Nexa Center for Internet & Society
Politecnico di Torino

Power Aware

Lights off, Brains on

Emanuele Mottola, Marco Pietro Abrate, Diego Mariani, Antonio Vetrò

Working paper nr. 1/2018

October 12th, 2018

Studying the Internet, exploring its potential & experimenting new ideas

Nexa Center
for Internet & Society

Via Pier Carlo Boggio 65/A, 10129 Torino, Italia
(http://nexa.polito.it/contacts-en)
+39 011 090 7217 (Telephone)
+39 011 090 7216 (Fax)
info@nexa.polito.it

Mailing address:

Nexa Center for Internet & Society
Politecnico di Torino - DAUIN
Corso Duca degli Abruzzi, 24
10129 TORINO
ITALY

The Nexa Center for Internet & Society is a research center affiliated to the Department of Control and Computer Engineering of Politecnico di Torino (http://dauin.polito.it).

Research reported in this report was supported by the University of California, Berkeley Foundation, in connection with the Siebel Energy Institute, under the "Power Aware" award.

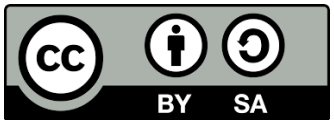


Table of contents

1 Deliverable 1....8
1.1 Focus of the Deliverable....9
1.2 External Factors....9
1.3 Dwelling Factors....11
1.4 Appliance Factors....15
1.5 Socio-economic factors....28
1.6 Behavioral Factors....34
1.7 Conclusion....36
1.8 Appendix A....36
1.9 Bibliography....37
2 Deliverable 2....44
2.1 Introduction....45
2.2 Ethics and Privacy....46
2.3 Bibliography....47
3 Deliverable 3....50
3.1 Introduction....51
3.2 Machine Learning....51
3.3 Supervised Learning....52
3.4 Unsupervised Learning....53
3.5 Classification of Machine Learning Algorithms....54
3.6 Clustering Classification....55
3.7 Conclusion....59
3.8 Bibliography....59
4 Deliverable 4....62
4.1 Introduction....63
4.2 Types of Cluster....63
4.3 Components of a Clustering Task....64
4.4 Density-based Clustering....65
4.5 Partitional Clustering....69
4.6 Hierarchical Clustering....70
4.7 Conclusion....72
4.8 Bibliography....72
5 Deliverable 5....74
5.1 Introduction....75
5.2 Cluster Validation....75
5.3 Components of a Clustering Task....76
5.4 Supervised cluster evaluation....81
5.5 Assessing the Significance of Cluster Validity Measures....83
5.6 Conclusion....83
5.7 Bibliography....84
6 Deliverable 6....86
6.1 Introduction....87
6.2 Requirements....87
6.3 Visual Encoding....88
6.4 Implementation....91
6.5 Conclusion....93
6.6 Bibliography....93
7 Deliverable 7....96
7.1 Introduction....97
7.2 Methodology....97
7.3 Results....98
7.4 Re-design....99
7.5 Conclusion....110
7.6 Appendix A....110
7.7 Appendix B....116

7.8 Bibliography.....119

Introduction - Power Aware

Collective Awareness Platforms (CAPs) aim to tackle issues in different focus areas (such as consumption, economy, open democracy, etc.) by harnessing partecipation of citizens on online platforms in order to create a social networking effect that would lead to shared knowledge and collective intelligence. The process of exchanging information among users, in fact, is finilized to find the best possible solution to a given challenge in order to induce social innovation and to encourage citizens towards engaging in more sustainable lifestyles.

Power Aware is a research project aimed at understating the key requirements for building up a CAP focused on the development of a network of knowledge shared among citizens about energy consumption issues and possible solutions. The project exploits ICT tools and data visualization in order to drive behavioral changes at individual and collective level, encouraging participation of users through competition and cooperation among them. Therefore, the goal of the platform is to engage citizens in changing their consumption patterns.

In particular, Power Aware is a web platform comparing energy consumption patterns for citizens with similar characteristics (e.g., house size, family composition, number and type of appliances) and recommending saving strategies to reduce power consumption at home. The comparison is made by means of interactive data visualizations that show to the citizens their power consumptions related to clusters of similar users. The web platform will offer additional features aimed at building a smart community, such as social tools and consumptions forecasting. The mission of Power Aware is to enhance citizens' awareness on key factors of energy costs and to reduce the energy use by changing the consumption habits.

1 Deliverable 1

List of Parameters for users profiling and clustering and their data type requirements

Authors: Emanuele Mottola – Marco Pietro Abrate
Project Manager: Diego Mariani
Scientific coordinator: Dr. Antonio Vetrò

1.1 Focus of the Deliverable

Power Aware needs to outline the parameters for the data set in order to group together users with similar power consumption behaviors. In fact, we want to investigate the *dwelling, appliance* and *socio-economic related factors* that have significant effects on domestic electricity demand. We will take into account, moreover, the influence that *lifestyle and habits* may have on it.

First of all, we will consider the external factors, which are related to the climatic zone of residence and which may influence everyday-life behaviors or appliances usage of the households.

1.2 External Factors

Geographic position (Zip Code) is a significant determinant of household electricity demand, contributing by up to 46% to the variability in consumption [1]. However, once Zip code is removed, other underlying drivers of electricity consumption are climate and weather variables.

1.2.1 Temperature

Temperature influences the use of space heating systems in cold seasons and air conditioning in warm ones. In fact, positive relationship between monthly residential electricity consumption and mean temperature was observed by Yan [2]. Atmospheric pressure, wind speed, humidity and precipitation are not directly involved in electricity demand, they are instead correcting terms for the influence of temperature (perceived temperature) [3-6].

In the following we will use **Heating Degree Days (HDD)** and **Cooling Degree Days (CDD)** as proxies for energy demand needed to heat or cool a home.

The most accurate calculation of HDD and CDD is done computing the difference between hourly data of outdoor air temperature and the base temperature [84]. The daily value of HDD and CDD is shown in equations (1) and (2).

$$HDD = \frac{\sum_{i=1}^{24} \left(T_{comfort} - T_i\right)}{24} \quad \text{with } (T_{comfort} - T_i) > 0, \tag{1}$$

where:

- $T_{comfort}$ is the base temperature to which the comfort inside a dwelling is achieved, typically 18 – 21 °C,

- T_i is the average temperature over one hour. It is calculated as $T_i = \frac{T_{max} + T_{min}}{2}$, where T_{max} is the maximum temperature in that hour, T_{min} the minimum one.
- 24 is the number of hours per day.

Similarly, CDDs is calculated as:

$$CDD = \frac{\sum_{i=1}^{24} (T_i - T_{comfort})}{24} \quad \text{with } (T_i - T_{comfort}) > 0, \tag{2}$$

where the parameters are the same of HDD, except $T_{comfort}$, which is set to 25 – 27 °C in summer.

Monthly and yearly values of HDD and CDD are calculated by the sum of daily values over the desired period of time.

The heating and cooling requirements for a given structure in a specific location are considered proportional to the number of HDD and CDD in that location [72]. Different sources [82, 83, 84] acknowledged that the electricity demand is linear to the number of HDD and CDD and in particular Moustris et al. [84] found in their study on seven cities in Greece that the annual energy consumption for heating purpose Q_h (KWh) is given by:

$$Q_h = \frac{U' \cdot AHDD \cdot 24}{\eta} \tag{3}$$

where:

- *U'* denotes the overall building heat loss coefficient, see Appendix A,
- *AHDD* denotes the annual sum of HDD (°C day),
- η is the coefficient of efficiency of the internal heat sources ($0<\eta<1$),
- 24 is the number of hours per day (to convert to hours).

To compute the annual energy consumption for cooling purpose Q_c (KWh), instead

$$Q_c = \frac{m \cdot C_p \cdot ACDD \cdot 24}{COP} \tag{4}$$

where:

- *m* denotes the mass flow rate of kilogram of air that is cooled per second (kg s^{-1}),
- C_p is the specific heat of air (kJ kg^{-1} K^{-1}),
- *ACDD* denotes the annual sum of CDD (°C day),
- *COP* is the coefficient of performance of the cooling unit, see chapter 4.5 ,
- 24 is the number of hours per day (to convert to hours).

To group together users more precisely, Power Aware will take into account the Zip Code of the city or district where they live and will calculate the corresponding HDD and CDD exploiting the service provided by [85].

- **Heating Degree Days (HDD):**

Data type: numeric integer	
Value	**Description**
0 – N	Number of HDDs

Table 1: Heating Degree Days

- **Cooling Degree days (CDD):**

Data type: numeric integer	
Value	**Description**
0 – N	Number of CDDs

Table 2: Cooling Degree Days

1.3 Dwelling Factors

Several dwelling factors have been studied in the literature and found to influence electricity consumption. These factors are: (i) dwelling type; (ii) dwelling age; (iii) number of bedrooms; (iv) total floor area; (v) presence of electric water heating systems; and (vi) presence of electric space heating systems [44].

1.3.1 Dwelling type

A large number of researches [8, 9, 13, 15-18, 23, 25, 27, 32, 33] have concluded that electrical energy demand of detached houses is higher than that of semi-detached ones. The consume decreases even more if we look at terrace houses and apartments [44].

Data type: categorical	
Value	**Description**
Detached house	A house that is not joined to any other house
Semi-detached house	A house that is joined to another house on one side by a shared wall
Terrace house	One of a row of similar houses joined together by their side walls
Flat / Apartment	Set of rooms for living in, on one or more floor and part of a larger building

Table 3: Dwelling type

According to the study of Leahy and Lyons [32], families residing in semi-detached and terrace houses use 6.9% less electricity per week than those in detached houses; families in apartments 10.7% less than who lives in detached ones.

The following table presents other studies underlining the relevance of this parameter. In particular they outline how much is the reduction in electricity consumption for dwelling type compared to detached houses.

Study	Dwelling type	Result (relative to detached house), KWh	Method	Relation
McLoughlin et al. [9]	Semi-detached	Coef. = -175.673** Std. Error = 34.170	Multiple linear regression	Linear
	Terrace	Coef. = -147.045** Std. Error = 45.923		
	Flat/Apartment	Coef. = -245.557* Std. Error=119.423		
Wiesmann et al. [13]	Semi-detached	Unstd β = 0.0046	OLS regression	Logarithmic
	Terrace	-		
	Flat/Apartment	-0.066* < Unstd β < -0.052*		

*Tabella 4: Influence of dwelling type. *p<0.05, **p<0.01. Note: [13] values refer to per capita consumption.*

1.3.2 Dwelling age

Dwelling age influences power consumption because newer houses are generally built with higher awareness in insulation and more efficient usage of appliances, lighting and air-conditioning than older ones [13, 16, 25, 26, 32]. On the other hand, some researches have found that electricity consumption of newer houses is higher because of a wider penetration of air-conditioning and other high-consumption appliances [11, 18, 40, 42] [44]. In both cases, either for positive or for negative correlation, this is a factor that affects the power demand.

Data type: numeric integer	
Value	**Description**
0 – N	Age of the building

Table 5: Dwelling age

The following table presents some results found in literature in order to define which may be the influence of this parameter on electricity consumption.

Study	Result	Method	Relation
Wiesmann et al. [13]	Unstd β = -0.015	OLS regression	Linear
Halvorsen, Larsen [40]	Coef. = 19 t-value = 10.19	Econometric analysis	Linear

Tabella 6: Influence of dwelling age. Note: [13] values refer to per capita consumption.

1.3.3 Number of bedrooms

Houses with more bedrooms have more appliances and a larger consumption of electricity for lighting [9, 11, 27, 33, 37]. Specifically, McLoughlin et al. [9] established that for each additional bedroom in Irish dwellings, total electricity consumption on average increased by 15.4% over a six months period; Hamilton et al. [33] found that electricity demand raised linearly from one to four bedrooms and that the increase from four to more than five bedrooms was 12% [44].

Data type: numeric integer	
Value	**Description**
0 – N	Number of bedrooms (no bedrooms: studio apartment)

Table 7: Number of bedrooms

The following table presents results by McLoughlin et al. [9] and, moreover, by Bedir et al. [8], which, instead, argues with results previously shown, presenting a negative relation between number of bedrooms and electricity demand.

Study	Result	Method	Relation
Bedir et al. [8]	β = -0.156*	Multiple linear regression	Linear
F. McLoughlin et al. [9]	Coeff= 349.036 Std. Error = 19.9782	Multiple linear regression	Linear

*Tabella 8: Influence of variable Number of Bedrooms, *p<0.05 .*

1.3.4 Total floor area

Floor area is one of the most trenchant variables in electricity consumption in residential buildings [8, 11-13, 15-17, 20, 22, 23, 25-29, 31, 35, 39-41]. In particular, its influence is stronger in houses where the main heating system is an electric boiler [26]; moreover, a larger floor area may bring a larger number of appliances and lights [17].

Data type: numeric integer	
Value	**Description**
1 – N	Number of m^2

Table 9: Number of m^2 of the house

According to Yohanis et al. [27], there is a linear relation between electricity demand and floor area, as shown in the following formula :

$$E = 49.439\,A + 233.19 \tag{5}$$

where:

- E is average annual electricity consumption in KWh.
- A is the total floor area in m^2.

The influence of this parameter on electricity consumption is outlined in the following table, where we take into account different studies from literature.

Study	Result	Method	Relation
Bedir et al. [8]	β = 0.334*	Multiple linear regression	Linear
Wiessmann et al. [13]	Unstd. β = 0.168***	OLS regression	Linear
Brounen et al. [16]	Unstd. β = 0.284**	Multiple regression	Logarithmic
Zhou and Teng [20]	Unstd. β = 0.1***	OLS regression	Logarithmic
Halvorsen, Larsen [40]	Coef. = 50 t-value = 34.87	Econometric analysis	Linear

*Tabella 10: Influence of total floor area. **p<0.01, *** p<0.001. Note:[13, 16] values refer to per capita consumption.*

1.3.5 Presence of electric hot water heating system

Many studies proved that the presence of an electric hot water heating system notably influences the electricity consumption in residential houses [9, 12, 15, 21, 24, 32, 39]. In particular, studying the behaviors in Norwegian dwellings, Larsen and Nesbakken [21] concluded that electricity consumption is on average 2684 kWh higher for the 80% of households taking showers and having an electric water heater than for other households. Furthermore, the electricity consumption for the 44% of households both taking baths and having an electric water heater is on average 1014 kWh higher than for other households.

Data type: categorical	
Value	**Description**
Yes	Presence of electric hot water heating systems
No	Absence of electric hot water heating systems

Table 11: Presence/Absence of Electric Hot Water Heating System

The influence of this parameter can be outlined in the following table:

Study	Result	Method	Relation
F. McLoughlin et al. [9]	Coef = 148.923 Std. Error = 29.504	Multiple linear regression	Linear

Ndiaye, Gabriel [24]	Coef = 5.892** t-value 5.410	Multiple linear regression	Linear
Tso, Yau [39]	Coef = 46.686* Std. Error = 3.225	Stepwise multiple regression	Linear

*Tabella 12: influence of presence of Electric Hot Water Heating System, *p<0.05, **p<0.0001*

1.3.6 Presence of electric space heating system

Many studies consistently agree that there is a significant and positive effect on electricity use because of the use of an electric space heating system [8, 11, 15, 21, 24, 26, 32, 40]. Larsen and Nesbakken [21] estimated the difference in electricity consumption between Norwegian households with electric under floor and/or central heating systems and other households. They found that families with an electric space heating system consumed on average 3700 kWh per year more electricity than the others [44].

Data type: categorical	
Value	**Description**
Yes	Presence of electric space heating systems
No	Absence of electric space heating systems

Table 13: Presence/Absence of Electric Space Heating System

The influence of this parameter can be outlined in the following table.

Study	Result	Method	Relation
Kavousian et al. [15]	Unstd. β = 0.032	Forward stepwise regression	Linear
Ndiaye, Gabriel [24]	Coef = 2.500* t-value = 5.480	Multiple linear regression	Linear

*Tabella 14: Influence of Electric space heating system, *p<0.0001*

1.4 Appliance Factors

Electrical appliances make an important contribute to electricity consumption. This is not only related to the number of appliances, but also concerns their power demand and frequency of use [44].

In this section we will outline the set of relevant appliances and average impact (in percentage) of each one of them on total electricity consumption, followed by an overview of their *power demand.* For those whose energy efficiency class is not known, we will consider the *installation year* in order to have at least a general idea of the possible consumption of the appliance.

Power Aware will consider the typical *usage frequency* of appliances to outline the electricity consumption of households. In fact, even if families use an efficient set of

appliances, power demand will be high if they overuse it. On the other hand, a wise usage of low efficiency machines may lead to restrained electricity consumption.

1.4.1 Fridge and freezer

The significant effect of refrigerators on electricity usage has been acknowledged [15, 17, 19-21, 40] as one of the most important predictors of electricity consumption compared to other appliances, accounting for 14.5 – 28 % according to Almeida et al. [51] and Bertoldi et al. [55]. In particular, it is useful to distinguish various types of cooling appliances, because they affect electricity demand in different ways. Zhou and Teng [20] demonstrated that Chinese households with a refrigerator had electricity consumption 22.2% higher than households without it. Moreover, Leahy and Lyons [32] found that households with a fridge-freezer used approximately 6.7% more electricity per week than households without such appliance (refrigerator only).

Hence, we will outline cooling appliances' characteristics affecting electricity consumption.

- **Number of cooling appliances**:

Data type: numeric integer	
Value	**Description**
0 – N	Number of cooling appliances

Table 15: Number of cooling appliances

- **Type of cooling appliances**:

Data type: categorical	
Value	**Description**
Fridge	A cooling appliance which holds the temperature between 3°C and 7°C
Fridge-freezer	A cooling appliance divided in: • fridge: temperature between 3°C and 7°C • freezer: temperature between -20°C and -15°C
Freezer / Chest-feezer	A cooling appliance which holds the temperature between -20°C and -15°C

Table 16: Type of appliance

- **Energy efficiency class**:

Data type: categorical	
Value	**Description**
Class A+++	EEI[1] < 22

1 EEI: Energy Efficient Index is an indication of the annual power consumption relative to a reference consumption that is based on the storage volume and the type of appliance.

Class A++	22 < EEI < 33
Class A+	33 < EEI < 42/44[2]
Class A	42/44 < EEI < 55
Class B	55 < EEI < 75
Class C	75 < EEI < 95
Class D	95 < EEI < 110

Table 17: cooling appliances energetic class [46]

- **Installation year**:

Data type: categorical	
Value	**Description**
0 – current year	Year in which the appliance was installed

Table 18: Installation year

1.4.2 Washing machine

The ownership of a washing machine has been studied as a positive factor of influence in different researches [21, 29, 37, 40], affecting electricity consumption for 7.2% according to Bertoldi et al. [55]. In particular, Larsen and Nesbakken [21] found that households with a washing machine used 2099 kWh more electricity per year than households without such an appliance. Additionally, Genjo et al. [29] determined that the size of the washing machine is a significant influential factor [44]. Finally, energetic class is an important parameter to take into account.

- **Number of washing machines**:

Data type: numeric integer	
Value	**Description**
0 – N	Number of washing machines

Table 19: Number of washing machines

- **Dimension**:

Data type: categorical	
Value	**Description**
Small	Load: 3 – 5 Kg

2 The boundary between the A+ and A classes is 44 up to 1 July 2014, and 42 after that date.

Medium	Load: 6 – 8 Kg
Large	Load: 9 Kg or more

Table20: Washing machines dimension

- **Energy efficiency class**:

Data type: categorical	
Value	**Description**
Class A+++	EEI < 46
Class A++	46 < EEI < 52
Class A+	52 < EEI < 59
Class A	59 < EEI < 68
Class B	68 < EEI < 77
Class C	77 < EEI < 87
Class D	EEI > 87

Table 21: Washing machines energy efficiency class [47]

- **Installation year**:

Data type: categorical	
Value	**Description**
0 – current year	Year in which the appliance was installed

Table 22: Installation year

Usage frequency

- **Number of weekly cycles:**

Data type: numeric integer	
Value	**Description**
1 – 15	From one to fifteen cycles a week

Table 23: Number of weekly cycles

- **Medium number of weekly cycles per occupant:**

Number of occupants	Medium number of weekly cycles
1	2
2	3
3	4
4	6
5 +	7

Table 24: Medium number of weekly cycles in function of the number of occupants

1.4.3 Dishwasher

The ownership of a dishwasher and its impact on electricity demand have been the focus of extensive researches and many authors found a significant relationship between the two [9, 17, 19, 21, 32, 40]. Leahy and Lyons [32], for example, established that having a dishwasher increased electricity consumption in Irish homes by over 10,5% per week [44]. The size, as well as the energetic class, are important features that must not be discarded when considering the electricity usage of a dishwasher.

- **Number of dishwashers**:

Data type: numeric integer	
Value	**Description**
0 – N	Number of dishwashers

Table 25: Number of dishwashers

- **Dimension of dishwasher**:

Data type: categorical	
Value	**Description**
Small	Six to nine covers
Medium	Ten to twelve covers
Large	Thirteen or more covers

Table 26: Dishwasher dimension

- **Energy efficiency class**:

Data type: categorical	
Value	**Description**
Class A+++	EEI < 50
Class A++	50 ≤ EEI < 56
Class A+	56 ≤ EEI < 63
Class A	63 ≤ EEI < 71
Class B	71 ≤ EEI < 80
Class C	80 ≤ EEI < 90
Class D	EEI > 90

Table 27: Dishwasher energetic class [48]

- **Installation year**:

Data type: categorical	
Value	**Description**
0 – current year	Year in which the appliance was installed

Table 28: Installation year

Usage frequency

- **Number of weekly cycles:**

Data type: numeric integer	
Value	**Description**
1 – 15	From one to fifteen cycles a week

Table 29: Number of weekly cycles

- **Medium number of weekly cycles per occupant:**

Number of occupants	Medium number of weekly cycles
1	3
2	4
3	5
4	5
5 +	6

Table 30: Medium number of weekly cycles in function of the number of occupants

1.4.4 Electric oven

Electric oven has a significant and positive effect on residential electricity demand, as different studies acknowledged [9, 15, 19, 32, 40][44]. Wood and Newborough [49], for example, found that the electricity consumption per operation is 1.2 – 2.0 kWh with a typical usage of 90 – 120 minutes, contibuting altogether for about 6.6% of yearly electricity demand [55]. In our analysis we will also take into account whether the oven is portable or not.

- **Number of ovens**:

Data type: numeric integer	
Value	**Description**
0 – N	Number of ovens

Table 31: Number of ovens

- **Oven dimension:**

Data type: categorical	
Value	**Description**
Small (portable)	12 l ≤ volume < 35 l
Medium	35 l ≤ volume < 65 l
Large	volume ≥ 65 l

Table 32: Oven dimension

- **Installation year:**

Data type: categorical

Value	Description
0 – current year	Year in which the appliance was installed

Table 33: Installation year

Usage frequency

- **Number of weekly usages:**

Data type: numeric integer	
Value	**Description**
1 – 15	From one to fifteen cycles a week

Table 34: Number of weekly oven usage

1.4.5 Air-conditioning

The significant and positive effect of this appliance on residential electricity consumption has been consistently reported by studies primarily based in locations with hot summers [10, 12, 20, 24, 36, 39, 44]. For example, Bertoldi et al. [55] observed that air conditioning consumed on average 4.7% of the electricity used in the residential sector in 2009 (together with dehumidifier and ventilation). We will also take into account that air conditioners can be used as heating systems during winter seasons (if provided with this feature).

In order to group users together more precisely, Power Aware is interested in different factors which influence the consumption:

- **Number of air conditioners:**

Data type: numeric integer	
Value	**Description**
0 – N	Number of air conditioners

Table 35: Number of air conditioners

- **Air conditioner power:**

Data type: categorical	
Value	**Description**
Small	Up to 9000 BTU[3]
Medium	From 9000 to 15000 BTU
Large	Over 15000 BTU

Table 36: Air conditioners power

- **Purpose of use:**

3 The **British thermal unit** (**Btu** or **BTU**) is a traditional unit of heat; it is defined as the amount of heat required to raise the temperature of one pound of water by one Fahrenheit degree.

