NTNU
Norwegian University of
Science and Technology

A DATA-DRIVEN APPROACH TO DETERMINE INPUT SCHEDULES FOR ENERGY SIMULATION OF RESIDENTIAL BUILDINGS

A statistical methodology to generate electric internal load and presence of occupancy schedules in a residential building in Milan (Italy)

Martina Ferrando

Master of Science in Civil and Environmental Engineering
Submission date: May 2018
Supervisor: Salvatore Carlucci, IBM
Co-supervisor: Francesco Causone, Politecnico di Milano

Norwegian University of Science and Technology
Department of Civil and Environmental Engineering

Acknowledgments

Firstly, I would like to thank Prof. Salvatore Carlucci (NTNU, Trondheim) and Prof. Francesco Causone (Politecnico of Milan) because of their enthusiasm and devotion to research. They always trusted me and assisted me with their experience.

I would also like to thank Dr Silvia Erba (Politecnico of Milan) who supported me with knowledge and kindness from the very beginning. Without her participation, the entire project could not have been successfully achieved. I would also like to acknowledge Eng. Amin Moazami (NTNU, Trondheim), who greatly helped me, being always resourceful and positive.

This work has been developed within the framework of the R&D project 269650 "Utvikling av metoder og system for automatisk effektkontroll i bolig”. I thank our colleagues from Sikom AS who provided useful discussion points.

I would like to thank Boris for having been by my side during these challenging and exciting years. A special thank goes to Debora, for her friendship and support.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study. Thank you.

Martina Ferrando

Contents

Abstract

Ensuring that the energy need predicted by energy modelling corresponds to the actual energy need, will be strictly important in the close future with the increase in the number of low-energy buildings and retrofitting projects and with the refining of regulations. Occupant behaviour is indicated as one of the biggest uncertainties in energy modelling. Thus, an improvement in the description of the occupant behaviour is needed. This thesis discusses one of the possible methods to analyse the monitored data and to apply a deduction approach to improve the energy model from the point of view of the occupant behaviour. A comprehensive data-driven approach for the assessment of the impact of occupant behaviour on the energy need is proposed. The presented methods are applied and validated through a case study regarding a residential building block in Milan, Italy.

In the first phase, a clustering methodology for creating five representative electricity daily load profiles is proposed. The implementation of machine learning techniques emerged from literature as the appropriate tool for the task. A two-level approach is used composed by a Self-Organizing Map, a Neural Network technique, coupled with the k-means algorithm, a classic machine learning method. The k-Nearest Neighbour algorithm is implemented to extend the results to the whole year. In the second phase, a detection method is proposed to estimate the presence of occupants in the household. The technique is based on the analysis of the electricity consumption data to detect the occupancy through the exploitation of the k-Nearest Neighbour algorithm. The extension to the whole year relies on the clustering obtained in the previous task. The resulted presences and electric load clusters emerged with different daily profiles, that can be ascribed to different types of families and habits. Highlighting three scenarios of energy spenders: energy-aware user, standard user and energy-intensive user. In the last phase, the schedules generated in the previous steps are used to assess the impact of the occupant behaviour on the heating energy need during a year and the modelled result is compared with the real registered data.

Keywords: Occupant behaviour, Energy modelling, Clustering techniques, Residential buildings, Energy needs

List of Figures

List of Tables

Definitions and Conventions

ISO 52000-1:2017 [1]

- DELIVERED ENERGY: energy, expressed per energy carrier, supplied to the technical building systems through the assessment boundary, to satisfy the uses taken into account or to produce the exported energy;
- ENERGY NEED FOR HEATING OR COOLING: heat to be delivered to or extracted from a thermally conditioned space to maintain the intended space temperature conditions during a given period of time;
- ENERGY SOURCE: source from which useful energy can be extracted or recovered either directly or by means of a conversion or transformation process. Example: oil or gas fields, coal mines, sun, wind, the ground (geothermal energy), the ocean (wave energy, ocean thermal energy), forests etc.;
- ENERGY USE (FOR SPACE HEATING OR COOLING OR DOMESTIC HOT WATER): energy input to the heating, cooling or domestic hot water system to satisfy the energy need for heating, cooling (including dehumidification) or domestic hot water respectively;
- ENERGY USE FOR LIGHTING: electrical energy input to a lighting system
- ENERGY USE FOR OTHER SERVICES: energy input to appliances providing services. Example: elevators, escalators, home appliances, TV, computers, etc.

