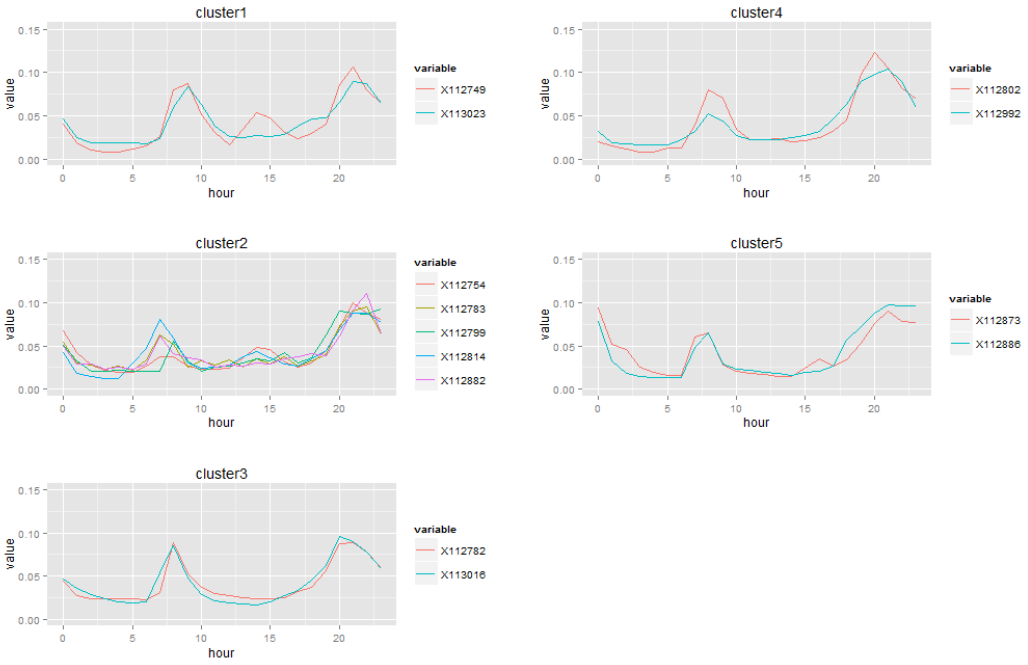
- It loses 1 member which fit better in the final cluster 1 (morning peaks).
    - X112817



K.5. Base Cluster 6 sub-division

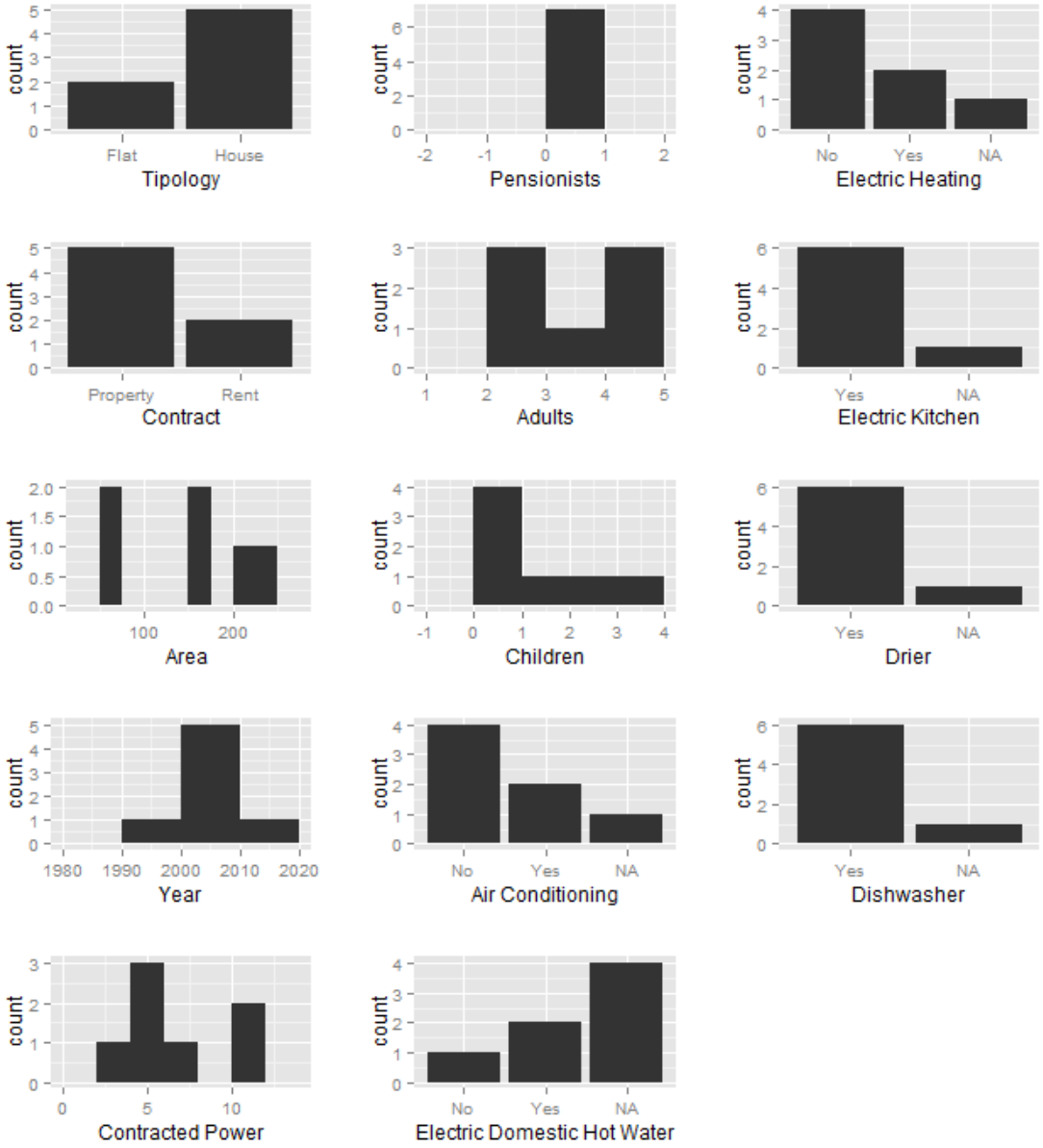
The base cluster 6 has 13 members, and the final cluster 6 ends up with 18 members.