Data type: categorical	
Value	**Description**
Heating	Used to heat the house
Cooling	Used to cool the house
Heating and Cooling	Used for both purposes

Table 37: Purpose of use of air conditioners

- **Energy efficiency class, except double ducts and single ducts:**

Data type: categorical		
Value	**Description (cooling) - SEER**	**Description (heating) - SCOP**
Class A+++	SEER[4] ≥ 8,50	SCOP[5] ≥ 5,10
Class A++	6,10 ≤ SEER < 8,50	4,60 ≤ SCOP < 5,10
Class A+	5,60 ≤ SEER < 6,10	4,00 ≤ SCOP < 4,60
Class A	5,10 ≤ SEER < 5,60	3,40 ≤ SCOP < 4,00
Class B	4,60 ≤ SEER < 5,10	3,10 ≤ SCOP < 3,40
Class C	4,10 ≤ SEER < 4,60	2,80 ≤ SCOP < 3,10
Class D	3,60 ≤ SEER < 4,10	2,50 ≤ SCOP < 2,80

Table 38: Energy efficiency class of air conditioners (not single or double duct) [50]

- **Energy efficiency classes for double ducts and single ducts:**

Energy Efficiency Class	**Double ducts**		**Single ducts**	
Value	**Description (cooling) - EER_{rated}[6]**	**Description (cooling) - COP_{rated}[7]**	**Description (cooling) - EER_{rated}**	**Description (cooling) - COP_{rated}**
Class A+++	EER≥4,10	COP≥4,60	EER≥4,10	COP≥3,60
Class A++	3,60≤ EER <4,10	4,10≤COP<4,60	3,60≤ EER <4,10	3,10≤COP<3,60
Class A+	3,10≤ EER <3,60	3,60≤COP<4,10	3,10≤ EER <3,60	2,60≤COP<3,10
Class A	2,60≤ EER <3,10	3,10≤COP<3,60	2,60≤ EER <3,10	2,30≤COP<2,60
Class B	2,40≤ EER <2,60	2,60≤COP<3,10	2,40≤ EER <2,60	2,00≤COP<2,30
Class C	2,10≤ EER <2,40	2,40≤COP<2,60	2,10≤ EER <2,40	1,80≤COP<2,00
Class D	1,80≤ EER <2,10	2,00≤COP<2,40	1,80≤ EER <2,10	1,60≤COP<1,80

Table 39: Energy efficiency class of single or double duct [50]

- **Installation year:**

Data type: categorical

4 The **Seasonal Energy Efficiency Ratio (SEER)** is the overall energy efficiency ratio of the unit, representative for the whole cooling season, calculated as the reference annual cooling demand divided by the annual electricity consumption for cooling; [50]

5 The **Seasonal coefficient of performance (SCOP)** is the overall coefficient of performance of the unit, representative for the whole designated heating season (the value of SCOP pertains to a designated heating season), calculated as the reference annual heating demand divided by the annual electricity consumption for heating; [50]

6 **Rated energy efficiency ratio (EER_{rated})** means the declared capacity for cooling [kW] divided by the rated power input for cooling [kW] of a unit when providing cooling at standard rating conditions; [50]

7 **Rated coefficient of performance (COP_{rated})** means the declared capacity for heating [kW] divided by the rated power input for heating [kW] of a unit when providing heating at standard rating conditions; [50]

Value	Description
0 – current year	Year in which the appliance was installed

Table 40: Installation year

Usage frequency

- **Number of weeks per month in summer:**

If the air conditioner is used to cool the ambient down, we are interested in the number of weeks per month households use the machine, especially in summer.

Data type: numerical integer			
June	**July**	**August**	**September**
0 – 4	0 – 4	0 – 4	0 – 4

Table 41: Number of weeks per month air conditioner is used (summer)

- **Number of hours per day in summer:**

Data type: numerical integer	
Value	**Description**
0 – 24	From zero to twenty-four hours per day

Table 42: Number of hours air conditioner is used per day

Otherwise, if the air conditioner is used to warm the house, we will take into account winter frequency usage.

- **Number of weeks per month in winter:**

Data type: numerical integer			
November	**December**	**January**	**February**
0 – 4	0 – 4	0 – 4	0 – 4

Table 43: Number of weeks per month air conditioner is used (winter)

- **Number of hours per day in winter:**

Data type: numerical integer	
Value	**Description**
0 – 24	From zero to twenty-four hours per day

Table 44: Number of hours air conditioner is used per day

1.4.6 Dehumidifier

Dehumidifiers account for a substantial portion of residential energy use and they consume on average 4.2 kWh/day: obviously this data depends on the daily number of hours it works [75]. The electricity consumption, moreover, is related to the capacity, defined as the number of litres / day the appliance can remove from the indoor ambient.

- **Number of Dehumidifier:**

Data type: numerical integer	
Value	**Description**
0 – N	Number of dehumidifiers

Table 45: Number of dehumidifiers

- **Capacity (litre/day):**

Data type: categorical	
Value	**Description**
Small	Capacity ≤ 12
Medium	12 < Capacity < 20
Large	Capacity ≥ 20

Table 46: Dimension of dehumidifier

Usage frequency

- **Number of hours per week:**

Data type: numerical integer	
Value	**Description**
0 – 168	From zero to twenty-four hours per week

Table 47: Weekly average usage

1.4.7 Entertainment appliances

In relation to entertainment appliances, several authors [9, 11, 15, 17, 19, 21][44] have observed that the ownership of television has a significant influence on electrical energy demand in residential buildings, ranged between 8 – 10 % [51,55] . In particular, the number of televisions as well as the technology used are important for our aim.

- **Number of TVs**:

Data type: numerical integer	
Value	**Description**
0 – N	Number of televisions

Table 48: Number of televisions

- **Type of TV:**

Data type: categorical	
Value	**Description**
LCD	Liquid-crystal display
CRT	Cathode ray tube
PLASMA	Flat panel with small cells containing electrically charged ionized gases

LED	Flat panel using light-emitting diodes as pixels
OLED	A LED panel using as emissive electroluminescent layer a film of organic compound that emits light in response to an electric current

Table 49: TV type

- **TV dimension:**

Data type: numerical integer	
Value	**Description**
1 – N	Screen diagonal length in inches

Table 50: TV dimension in inches

- **Energy efficiency class**:

Data type: categorical	
Value	**Description**
Class A++	0,10 ≤ EEI < 0,16
Class A+	0,16 ≤ EEI < 0,23
Class A	0,23 ≤ EEI < 0,30
Class B	0,30 ≤ EEI < 0,42
Class C	0,42 ≤ EEI < 0,60

Table 51: TV energetic class [74]

- **Installation year:**

Data type: categorical	
Value	**Description**
0 – current year	Year in which the appliance was installed

Table 52: Installation year

Usage frequency

- **Number of hours per day the TV is on:**

Data type: numerical integer	
Value	**Description**
0 – 24	From zero to twenty-four hours per day

Table 53: Number of hours TV is on

1.4.8 Cooking appliances

Energy used for cooking is accounted for about 8 – 11 % of the overall consumption in a typical centrally-heated house [51, 56, 58, 59]. Furthermore, the cooker is one of the least energy efficient and most expensive to “run” of all domestic appliances [57, 61].

However, an induction hob can save up to 59% of the energy when compared to a conventional electric cooker [63] and, according to Singer et al. [60], it is also 50% cheaper to “run” than other electric hobs. Moreover, an induction hob has been fitted to a hospital kitchen, which provides nearly 1200 meals per day and it reduced the average daily electricity consumption from 1100 kWh to 650 kWh, during winter months [62].

To conclude, cooking appliances will be taken into account due to their penetration in dwellings and to the differences in efficiency.

- **Number of electric cooking appliances**:

Data type: numerical integer	
Value	**Description**
0 – N	Number of electric cooking appliances

Table 54: Presence/Absence of electric cooking appliances

- **Type of cooking appliances:**

Data type: categorical	
Value	**Description**
Standard	Non-induction electric hobs
Induction	Induction electric hobs

Table 55: Cooking appliances type

1.4.9 IT equipment

The ownership of IT appliances, such as desktop computers and laptops, has a significant effect as it has been demonstrated by different studies [9, 11, 17, 20]. In particular, according to IEA [52], the energy consumption of information and communication technologies (ICT) and consumer electronics (CE) has risen considerably in recent years and now accounts for approximately 7 – 12 % of global residential electricity consumption [51, 55].

- **Number of devices per household:**

Data type: numerical integer	
Value	**Description**
0 – N	Number of devices

Tabella 56: Number of IT equipment

Usage frequency

- **Number of hours a week the IT equipment devices are on:**

Data type: numerical integer

Value	Description
0 – 168	From zero to 168 hours a week

Table 57: Number of hours IT equipment devices are on

1.4.10 Lightning

Residential electricity consumption for lighting amounts on average 9-18% on the electricity demand [51, 53 – 55]. Bedir et al. [8] and Kavousian et al. [15] found that the usage of energy efficient bulbs implicates lower electricity demand in Netherlands and USA. Hence, the parameter we are interested in is the percentage of low energy bulbs (LED or CFL) installed in the house.

Data type: categorical	
Value	**Description**
0	No low energy bulbs installed
25	A quarter of bulbs installed are low energy
50	Half bulbs installed are low energy
75	Three quarter of bulbs installed are low energy
100	All bulbs installed are low energy

Table 58: Percentage of low energy bulbs installed

1.4.11 Others

Power Aware, as a first approach, will not take into account appliances like vacuum-cleaner and hair-dryer because even if the penetration of such appliances in houses is high, the time of usage is limited and the consequent electricity demand is low.

1.5 Socio-economic factors

Taking into account a variety of studies on the correlation between socio-economic features and electricity consumption, we have identified a range of the most important factors: (i) number of occupants; (ii) family composition, including presence and number of children, presence and number of teenagers, presence and number of adults; (iii) age of household responsible person (HRP); (iv)employment status; (v) household income [44] and (vii) occupancy level .

1.5.1 Number of occupants

The number of occupants effect on electricity consumption in residential buildings has been extensively studied. Most of the previous researches that have examined the matter concluded that there is a significant positive relationship between the household size and

the domestic electricity usage, suggesting that the higher the number of people living in a dwelling, the more the electricity consumed [8, 12-18, 20, 23, 24, 27, 29, 30, 32, 36, 39-41] [44].

Data type: numeric integer	
Value	**Description**
1 – N	Number of occupants

Table 59: Number of occupants

In literature, some researches outline the relation between electricity consumption and number of occupants as linear, certains logarithmic, and others neither the first nor the second. Results are presented in the table below.

Study	Result	Method	Relation
Bedir et al. [8]	$\beta = 0.334$ with p-value<0.01	Multiple linear regression	Linear
Kavousian et al. [15]	$0.27 < \beta < 0.311$ p-value < 0.01	Multiple regression	Proportional to the square of number of occupants.
Zhou, Teng [20]	$0.079 <$ Unstd $\beta < 0.085$ with p-value < 0.01	OLS analysis	Logarithmic
Ndiaye, Gabriel [24]	$\beta = 0.5655$ with p-value = 0.0376 and t-value = 2.13	Latent root regression	Linear
Lam [30]	$\beta = 0.59$	Multiple log-linear regression analysis	Log-linear

Tabella 60: Summary of number of occupants - electricity relationships in literature. Note: [24] value computed per square foot.

1.5.2 Family composition

The presence of children and its influence on electricity consumption was shown to be significant by McLoughlin et al. [9] who determined that adults living with children in Ireland consumed considerably more electricity than those living alone or with other adults.

Brounen et al. [16] revealed that households with children, in the Netherlands, consumed almost one-fifth more electricity than families without children, and this effect was stronger when the age of the children increased. The authors believed that this was because older children and teenagers watched more television, used personal computers, and were frequent users of gaming devices. Similar results were published in Wiesmann et al. [13] (Portugal) and Nielsen [31] (Denmark).

To conclude, we are interested in the number of adults, teenagers, children and infants living in the house.

- **Number of adults:**

Data type: numeric integer	
Value	**Description**
1 – N	Number of persons older than 19

Table 61: Number of adults in the household

- **Number of teenagers:**

Data type: numeric integer	
Value	**Description**
0 – N	Number of persons between 12 and 18 years old

Table 62: Number of teenagers in the household

- **Number of children:**

Data type: numeric integer	
Value	**Description**
0 – N	Number of persons between 5 and 12 years old

Table 63: Number of children in the household

- **Number of infants:**

Data type: numeric integer	
Value	**Description**
0 – N	Number of persons between 0 and 4 years old

Table 64: Number of infants in the household

According to Brounen et al. [16], once these parameters are acquired, it is useful to convert responses into dummy variables in order to weight their influence. The authors outlined the following formula to calculate the use of energy per capita:

$$\log\left(E_{per\,capita}\right)=\alpha+\sum_{i=1}^{5}\delta_i D_i+\epsilon \tag{6}$$

where α represents the consumption per capita of households without children and ε is the error term. The other variables, δ_i and D_i, represent, respectively, whether the family composition falls in one of the category listed in the following table ($\delta_i = 1$) or not ($\delta_i = 0$) and the weight that occurrence will have on electricity consumption per capita ($\delta_5 = N$ where N is the number of people younger than eighteen years old).

So, for example, let's suppose a family is made up by two parents and two children, one is 3 years old, the other is 7. The per capita electricity comsumption is given by:

$$\log\left(E_{per\,capita}\right)=6.148+(1)\cdot(-0.054)+(1)\cdot(0.068)-(2)\cdot(0.194) \quad .$$

Index (i)	Description	D_i [std. errors]
i=1	Single household	0.367* [0.005]
i=2	Family with children < 4 year	-0.054* [0.007]
i=3	Family with children 5-12 years	0.068* [0.009]
i=4	Family with children > 12 years	0.143* [0.008]
i=5	Number of people younger than eighteen years old in the household	-0.194* [0.004]
-	Constant (α)	6.148* [0.032]

*Tabella 65: Regression estimates for electricity consumption per capita, Brounen et al. [16].*Significance at 0,01 level.*

1.5.3 Age of HRP (Household Reference Person)

Literature agrees that the primary occupant generally dictates the household behaviors and he consequently has an influence over consumption [8, 9, 15, 16, 18, 22, 27, 32]. In particular, McLoughlin et al. [9] found that electricity demand in Irish households with younger HRPs is significantly lower when compared to other age categories such as 36 – 55, and 56 plus. Authors explain it asserting that younger HRPs have lower occupancy patterns and tend to have less children living at home than middle aged HRPs. Moreover, Kavousian et al. [15] also found lower consumption among older HRPs compared to middle aged ones suggesting that older household members are more conscious about electricity consumption and tend to use less electric gadgets. Finally, Yohanis et al. [27] explain middle aged HRPs high consumption because of higher household incomes, bigger houses and a broad range of appliances [44].

Data type: categorical	
Value	**Description**
19 – 30	HRP age between 19 and 30
31 – 45	HRP age between 31 and 45
46 +	HRP age over 46

Table 66: Age of HRP [44]

The influence of this parameter can be outlined in the following table:

Study	Result	Method	Relation
Brounen et al. [16]	D_i = 0.001[8], p-value <0.001	Linear regression	Logarithmic

8Coefficient to be inserted in the formula explained in chapter 5.3

Tiwari [18]	Unstd β = 0.008 , t-value = 67.41	Ridge regression	Linear

Tabella 67: Influence of age of HRP parameter.

1.5.4 Employment status

Employment status impacts upon the timing of electricity loads, Anderson et al. [73]. Yohanis et al. [27] found that homes where all occupants worked or attended school during the day had peak consumptions in the morning (prior to working hours) and in the evening.

Hence, Power Aware is interested in the employment status of every member of the household.

Data type: categorical	
Value	**Description**
Employed full time	Occupant working full time
Employed part time	Occupant working part time
Unemployed	Occupant without a job
Retired	Occupant having withdrawn from a job or career
Student	Occupant is a student

Table 68: Employment status

Throught a multilevel regression modelling, Anderson et al. [73] found a linear relation between mean daily electricity consumption and employment status

$$\text{Unst } \beta = 0.11 \text{ and Z-score} = 2.27.$$

1.5.5 Household income

Many studies determined that household income is proportional to the electricity consumption [8, 10, 13, 17-20, 23, 25, 27, 29, 30, 34-36, 38, 41, 43]. Specifically, Yohanis et al. [27] explained that a higher income generally implicates a higher number of appliances, a larger home and greater number of occupants and a consequent positive effect on electricity demand.

Data type: categorical	
Value	**Description**
Less than $20,000	Income lower than $20,000
$20,000 to $39,999	Income between $20,000 and $39,999
$40,000 to $59,999	Income between $40,000 to $59,999
$60,000 to $79,999	Income between $60,000 to $79,999
$80,000 to $99,999	Income between $80,000 to $99,999
$100,000 to $119,999	Income between $100,000 to $119,999
$120,000 to $139,999	Income between $120,000 to $139,999
$140,000 or more	Income over $140,000

Table 69: household income rank[81]

The influence of this parameter on electricity consumption is outlined in the following table, where we take into account different studies from literature.

Study	Result	Method	Relation
Sanquist et al. [10]	β = 0.11 (year 2005) β = 0.07 (year 2001)	Multiple regression analysis	Logarithmic
Wiesmann et al. [13]*	Unstd β = 0.212 (top-down approach) Unstd β = 0.128 (bottom-up approach), p-value<0.001	OLS regression	Linear
Zhou, Teng [20]	0.14 < Income elesticity < 0.34	Econometric approach	Less than proportional
Genjo et al. [29]	β = 0.26	Multiple linear regression	Linear
Lam [30]	Unstd β = 0.533	Multiple linear regression	Logarithmic

*Tabella 70: Influence of Income on electricity consumption. * per capita electricity consumption.*

1.5.6 Occupancy level

Occupancy level can be expressed as the number of hours per occupant spent at home. This parameter depend on different factors: kind of job, hobbies, social life. Seryal and Kissock [45], studying the power consumption behaviors of nearly 1,700 students, have shown that the electricity usage in student's houses during summer was drastically lower than in other seasons because of an obviously lower occupancy level.

It is clear that this factor is strongly correlated to the employment status of the household and it indirectly influences other important parameters such as time of use of appliances (i.e., lighting, space heating, air conditioning).

- **Number of hours out of the house per week on average:**

Data type: numerical integer	
Value	**Description**
0 – 168	Hours out of the house

Table 71: Occupancy level measured as number of hours out of the house

In order to quantify how much the electricity consumption varies according to the presence of occupants in the dwelling, Yohanis et al. [27] defined the concept of *average daily base-level consumption*: it is the daily average electricity demand of a house when no occupant is in and the only components of consumption are the necessary constant use of appliances (such as fridges and freezers) and the losses due to appliances on stand-by. Authors compared this value with the average daily consumption of the same houses (with

routine occupancy level of households) and found that the first was one fifth of the second, according to Fig. 1.

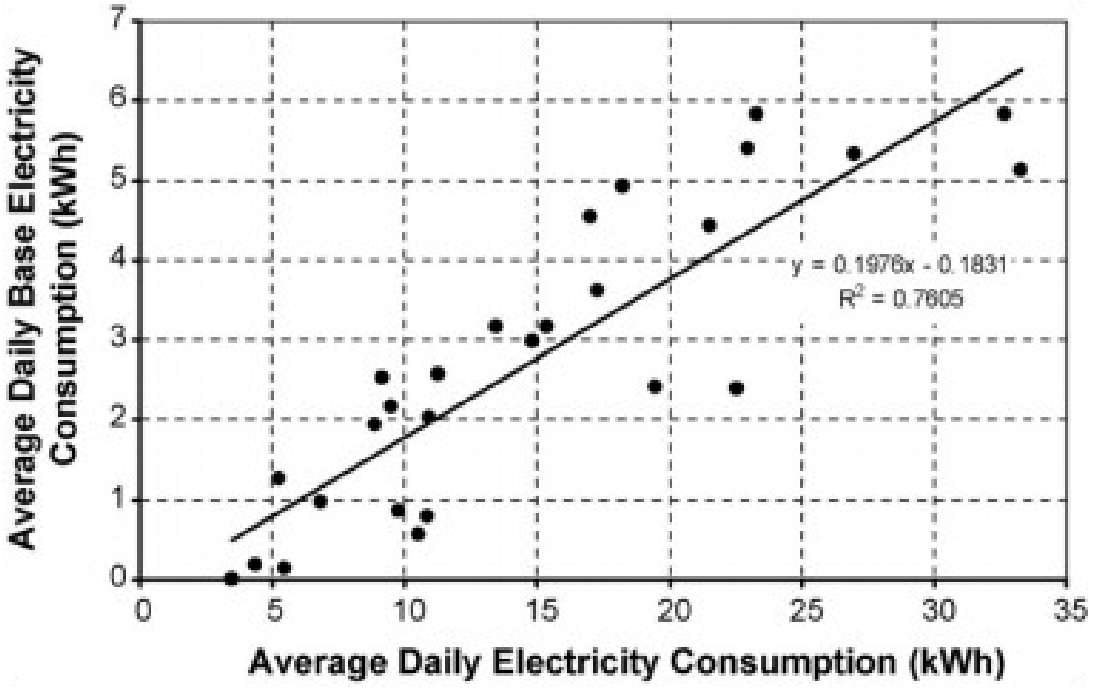


Figure 1: Average daily base-level consumption correlated with average daily consumption

1.6 Behavioral Factors

The following parameters are not to be considered as classification factors during the application of the clustering algorithm, because they do not directly influence the power consumption of a household. They are optional and can be used to give some tips to users in order to change their behaviors in a positive way.

1.6.1 Vampire power

According to Almeida et al. [51], standby consumption, which represents about 11% of the total consumption, is embedded in all end-uses, but is mostly concentrated in office equipment (i.e., information and communication technologies, including internet connection) and in entertainment devices (i.e., consumer electronics).

A monitoring study of the electricity consumption in 72 dwellings carried out in the UK over a two years period showed that the mean annual electricity consumption of the households increased significantly by 4.5% between the first year and the second year of the monitoring. The overall increase was attributed to a 10% increase in the consumption of standby appliances [77]. According to the EcoDesign EuP Lot 6 study [78], the business

as usual standby consumption by 2020 will amount to around 100 TWh without implementing measure.

1.6.2 Daylighting

The usage of daylight instead of artificial light is related to different factors mainly dependent on the dwelling physical structure: the building orientation influences the time the dwelling is exposed to sun light; the number of windows and their area affect the light penetration; shading devices (i.e. balconies, trees) shade the windows from direct sun but allow diffuse daylight to be admitted. Mardaljevic et al. [79] analysed the potential for increased daylight provision for a house with or without skylight to save electric lighting energy at eight European locations. The study shows that householders who use daylight as much as possible are estimated to reduce the need for artificial lighting up to 16 – 20% [80], depending on dwelling factors above mentioned.

1.6.3 Set point temperature

The set point temperature is the temperature at which users set a thermostat or air conditioner to reach thermal comfort. The acceptance of a higher indoor temperature in summertime conditions and a lower one during winter can lead to a reduction in energy consumption. For example, as found by Al-Sanea et al. [71], changing yearly-fixed thermostat settings such as 21 – 24.1°C (winter/summer) to optimised monthly fixed settings like 20.1 – 26.2°C introduces a yearly energy reduction of 26.8 – 33.6%.

It is not easy to understand which is the temperature range in which thermal comfort is granted. For instance, according to Yang et al. [70], for an indoor relative humidity of 50%, the temperature range is approximately from 20°C (lower limit in winter) to just over 27°C (upper limit in summer) and may vary of 1 – 2°C in case of higher or lower humidity.

1.6.4 Rebound effect

The rebound effect is not properly a behavioral factor influencing electricity consumption. It can more accuratley be described as a side-effect of the introduction of policy, market and/or technology interventions aimed at environmental efficiency improvements, especially where the efficiency gains bring reduced costs in electricity consumption [64]. In such cases, energy saving generated through energy efficiency measures is taken back by consumers in the form of higher consumption, either by increasing the quantity of energy used (for instance to increase their comfort level) or due to a higher quality of energy service [65].

There are three types of rebound effect [65]:

- **direct**: where increased efficiency and associated cost reduction for a product/service results in its increased consumption because of cheaper prices;
- **indirect:** where savings from efficiency cost reductions enable more income to be spent on other electric products and services;
- **economy-wide:** where more efficiency drives economic productivity overall, resulting in more economic growth, and hence additional consumption at a macroeconomic level.

The *direct* rebound effect is the only one we are interested in, because *indirect* and *economy-wide* ones are difficult to be estimated and fall outside our domain.

In literature there is a disagreement about the magnitude of the direct rebound effect. Some authors [66,67] assess that the extent of the take-back effect is so big to erode most of the energy savings. Although, according to Geller et al. [68], Greening et al. [69] and the empirical evidences they overviewed, the size of the direct rebound effect is not so high to neutralize the efforts performed with energy efficiency measurements. In the following, a table describes the estimated size of rebound effect by technology:

End Use	Size of rebound-effect
Space heating	10 – 30%
Space cooling	0 – 50%
Water-heating	10 – 40%
Lighting	5 – 12%
Appliances	0%

Table 72: Summary of Empirical Evidence of the Rebound Effect in the United States [69]

1.7 Conclusion

Literature review identified four main sectors which will be used as basis to compare energy consumption patterns for users.

Power Aware will first take into account the geographic position of users' house, as this highly affects variance of electricity demand. Secondly, we will consider the number and the characteristics of the set of appliances owned since they are strong discriminating factors, impacting directly on the energy request due to their dimension and efficiency. Finally, we will consider dwelling and socio-economic related factors which indirectly influence energy demand, for example affecting usage frequency.

In the following deliverables we will define the algorithms and techniques to group users with similar characteristics together.

1.8 Appendix A

Overall building heat loss coefficient $U' = \frac{A \cdot U + \frac{1}{3} \cdot N \cdot V}{1000}$ [84], where:

- A is the component area (m^2),
- U is the fabric value ($W\ m^{-2}\ K^{-1}$),
- N is the air infiltration rate in air changes per hour (h^{-1}),
- V is the volume of the space (m^3).