EnergyPlus Manual [2]

- THERMAL ZONE: space or a group of spaces having similar conditioning requirements, such as the same thermal setpoints or a single thermal controlling device. It is the thermal unit used in modelling in the software EnergyPlus.
- SCHEDULE: it defines the presence of occupancy, equipment, lighting or HVAC operation, heating and cooling temperature setpoints, transparency and the activation of some components (such as the shading devices). It can be defined directly in the software if it is a series of defined repeated values or can be derived from an external file that assigns to each time-step a specific value.

Comma “,” is used as decimal separator in the whole thesis.

1. Introduction

1.1. Buildings and energy use

All over the world, the building sector and its industry are one of the major users in terms of energy and materials [3]. In the European Union, the residential building sector counts for the 25,4 % of the total yearly energy use [4]. Particularly, in Italy (in 2015 [5]), the residential buildings used 27,9 % of the total final consumptions (Figure 1.1). Therefore, in the last decades, the interest in reducing the impact of the building sector is greatly increased [6]. Specifically, the building industry is coping with the reduction of the energy need. As a matter of fact, the used energy is mainly exploited to ensure adequate comfort levels for the occupants, through heating and cooling systems [7]. Scientists, worldwide, are aware that a change in our way of energy usage is inevitable in the next future. The Intergovernmental Panel on Climate Change's (IPCC) in the Fourth Assessment Report (2007) showed that the potential of emissions' reduction in the residential and commercial fields by 2020 is about 29 % [8].

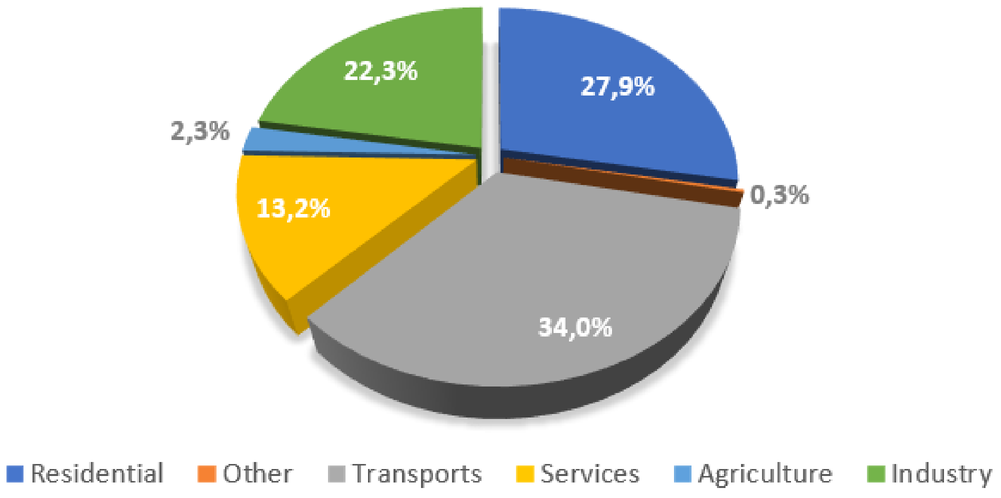


Figure 1.1: Final Energy uses for sector (%) in Italy in 2015. From ref. [5]

In the last years, also authorities started to face the problem. Governments established new policies to decrease the energy use in the building sector [9]. These regulations have the purpose of guaranteeing adequate conditions in the indoor environments but also of increasing the energy performance of future and existing dwellings [3]. At the same time, it is evident a growing public consciousness about global warming and sustainability issues. Governments, economies, and businesses are preparing for a carbon-constrained future [10]. Nevertheless, further efforts need to be made to achieve appreciable results in the environmental sustainability, and to sensitize the users [6].