- It loses 2 members to the final cluster 2 (night peak).
    - X112873, X112886
- It loses 1 member that fits better to final cluster 4 (evening peak)
    - X112992

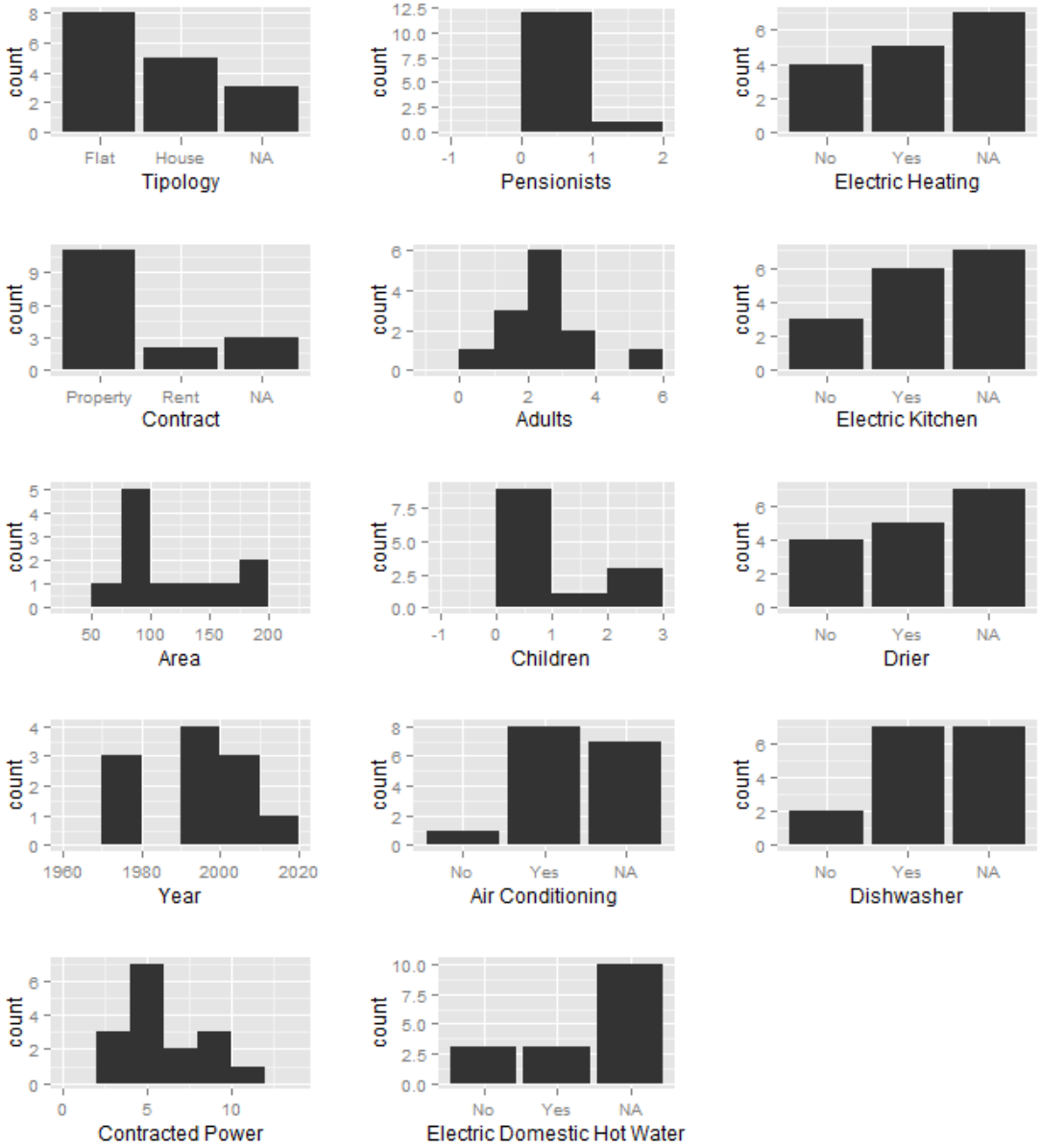


Appendix L: House features analysis