1.9 Bibliography

[1] A. Kavousian, R. Rajagopal, M. Fischer. A Method to Analyze Large Data Sets of Residential Electricity Consumption to Inform Data-Driven Energy Efficiency, CIFE Working Paper # WP1 30 June 2012.

[2] Yan Y. Y., 1998: Climate and residential electricity consumption in Hong Kong. Energy, 23, 17–20.

[3] Engle, R. F., C. Mustafa, and J. Rice, 1992: Modelling peak electricity demand. J. Forecasting, 11, 241–251.

[4] Li, X., and D. J. Sailor, 1995: Electricity use sensitivity to climate and climate change. World Res. Rev., 7, 334–346.

[5] Cancelo, J. R., and A. Espasa, 1996: Modelling and forecasting daily series of electricity demand. Invest. Econ., 20, 359–376.

[6] Chen, D., Wang, X., Ren, Z., 2012. Selection of climatic variables and time scales for future weather preparation in building heating and cooling energy predictions. Energy Buildings 51, 223–233.

[7] Michael C. Baechler, Jennifer Williamson, Theresa Gilbride, Pam Cole, and Marye Hefty and Oak ridge National Laboratory, Pat M. Love, Guide to determining Climate Regions by County, Pacific Northwest National Laboratory, August 2010

[8] Bedir M, Hasselaar E, Itard L. Determinants of electricity consumption in Dutch dwellings. Energy Build 2013;58:194–207.

[9] McLoughlin F, Duffy A, Conlon M. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: an Irish case study. Energy Build 2012;48:240–8.

[10] Sanquist TF, Orr H, Shui B, Bittner AC. Lifestyle factors in U.S. residential electricity consumption. Energy Policy 2012;42:354–64.

[11] Baker KJ, Rylatt RM. Improving the prediction of UK domestic energy-demand using annual consumption-data. Appl Energy 2008;85(6):475–82.

[12] Tso GKF, Yau KKW. Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks. Energy 2007;32 (9):1761–8.

[13] Wiesmann D, Lima Azevedo I, Ferrão P, Fernández JE. Residential electricity consumption in Portugal: findings from top-down and bottom-up models. Energy Policy 2011;39(5):2772–9.

[14] Druckman A, Jackson T. Household energy consumption in the UK: a highly geographically and socio-economically disaggregated model. Energy Policy 2008;36(8):3177–92.

[15] Kavousian A, Rajagopal R, Fischer M. Determinants of residential electricity consumption: using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. Energy 2013;55:184–94.

[16] Brounen D, Kok N, Quigley JM. Residential energy use and conservation: economics and demographics. Eur Econ Rev 2012;56(5):931–45.

[17] Bartiaux F, Gram-Hanssen K. Socio-political factors influencing household electricity consumption: a comparison between Denmark and Belgium. In: Proceedings of the ECEEE 2005 Summer Study, European Council for an Energy Efficient Economy; 2005. 1313–1325.

[18] Tiwari P. Architectural, demographic, and economic causes of electricity consumption in Bombay. J Policy Model 2000;22(1):81–98.

[19] Parti M, Parti C. The total and appliance-specific conditional demand for electricity in the household sector. Bell J Econ 1980;11(1):309–21.

[20] Zhou S, Teng F. Estimation of urban residential electricity demand in China using household survey data. Energy Policy 2013;61:394–402.

[21] Larsen BM, Nesbakken R. Household electricity end-use consumption: results from econometric and engineering models. Energy Econ 2004;26(2):179–200.

[22] Filippini M, Pachauri S. Elasticities of electricity demand in urban Indian households. Energy Policy 2004;32(3):429–36.

[23] Gram-Hanssen K, Kofod C, Petersen KN. Different everyday lives: different patterns of electricity use. In: Proceedings of the ACEEE 2004 Summer Study, American Council for an Energy Efficient Economy; 2004. 7:74–85.

[24] Ndiaye D, Gabriel K. Principal component analysis of the electricity consumption in residential dwellings. Energy Build 2011;43(2–3):446–53.

[25] Wyatt P. A dwelling-level investigation into the physical and socio-economic drivers of domestic energy consumption in England. Energy Policy 2013;60:540–9.

[26] Bartusch C, Odlare M, Wallin F, Wester L. Exploring variance in residential electricity consumption: household features and building properties. Appl Energy 2012;92:637–43.

[27] Yohanis YG, Mondol JD, Wright A, Norton B. Real-life energy use in the UK: how occupancy and dwelling characteristics affect domestic electricity use. Energy Build 2008;40(6):1053–9.

[28] Parker DS. Research highlights from a large scale residential monitoring study in a hot climate. Energy Build 2003;35(9):863–76.

[29] Genjo K, Tanabe S, Matsumoto S, Hasegawa K, Yoshino H. Relationship between possession of electric appliances and electricity for lighting and others in Japanese households. Energy Build 2005;37(3):259–72.

[30] Lam JC. Climatic and economic influences on residential electricity consumption. Energy Conserv Manage 1998;39(7):623–9.

[31] Nielsen L. How to get the birds in the bush into your hand: results from a Danish research project on electricity savings. Energy Policy 1993;21(11):1133–44.

[32] Leahy E, Lyons S. Energy use and appliance ownership in Ireland. Energy Policy 2010;38(8):4265–79.

[33] Hamilton IG, Steadman PJ, Bruhns H, Summerfield AJ, Lowe R. Energy efficiency in the British housing stock: energy demand and the Homes Energy Efficiency Database. Energy Policy 2013;60:462–80.

[34] Santamouris M, Kapsis K, Korres D, Livada I, Pavlou C, Assimakopoulos MN. On the relation between the energy and social characteristics of the residential sector. Energy Build 2007;39(8):893–905.

[35] Summerfield AJ, Lowe RJ, Bruhns HR, Caeiro JA, Steadman JP, Oreszczyn T. Milton Keynes Energy Park revisited: changes in internal temperatures and energy usage. Energy Build 2007;39(7):783–91.

[36] Cramer JC, Miller N, Craig P, Hackett BM. Social and engineering determinants and their equity implications in residential electricity use. Energy 1985;10(12):1283–91.

[37] Carter A., Craigwell R., Moore W., Price reform and household demand for electricity. J Policy Model 2012;34(2):242–52.

[38] Louw K, Conradie B, Howells M, Dekenah M. Determinants of electricity demand for newly electrified low-income African households. Energy Policy 2008;36(8):2812–8.

[39] Tso GKF, Yau KKW. A study of domestic energy usage patterns in Hong Kong. Energy 2003;28(15):1671–82.

[40] Halvorsen B, Larsen BM. Norwegian residential electricity demand—a micro-economic assessment of the growth from 1976 to 1993. Energy Policy 2001;29 (3):227–36.

[41] Haas R, Biermayr P, Zoechling J, Auer H. Impacts on electricity consumption of household appliances in Austria: a comparison of time series and cross-section analyses. Energy Policy 1998;26(13):1031–40.

[42] Chong H. Building vintage and electricity use: old homes use less electricity in hot weather. Eur Econ Rev 2012;56(5):906–30.

[43] Munley VG, Taylor LW, Formby JP. Electricity demand in multi-family, renter occupied residences. Southern Econ J 1990;57(1):178–94.

[44] Rory V. Jones, Alba Fuertes, Kevin J. Lomas. The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings Renewable and Sustainable Energy Reviews 43 (2015) 901–917.

[45] J. Seryak, Dr. Kelly Kissock, Occupancy and behavioral effects on residential energy use, Department of Mechanical and Aerospace Engineering, University of Dayton, 300 College Park.

[46] COMMISSION DELEGATED REGULATION (EU) No 1060/2010 of 28 September 2010 supplementing Directive 2010/30/EU of the European Parliament and of the Council with regard to energy labelling of household refrigerating appliances. EU (2010)

[47] COMMISSION DELEGATED REGULATION (EU) No 1061/2010 of 28 September 2010 supplementing Directive 2010/30/EU of the European Parliament and of the Council with regard to energy labelling of household washing machines Text with EEA relevance. EU (2010)

[48] COMMISSION DELEGATED REGULATION (EU) No 1059/2010 of 28 September 2010 supplementing Directive 2010/30/EU of the European Parliament and of the Council with regard to energy labelling of household dishwashers

[49] G. Wood, M. Newborough, Dynamic energy-consumption indicators for domestic appliances: environment, behaviour and design, Energy and Buildings, volume 35, Issue 8, 2003

[50] COMMISSION DELEGATED REGULATION (EU) No 626/2011, 4 May 2011, supplementing Directive 2010/30/EU of the European Parliament and of the Council with regard to energy labelling of air conditioners.

[51] Aníbal de Almeida, Paula Fonseca, Barbara Schlomann, Nicolai Feilbergd, Characterization of the household electricity consumption in the EU, potential energy savings and specific policy recommendations, Energy and Buildings, Volume 43, Issue 8, August 2011, Pages 1884-1894

[52] IEA, Cool Appliances, Policy strategies for Energy Efficient Homes, OECD/IEA 2003, 2009

[53] Bertoldi P, Atanasiu B. Residential lighting consumption and saving potential in the enlarged EU. In: 4th international conference on energy efficiency in domestic appliances and lighting (EEDAL).

[54] EIA, https://www.eia.gov/tools/faqs/faq.php?id=96&t=3, Estimated U.S. residential sector electricity consumption by end use, 2016

[55] P. Bertoldi, B. Hirl, N. Labanca, Energy efficiency status report 2012, Electricity consumption and efficiency trends in the EU-27

[56] Probert, D., Newborough, M. (1985). Designs, thermal performances and other factors concerning cooking equipment and associated facilities, Applied Energy: an international journal, Volume 21, Issues 2-3, p. 81-222, London : Elsevier Applied Science.

[57] Lawson, F. R. (1983). Energy use in the catering industry of Great Britain, Energy Technology Support Unit, Harwell, Oxon.

[58] BRE, Energy conservation: a study of energy consumption in buildings and possible means of saving energy in houses, Working Party Report, Current Paper 56, Building Research Establishment, Garston, 1975.

[59] Vale, R., Vale, B. (1980).The self-sufficient house, Book Club Associates, London, p. 113.

[60] Singer, D. D., Smart, G. A., Hunt, R. O. (1976). Energy consumption in kitchens and catering establishments, Food Paper 3, Ministry of Agriculture, Fisheries and Food, UK.

[61] Erickson, R. C. (1977). Energy efficiency program for kitchen ranges and ovens, prepared for the Assistant Secretary for Conservation and Solar Applications, Dept. Of Energy, Washington, DC.

[62] Anon., Induction Reduction, Catering (November 1984), p. 99.

[63] The English Induction Cooking Company Ltd, Charlton, London, private communication, 1984.

[64] European commission DG ENV, D. Maxwell, P. Owenm, L. McAndrew, S. Mudgal, F. Cachia, K. Muehmel, A. Neubauer, J. Tröltzsch, Addressing the rebound effect, 2011

[65] European Environment Agency, A.-D. Barbu, N Griffiths, G. Morton, Achieving energy efficiency through behaviour change: what does it take?, 2013

[66] Khazzoom, J.D. 1987. Energy Savings Resulting from the Adoption of More Efficient Appliances. The Energy Journal 8(4): 85-89.

[67] Inhaber, H. 1997. Why Energy Conservation Fails. Westport, CT: Quorum Books.

[68] Geller H., P. Harrington, M.D. Levine, A.H. Rosenfeld and S. Tanishima. 2005. “Policies for Increasing Energy Efficiency: Thirty Years of Experience in OECD Countries.” Manuscript accepted for publication in Energy Policy.

[69] Greening, L.A., D.L. Greene and C. Difiglio. 2000. “Energy efficiency and consumption – the rebound effect – a survey.” Energy Policy 28(6-7): 389-401.

[70] Liu Yang, Haiyan Yan, Joseph C. Lam, Thermal comfort and building energy consumption implications – A review, Applied Energy 115 (2014): 164-173.

[71] Al-Sanea SA, Zedan MF. Optimized monthly-fixed thermostat-setting scheme for maximum energy-savings and thermal comfort in air-conditioned spaces. Appl Energy 2008;85:326–46.

[72] https://www.eea.europa.eu/data-and-maps/indicators/heating-degree-days

[73] Ben Anderson, Sharon Lin, Andy Newing, AbuBakr Bahaj, Patrick James, Electricity consumption and household characteristics: implications for census taking in a smart metered future, Computers, Environment and Urban Systems 63 (2017) 58-67

[74] COMMISSION DELEGATED REGULATION (EU) No 1062/2010 of 28 September 2010 supplementing Directive 2010/30/EU of the European Parliament and of the Council with regard to energy labelling of televisions

[75] Lauren Mattison and Dave Korn,, Dehumidifiers: A Major Consumer of Residential Electricity, The Cadmus Group, Inc., May 9 2012

[76] K.J. Baker, R.N. Rylatt, Improving the prediction of UK domestic energy-demand using annual consumption-data, Applied Energy 85 (2008) 475-482.

[77] S. Firth, K. Lomas, A. Wright, R. Wall, Identifying trends in the use of domestic. Appliances from household electricity consumption measurements, Energyand Buildings 40 (2008) 926–936

[78] EuP Lot 6, Report for tender no. TREN/D1/40 Lot 6 – EuP Lot 6 – Task 7 + Task 8, 2007

[79] J. Mardaljevic, M. Andersen, N. Roy, J.Christoffersen, Daylighting metrics for residential buildings, 27th Session of the CIE, Sun City, South Africa, July 11-15, 2011.

[80] http://www.velux.com/deic/daylight/benefits-of-daylight

[81] www.eia.gov/consumption/residential/data, RECS 2015

[82] https://en.wikipedia.org/wiki/Heating_degree_day

[83] BizEE Energy Lens - Energy management made easy, Degree Days - Handle with Care! , http://www.energylens.com/articles/degree-days

[84]K. P. Moustris, P. T. Nastos, A. Bartzokas, I. K. Larissi, P. T. Zacharia, A. G. Paliatsos, Energy consumption based on heating/cooling degree days within the urban environment of Athens, Greece, Theoretical and Applied Climatology, November 2014

[85] http://www.degreedays.net/

2 Deliverable 2

List of data privacy issues (according to European laws) and countermeasures needed

Author and Project Manager: Diego Mariani
Scientific coordinator: Dr. Antonio Vetrò

2.1 Introduction

Since the project will require the collection of users' behaviour and socio-economic data, Power Aware will have to comply with all European and national legislation and directives relevant to the Country where the data collection is taking place. The collection, processing and transmission of personal data will therefore be analysed under the principles of:

1. The Universal Declaration of Human Rights and the Convention 108 for the Protection of Individuals with Regard to Automatic Processing of Personal Data;
2. Directive 95/46/EC & Directive 2002/58/EC of the European parliament regarding issues with privacy and protection of personal data and the free movement of such data;
3. The local national laws applying their provisions. Any additional regulations at national level that do not fall under the Directive and apply to data protection or any other sensitive information will also be taken into account for Power Aware action development.

Data managed during the project will be processed only under the following preconditions which need to be met (Art. 7, Directive 95/46/EC):

- When the data subject has given her/his consent
- When the processing is necessary for the performance of or the entering into a contract
- When processing is necessary for compliance with a legal obligation
- When processing is necessary in order to protect the vital interests of the data subject.

To this end, personal data managed by the project will be anonymized and stored in a form which does not permit identification of users. Data processing in the project will be compliant with the purposes for which the data were collected or for which they are further processed. Power Aware will establish a data management framework that guarantees security of collected personal data from potential abuse, theft, or loss. Behavioural and users socio-economic profiles developed in the course of the project will be anonymized. A Data Management Plan (DMP) will detail what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated & preserved.

2.2 Ethics and Privacy

The project is aware of the importance of privacy protection in monitoring of everyday life and energy use, ensuring privacy will be emphasized throughout the project.

In 2012 the European Data Protection Supervisor recommended the need to assure "*direct access to consumers to their energy usage data, as well as disclosure to them of their individual profiles and the logic of any algorithms used for data mining and information on remote on/off functionality*". This is one of Power Aware's objectives.

2.2.1 Law compliance

Power Aware must be compliant with European and national laws. Main rules and legal domains relevant for Power Aware are defined by:

- Regulation (EU) 2016/679, Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, General Data Protection Regulation (GDPR), – in particular Articles 4 and 7 on consent, Article 22 on automated decision-making and profiling, Article 25 on privacy by design, Article 33 on data breaches, Article 18 on data portability, …) that will be applied from May 25, 2018;

- Regulation (EU) 1227/2011 on wholesale energy market integrity and transparency (REMIT) (and the Proposal for a Regulation on the Governance of the Energy Union, Proposal com (2016) 759 final 2;

- Directive 2006/32/EC on energy end-use efficiency and energy services.

For a proactive and effective protection of privacy and personal data, the projetc will take into consideration the recommendation made by the Article 29 Working Party and by the European Data protection Supervisor (EDPS). In particular:

- Working Party Art. 29, Opinion 12/2011 on smart metering (wp 183)

- Working Party Art. 29, Opinion 07/2013 on the Data Protection Impact Assessment Template for Smart Grid and Smart Metering Systems ('DPIA Template') prepared by Expert Group 2 of the Commission's Smart Grid Task Force (wp209)

- Working Party art. 29, Opinion 8/2014 on the recent developments of Internet of Things (wp223)

- European Data Protection Supervisor, Opinion on the Commission Recommendation on preparations for the roll-out of smart metering systems (adopted on June 8, 2012)

- European Data Protection Supervisor, Opinion 8/2016 on coherent enforcement of fundamental rights in the age of big data.

2.2.2 Ethical principles

Power Aware will be compliant with the requirements outlined in the Horizon 2020 (The EU framework programme for research and innovation) Guide “How to complete your ethics Self-Assessment” (http://ec.europa.eu/research/participants/portal/doc/call/h2020/h2020-msca-itn-2015/1620147-h2020_-_guidance_ethics_self_assess_en.pdf). With particular reference to the Personal Data section (pages 17 to 22).

2.2.3 Data management plan

Following the Guidelines on FAIR Data Management in Horizon 2020 (http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf), Power Aware will draft a Data Management Plan.

2.2.4 Privacy Law issues to be stressed

Power Aware project will analyze and take care of the following issues:
- freely given and valid consent to data processing
- data quality (inter alia purpose limitation principle, data minimization)
- data integrity
- restriction to access to data by third parties
- data security
- profiling and risks to private life
- anonymization (possibility)
- privacy by design
- transparency of the data processing
- auditability
- data subject full control of data throughout the entire data lifecycle

2.2.5 Other legal constraints and issues

For a fair data processing, Power Aware will also have to respect all Intellectual Property Rights limitations, with particular focus on data ownership (private data from energy companies, personal data from users, etc.).

2.3 Bibliography

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

[2] Regulation (EU) No 1227/2011 of the European Parliament and of the Council of 25 October 2011 on wholesale energy market integrity and transparency

[3] Working Party Art. 29, Opinion 12/2011 on smart metering (wp 183)

[4] Next privacy. Il futuro dei nostri dati nell'era digitale – L. Bolognini, D. Fulco, P-. Paganini – ETAS (2010)

[5] La nuova disciplina della privacy : commento al D. lgs. 30 giugno 2003, n.196 / diretto da Salvatore Sica e Pasquale Stanzione; con la collaborazione di Domenico Apicella [et al.] - Zanichelli (2004)

3 Deliverable 3

Model for the classification of clustering and machine learning algorithms

Authors: Marco Pietro Abrate - Emanuele Mottola
Project Manager: Diego Mariani
Scientific coordinator: Dr. Antonio Vetrò

3.1 Introduction

To perform any comparison among users, a grouping operation is always required. It is therefore necessary to adopt a mechanism to do this operation in the best possible way, using knowledge from both a topic of artificial intelligence (Machine Learning) and algorithms of data elaboration (Data Mining).

Machine Learning aims at programming computers to optimize a performance criterion using example data or past experience (Alpaydin, 2004), while Data Mining, on the other hand, can be defined as a non-trivial extraction of implicit, previously unknown and potentially useful information from data (Honkela, 1997).

This deliverable will mainly deal with the selection of the most efficient machine learning or data mining algorithms in order to deliver the most fitting result for our problem: how to group together users with similarities.

Before starting the analysis, it is necessary to take into account the type of data the project deals with. We can assume that Power Aware will have access to the set of parameters outlined in Deliverable 1, although such list should not be considered as a thorough set.

3.2 Machine Learning

Algorithms can certainly help in solving problems on a computer. For some tasks, however, an algorithm does not exist yet, for example, to tell spam emails from legitimate emails. We know what the input is: an email document. We know what the output should be: a yes/no answer indicating whether the message is spam or not but we do not know how to transform the input into the output, and we would like the computer to extract automatically the algorithm for this task.

We may not be able to identify the process completely (e.g. to recognise spam emails), but it is possible to construct a good and useful approximation and to detect patterns or regularities. This is the niche of machine learning (Alpaydin, 2004).

Machine learning is also used to extract further knowledge from a data set revealing relations between data structure, or useful information to perform *descriptive* or *predictive* activities (i.e. Marketing).

The machine learning area can be divided into four approaches (Alpaydin, 2004):

- Supervised learning
- Semi-supervised learning
- Unsupervised learning
- Reinforcement learning

Name	Goal	Labeled input	Application
Supervised learning	Learn a mapping function from input to output	Yes	Spam detection, speech pattern and handwriting recognition
Semi-supervised learning	Cost-efficient mixture of supervised and unsupervised learning	Mixture (more unlabeled)	Webpage classification, object recognition
Unsupervised learning	Finding regularities in the input from unlabeled data	No	Customer segmentation, search result grouping
Reinforcement learning	Finding an efficient way to achieve reward with trial and error	-	Chess game

Table 73: Machine learning algorithms classification (Alpaydin, 2004).

Semi-supervised learning represents an efficient way to deal with a big dataset where a lot of data are unlabeled and only few of them are labeled. This type of algorithm uses regression, clustering and classification (Chapelle, Schölkopf, Zien, 2006).

Reinforcement learning concerns the ability of software agents to make actions in order to maximize a numerical reward signal (Barto, 1998).

3.3 Supervised Learning

Supervised learning is the task of inferring a function from labeled training data (Mohri, Rostamizadeh, Talwalkar, 2012) that consist in a set of training examples. In supervised learning, each example pairs an input object (typically a vector) and a desired output value. A supervised learning algorithm analyzes the training data and produces a function which can be used for mapping new examples. For instance, if the problem is to decide if a car is a sport car or not, the output is either 0 (not a sport car) or 1 (a sport car). The algorithm should find patterns in the input data in order to create a general rule (function) for classifying the next items, like:

if (horse_power > 300 *and* acceleration_to_100km/h < 8s) *then* 1 *else* 0

Since a training set of labeled data is required for such an approach (Alpaydin, 2004), the main issues in connection of the usage of supervised learning are:

- *Bias-variance tradeoff*: a learning algorithm is *biased* for a particular input if, when trained on each of these data sets, it is systematically wrong when predicting the

correct output for ***x***. A learning algorithm has high *variance* for a particular input ***x*** if it predicts different output values when trained on different training sets. If the learning algorithm has low bias (flexible), it will fit each training data set differently, and hence have high variance. A key aspect of many supervised learning methods is that they are able to adjust this trade-off between bias and variance.

- *Function complexity*: if the true function is simple, then an inflexible learning algorithm with high bias and low variance will be able to learn it from a small amount of data. But if the true function is highly complex (e.g. involving complex interactions among many different input features), then it will only be learned from a very large amount of training data and by using a flexible learning algorithm with low bias and high variance (James, 2003).
- *Dimensionality of the input space*: if the input feature vectors have very high dimension, the learning problem can be difficult even if the true function only depends on a small number of those features. This is because the many "extra" dimensions can confuse the learning algorithm and cause it to have high variance (Brodely, Friedl, 1999).
- *Noise*: If the desired output values are often incorrect (because of human error or sensor errors), then the learning algorithm should not attempt to find a function that exactly matches the training examples. Attempting to fit the data too carefully leads to *overfitting*.

3.4 Unsupervised Learning

Consider a machine which receives some sequence of input data x_1, x_2, x_3, ..., where x_t is the sensory input at time t. In unsupervised learning the machine simply receives the inputs but obtains neither supervised target outputs (supervised learning) nor rewards from its environment (reinforcement learning).

Unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise. Two very simple classic examples of unsupervised learning are clustering and dimensionality reduction (Ghahramani, 2004).

3.4.1 Cluster Analysis

A particular technique of unsupervised learning is cluster analysis which groups data objects based only on information found in the data that describe the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering (Tan, Steinbach, Kumar, 2006).

3.5 Classification of Machine Learning Algorithms

In general, algorithms can be classified by the following criteria (Dietterich, Kong, 1995):

- Goal,
- Type,
- Complexity.

These criteria can also be applied to machine learning algorithms.