1.2.European Union's Directives

The European Union has designed the Energy Efficiency Directive [11] (firstly in 2004 and lastly updated in 2012) that mandates legally binding measures to foster energy efficiency and that defines guidelines to help the EU countries to reach their 20 % energy efficiency target by 2020, incrementing the efficiency at different levels of energy production and consume. The policy requirements are minimum targets and each member state, by submitting a National Energy Efficiency Action Plan (NEEAP), can introduce more stringent measures. Annually, each member state has to report the progress towards the national standard. Regarding the buildings, the minimum specified requirements [11] are:

- EU countries have to apply energy efficient renovations to at least 3 % of buildings owned and occupied by central government,
- EU governments must only purchase buildings which are highly energy efficient,
- EU countries have to draw-up long-term national building renovation strategies which can be included in their National Energy Efficiency Action Plans.

European Union has implemented also the Energy Performance of Buildings Directive (EPBD), the first version is dated 2002, then replaced by a new version in 2010, and with a proposed update of 2016 [12]. This legislative instrument, together with the EU Building Stock Observatory [13], has the aim to track the energy performance of buildings and promote the increase of efficiency and the use of smart technology in existing and future buildings. Also, in this case, the directive shows only the minimum requirements that each EU-member state must actuate. It underlines [11] that:

- energy performance certificates are to be included in all advertisements for the sale or rental of buildings,
- EU countries must establish inspection schemes for heating and air conditioning systems or put in place measures with equivalent effect,
- all new buildings must be nearly zero energy buildings by 31st December 2020 (public buildings by 31st December 2018),
- EU countries must set minimum energy performance requirements for new buildings, for the major renovation of buildings, and for the replacement or retrofit of building elements (heating and cooling systems, roofs, walls and so on),
- EU countries have to draw up lists of national financial measures to improve the energy efficiency of buildings, and
- EU countries must draw-up long-term national building renovation strategies which can be included in their National Energy Efficiency Action Plans.

Italy, as EU-member, approved the directives and undertook the commitments for 2020.

1.3.Building simulation modelling and the "performance gap"

According to the objectives expressed by European regulations, advanced building simulation models are becoming increasingly important to support new constructions or renovation design. They are a cost-effective method to assess the impact of buildings and they can be used in the design phase to optimize interventions, in terms of energy, costs and comfort [14] leading towards high-performance and sustainable buildings [15]. Building performance simulation (BPS) can be seen as an economical and technically flexible tool to study a real scale building [7]. Although the aim is to recreate as faithfully as possible a real construction, the model will remain different from the actual physical object. For this reason, the building performance simulator is meant to reproduce a building behaviour with a controlled deviation from reality [6].

The common strategy concerns in modelling a building, and through the BPS software optimizing it to fulfil the regulations. However, the energy demand in buildings does not always decrease after the application of improvements on the building envelope and system as suggested by BPS [8]. Technology alone does not assure energy reduction in buildings, leading to a disparity between predicted and real energy saving in buildings [3,8,9]. A wide number of researchers observed this mismatch between the predicted and the actual energy need of modelled buildings. The phenomenon is called “performance gap” [14]. According to Carlucci et al. [16], this disparity is mainly attributable to: lack of information about the performance of building components that can change during time; inappropriate description of the occupancy behaviour; failure of the building systems or inadequate maintenance or operation; disparities between the design and the construction; lack or inappropriate post-occupancy evaluation; limitations of the quality and spatial density of the used weather data; intrinsic limitations and uncertainties of the energy simulation software.

Studies confirm that Occupant Behaviour (OB) has a key role in the variation in energy need of dwellings [3]. As a matter of fact, neither cities nor buildings use energy, but the occupants do [9]. The software used in modelling usually lack the capacity to describe the individual actions of the building users. The major BPS software (e.g., EnergyPlus, IES, eQuest, TRNSYS, etc.) usually assume generic profiles, which are not able to reproduce the uncertainty and fortuity of occupants’ behaviour or facility managers characteristics over time [17]. The OB field is very complex to be investigated and studied; indeed, social, cultural, and economic factors provide a further significant contribution defining occupants’ attitudes towards energy usage in buildings [6]. Wei et al. [18] highlighted that up to twenty-seven factors can affect the occupants’ space-heating behaviour in residential buildings, grouped into four main categories:

- environmental factors,
- building and system-related factors
- occupant related factors
- and other factors.