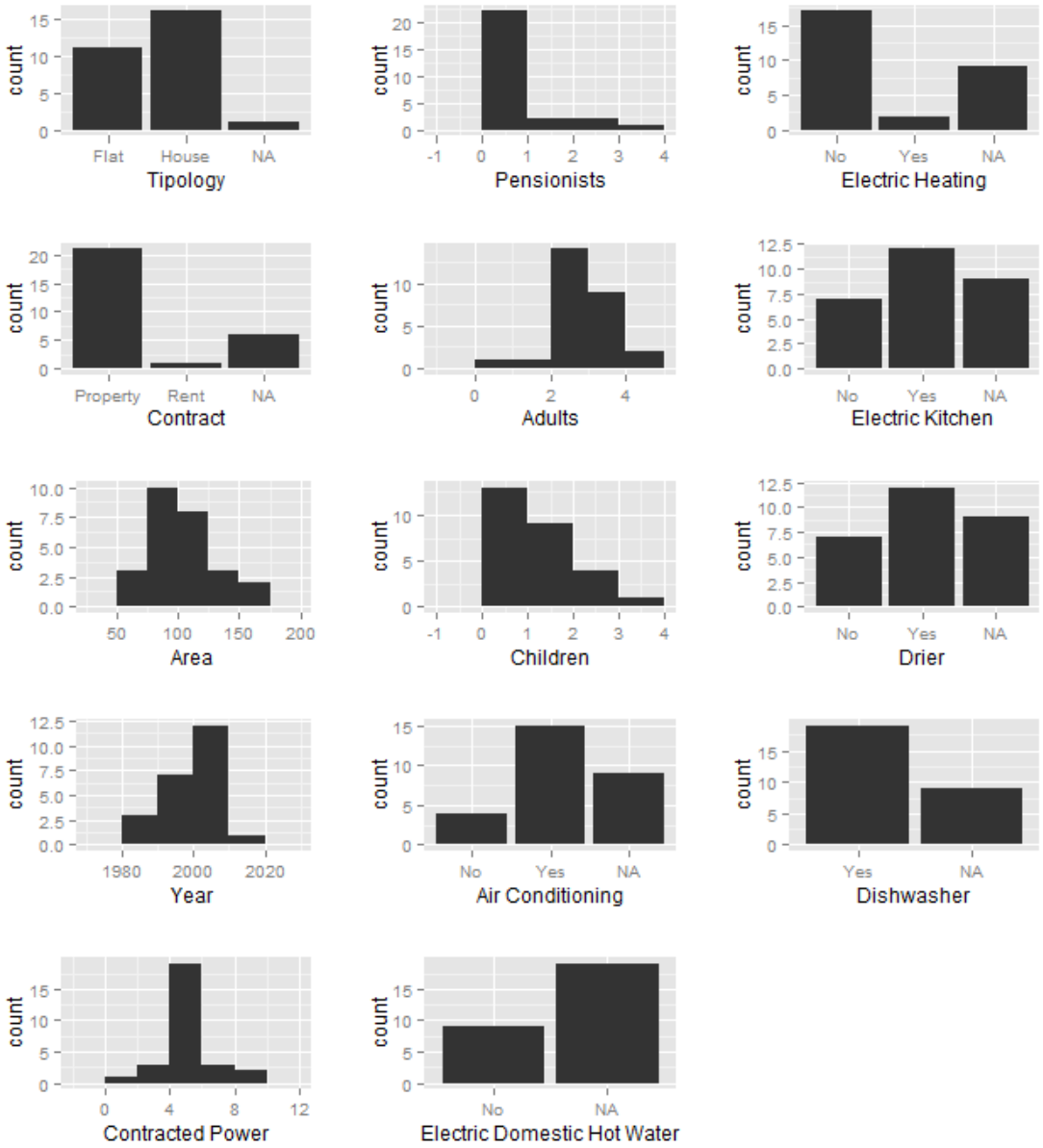
L.1. Cluster 1 features



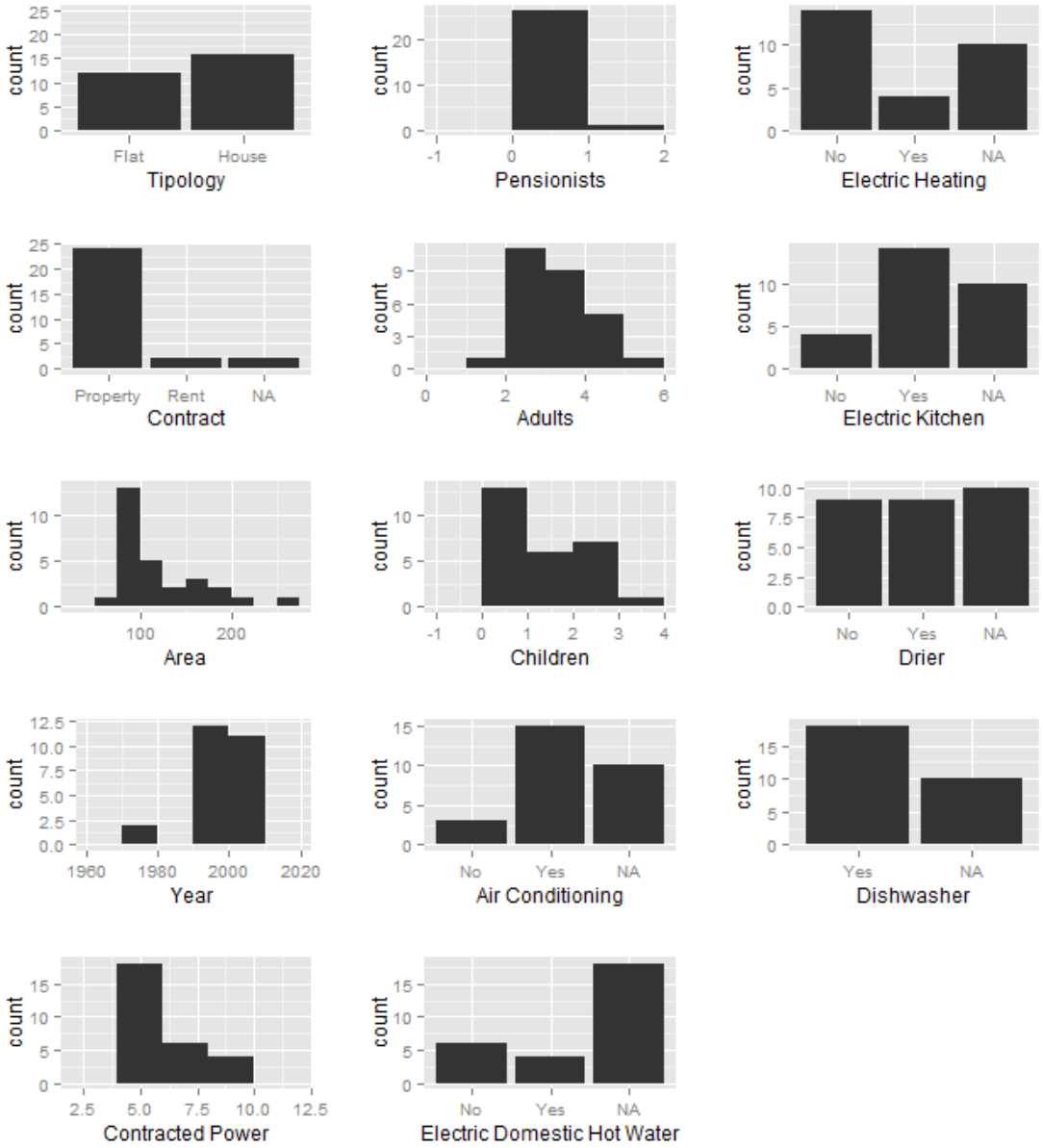
L.2. Cluster 2 features