The *goal* can be related to the kind of question the researcher is asking (relational, casual or descriptive [Jha, 2014]). Furthermore, the *type* of the algorithm can be directly associated with the machine learning method it satisfies. This could be regression, classification, clustering or reinforcement learning. The *complexity* describes the cost of an algorithm (Saake, Sattler, 2014) and, in this context, represents the runtime.

Additionally, more criteria can be added to this classification:

- Learning bias (see chapter 3),
- Knowledge representation.

K*nowledge representation* means how the output is presented: as a decision tree (classification), a dendogram (hierarchical clustering), a clique in a graph (clustering), etc..

Lastly, we shall consider the requirements and the suitability concerning the *input data*. Some algorithms require a specific size of the data set (see Table 2 for a classification) to work correctly and some others are best suited for specific features of the input vector (e.g. numerical/categorical data).

Small database	**Medium database**	**Large database**
Fits in memory	Fits in a single server	Spread over multiple servers
No DBA[9] required	1 DBA required	2+ DBA required
$<10^5$ records	10^5 – 10^7 records	$>10^7$ records
<10 GB of data	10GB – 40GB of data	>40GB of data

Table 74: Data set sizes (ScaleDB. 2014)

The following table shows a scheme of the above algorithm classification criteria.

9 DBA: Database Administrator

Criterion	Values
Goal (question type)	Relational, casual, descriptive question
Machine learning type	Clustering, classification, regression ...
Complexity	Big-O notation
Learning bias	Numerical (statistical bias) or descriptive
Knowledge representation	Graph, table, dendogram, logical rule, tree
Suitability of the data input	Descriptive

Table 75: Classification scheme

To choose between supervised or unsupervised learning algorithms a researcher has to identify the characteristics of the data set and define a question type. Then, a simple rule can be applied: if inputs are unlabeled or the question is descriptive, then unsupervised learning is the most suitable choice. Otherwise, he should opt for supervised learning.

At this point of our analysis, it is clear that supervised learning will not serve our goal. The data used in Power Aware, gathered from families who consume electricity, is not labeled and it can not be considered as a training set for other samples. Hence, the most reasonable way to use these information in order to get some useful description of the input space seems to be unsupervised learning.

More specifically, the purpose is to group different people so that they can relate to each others and compare their power consumption habits. Thus, the branch of unsupervised learning which fits best is actually *clustering.*

In the following paragraph we will give an overview of clustering techniques that might be fit to our project.

3.6 Clustering Classification

Different approaches to clustering data can be explained with the help of the hierarchy shown in Figure 1.

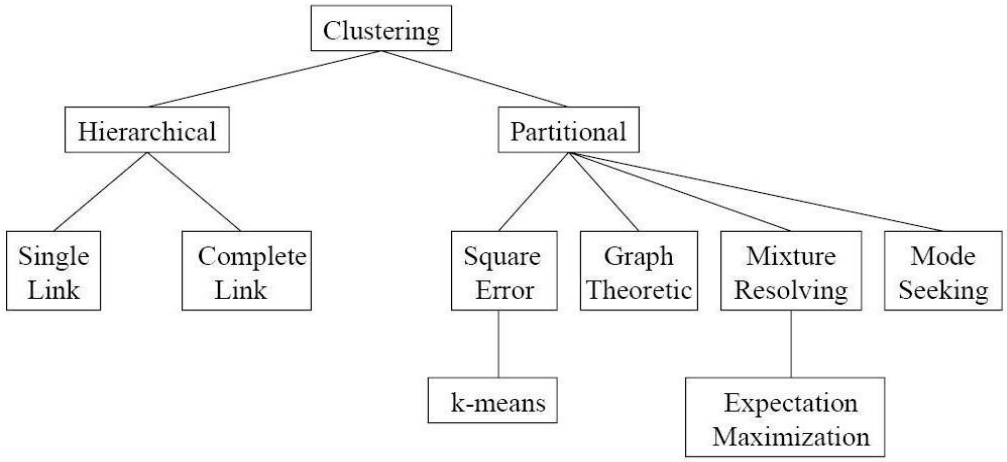


Figure 2: A taxonomy of clustering approaches (Jain, Murty, Flynn, 1999)

Partitional versus Hierarchical - A partitional clustering (unnested) is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is only in one subset. If we permit clusters to have subclusters, then we obtain a hierarchical clustering (nested), which is a set of nested clusters that are organized as a tree. Each node (cluster) in the tree, except for the leaf nodes, is the union of its children (subclusters), and the root of the tree is the cluster containing all the objects.

Agglomerative versus Divisive - An agglomerative approach begins with each pattern in a distinct (singleton) cluster, and successively merges clusters together until a stopping criterion is satisfied. A divisive method begins with all patterns in a single cluster and performs splitting until a stopping criterion is met.

Monothetic versus Polythetic - This aspect relates to the sequential or simultaneous use of features (of the input vectors) in the clustering process. Most algorithms are polythetic: all features enter simultaneously into the computation of distances between patterns, and decisions are based on those distances.

Hard versus fuzzy - A hard clustering algorithm allocates each pattern to a single cluster during its operation and in its output. A fuzzy clustering method assigns degrees (percentages) of membership in several clusters to each input pattern. A fuzzy clustering can be converted to a hard clustering by assigning each pattern to the cluster with the largest measure of membership (Jain, Murty, Flynn, 1999).

3.6.1 Clustering Selection

The fruitful literature of cluster analysis provides plenty of methods from which to choose but describes and distinguishes them primarily by their under-the-hood operation rather than by their qualitative results. However, clustering algorithms are often based on heuristics and justified by practical performance considerations rather than by formally proven guarantees, due to the difficulty of clustering problems (Ovelgönne, Geyer-Schulz, 2012).

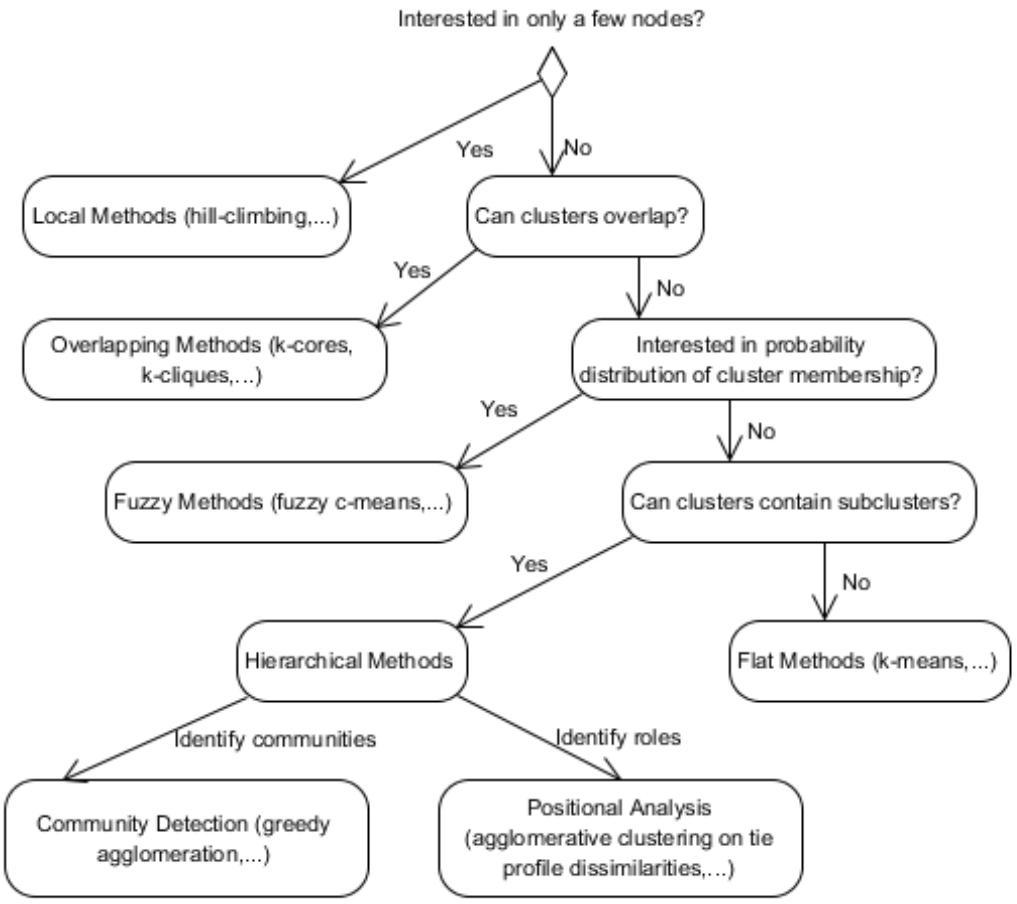


Figure 3: Clustering guideline (Pimentel, 2014)

Below we show a procedure in the attempt to summarize a general taxonomy for clustering algorithms (Pimentel, 2004).The series of questions represented highlights key attributes of clustering results that are well-known a priori and allows a researcher who plans to use clustering to eliminate wide class of inappropriate methods immediately.

A *priori clustering selection* works on the assumption that the best way to choose an algorithm is according to a decision tree as shown below.

In contrast in their 2011 paper, Justin Grimmer and Gary King introduce an idea that will be referred to as posteriori clustering selection. A *posteriori cluster selection* works by presenting an extremely long list of clusterings (ideally, all of them) as performed on the data set and letting the researcher choose the best one for his or her purposes (Grimmer, King, 2011). Here the choice among a wide range of methods remains but instead of having to choose between methods based on algorithm definitions, the researcher is able to determine the best method by examining the resulting clusters in the data according to his or her particular interest.

In order to answer the questions posed by the diagram, one must merge the a priori clustering classification with the techniques mentioned above.

In our data analysis we are neither interested in only a few nodes, because nobody has to be excluded, nor in an overlapping approach, because users belong to exclusive, well separated classes. Algorithms which involve a probability distribution are not taken into account since we are focused on a descriptive procedure for now.

Subclusters might be useful for some of the intentions of Power Aware, although the decision of which specific algorithm should be chosen will be argued further in next deliverables.

3.7 Conclusion

The branch of machine learning that will interest the project is unsupervised learning and, in particular, clustering algorithms as the most useful techniques to group people with similar characteristics together. Hence, in Deliverable 4, we will provide an overview of the main approaches to clustering.

3.8 Bibliography

Alpaydin, E. (2004). Introduction to machine learning: MIT press.

Barto, A. G. (1998). Reinforcement learning: An introduction: MIT press.

Brodely, C.E., Friedl, M.A. (1999). Identifying and Eliminating Mislabeled Training Instances, Journal of Artificial Intelligence Research 11, 131-167.

Chapelle, O., Schölkopf, B., Zien, A. (2006). Semi-supervised learning (Vol. 2): MIT press Cambridge.

Dietterich, T. G., Kong, E. B. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms: Technical report, Department of Computer Science, Oregon State University.

Ghahramani, Z. (2004). Unsupervised learning Advanced Lectures on Machine Learning (pp. 72-112): Springer.

Grimmer, J., King, G. (2011). General Purpose Computer-Assisted Clustering and Conceptualization. PNAS 108, 7: 2643-2650

Honkela, T. (1997). Data Mining and Document Modeling. Neural Networks Research Centre Helsinki University of Technology Tue Aug, 5.

Jain, Murty, Flynn (1999). Data clustering: a review.

James, G. (2003). Variance and Bias for General Loss Functions, Machine Learning 51, 115-135.

Jha, A. (2014). A Generic Knowledge-based Learning Tool and its Application to Fitness Training. (Master), Technische Universität München, Munich.

Mohri, M., Rostamizadeh, A., Talwalkar, A. (2012). Foundations of Machine Learning.

Ovelgönne, M, Geyer-Schulz, A. (2012). An Ensemble Learning Strategy for Graph Clustering. In Graph Partitioning and Graph Clustering, 187-206.

Pimentel, S. D. (2014). Choosing a Clustering: An A Posteriori Method for Social Networks. Journal of Social Structure, 15.

Rice, J. R. (1975). The algorithm selection problem.

Saake, G., Sattler, K.-U. (2014). Algorithmen und Datenstrukturen.

ScaleDB. (2014). Large Database, from http://www.scaledb.com/large-database.php

Tan, P. N., Steinbach, K., & Kumar, V. (2006). Data Mining Cluster Analysis: Basic Concepts and Algorithms: Boston: Pearson Addison Wesley.

Vapnik, V. N. (2000). The Nature of Statistical Learning Theory (2nd Ed.), Springer Verlag.

Wolpert, D. H., Macready, W. G. (1995). No free lunch theorems for search: Technical Report SFI-TR-95-02-010, Santa Fe Institute.

Xu, L., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2008). SATzilla: Portfolio-based Algorithm Selection for SAT. J. Artif. Intell. Res.(JAIR), 32, 565-606.

4 Deliverable 4

Classification of clustering algorithms from literature according to deliverable 3

Authors: Marco Pietro Abrate – Emanuele Mottola
Project Manager: Diego Mariani
Scientific coordinator: Dr. Antonio Vetrò

4.1 Introduction

In the previous chapter we stated that the project will deal with unlabeled data and, consequently, that clustering appears to be the best technique to elaborate them. Cluster analysis is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) based on similarity. Intuitively, patterns within a cluster are more similar to each other than they are to a pattern belonging to a different cluster (Jain, Murty, Flynn, 1999).

To better understand the difficulty of deciding what constitutes a cluster, consider Figure 1, which shows twenty points and three different ways of dividing them into clusters (Tan, Steinbach, Kumar, 2006).

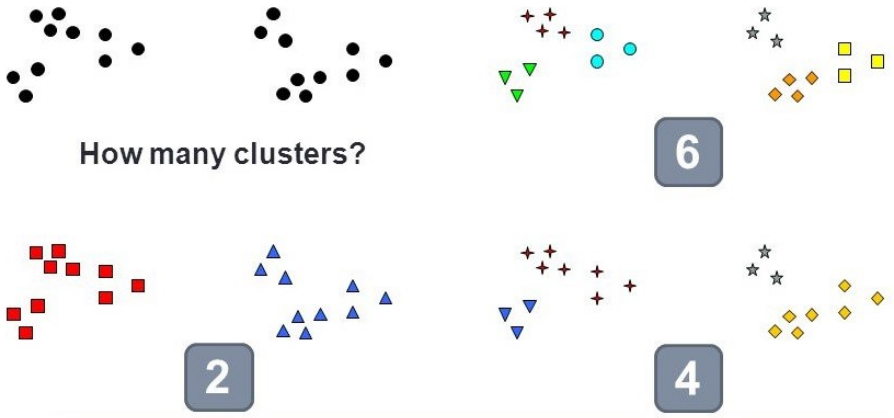


Figure 4: Different ways of clustering (Tan, Steinbach, Kumar, 2006)

This illustrates how a cluster can be best defined taking into account the nature of data and the desired results.

4.2 Types of Cluster

Clustering techniques can be classified by the type of clusters they generate. Hence, there are several different notions of a cluster that prove useful in practice. Here are some examples (Tan, Steinbach, Kumar, 2006):

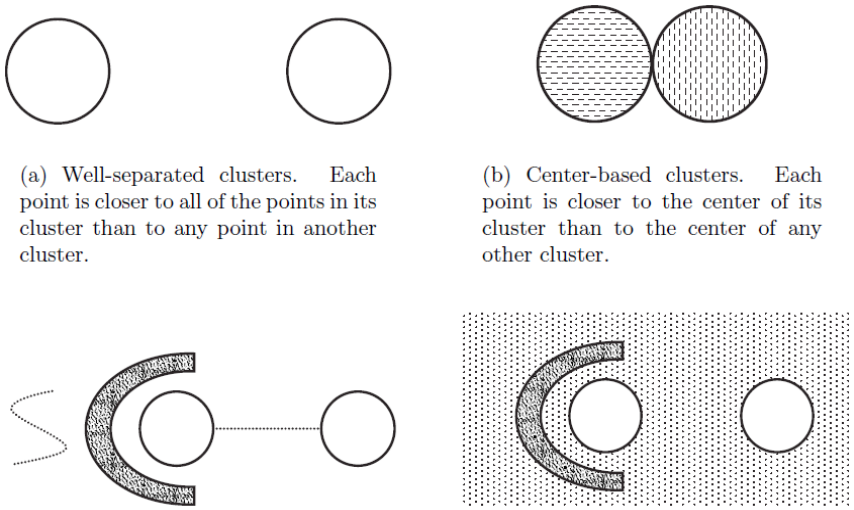
Well-Separated Figure 2(a). A cluster is a set of objects in which each object is closer (or more similar) to every other object in the cluster than to any object not in the cluster. This definition of a cluster is satisfied only when the data contains natural clusters that are quite far from each other.

Prototype-Based Figure 2(b). A cluster is a set of objects in which each object is closer (more similar) to the prototype that defines the cluster than to the prototype of any other

cluster. For data with continuous attributes, the prototype of a cluster is often a centroid, i.e., the average (mean) of all the points in the cluster.

Graph-Based Figure 2(c). If the data are represented as a graph, where the nodes are objects and the links represent connections among objects, then a cluster can be defined as a *connected component*. For example, a group of objects that are connected to one another, but that have no connection to objects outside the group. An important example of graph-based clusters are **contiguity-based clusters**, where two objects are connected only if they are within a specified distance of each other. This implies that each object in a contiguity-based cluster is closer to some other object in the cluster than to any point in a different cluster.

Density-Based Figure 2(d). A cluster is a high density region of objects that is surrounded by a region of low density. A density-based definition of a cluster is often employed when the clusters are irregular or intertwined, and when *noise* and *outliers* are present.



(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.

(d) Density-based clusters. Clusters are regions of high density separated by regions of low density.

Figure 5: (Tan, Steinbach, Kumar, 2006)

4.3 Components of a Clustering Task

Typical pattern clustering activity involves the following steps (Jain, Dubes, 1988):

1. pattern representation (feature selection and extraction),
2. definition of a pattern proximity measure appropriate to the data domain,
3. clustering or grouping,
4. data abstraction (if needed), and

5. assessment of output (if needed).

Pattern representation refers to the number of classes, the number of available patterns, and the number, type, and scale of the features available to the clustering algorithm.

Feature selection is the process of identifying the most effective subset of the original features to use in clustering. *Feature extraction* is the use of one or more transformations of the input features to produce new salient features.

Pattern proximity is usually measured by a distance function defined on pairs of patterns. A simple distance measure like Euclidean distance can often be used to reflect dissimilarity between two patterns.

The *grouping* step can be performed in a number of ways. The output clustering can be hard or fuzzy. *Hierarchical clustering* algorithms produce a nested series of partitions based on a criterion for merging or splitting clusters based on similarity. *Partitional clustering* algorithms identify the partition that optimizes (usually locally) a clustering criterion.

Data abstraction is the process of extracting a simple and compact representation of a data set.

Cluster validity analysis is the assessment of a clustering procedure's output. Often this analysis uses a specific criterion of optimality.

In Power Aware patterns consist of input with different features (see Deliverable 1) which are typically represented as multidimensional vectors (Duda, Hart, 1973). Data can be rendered through numerical values: categorical values, in fact, are converted to numerical ones in order to allow the definition and application of a distance function in the data space to measure the proximity among the points.

The grouping process will be discussed in the following paragraphs, but a fuzzy clustering will not be adopted for sure.

A representation of the provided data set can be obtained by plotting the normalized (from 0 to 1) patterns on a 2D or 3D Cartesian plane, in order to have a complete visual overview.

Cluster validity will be further discussed in the following deliverable.

4.4 Density-based Clustering

Density-based clustering locates regions of high density that are separated from one another by regions of low density by using several methodologies.

DBSCAN is a simple and effective density-based clustering algorithm and it follows a *center-based* approach. This algorithm is very effective in performing a first evaluation of

the data available, because it points out different density region, which are naturally present in the patterns, and discards noise elements. In this technique, density is estimated for a particular point in the data set by counting the number of points within a specified radius *Eps*.

The center-based approach allows us to classify a point in three different ways (Tan, Steinbach, Kumar, 2006):

Core points: These points are in the interior of a density-based cluster. A pattern is named as core if the number of points within a given neighbourhood around it, as determined by the distance function and the radius *Eps,* exceeds a certain threshold *minPts*.

Border points: A border point is not a core point, but falls within the neighbourhood of a core point.

Noise points: A noise point is any pattern that is neither a core point nor a border point.

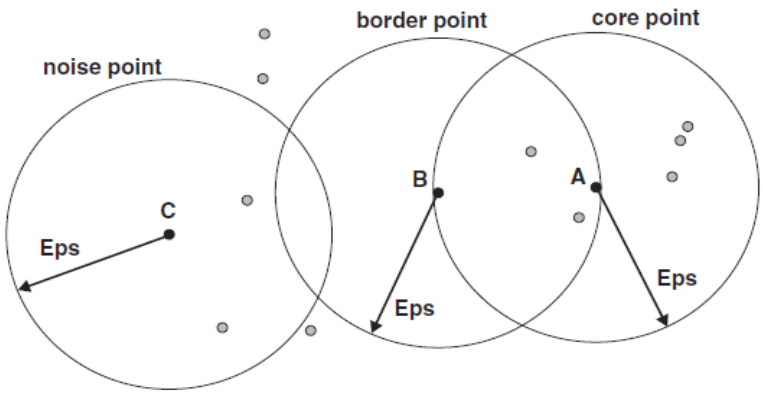


Figure 6: Core, border and noise points (Tan, Steinbach, Kumar, 2006)

4.4.1 DBSCAN Algorithms

Basic steps of the algorithm are:

1. Label all points as core, border, or noise points
2. Eliminate noise points
3. Put an edge between all core points that are within *Eps* of each other
4. Make each group of connected core points into a separate cluster
5. Assign each border point to one of the cluster of its associated core points.

The basic time complexity of this algorithm is $O(m^2)$, where m is the number of points. The space requirement, even for high-dimensional data, is $O(m)$ (Tan, Steinbach, Kumar, 2006).

4.4.2 Choice of Eps and minPts

There is not a fixed algorithm used to choose these two parameters. The basic approach is to look at the behaviour of the distance from a point to its k^{th} nearest neighbour, which we will call the k-dist.

If we compute the k-dist for all the data points for some k, sort them in increasing order (Figure 3, x axis of the graph), and then plot the sorted values of the k-dist (Figure 3, y axis), we expect to see a sharp change in the value of k-dist that corresponds to a suitable value of Eps. If we select this distance as the Eps parameter and take the value of k as the MinPts parameter, then points for which k-dist is less than Eps will be labeled as core points, while other points will be labeled as noise or border points.

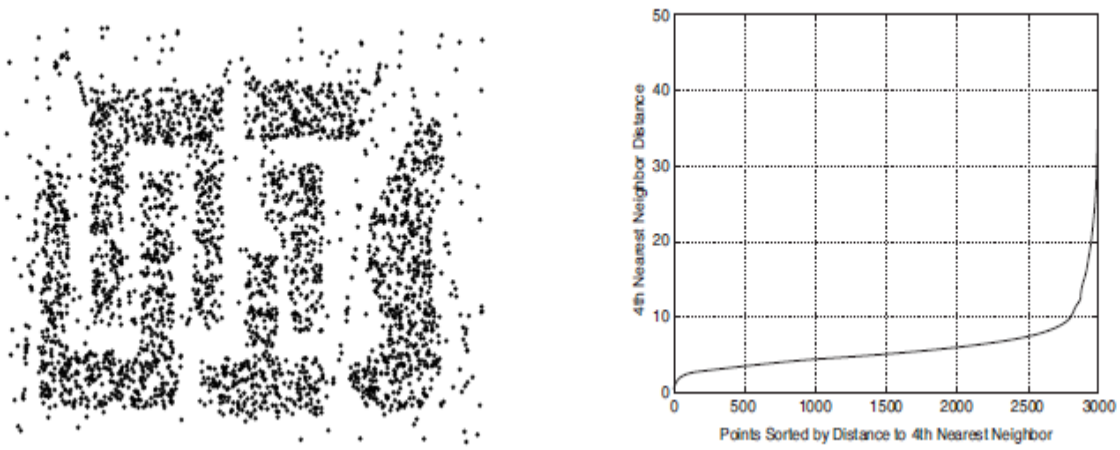


Figure 7: Sample data and k-dist graph (Tan, Steinbach, Kumar, 2006)

The value of Eps, determined in this way, depends on k, but does not change dramatically as k changes. If the value of k is too small, then even a small number of closely spaced points that are noise or outliers will be incorrectly labeled as clusters. If the value of k is too large, then small clusters (of size less than k) are likely to be considered as noise. The original DBSCAN algorithm used a value of k = 4, which appears to be a reasonable value for most two-dimensional data sets (Tan, Steinbach, Kumar, 2006).

As we can see in Figure 4, the sharp of the graph changes around a distance equal to 10. For this reason it is taken Eps = 10. Result in Figure 5.