Deuble et al. [10] focused the attention on the importance to shift from conceptualizing the occupant as a passive recipient of indoor conditions to the inhabitant who may play a more active role in the maintenance and performance of the building, especially of the high-efficiency ones.

In the retrofitting field, this issue is widely evident and studied. The energy usage of existing buildings is supposed to decrease after an energy refurbishment. But changes in the behaviour of occupants can cause a significant gap between the actual and estimated energy need [19]. There is an increasing interest in the effectiveness of energy upgrades because of the shortfall with the reality, and this deficit is called 'rebound effect' [20]. R. Galvin in his book [21], underlines the difference between what "energy service" and "energy efficiency" mean. The first is a term that defines the benefits people enjoy, and the energy is the product that is consumed to provide these benefits. The energy efficiency is how much the consumption of fuel is turned into benefits for the owner. Especially after a retrofit, a large part of the energy efficiency increment is used to raise the level of energy services. The Figure 1.2 [21] shows the scheme of rebound effect: the improvements in a retrofit increase the energy efficiency (in yellow) that should bring to a fall in energy use (in light blue), part of this theoretical fall is used to increase the energy services (in grey) and so the result is a smaller fall in the energy use (in purple) respect to the theoretical one.

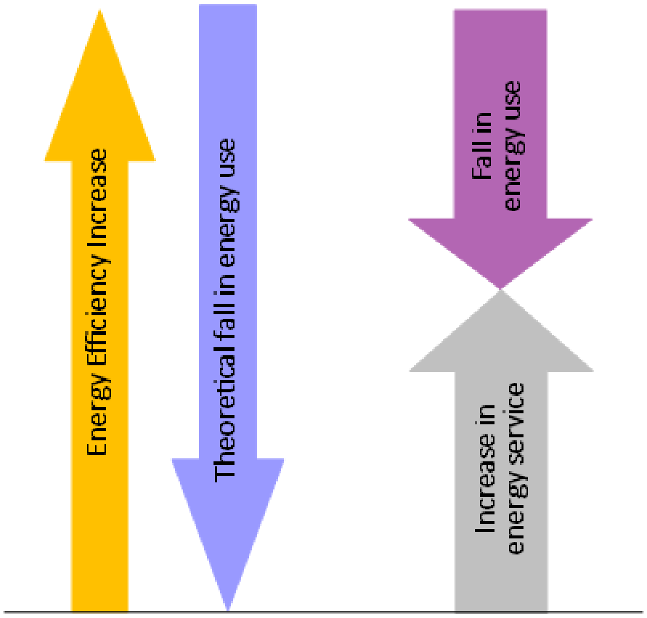

Energy Efficiency Increase
Theoretical fall in energy use
Fall in energy use
Increase in energy service

Figure 1.2: Schematic of the rebound effect. From ref. [21]

To explain the causes of this effect, Galvin states that, supposing a detailed and good optimization and modelling, the problem of rebound effect can be decreased or avoided with education of the owners. This is a very important topic in the perspective of energy efficiency goals of 2020 [11], indeed the slowing down of the energy savings can cause the failure of the objectives of the national energy policies [19]. Also in the Italian annual report on the energy efficiency of April 2017 [5], the ministry of economic development stresses that the retrofitting projects are fundamental to increase the efficiency of the built environment. However, the rebound effect, if not adequately considered, can jeopardize the expected energy savings.