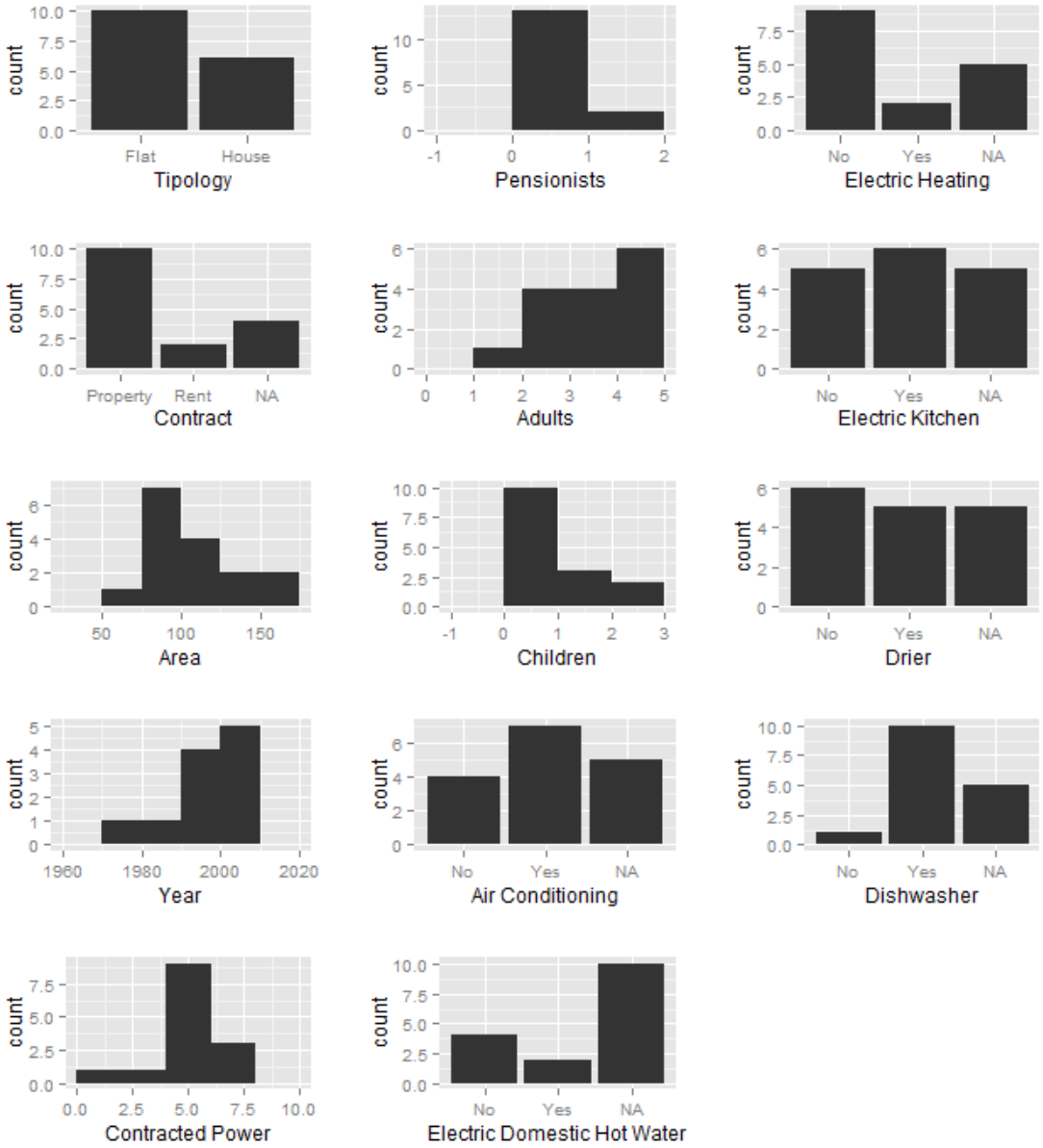
L.3. Cluster 3 features



L.4. Cluster 4 features



L.5. Cluster 5 features



L.6. Cluster 6 features