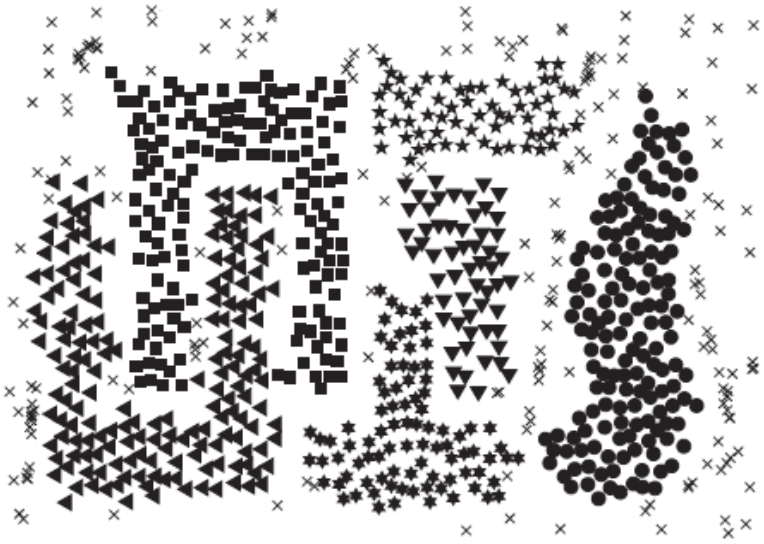


Figure 8: Result of DBSCAN with Eps=10 and minPts=4 (Tan, Steinbach, Kumar, 2006)

In Power Aware the same task can be performed, so to obtain suitable values of the radius (Eps) and the minimum number of points (minPts) in order to set the parameters for the elaboration with DBSCAN.

After executing this density-based algorithm, which leads to the decision of the correct number of clusters, it is necessary to apply some partitional or hierarchical clustering techniques.

4.5 Partitional Clustering

Partitional techniques usually produce a well defined number of clusters by optimizing a *criterion function* defined either locally (on a subset of the input vectors) or globally (over all the input vectors). Therefore, in practice, the algorithm is typically run multiple times with different starting states. The most intuitive and frequently used criterion function to identify the best configuration is the *squared error criterion*, which tends to work well with isolated and compact clusters. (Jain, Murty, Flynn, 1999).

These techniques are used after the execution of a pre-clustering algorithm like the before mentioned DBSCAN, so that the number of clusters (K in the K-means algorithm) is already known.

The K-means is the simplest and most commonly used algorithm employing a squared error criterion.

4.5.1 K-means

This algorithm defines a prototype in terms of a *centroid*, which is usually the mean of a group of points, and it is typically applied to objects in a continuous n-dimensional space, like the one of Power Aware.

It starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centroid until a convergence criterion is met (Tan, Steinbach, Kumar, 2006).

For example, the basic algorithm is:

1. Select K (decided by the user) points as initial centroids
2. **repeat**
3. Form K clusters by assigning each point to its closest centroid
4. Recompute the centroid of each clustering
5. **until** Centroids do not change.

The k-means algorithm is popular because it is easy to implement and its time complexity is $O(n)$, where n is the number of patterns. A major problem with this method is that it is sensitive to the selection of the initial partition and may converge to a local minimum if the initial partition is not properly chosen. For this reason it is necessary to compute the algorithm multiple times, in order to identify the best initial partition to be used.

4.5.2 Choosing the number of clusters

Given any k integer number, partitional clustering will always find k centroids, whether they really are meaningful groups or they are imposed by the method we use. There are various ways to fine-tune k (Alpaydin, 2004), beyond DBSCAN:

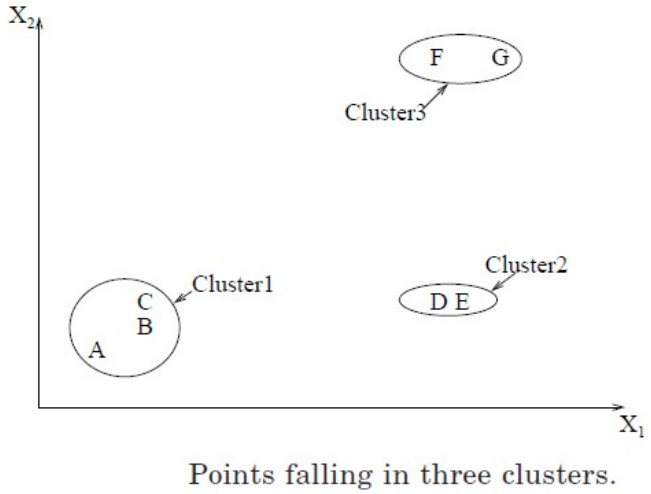
- In some applications , like in colour quantization, k is already defined.
- Plotting the data in two dimensions using *PCA* (Principal Component Analysis) may be used in uncovering the structure of data and the number of clusters.
- An incremental approach may be used, setting a maximum allowed distance between a centroid and an input vector.
- Validation of the groups can be done manually by checking whether clusters actually code meaningful groups of the data.

In the case of Power Aware it could be possible that the dataset is already naturally organised in different groups, which would lead to the clusters. This can be even more clear after the visualization of the points in a plot, as mentioned before (see chapter 3, Data abstraction).

4.6 Hierarchical Clustering

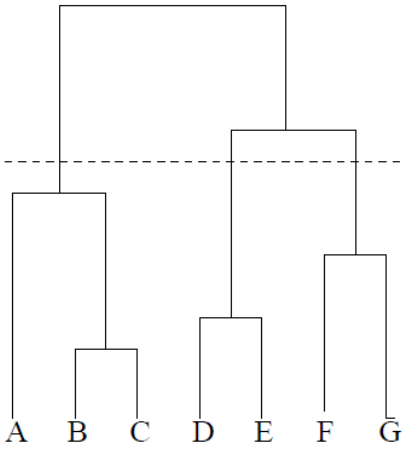
Another technique, different from K-means, is hierarchical clustering. There are two basic approaches for generating a hierarchical clustering: divisive and agglomerative. The last one is the most used because of its low complexity, $O(n^2)$ instead of $O(n^3)$.

This clustering approach refers to a collection of closely related clustering techniques that produce a hierarchical clustering by starting with each point as a singleton cluster and then repeatedly merging the two closest clusters until a single one remains (Tan, Steinbach, Kumar, 2006). The following figures (Figure 6, Figure 7) show two representations of the possible results obtained using a hierarchical clustering on seven patterns labelled A, B, C, D, E, F and G.



Points falling in three clusters.

Figure 10: (Jain, Murty, Flynn, 1999)



The dendrogram obtained using the single-link algorithm.

Figure 9: (Jain, Murty, Flynn, 1999)

A hierarchical algorithm yields a *dendogram* representing the nested grouping of patterns and similarity levels at which groupings change. The dendogram can be broken at different levels in order to obtain different clusterings of data. The same can be done in Power Aware, exploiting the K number of clusters already obtained (see 5.2).

Most hierarchical clustering algorithms are variants of *single-link*, *complete-link* or *group average* (see 6.2).

4.6.1 Basic Agglomerative Algorithm

Main steps of the algorithm are:

1. Compute the proximity matrix
2. **repeat**
3. Merge the closest two clusters
4. Update the proximity matrix

5. **until** Only one cluster remains.

4.6.2 Defining Proximity Between Clusters

It is the definition of cluster proximity that differentiates the various agglomerative hierarchical techniques. This requires the use of a similarity, or equivalently a distance, measure defined between instances. Generally Euclidean distance is used, which is a special case of the *Minkowski distance* (with p=2):

$$dist_m(\boldsymbol{x}^r, \boldsymbol{x}^s) = [\sum_{j=1}^{d} \left| x_j^r - x_j^s \right|^p]^{1/p}$$

where x is the input vector, d is the number of features in the vector and x is the single feature. The letters r and s are used to differentiate the two vectors. At each iteration of an agglomerative algorithm, we choose the two closest groups to merge.

In *single-link*, this distance is defined as the smallest distance between any possible pair of elements of the two groups G_i and G_j:

$$dist(G_i, G_j) = \min_{\boldsymbol{x}^r \in G_i,\, \boldsymbol{x}^s \in G_j} d(\boldsymbol{x}^r, \boldsymbol{x}^s)$$

while in *complete-link* the distance between two groups is taken as the largest distance between any possible pairs:

$$dist(G_i, G_j) = \max_{\boldsymbol{x}^r \in G_i,\, \boldsymbol{x}^s \in G_j} d(\boldsymbol{x}^r, \boldsymbol{x}^s)$$

In *group average* the distance is expressed as the average pairwise proximity among all pairs of points in the different clusters,

$$dist(G_i, G_j) = \frac{\sum_{\boldsymbol{x} \in C_i, \boldsymbol{y} \in C_j} d(\boldsymbol{x}^r, \boldsymbol{y}^s)}{m_i \cdot m_j}$$

where m_i and m_j are the sizes of G_i and G_j.

In Power Aware we can adopt the group average approach which is a good compromise between single and complete link, in fact it is neither sensitive to noise and outliers like the single one nor tends to break large clusters like the complete one. Its only limitation consists in showing bias towards globular clusters, but in that case Power Aware would use DBSCAN and K-means algorithms.

4.7 Conclusion

According to what we stated in this chapter, the following approach seems to be the most appropriate considering our goals:

1. Pre-clustering evaluation with DBSCAN and visual representation of the points, in order to get an approximation of the best number of clusters.
2. Clustering with K-means and hierarchical algorithms, in order to have different prospectives of the result.
3. Evaluation of the obtained clusters and, where not satisfying, return to the pre-clustering process.

4.8 Bibliography

Alpaydin, E. (2004). Introduction to machine learning: MIT press.

Duda, R. O., Hart, P. E. (1973). Pattern Classification and Scene Analysis: John Wiley and Sons, Inc., New York, NY.

Jain, A.K., Dubes, R.C. (1988). Algorithms for Clustering Data: Prentice-Hall advanced reference series.

Jain, Murty, Flynn (1999). Data clustering: a review.

Tan, P. N., Steinbach, K., & Kumar, V. (2006). Data Mining Cluster Analysis: Basic Concepts and Algorithms: Boston: Pearson Addison Wesley.

5 Deliverable 5

Evaluation of the algorithms

Authors: Marco Pietro Abrate – Emanuele Mottola
Project Manager: Diego Mariani
Scientific coordinator: Dr. Antonio Vetrò

5.1 Introduction

In the previous Deliverable, we indicated density-based, partitional and hierarchical clustering algorithms as best fit for our project. The purpose of this chapter is to study the validation process of these techniques and to evaluate the goodness of the results obtained from data elaboration in the project.

A key motive for using the cluster validation is that almost every algorithm will find groups in a data set, even if that data set has no natural cluster structure. In fact, although we assumed a predefined structure in the project data set, the same should not be taken for granted.

5.2 Cluster Validation

The following is a list of several important issues for cluster evaluation:

- Determining the *clustering tendency* of a set of data.
- Evaluating how well the results of a cluster analysis fit the data without reference to external information.
- Comparing the results of a cluster analysis to externally known results, such as externally provided class labels.
- Comparing two sets of clusters to determine which is better.

The evaluation measures, or indices, that are applied to judge various aspects of cluster validity are traditionally classified into the following three types:

Unsupervised: It measures the goodness of a clustering structure without considering external information. These measures are often divided into two further classes: *cluster cohesion*, which determine how closely related the objects in a cluster are and *cluster separation*, which determine how distinct or well-separated a cluster is from the others.

Supervised: It measures the extent to which the clustering structure discovered by clustering algorithm matches some external structure provided by the researcher. An example is *entropy*, which measures how well cluster labels match externally supplied class labels.

Relative: It compares different clusterings or clusters. A relative evaluation is a supervised or unsupervised evaluation that is used for the purpose of comparison. Hence, relative measures are not actually a separate type of cluster evaluation measure, but are instead a specific use of such measures.

In general, we can express overall cluster validity for a set of K clusters as a weighted sum of the validity of individual clusters:

$$overall\,validity = \sum_{i=1}^{K} w_i \; validity(C_i)$$

The *validity function* can be cohesion, separation, or a combination of both. In some case, the weights are simply 1 or the size of the cluster, while in other cases they reflect a more complicated property.

5.3 Components of a Clustering Task

5.3.1 Cohesion and separation

In order to explain cohesion and separation, it is useful to interpret the clustering result as a graph.

The cohesion of a cluster can be defined as the sum of the weights of the links in the proximity graph that connect points within the cluster. The proximity graph has data objects as nodes and the weights of the links are the proximities between two data objects: this proximity can be a similarity, a dissimilarity or a simple function of these quantities (Tan, Steinbach, Kumar, 2006).

$$cohesion(C_i) = \sum_{\substack{x \in C_i \\ y \in C_i}} proximity(x, y)$$

Likewise, the separation between two clusters can be measured by the sum of the weights of the links from points in one cluster to points in the other cluster.

$$separation(C_i, C_j) = \sum_{\substack{x \in C_i \\ y \in C_j}} proximity(x, y)$$

Another approach to calculate cohesion and separation is the prototype-based one.

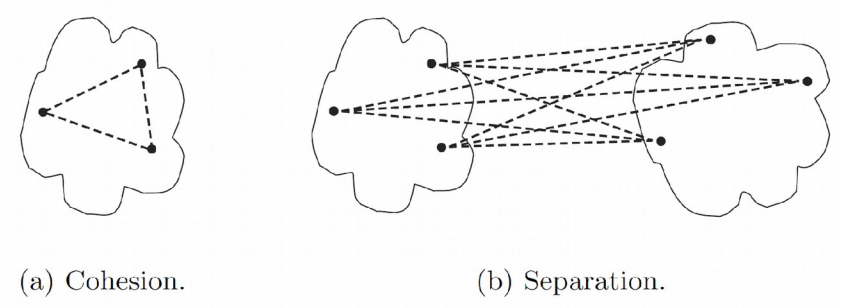


Figure 1: (Tan, Steinbach, Kumar, 1999)

In this case, the cohesion of a cluster can be defined as the sum of the proximities with respect to the prototype (centroid or medoid).

$$cohesion(C_i) = \sum_{x \in C_i} proximity(x, c_i)$$

where c_i is the prototype of the cluster C_i.

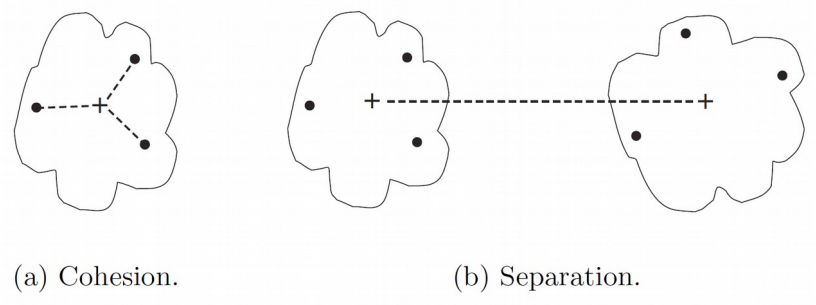
This is the cluster SSE (Sum of Squares Error) if we let proximity be the squared Euclidean distance.

Similarly, the separation between two clusters can be measured by the proximity of the two cluster prototypes. Two different equations are given.

$$separation(C_i, C_j) = proximity(c_i, c_j)$$

$$separation(C_i) = proximity(c_i, c)$$

where c_i is the prototype of cluster C_i and c is the overall prototype.



(a) Cohesion. (b) Separation.

Figure 2: (Tan, Steinbach, Kumar, 1999)

When proximity is measured by Euclidean distance, the traditional separation is the between clusters sum of squares (SSB). By summing the SSB over all clusters, we obtain the total SSB.

$$total\,SSB = \sum_{i=1}^{K} m_i \; dist_E(c_i, c)^2$$

where m_i is the size of the i^{th} cluster.

The Silhouette coefficient for individual objects and clusters

The popular way to calculate the silhouette coefficient combines both *cohesion* and *separation*. Here it is explained how to compute this coefficient for an individual point (distances are used here, but an analogue approach can work for similarities as well):

1. For the i^{th} object, calculate its average distance to all other objects in its cluster. Call this value A_i.
2. For the i^{th} object and any cluster not containing the object, calculate the object's average distance to all the objects in the given cluster. Find the minimum such value with respect to all clusters and call this value B_i.
3. For the i^{th} object, the silhouette coefficient is $S_i = (B_i - A_i) / \max(A_i, B_i)$.

This value can vary between -1 and 1. A negative value is undesirable because this corresponds to a case in which A_i is greater than B_i. The silhouette coefficient should be positive and the more A_i is close to 0 the better, since the coefficient assumes its maximum value of 1 when $A_i = 0$. We can compute the average silhouette coefficient of a cluster by simply taking the average of the silhouette coefficients of points belonging to the cluster (Kaufman, Rousseeuw, 1990).

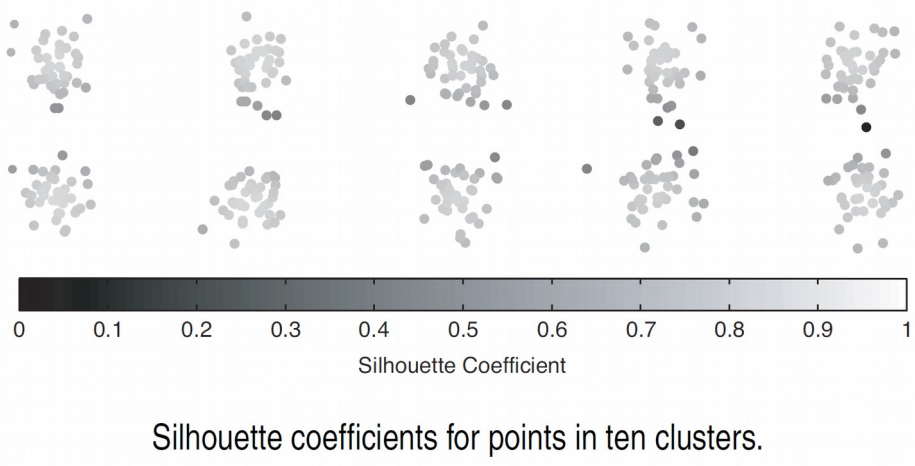


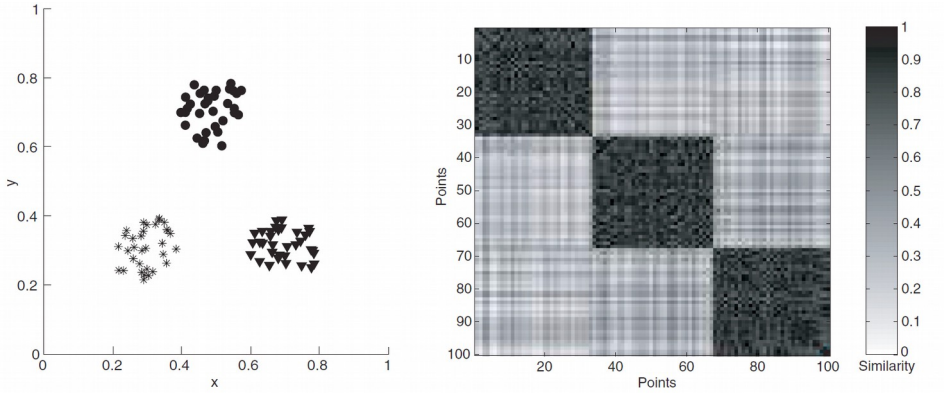
Figure 3: (Tan, Steinbach, Kumar, 1999)

5.3.2 Proximity Matrix

Given the similarity matrix for a data set and the cluster labels from a cluster analysis, we can evaluate the goodness of the clustering by looking at the *correlation* between the similarity matrix and an ideal version of the similarity matrix based on the labels previously found thanks to the clustering algorithms. The ideal similarity matrix is constructed by creating a matrix that has one row and one column for each data point and assigning a 1 to an entry if the associated pair of points belongs to the same cluster. All other entries are

0. Thus, if we sort the rows and columns so that all objects belonging to the same cluster are together, then an ideal similarity matrix has a *block diagram* structure, as it is shown in the figure below.

High correlation between the ideal and actual similarity matrices indicates that the points that belong to the same cluster are close to each other, while low correlation indicates the opposite. (Kaufman, Rousseeuw, 1990)



(a) Well-separated clusters.

(b) Similarity matrix sorted by K-means cluster labels.

Figure 4: (Tan, Steinbach, Kumar, 1999)

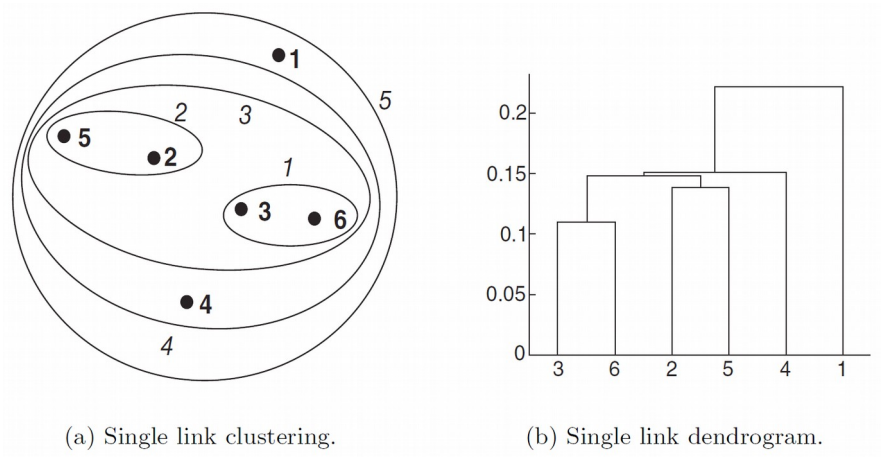
5.3.3 Hierarchical Clustering

The cophenetic distance is a popular measure for evaluating hierarchical clustering algorithms. This distance between two objects is the proximity at which an *agglomerative* hierarchical clustering technique puts the objects in the same cluster for the first time.

For example, Table 1 represents the cophenetic distance matrix for the single-link clustering shown in Figure 5.

Point	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*
P1	0	0.222	0.222	0.222	0.222	0.222
P2	0.222	0	0.148	0.151	0.139	0.148
P3	0.222	0.148	0	0.151	0.148	0.110
P4	0.222	0.151	0.151	0	0.151	0.151
P5	0.222	0.139	0.148	0.151	0	0.148
P6	0.222	0.148	0.110	0.151	0.148	0

Table 76: (Tan, Steinbach, Kumar, 1999)



(a) Single link clustering. (b) Single link dendrogram.

Figure 5: (Tan, Steinbach, Kumar, 1999)

The CoPhenetic Correlation Coefficient (CPCC) is the correlation between the entries of this matrix and the original *dissimilarity matrix* and it is a standard measure of how well a hierarchical clustering fits the data. One of the most common uses of this measure is to evaluate which type of hierarchical clustering is best for a particular type of data (Tan, Steinbach, Kumar, 2006).

5.3.4 Unsupervised evaluation in Power Aware

The previously mentioned methods make up the set of evaluation techniques that can be used in the project.

For the measurement of cohesion and separation it is necessary to find a suitable function for proximity, even if the most reasonable choice would be the simple Euclidean distance between points (dissimilarity). On the base of these two values the silhouette coefficient can be calculated in order to have another evaluation of the goodness of the clustering. Furthermore, similarity matrix is an optimal method of validation since it provides a comparison between the ideal values and the real ones. It can also be useful for the hierarchical algorithm because it leads to the CoPhenetic Correlation Coefficient (CPCC).

In the present state of knowledge, the unsupervised evaluation techniques are the most appropriate for the aim of Power Aware. Nonetheless, hereafter it is presented a brief overview of supervised validity methods because they might be useful in the future in case some previously clustered data will be available.

5.4 Supervised cluster evaluation

When we have external information about data, like class labels for the objects, the usual procedure is to measure the degree of correspondence between the cluster labels and the class labels. But why do we need to perform a clustering if we already have the class labels? Because clustering technique must be compared with the “ground truth” and/ or the validity of a manual classification process must also be evalueted.

5.4.1 Classification-oriented measures of cluster validity

There are a number of measures that are commonly used to evaluate the performance of a classification model. The same criteria can be used to evaluate cluster validity (Tan, Steinbach, Kumar, 2006):

Precision The portion of a cluster that consists of objects of a specified class. The precision of cluster *i* with respect to class *j* is the probability that a member of cluster *i* belongs to class *j* and it is computed as $p_{ij} = m_{ij} / m_i$ where m_i is the number of objects in cluster *i* and m_{ij} is the number of objects of class *j* in cluster *i*.

$$precision(i, j) = p_{ij} = \frac{m_{ij}}{m_i}$$

Entropy The degree to which each cluster consists of objects of a single class. The entropy of each cluster *i* is calculated as

$$e_i = -\sum_{j=1}^{L} p_{ij} \log_2(p_{ij})$$

where *L* is the number of classes. The total entropy for a set of clusters is computed as the sum of the entropies weighted by the size of each cluster

$$e = -\sum_{i=1}^{K} \frac{m_i}{m} e_i$$

where *K* is the number of clusters and *m* is the total number of data points.