1.4. Occupancy Behaviour

The hypothesis that people in buildings can affect the thermal and energy performance and the consequent emissions is widely studied nowadays [22,23]. People influence energy need passively through their presence, and actively through interactions with the systems, appliances, and devices [6]. The presence of people in a space changes the internal gains: the higher is the number of people, the higher will be the sensible and latent heat produced by them. Whilst, the interactions include settings with operable windows and blinds, thermostats, plug-in appliances and lights [7], that can change respectively the solar gains, the set point for the heating or cooling systems and the internal gains due to electric devices in the space.

The decrease of the interactions between buildings and people cannot be a solution in the close future. User influence is thought to be fundamental for the well-being of the dwellers, especially in housing. Modern buildings designs have to consider this issue and not to avoid it [15]. It is widely believed that occupants prefer a high degree of adaptive opportunities [13], as can be provided within naturally ventilated buildings as opposed to centrally controlled air-conditioned ones [10]. Schakib-Ekbatan et al. [24], underlines that the “desire of control” has a strong impact on the well-being of occupants. Moreover, understanding how the building is supposed to work and how it consumes energy leads to happier and more careful in energy-related actions users [9,10].

Studies have confirmed that behaviour of occupants plays an important role in the differences in energy need, even if the extent of such influence should be studied deeper; indeed, the variation in energy use is still large for dwellings with the same characteristics [3]. Carlucci et al. [6] observed that virtuous behaviours can reduce sensibly the energy need, for both heating and cooling, in poorly insulated buildings. Nevertheless, in highly insulated buildings, the people’s actions can result in a negative effect, increasing the heating or the cooling demand. In any case, it is established that the occupancy behaviour is fundamental in the final consumption of a building. However, the simplistic approaches commonly used by designers in building simulation software are inadequate to study the interaction between users and dwellings [15]. They use similar energy need patterns and profiles that remain static all the yearlong and do not account for occupancy-related behaviour [14].

Moreover, with the increase of efficiency and optimization of building’s characteristics, the sensibility towards occupants and their behaviour on the overall energy need will grow [3,15]. For example, if the electric consumption is assumed to be constant before and after a building retrofit, it means that, after the renovation, the electric consumptions will count more on the final energy demand (due to the decrease of energy spent for heating and cooling). This shows the higher role played by OB in low energy buildings compared to existing and low-efficiency ones. Another consideration can be done on the increased airtightness in modern dwellings: the OB will have a greater effect on the air change rate due to opening and closing of windows and, by consequence, on the energy need [25].

Building’s regulations are becoming stricter, it follows that highly optimized building will be more and more widespread. This fact leads to the emphasis on better knowledge about occupancy behaviour models [6]. It is important not only to model more realistic scenarios but also quantify the impact of them on the energy need.

Concluding, it is clear that OB has a significant impact on energy usage in buildings and represents one of the biggest uncertainties in modelling [16,17,26]. There is therefore a strong need to study not only how the characteristics of the buildings (called by Schweiker et al. [8] the hardware), such as envelopes and the efficiency of mechanical heating and cooling systems, etc, affect the building performance, but also, how the human behaviour (called the software [8]) does it too.

Employing different and more accurate occupancy profiles in the building performance simulators can increase the reliability of the results, or at least, the calculation of their uncertainty bands. On the other hand, this approach makes the modelling process more complex and time-consuming, and this should be taken into account [27]. Simulations to recognize the virtuous behaviours should be investigated in detail. The study of Fabi et al. [9], completed on the university campus of Politecnico of Turin, is quite relevant. They demonstrated that the "informed user" is able to affect the annual primary energy of the till -29 %. Also, the study of Deuble et al. [10] suggests that the people leaving in low energy buildings are tolerant about the internal comfort because they are aware of energy need to maintain certain internal levels, and states that "'green' buildings work best with 'green' occupants".

1.5.IEA and Annex 66

The International Energy Agency (IEA) is focusing on the Occupancy Behaviour to stress how the phenomenon gained importance in the last years [28]. In 1977, the Agreement on Energy in Building and Communities has been created to manage research in building energy efficiency. Their priority research themes are:

- integrated planning and building design,
- building energy systems,
- building envelope,
- community scale methods,
- and real building energy use.