Purity Another measure of the extent to which a cluster contains objects of a single class. Using the previous terminology

$$p_i = \max_j p_{ij}$$

$$purity = -\sum_{i=1}^{K} \frac{m_i}{m} p_i$$

Recall The measure of the number of objects of a class contained in a cluster. The recall of cluster *i* with respect to class *j*, where m_j is the number of objects in class *j*, is

$$recall(i, j) = \frac{m_{ij}}{m_j}$$

Here is an example where K-means was used with the cosine similarity in order to cluster 3204 newspaper articles from *Los Angeles Times*. These articles come from six different classes, the results are shown in the Table 2.

Cluster	*Enter-tainment*	*Financial*	*Foreign*	*Metro*	*National*	*Sports*	*Entropy*	*Purity*
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Table 77: (Tan, Steinbach, Kumar, 1999)

As it can be seen, if a cluster (e.g. the 3rd) contains a lot of objects of a specified class and very sparse objects of other classes, the entropy tends to 0 and the purity tends to 1. In the other cases, there are higher values of entropy and lower values of purity, because the separation of the objects is less precise.

5.4.2 Cluster validity for hierarchical clustering

Supervised evaluation of hierarchical clustering is more difficult considering that a preexisting hierarchical structure does not often exist. However, it is possible to evaluate a hierarchical clustering in terms of a set of class labels. The key idea is to evaluate whether a hierarchical clustering contains, for each class, at least one cluster that is relatively pure and includes most of the objects of that class. In order to do so, the F-measure is computed for each cluster. The F-measure of cluster i with respect to class j is

$$F(i, j) = \frac{2 \cdot precision(i, j) \cdot recall(i, j)}{precision(i, j) + recall(i, j)}$$

For each class, the maximum F-measure attained for any cluster is taken. Finally, an overall F-measure is calculated by computing the weighted average of all per-class F-measure, where the weights are based on the class size. (Stein, Eissen, Wißbrock, 2003)

$$F = \sum_j \frac{m_j}{m} \max_i F(i, j)$$

5.5 Assessing the Significance of Cluster Validity Measures

Cluster validity measures, as already discussed, are intended to assess the goodness of the clusters obtained from a particular algorithm. However, researchers always face the problem of interpreting the significance of the numbers acquired, a task that can prove even more difficult.

The minimum and maximum values of cluster evaluation measures may provide some guidance. For example, a purity of 0 is always bad, while a purity of 1 is good, as long as we trust the class labels. The same reasoning can be done for the silhouette coefficient, because an interval is provided. Likewise, an entropy of 0 is good, as is a SSE of 0.

Sometimes, however, there may not be a minimum or maximum value, or the scale of the data may affect the interpretations. In some cases, we can use an absolute standard but in others, we must follow a different approach such as to interpret the value of our validity measure in statistical terms. More accuratley, to judge how likely it is that the observed value may be achieved by random chance. The measure is good if it is unusual, namely, if it is unlikely to be the result of random chance (Jain, Dubes, 1988).

5.6 Conclusion

In this chapter we have shown an overview of cluster evaluation techniques.

Unsupervised proved to be the most suitable validation process for the clustering results of Power Aware in order to assess the goodness of the groups of people we will obtain:

- cohesion and separation are usefull to evaluate the precision of a cluster and the dispersion of the points in the data space, respectively.
- Silhouette coefficient is important because it has rigid bounds which help to decide if a result is useful or not.
- Proximity Matrix can be compared to the ideal one to get the correlation. The higher the correlation value, the better.
- CoPhenetic Correlation Coefficient (CPCC) is used to evaluate hierarchical clustering results.

Supervised approaches should be considered only if class-related data are used. At present, however, Power Aware is not concerned with this type of evaluation.

5.7 Bibliography

Jain, A.K., Dubes, R.C. (1988). Algorithms for Clustering Data: Prentice-Hall advanced reference series.

Kaufman, L., Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics. John Wiley and Sons, New York.

Stein, B., Eissen, S. M., Wißbrock, F. (2003). On Cluster Validity and the Information Need of Users: 3rd IASTED Int. Conference on Artificial Intelligence and Applications. Benalmádena, Spain, September 2003. ISBN 0-88986-390-3, ACTA Press

Tan, P. N., Steinbach, K., & Kumar, V. (2006). Data Mining Cluster Analysis: Basic Concepts and Algorithms: Boston: Pearson Addison Wesley.

6 Deliverable 6

Elaboration of a visualization for users comparison and power forecasting

Author: Emanuele Mottola
Project Manager: Diego Mariani
Scientific coordinator: Dr. Antonio Vetrò

6.1 Introduction

To encourage final users to cut their consumption, Power Aware will adopt visualizations in order to facilitate the understanding of data sets and to convey information with a greater effectiveness compared to raw data and text. In fact, according to Medina [6], the percentage of a read information a person preserves after three days amounts to 10% if only text encodes it, while it goes up to 65% when text and visual encoding are used together.

In this chapter we describe the requirements visalizations should fulfill, we discuss the visual properties adopted and we provide details on design choices exploited to create them. Finally, snapshots from actual implementation are shown.

6.2 Requirements

Power Aware wants to achieve three main goals through visualizations: (1) comparing energy consumption patterns for citizens with similar characteristics, (2) to generate power forecasting and, finally, (3) learning the actual effects of adding and/or removing parameters on cluster generation.

6.2.1 User comparison

In order to increase the awareness over their consumptions, users will be compared to a cluster of reference. The platform will set the cluster up based on the parameters described in Deliverable 1, exploiting the techniques examined in Deliverables 3 and 4. Hence, the visualizations should mainly allow the users to make a *comparison* between their own consumption and the cluster .

6.2.2 Power forecasting

Users will be able to see a representation of their future energy consumption, computed through machine learning techniques outlined in previous deliverables. This requirement allows us to underline another important task visualizations must carry out which is *to show the energy consumption over time.* Moreover, this task will underline the trend of user consumption, showing current and past energy demand too.

6.2.3 Changing set of parameters

In order to let the user understand how much a given parameter affects energy consumption, the platform will allow to add and/or remove it. This action is expected to produce a variation in the reference group users will be compared to. Therefore, the visualizations must support *the ability to be responsive to modifications* of the set of attributes used to compute the cluster.

6.3 Visual Encoding

In the following praragraphs design choices undertaken to create visualizations are outlined and visual properties are identified and ranked to accomplish our purposes.

6.3.1 Design choice

The first requirement that visualizations must accomplish is the *comparison.* Gleicher et al. [3] developed a model in which they propose a taxonomy that divides the space of comparative designs into three general categories, based on how relationships between the related parts of different objects are encoded. These three categories are *juxtaposition*, *superposition* and *explicit encoding of relationships*.

- **Juxtaposition** shows the objects to be compared separately. This separation can occur either in time or in space and in both cases it relies on memory for comparison. The key element in a juxtaposition design is that individual objects are shown independently. It is easy to implement and can be applied to any visual representation. A key challenge in juxtaposition design is that, since the objects are separated, it may be difficult for a viewer to see the relationships between them [3].
- **Superposition** shows the objects to be compared in the same space. Such designs can be referred to as *overlay* designs as they usually involve overlaying one object over another. As with juxtaposition designs, the display of the objects are independent although, sometimes, small adjustments may be made to improve clarity and avoid occlusions. Superposition is commonly used for situations in which compared objects are similar enough that they can be viewed on the same plane for the purpose of detecting similarities and differences [3].
- In **Explicit encoding** the objects themselves are not visualized; but rather, a new object is presented visually, determined (usually computed) as the relationship between the original objects. We might think of such designs as a *replacement* of the original objects with the new object that represents the relationships [3].

The design type that our visualization will exploit is superposition. Even if juxtaposition is easy to implement and can be adopted for a wide range of visualizations, it is based on a heavy usage of visual memory which makes the comparison difficult. Explicit encoding, on the other hand, would be misleading for our purpose for it uses replacement of original objects with a new object representing the relationship: in this scenario the user must not only be aware of the difference between his consumption and the reference group , he must also know to what value his consumption amounts. Hence, the best design choice for our purpose seems to be the superposition, through the definition of a common space for our measures represented on the same chart (exploiting a common axis) [3].

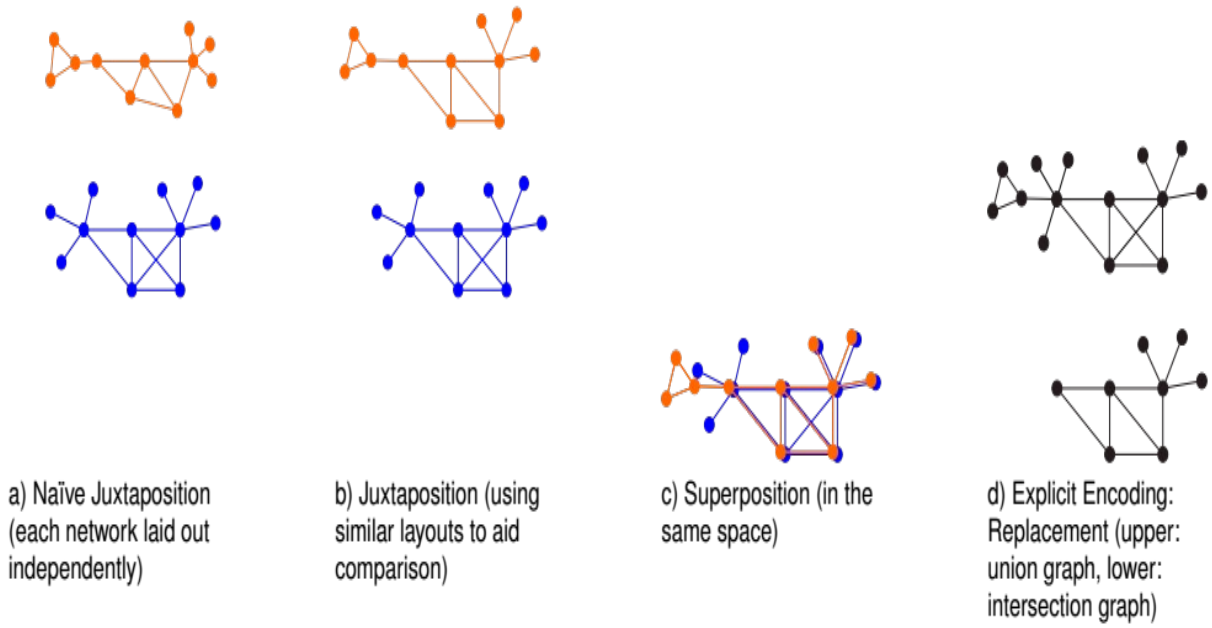


Figure 11: Two networks (illustrated as node-link diagrams) are compared using designs defined in the taxonomy [3]

6.3.2 Visual properties

Cleveland and McGill [1] and Healey and Enns [2] agree on the identification of a limited set of visual features that are detected very rapidly by low-level, fast-acting visual processing and which are used in order to convey quantitative information. The value of identifying basic elements is that we thus have avaible a framework to organize knowledge and predict behavior of how this basic elements are decoded by users.

Table 1 shows an overview of the main ones [4].

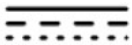
Example	Description	Example	Description
	Position, placement		Saturation, brightness
1,2,3; A,B,C	Text labels		Color
	Length		Shape, icon
	Size, area		Pattern texture
	Angle		Enclosure, connection
	Pattern, density		Line pattern
	Weight, boldness		Line endings

Table 78: Visual properties [4]

We need to understand which is the most useful and effective visual property to encode electricity consumption in order to allow comparison. Cleveland and McGill [1] ranked the visual properties shown above on the basis of the accuracy according to which people can extract quantitative information decoding them. However, accuracy does not mean to convey numbers with as many decimal places as possible, for in that case we would use tables. A graph is accurate if it enables one to take in the quantitative information, organize it, and see patterns and structures not readily revealed by other means of studying the data.

The following list shows some of the elementary tasks ordered from the most accurate to the least one:

1. Position along a common scale
2. Position along nonaligned scales
3. Length
4. Angle and slope
5. Area
6. Volume, density, and color saturation
7. Color hue

Position along a common scale is the visual property which conveys quantitative information in the most accurate way. So, Power Aware will exploit it for its visualizations.

6.4 Implementation

To fulfill the requirements outlined, our choice fell on grouped bar chart, which allows us to implement design choices and visual encoding mentioned in the previous paragraphs. An example of visualization is shown in Figure 2, in which the user is compared to reference group clustered throught the K-means algorithm as shown in deliverable 4, using

as attributes the Total floor area and the Number of occupants (deliverable number one). Data represented in this and the following figures are examples.

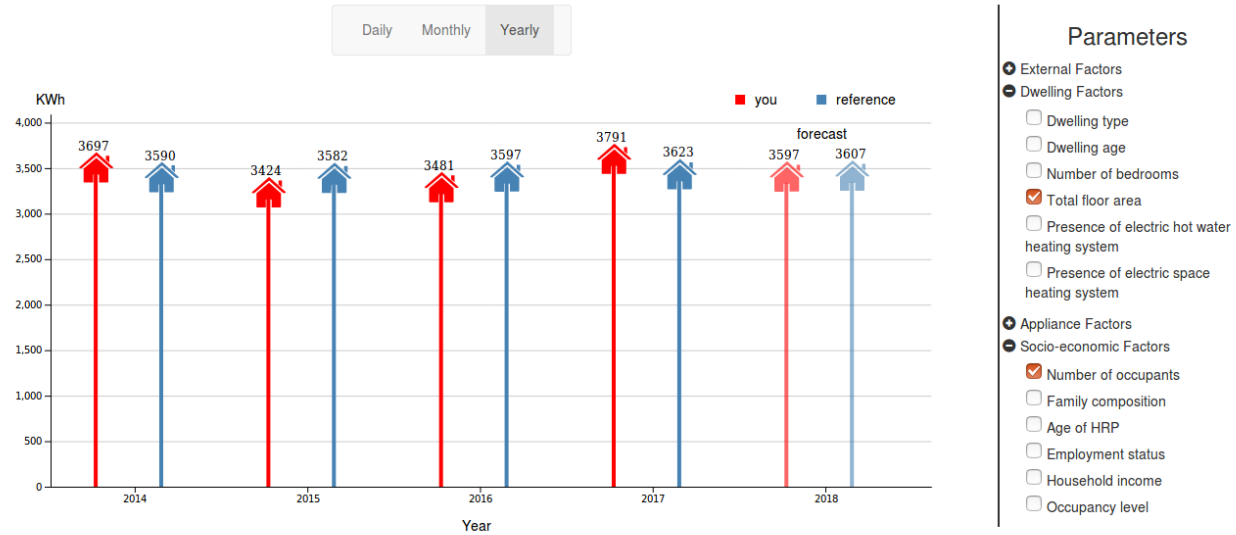


Figure 12: Snapshot of visualization, yearly consumption. Algorithm: K-means. Attributes: Total Floor Area and Number of Occupants

Power Aware wants to allow the users to compare their consumption with the cluster of reference over time. So, Time (in Figure 2, Years) is the dimension we assigned to the x-axis. This dimension is also adjustable in Months (Figure 3) or Days, if interested in electricity consumption at that level of detail. This configuration favors to show the power forecast: estimated values are colored with the same hues of the actual ones, but with an higher trasparency. The other dimension is, instead, electricity consumption (in KWh), set on the y-axis.

The comparison between user and cluster demand is essentially performed throught the position on common scale of the shapes (the houses), standing for electricity demand.

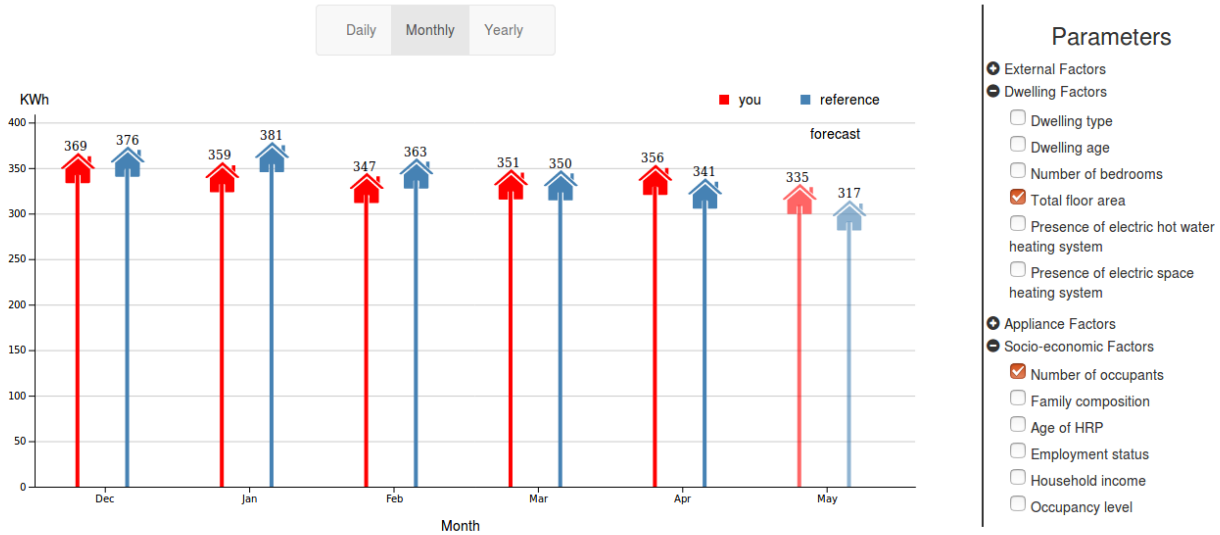


Figure 13: Snapshot of visualization, monthly consumption. Algorithm: K-means. Attributes: Total Floor Area and Number of Occupants

In the implementation of visualizations, one problem in particular arised: the way to symbolize the cluster. Since using all the values composing the reference group would result too confusing, we chose to represent all the values throught a statistic, the median, which is robust to outliers and noise.

The last requirement we had to fulfill was the ability of visualizations to be responsive to modification in the set of parameters. To accomplish that, the visualization will be dynamic, in order that, depending on the attributes, values of the new cluster are shown. Figure 4 is an example of how cluster representation changes according to the new set of parameters. In particular, it is re-computed using the K-means algorithm on the attributes Number of bedrooms and Number of occupants.

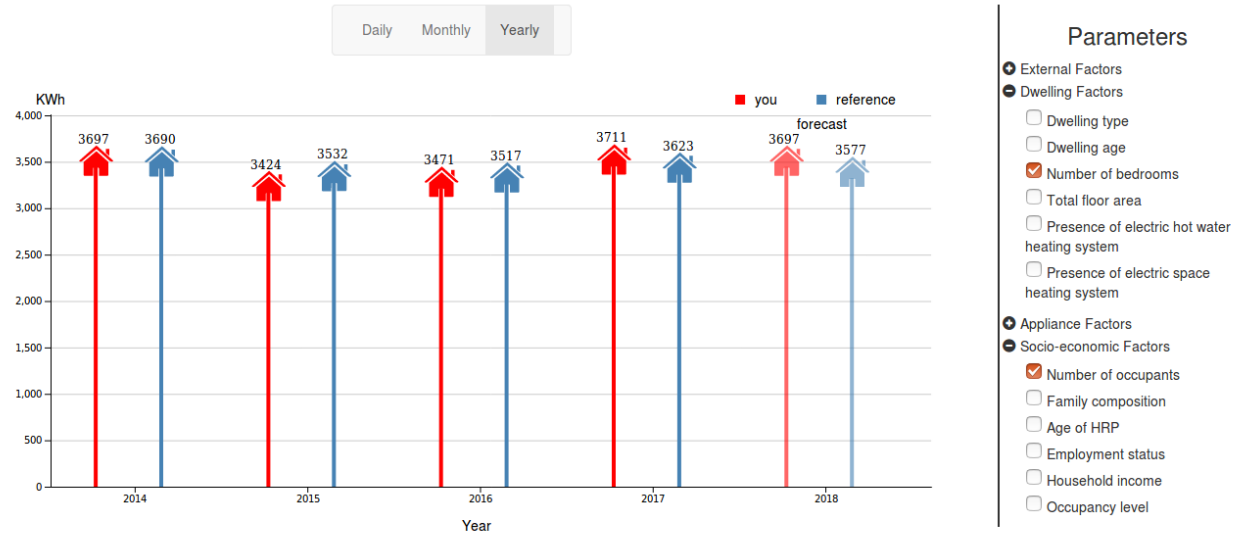


Figure 14: Snapshot of visualization, yearly consumption. Algorithm: K-means. Attributes: Number of bedrooms, Number of occupants.

6.5 Conclusion

To conclude, visualizations presented in this deliverable are created considering superposition design and exploiting position along a common axis as key visual property to encode quantitative information about electricity consumption.

In the next deliverable, we will evaluate the effectiveness of these visualizations with a pool of users in order to identify possible issues.

6.6 Bibliography

[1] W. S. Cleveland, R. McGill, Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods, Journal of the American Statistical Association, Vol. 79, No.387 (Sep. 1984), 531-554

[2] Christopher G. Healey, Senior Member, IEEE, and James T. Enns, Attention and Visual Memory in Visualization and Computer Graphics, IEEE Transactions on visualization and computer graphics, Vol. 18, No.18, July 2012

[3] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, J.C. Roberts, Visual comparison for information visualization, Information Visualization, Volume: 10 issue: 4, page(s): 289-309, September 7, 2011

[4] Noah Iliinsly, Julie Steele, Designing data visualizations, O'Reilly, 2011

[5] Flowing Data, http://flowingdata.com/2010/03/20/graphical-perception-learn-the-fundamentals-first/

[6] John Medina, Brain Rules, 2008

7 Deliverable 7

Evaluation of the visualization with users

Author: Emanuele Mottola
Project Manager: Diego Mariani
Scientific coordinator: Dr. Antonio Vetrò

7.1 Introduction

In order to understand whether the work done in the previous chapter is effective, it is necessary to take into account which are the outcomes the users get from the visualization, so to check if they match with what we expected. Hence, in this chapter we will deal with the evaluation of the visualization designed. Moreover, we will develop a re-design based on the results of the evaluation.

7.2 Methodology

Generally, there are two approaches used to evaluate the product of an activity: quantitative and qualitative evaluation.

Quantitative methods produce essentially quantitative data or “hard numbers” which are analysed through statistical analysis and can be collected by surveys or questionnaires, pretests and post-tests, observation and document analyses. The strengths of these methods are the precision and consistency that the results produce. On the other hand, this kind of evaluation does not allow an in-depth understanding of the context in which the work is carried out.

Qualitative methods, instead, supply data essentially in text form which are analysed by examining, comparing, contrasting and interpreting patterns. They are collected using think-aloud technique, focus groups, observations or interviews. Qualitative approach is useful to explore complex issues, in order to understand the “why” and “how” over the “what”. Anyway, limitations in the adoption of such methodologies is the lack of generazability and the difficulty and complexity of data analysis and interpretation [1 - 2].

In order to enrich the results of our evaluation, we decide to opt for a mixed approach to gather the benefits of both quantitative and qualitative methods. In fact, the evaluation process we adopt is an extended version of the think-aloud technique: at the end of each think-aloud session we ask a question to the user, in order to gather quantitative data over the qualitative ones. In this way, we are able to access the cognitive processes and mental behavior of the users, which give us deeper insight into user thinking.

The evaluation is divided in three tasks:

- **First task.** It is aimed to highlight confusing or useless elements of the visualization. Moreover, we are interested in figuring out whether the users easily infer the difference between their consumption and the reference group one and if they are able to quantify it.
- **Second task.** It is oriented to evaluate whether the visualization is able to convey the learning aspects of adding or removing a parameter from the list used to generate the reference group to the user. The dynamic changes of the graph should

lead the user to comprehend how much a given parameter affects group reference consumption.

- **Third task.** Finalized to understand how the user perceives the projection of the current timeslot compared to the past ones.

The complete description of the evaluation test is presented in Appendix A.

7.3 Results

The results we are going to present are divided in *qualitative* and *quantitative* results. We obtained them by interviewing thirty persons of different age and education level.

7.3.1 Qualitative Results

In the following we will outline, task by task, the feedbacks we got by the users.

- **First task.** We can summarize qualitative results of the first task in three main points.
  (1) At the beginning of the interview, some users (23,3%, 7/30) tell they do not get the point of the visualization immediately. In fact, they do not understand the main aim of the visualization is the comparison between their and the reference group's consumption. Some of them explain the understanding would be more quickly if the legend was clearer.
  (2) 20% of users perceive the house shapes and the labels near them as confusing or useless elements: the main reason they adduct is the fact that they do not understand which part of the house shape would indicate the value of the consumption.
  (3) On average, users take some seconds (~10-15) to answer quantitative questions about user-reference consumption difference. Many users employ that time to calculate it exploiting data labels: hence, the absence of an explicit element representing that difference may be one of the causes of that latency.
- **Second task.** Users find the transition of the bars length intuitive to quantify and highlight the influence of a set of parameters on group reference consumption.
- **Third task.** A lot of users (64%, 19/30) do not note the transparency element since the beginning of the interview and do not understand it is a forecast. Once the interviewer focuses their attention on it, they suggest to design it in a more intuitive way, since it is not clear.

7.3.2 Quantitative Results

The full list of answers provided by users can be found in Appendix B, in Figure 28.

As we did in chapter 3.1, in the following we will comment the results of the quantitative analysis.

- **First task.** (1) Visualization is not able to communicate directly to the users the electricity amount they consumed more or less than the reference. In fact, answers 1.1 – 1.4 (Figure 29-32) show that on average 75% of users are able to infer correctly how much they consume compared to the reference. (2) Instead, it is easy for users to understand when their consumption is “more or less the same” of the reference one (~98%, Figure 33-34).
- **Second task.** As already said in the chapter on qualitative analysis, users are able to understand clearly and correctly how much a set of parameters influences power consumption. In fact, about 90% of users answer correctly to question 2.1-2.2 (Figure 35-36).
- **Third task.** 80% of users (Figure 37) find the prediction increased with respect to the average of the consumption of the previous years, as expected: hence, the visualization correctly communicates the future trend consumption .

7.3.3 Summary

In conclusion, the grouped bar chart is generally able to communicate to the users which comparison element is behind the visualization, even if some misunderstandings was caused by the confusing legend. There are some elements, like the house shapes and labels, which do not help to increase the clarity of the graph. The use of transparency does not seem to be enough intuitive in order to indicate the forecast consumption. On the other hand, the use of transition of bars length to denote the influence of a set of parameters on reference consumption is successful. In the following chapter we will present the re-design of the visualization according to the hints got by the users we interviewed.

7.4 Re-design

Since the comparison element between user and group-reference consumption is generally perceived by users, the grouped bar chart is the graph type we will use in this re-design too. An example of visualization is shown in Figure 15, in which the user is compared to reference group clustered through the K-means algorithm as shown in chapter number four, using as attributes the Dwelling age and the Age of HRP (chapter number one). Data represented in this and the following figures are examples.

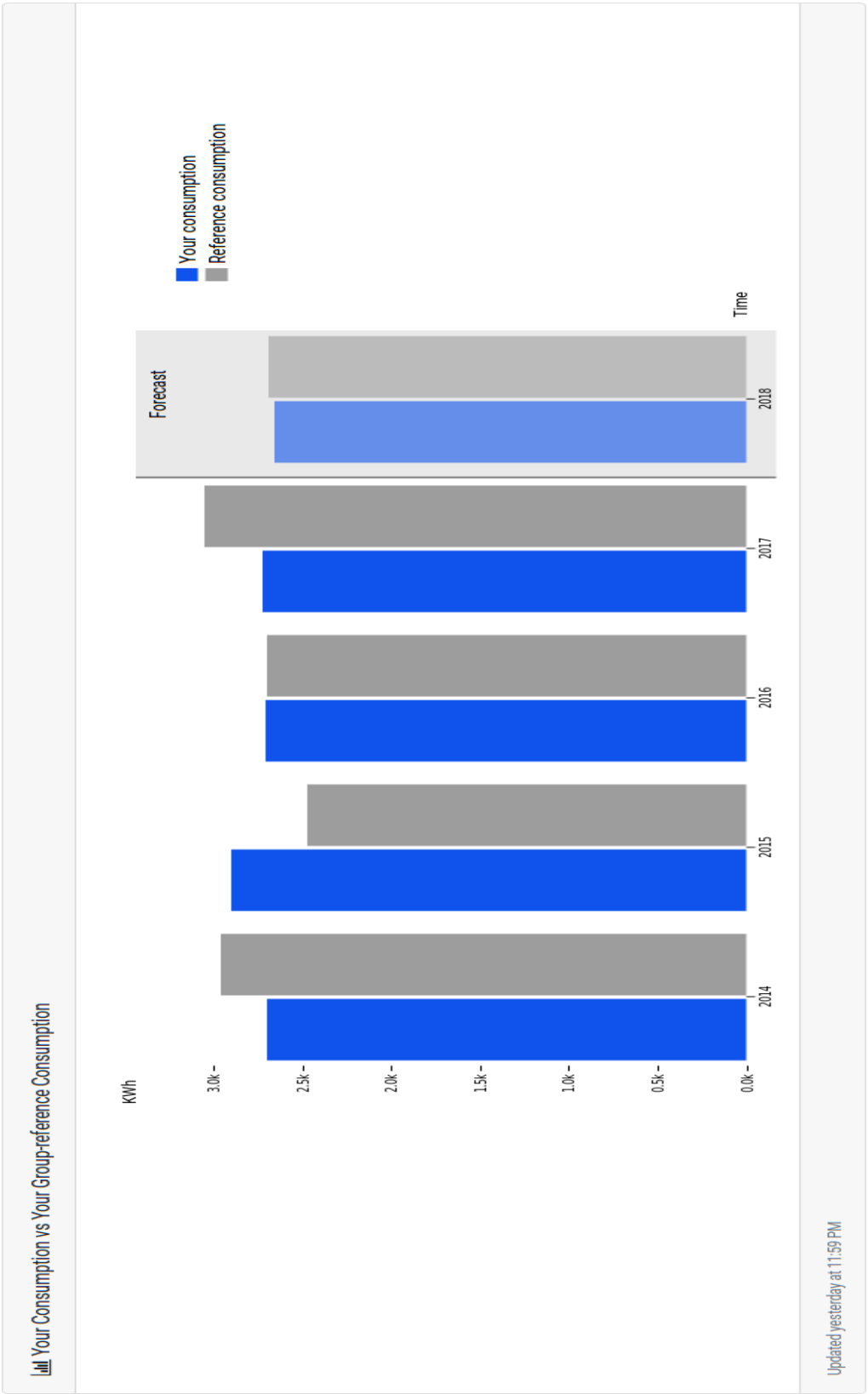


Figure 15: Snapshot of visualization, yearly consumption. Algorithm: K-means. Attributes: Total Floor Area and Number of Occupants

Axis dimensions are the same of the previous version. We decide to eliminate the house shapes and labels, which are seen as confusing elements, and to increase the width of each bar, in order to place the bars belonging to the same timeslot closer. In this way, the visual element which encodes the comparison of electricity demand is no more the position of shapes on common scale, but the length of the bars.

To help users quantifying the amount of electricity they consume compared to the reference group, we introduce visual elements, represented by coloured bars, which fill the difference between user and reference group electricity demand. These bars appear only when the mouse is over the graph and they are combined with percentages indicating the amount. Figure 16 is an example of the visualization. Algorithms and parameters are the same of Figure 15.

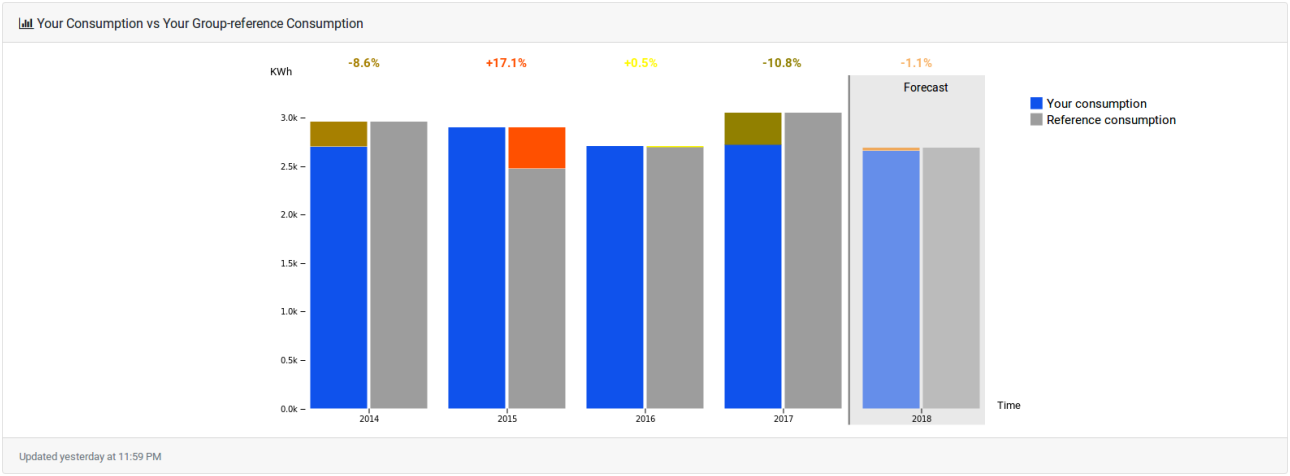


Figure 16: Snapshot of visualization with difference highlighted

We choose colors to fill the above mentioned elements according to a gradient bar, placed above the grouped bar chart, created to indicate the trend of user consumption over all the time periods. Figure 17 is an example of this gradient bar.

Figure 17 shows that the bar is gradually coloured with different shapes of green, yellow

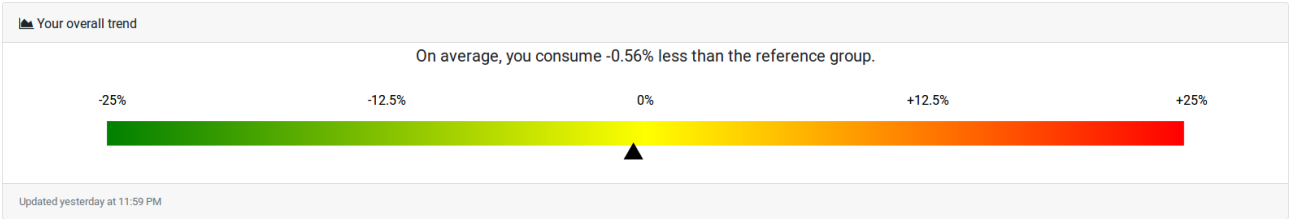


Figure 17: Gradient bar

and red. The percentages above the gradient bar are examples, they could be changed according to real data. The black triangle indicates the average of all the percentage differences illustrated in Figure 16. The lower the user electricity demand compared to the

reference one, the closer the triangle is to green color; the higher, the closer the triangle is to red color.

In the interviews, users explain that they do not understand the last group of bars is a forecast because the transparency is not so visible. In order to highlight it, we modify the background color, changing it from white to lightgrey, and maintain the “Forecast” label. Final result is shown in Figure 15.

To conclude, even if users are able to understand the influence of the set of parameters they selected or deselected to create the new group reference, we think that adding a visual notification of the actual influence can be even more helpful. Hence, we add a flash notification made of the influence percentage inside a circle. This notification lasts ten seconds and then disappears. Figure 18 represents the final version of our visualization, in which the user is compared to the reference group clustered through the K-means algorithm as shown in chapter number four, using as attributes the dwelling type, the dwelling age, the number of bedrooms, the total floor area, the fridge and freezer, the washing machine, the electric oven, the air-conditioning, the dehumidifier, the entertainment appliances and the lighting (chapter number one).

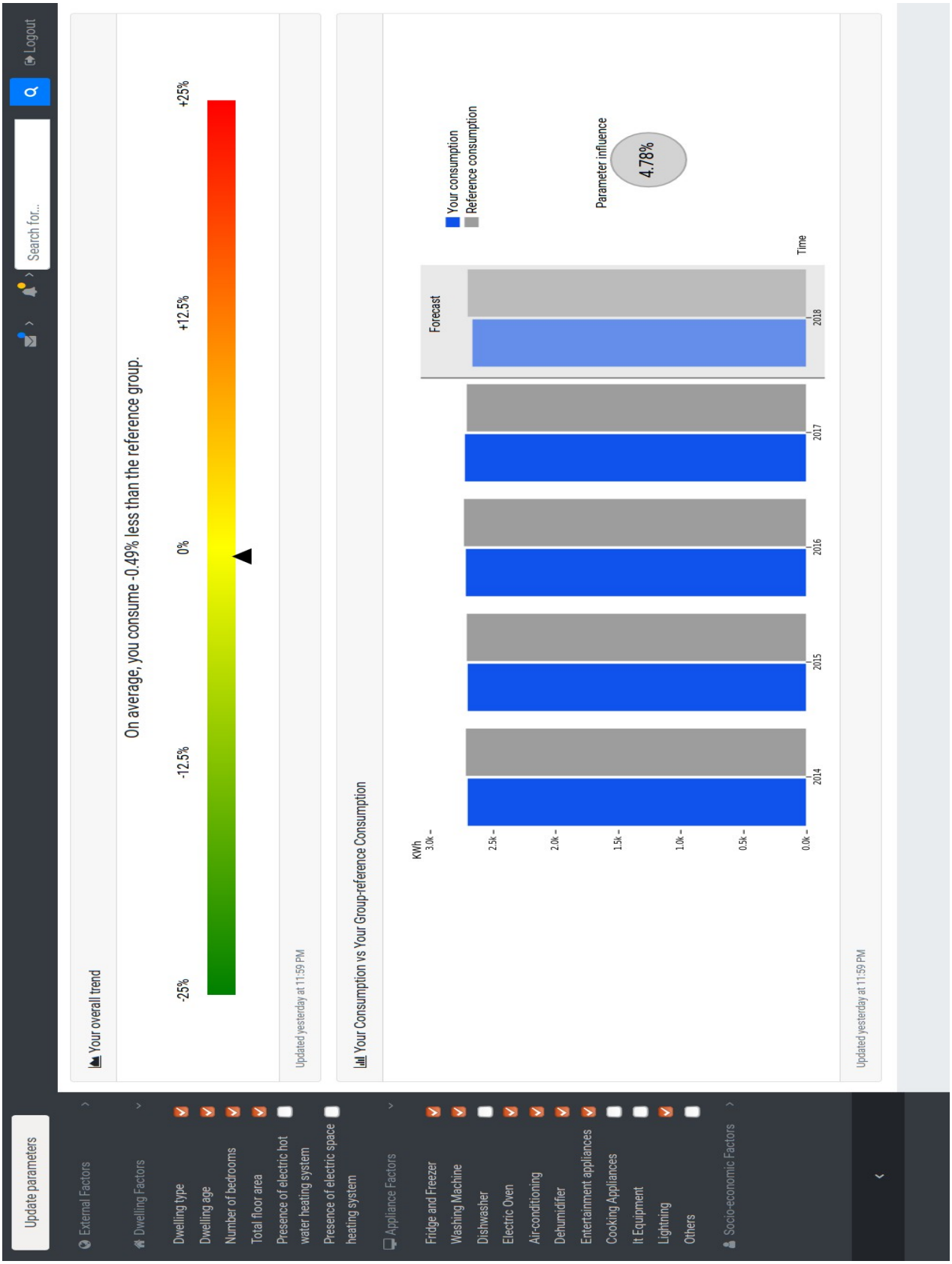


Figure 18: Snapshot of visualization.

7.5 Conclusion

In conclusion, qualitative and quantitative results constituted important hints to improve the visual effectiveness of elements encoding the comparison between user and reference-group electricity demand and the forecast. In fact, the main goal of re-design is to make those elements more explicit .

7.6 Appendix A

7.6.1 Requested tasks

First task. Objective: perception evaluation of the comparison between the user power consumption and reference consumption.

The interviewer shows the user six graphs (graph_1, graph_2, graph_3, graph_4, graph_5, graph_6), organized in three categories.

The first category (graph_1 and graph_2) shows user consumption higher than the reference one (about 15% more). Using the think-aloud technique, the user expresses his thoughts concerning the visualization aloud for each graph. When the user stops speaking, the interviewer asks how the user perceives his own consumption compared to the reference:

a) about 25% lower

b) about 15% lower

c) more or less the same

d) about 15% higher

e) about 25% higher

The question is posed twice to the user, one for each graph. We expect the user answers “about 15% higher” to both questions.

Figure 19: graph_1

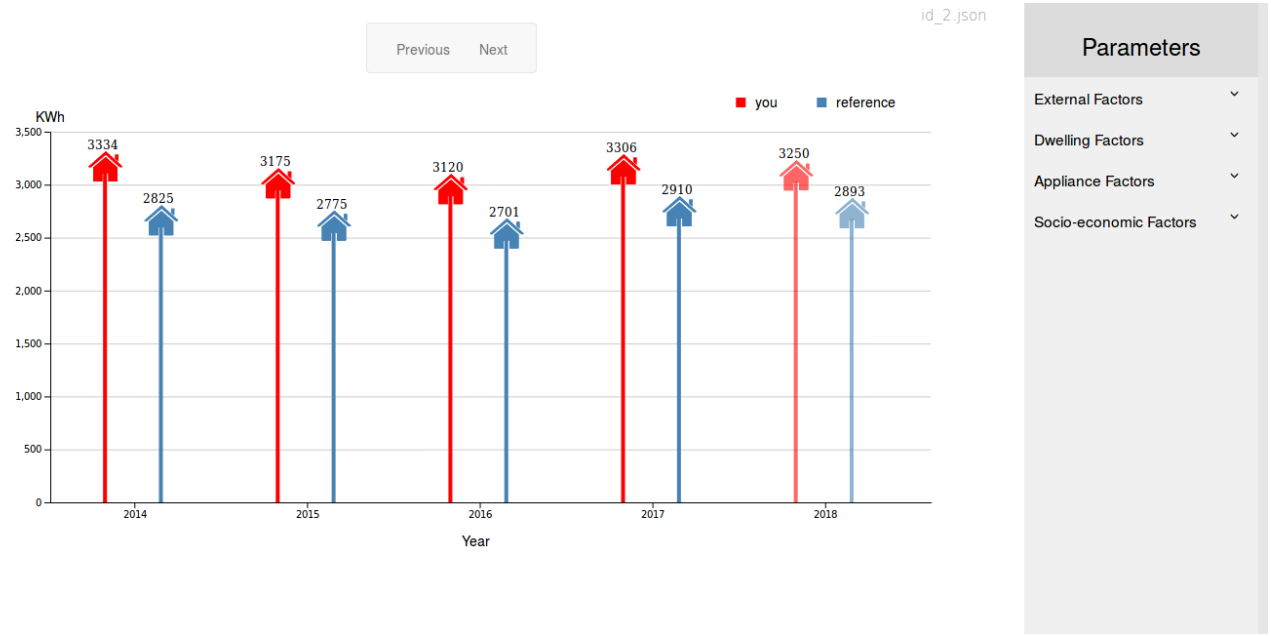


Figure 20: graph_2

The second category (graph_3 and graph_4) shows user consumption lower than the reference one (about 15% less). Using the think-aloud technique, the user expresses his thoughts concerning the visualization aloud. When the user stops speaking, the interviewer asks how the user perceives his own consumption compared to the reference:

a) about 25% lower

b) about 15% lower

c) more or less the same

d) about 15% higher

e) about 25% higher

The question is posed twice to the user, one for each graph. We expect user answers “about 15% lower” to both questions.

Figure 22: graph_4

The third category (graph_5 and graph_6) shows user consumption similar to the reference one (range +-3%). Using the think-aloud technique, the user expresses his thoughts concerning the visualization aloud. When the user stops speaking, the interviewer asks how the user perceives his own consumption compared to the reference:

a) about 25% lower

b) about 15% lower

c) more or less the same

d) about 15% higher

e) about 25% higher

The question is posed twice to the user, one for each graph. We expect user answers “more or less the same” to both questions.

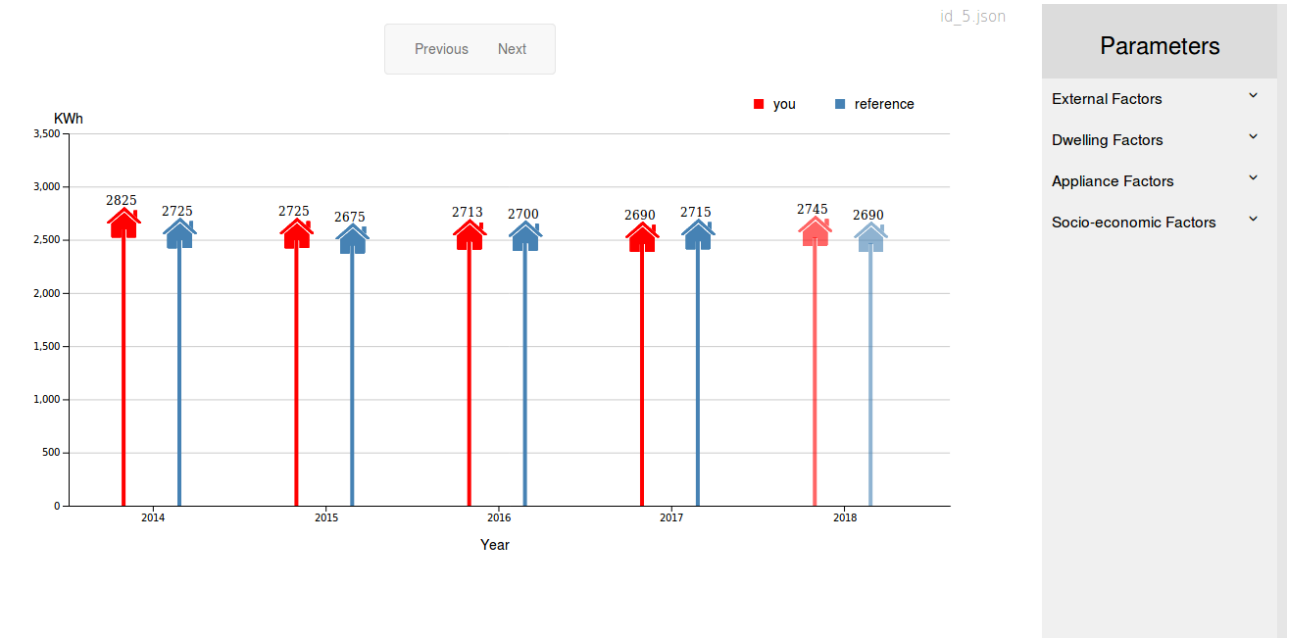


Figure 23: graph_5

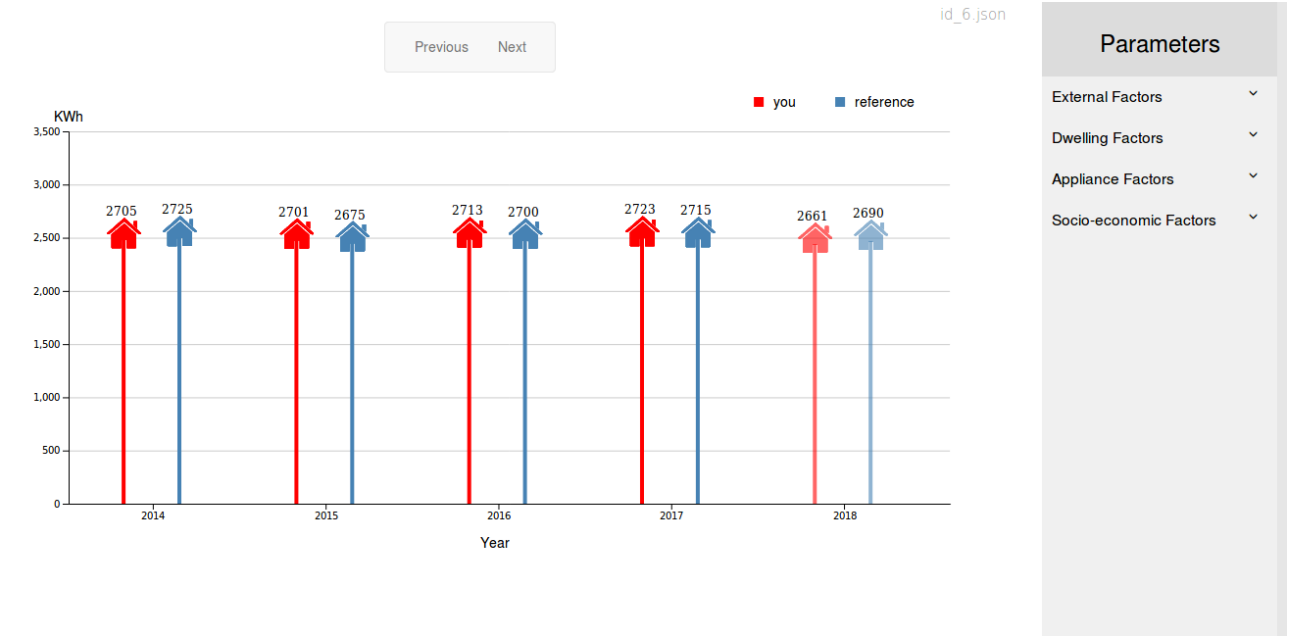


Figure 24: graph_6

Second task. Objective: evaluation of the learning aspect in adding or removing a parameter.

1. The interviewer shows graph_7 to the user. The user adds or removes a parameter of his choice. As consequence, this causes an increase in consumption by 5% of the reference group (graph_8). Using the think-aloud technique, the user expresses his thoughts aloud concerning the effects of adding/removing that parameter. When the user stops speaking, the interviewer asks how the user perceives the current situation compared to the previous one:

a) the parameter affects group reference consumption decreasing it by 15%.

b) the parameter affects group reference consumption decreasing it by 5%.

c) the parameter does not affects group reference consumption.

d) the parameter affects group reference consumption increasing it by 5%.

e) the parameter affects group reference consumption increasing it by 15%.

We expect user answers “the parameter affects group reference consumption increasing it by 5%”.