The IEA developed the Annex 66 project, from November 2013 to June 2017. This project aim was "to set up a standard occupant behaviour definition platform, establish a quantitative simulation methodology to model occupant behaviour in buildings and understand the influence of occupant behaviour on building energy use and the indoor environment" [28]. Twenty-four countries and fifty-five organizations participated in the project. The scope was to better understand how people act and how OB can be introduced in the methodologies and in the simulation tools. The final hope is to "bridge the gap" between the built environment and the occupant behaviour (Figure 1.3).

Yan et al. [29] explained how the Annex 66 will help to establish a scientific methodology framework for the simulation of occupant behaviour, comprehending data collection, modelling and evaluation, and software integration. Based on previous related studies [30], Yan et al. grouped occupant models into three types:

- adaptive behaviour models,
- non-adaptive behaviour models,
- and occupancy models.

In the first type, models predict the likelihood of an adaptive action to a given state or variable (for example, the probability of the use of the shading devices at a given outdoor illuminance). In the second typology, the models predict the lifetime of an occupant action or of a state of a building component with which the occupants interact (for example, the lifetime of the state of the lights, on or off, before it is changed). Finally, in the third category, the models predict the presence of people. It can be forecasted as timing or frequency of entrances and exits or as lifetime of an uninterrupted occupancy/vacancy state.

Figure 1.3: Annex 66 - bridging the gap. From ref. [28]

1.6. Occupant behaviour in modelling

Occupant behaviour contributes significantly to the energy need of a building but studying it in depth is not a simple task. The occupancy issue requires a multidisciplinary approach, between engineering and social sciences to be completely understood. A wide range of driving forces has a significant influence on the actions and presence of people in dwellings. IEA, in the final report of Annex 53 [25], grouped these forces into five main groups: biological, psychological, social, time, and physical parameters. It can be inferred that to create a model that assembles all these characteristics can be complex. Nowadays, six main typologies of occupant models are available to describe the phenomenon [6]: psychological models, average value models, deterministic models, probabilistic models, agent-based models, and action-based models. These methods can be used for mainly two aims [25]: modelling the OB to understand the behaviour itself or to reveal the relationships with the energy demand. For the first aim, some examples are the theory of planned behaviour [31] or the MODE model of attitude-behaviour processes [32]. The second aim is the one of more interest concerning the energy field, and different methods can be used to achieve it, the list is reported in Figure 1.4 (according to [25]).

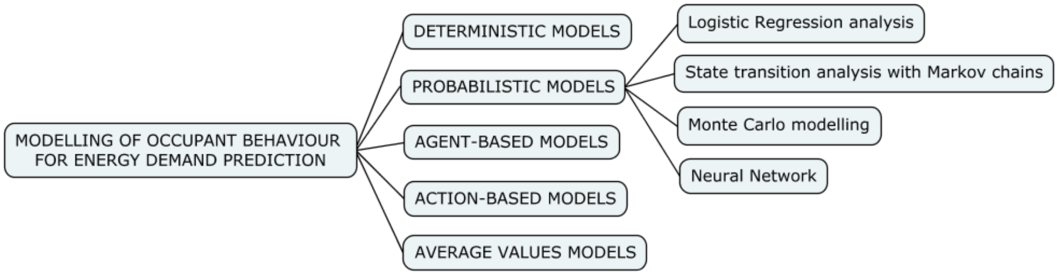


Figure 1.4: Scheme of the main methodology to model occupancy behaviour to predict energy demand in buildings. From ref. [25]

Deterministic models use predefined typology of occupancy as input in computer simulations. The occupancy by nature is related to randomness elements, and the building simulation tools are not capable to include it in the deterministic equations of thermodynamics. The result is that usually the occupants are modelled with predefined fixed schedules or with rules (e.g. lights are turned on if the illuminance level is below a certain threshold).

Probabilistic models use equations or algorithms to predict the probability of a state or action. These methods look at the occupants as if not all the actions were determined by an external or internal stimulus but including random behaviours. They account stochastic factors that have probabilities as results and not “single values”. Different methodologies can be implemented to develop this type of models, such as logistic regressions analysis, statetransition analyses with Markov chains, Monte Carlo modelling, and Artificial Neural Networks (ANN).