2. The interviewer shows again graph_7 to the user. The user adds or removes another parameter of his choice (different from the previous one). As consequence, this causes an increase in consumption by 15% of the reference group (graph_9). Using the think-aloud technique, the user expresses his thoughts aloud concerning the effects of adding/removing that parameter. When the user stops speaking, the interviewer asks how the user perceives the current situation compared to the previous one:

   a) the parameter affects group reference consumption decreasing it by 15%.

   b) the parameter affects group reference consumption decreasing it by 5%.

   c) the parameter does not affects group reference consumption.

   d) the parameter affects group reference consumption increasing it by 5%.

   e) the parameter affects group reference consumption increasing it by 15%.

We expect user answers “the parameter affects group reference consumption increasing it by 15%”.

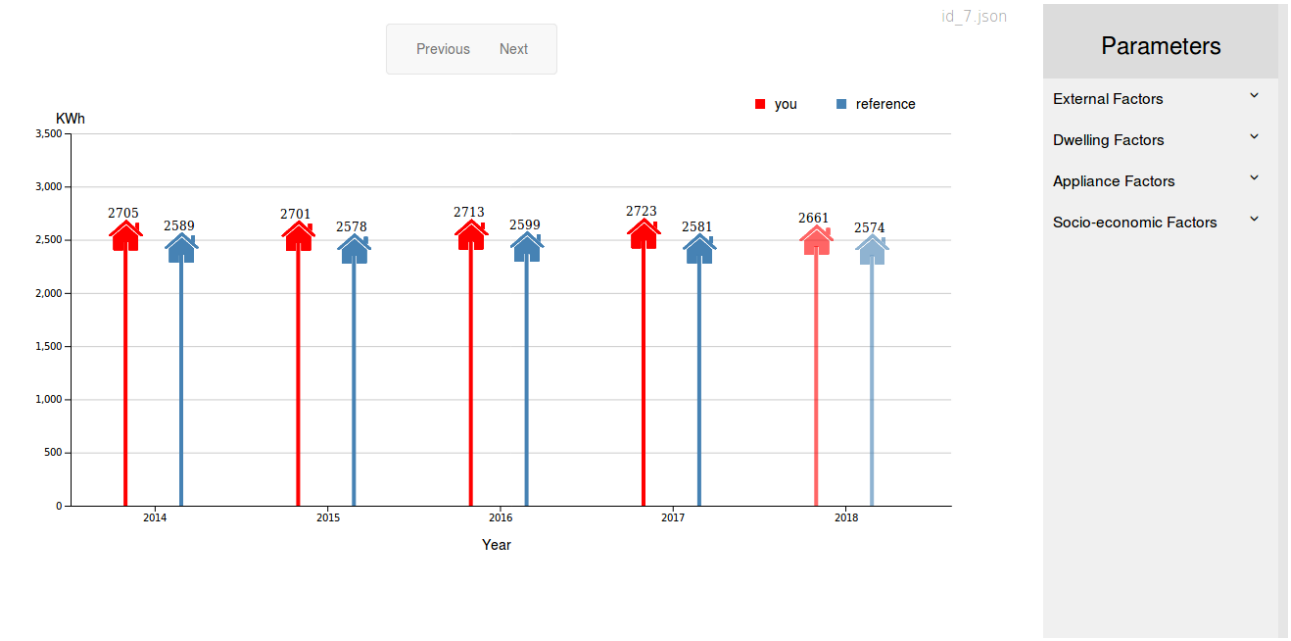
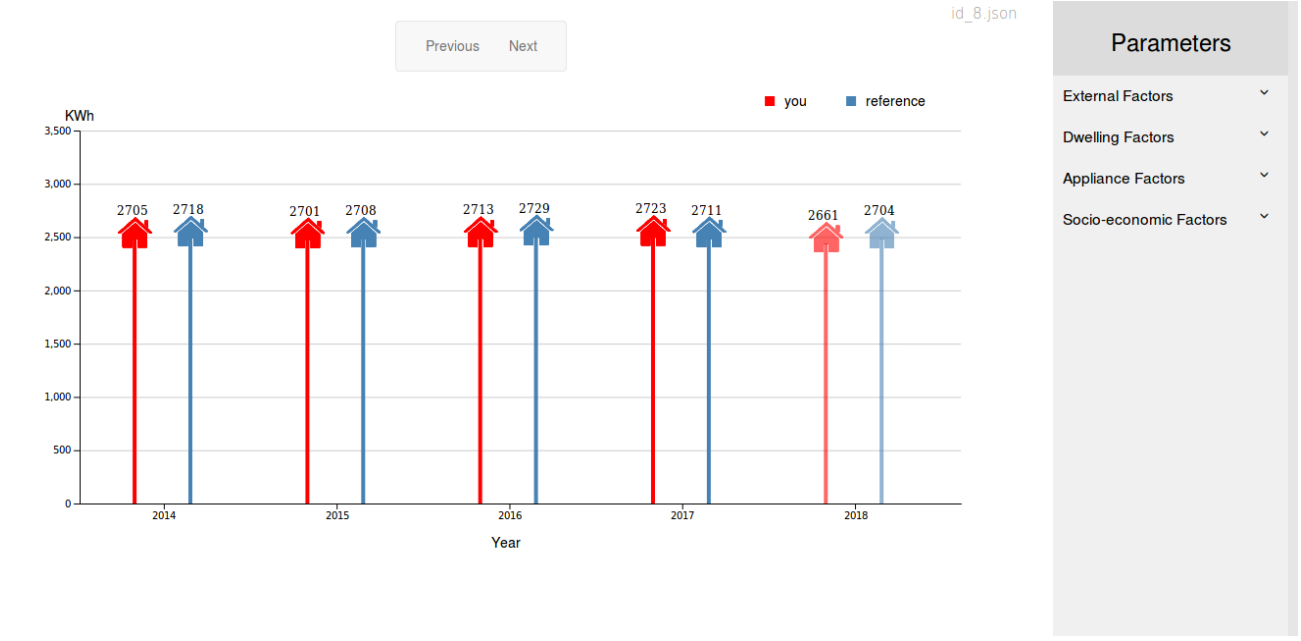
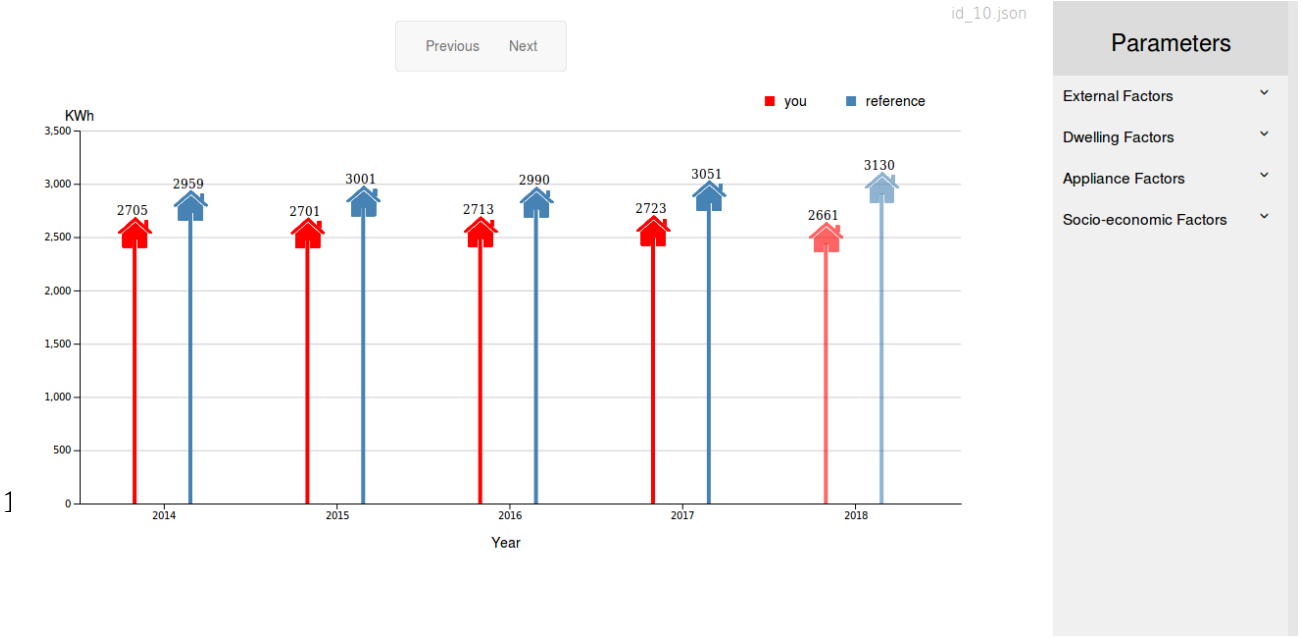


Figure 25: graph_7





1

Figure 27: graph_9

Third task. Objective: forecasting evaluation.

Using graph_9, the interviewer shows the forecasting element to the user. The interviewer asks the user whether the forecast, compared to the average of all the years, is:

a) decreased by 15%.

b) decreased by 5%.

c) more or less the same

d) increased by 5%

e) increased by 15%

We expect user answers “increased by 5%”.

7.7 Appendix B

Table of the answers provided by users.

	1-1	1-2	1-3	1-4	1-5	1-6	2-1	2-2	3-1
User1	d	e	b	b	c	c	d	e	c
User2	d	d	b	b	c	c	d	e	d
User3	d	e	a	b	c	c	c	e	d
User4	e	e	a	a	c	c	d	e	d
User5	d	e	b	b	c	c	d	e	d
User6	d	e	b	b	c	c	e	e	e
User7	d	d	a	a	c	c	d	e	d
User8	d	d	b	b	c	c	d	e	c
User9	d	d	b	b	c	c	d	e	c
User10	c	c	b	b	c	c	d	e	d
User11	e	e	b	c	c	c	d	e	d
User12	d	e	b	c	d	c	d	e	d
User13	d	d	b	b	c	c	d	e	d
User14	d	d	b	c	c	c	d	e	d
User15	e	d	a	b	c	c	d	e	d
User16	d	d	b	a	c	c	d	e	d
User17	d	e	a	b	c	c	e	e	e
User18	d	d	b	b	c	c	d	e	d
User19	d	d	b	b	c	c	d	e	d
User20	d	e	b	b	c	c	d	e	d
User21	d	d	b	b	c	c	d	e	d
User22	e	d	b	b	c	c	d	e	d
User23	d	d	b	b	c	c	d	e	d
User24	d	d	b	b	c	c	d	e	d
User25	d	d	b	b	c	c	d	d	d
User26	d	d	a	a	c	c	d	e	d
User27	d	d	b	b	c	c	d	e	d
User2									c
User2									d
User3									d

Figure

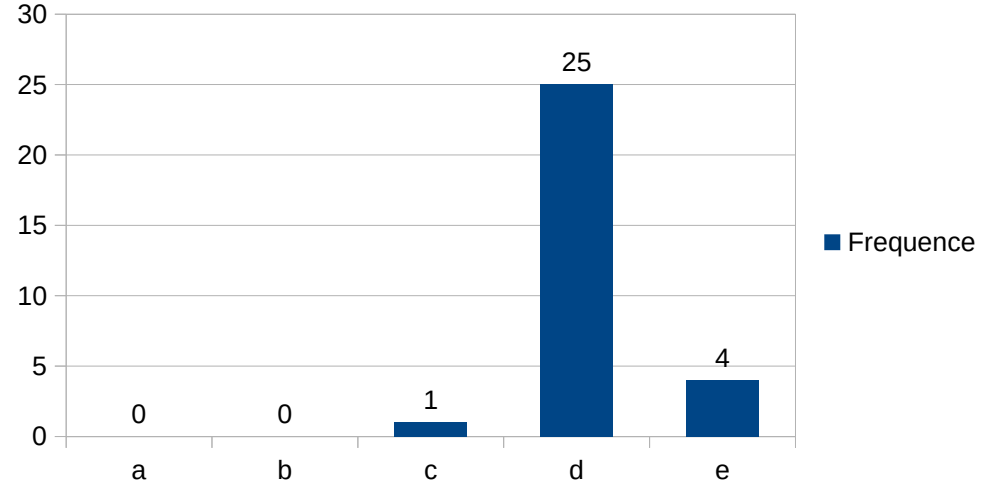


Figure 29: Summary of answers to question 1-1

Answers to question 1-1.

Answers to question 1-2.

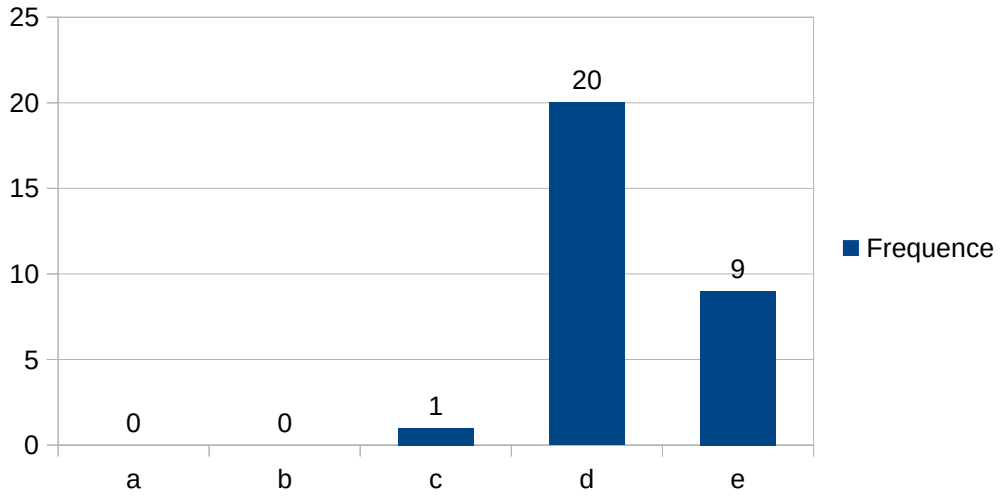


Figure 30: Summary of answers to question 1-2

Answers to question 1-3.

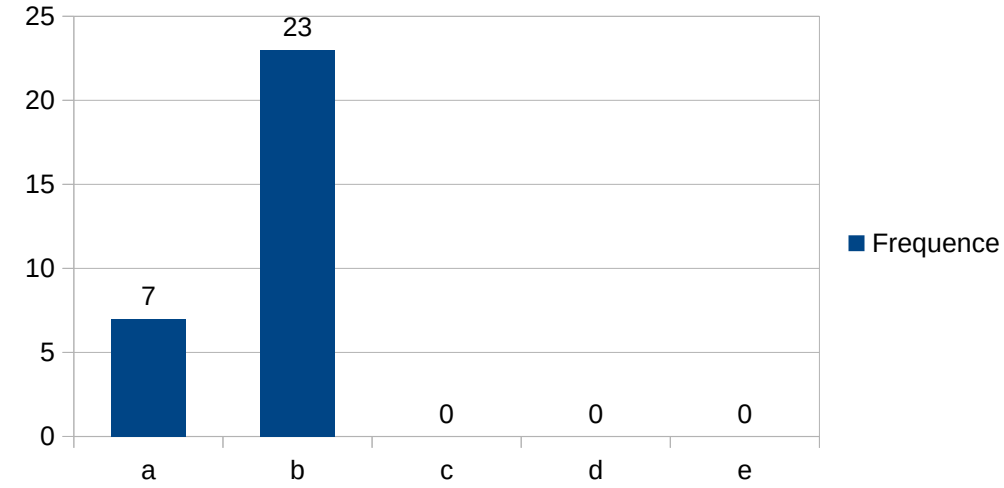


Figure 31: Summary of answers to question 1-3

Answers to question 1-4.

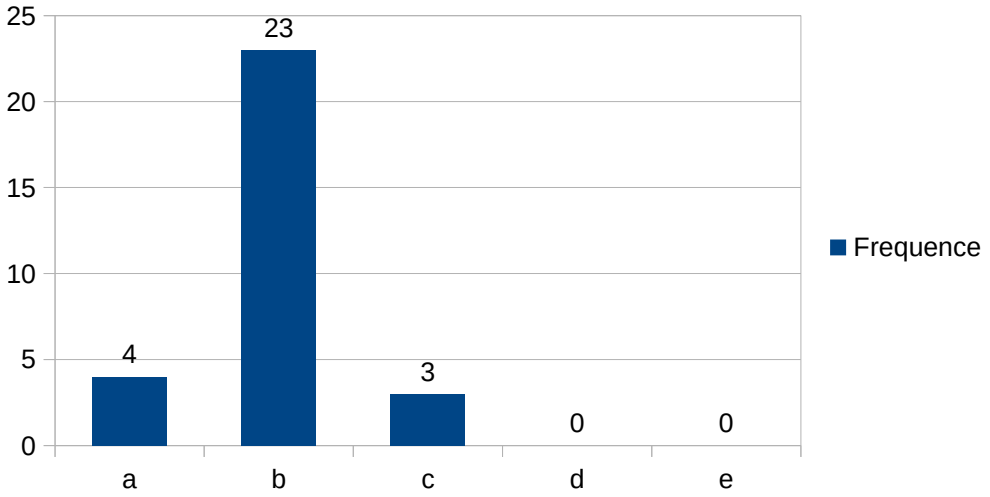


Figure 32: Summary of answers to question 1-4

Answers to question 1-5.

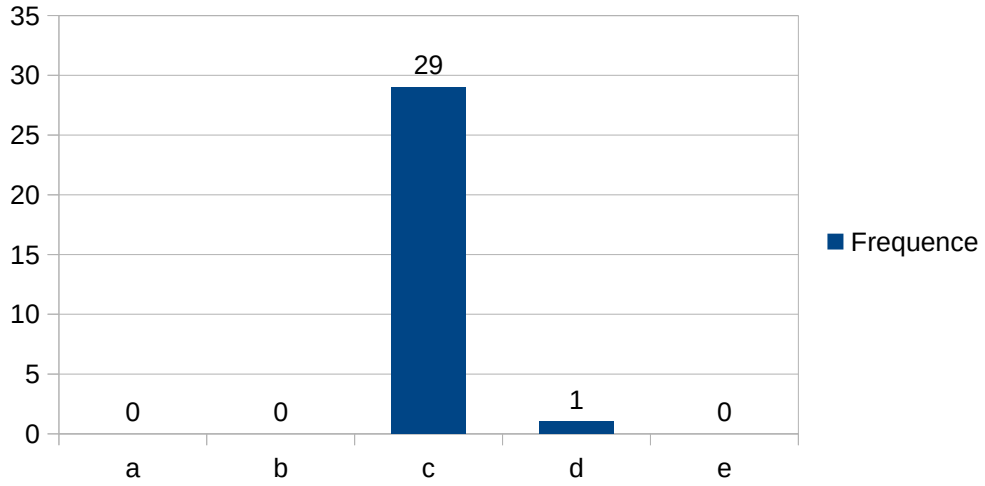
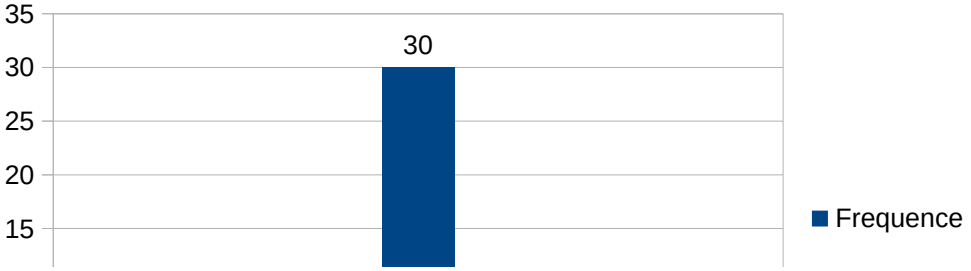


Figure 33: Summary of answers to question 1-5

Answers to question 1-6.



Answers to question 2-1.

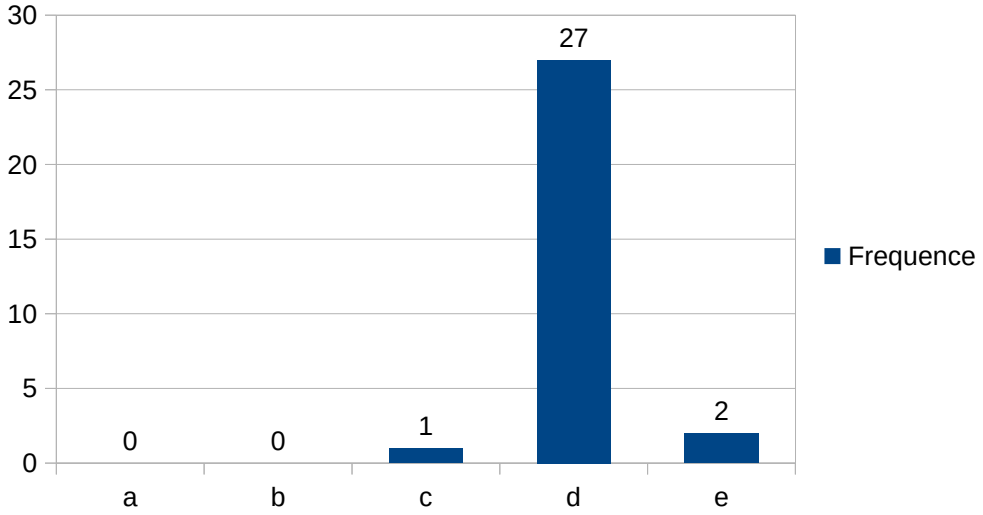
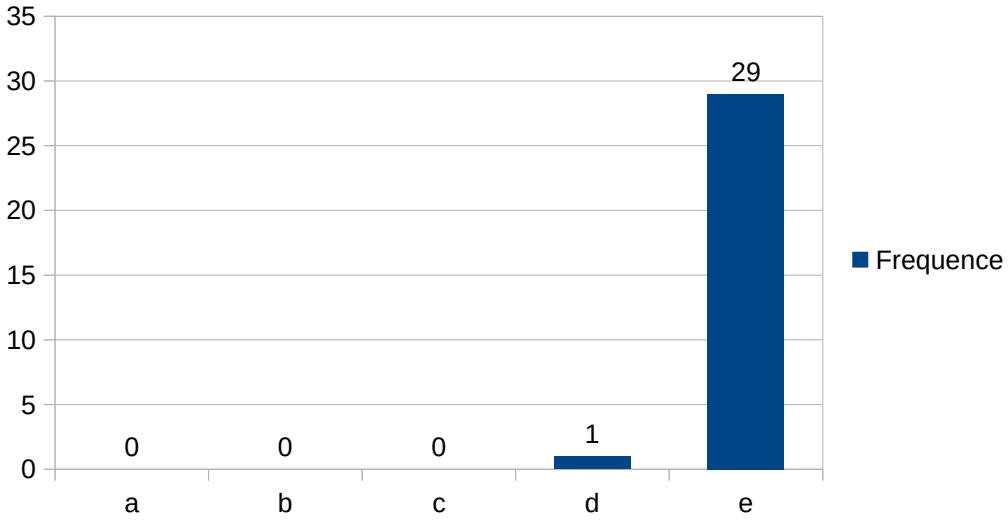
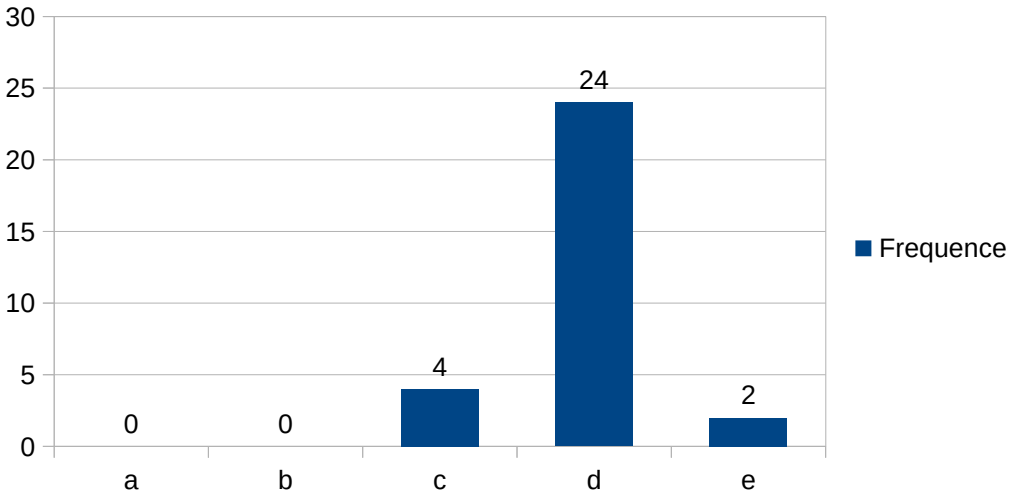


Figure 35: Summary of answers to question 2-1

Answers to question 2-2.



Answers to question 3-1.



7.8

Figure 37: Summary of answers to question 3-1

Bibliography

[1] Agency for Toxic Substances and Disease Registry – Principles of community Engagement. Chapter 7.

[2]https://classroom.synonym.com/difference-between-qualitative-quantitative-evaluation-8281411.html