Agent-based models allow representing occupants as individuals with autonomous attributes but also simulating the changes of these attributes in time. In these methods, each agent (that can be very different objects, from an individual human being to a component of the energy network) is in a specific state at the beginning of the simulation and it can change state over time due to interactions with other agents. Usually, an agent-based model is used in co-simulation with a building model.

Action-based models define occupancy as actions that can change the state of the location in all the ways (operating with windows, lights, air-conditioning, etc.). The movement in the building spaces and the control actions (opening/closing of windows, turning on/off of lights, of heating and cooling systems, or of the electric appliances) become through patterns, the description of the occupants and their behaviours.

The average value models exploit the parameters of occupancy that influence the total energy use of the building for a specific time-step (daily, weekly, monthly, etc.). They use different analysis methodology (cluster analysis, crowd-sourced inventory databases, genetic algorithm or others) to create a simplified OB model regarding internal gains, domestic hot water use, lighting, appliances, heating and cooling or ventilation rates. They are suitable to be applied to large-scale residences with a large number of very similar or identical flats. The inputs data on occupancy can come from different sources, for example, statistical databases, time-use surveys, monitoring data of case study buildings, questionnaires and surveys, crowd-sourced inventory data, personal observations. This thesis will be focused on one methodology concerning average values but taking some tools from probabilistic models such as the ANN.

1.7.Conclusion

Ensuring that the energy use predicted in the design phase corresponds to the actual energy use, will be strictly important in the close future with the increase in the number of low-energy buildings and retrofitting projects and with the refining of regulations. Thus, an improvement in the description of the occupant behaviour is needed.

This thesis discusses one of the possible methods to analyse the monitored data and to apply a deduction approach to improve the energy model from the point of view of the occupancy behaviour. The presented methodology helps in the creation of yearly schedules for internal gains by appliances and presence of people for the energy modelling based on real monitored data. Differently from the most used deterministic approaches, this is a stochastic methodology that is able to include the intrinsic variability of occupants in the residential buildings.

The input data is the hourly electric consumption per single flat, registered in some periods of the year. Using these data, the research defines a methodology to improve buildings' occupancy assessment in energy modelling and defines average schedules for electric demand and users' presence, which can be replicated in the context of residential housing. The process can be divided into three tasks:

1. Generation of yearly schedules of electric demand starting from incomplete 15-min real registrations,
2. Detection of yearly occupancy schedules from electricity consumption data,
3. Assessment of the impact of different occupancy (considering both presence and electric consumptions) on the overall energy demand of the building.

2. State of the Art

This thesis has the final aim of assessing the impact of occupants related schedules on the energy use of a residential building. To achieve the goal, a process made of subtasks has been established. The following state of art illustrates the approaches found in the literature to solve the problems faced in this work. It gives just a general overview of the assessment of occupants' impact on energy modelling and then it goes into details in the review of the methods for clustering electric load data in buildings and deriving the potential presence of occupants from them.

2.1. Introduction

The uncertainties such as weather datasets, internal gains, efficiencies, occupants, etc., affect the results of energy modelling of buildings. To deal with this issue a single and unique output from an energy software is no more seen as an adequate and sufficient result. The tendency is to give a range of results, in which the actual energy demand will be, counting all the uncertainties.

The traditional approach to energy modelling is the creation of a model with a deterministic set of inputs and the calculation of the energy use of this model. However, to better face the problem it is necessary to develop a set of models with different variables as inputs and calculate all the energy outputs of these models.

Sun et al. [26], proposed a methodology to estimate the performance of energy conservation measures (ECM) that are influenced by uncertainties. The focus is mainly on occupants' behaviour that is indicated as one of the biggest uncertainties that affect the effectiveness of building retrofits. Three types of OB style (austerity, normal and wasteful) were defined to represent different levels of energy consciousness in terms of the control of HVAC, window, lights and plug-in equipment. These behaviour's styles are chosen to be representative of the extreme energy savers and spenders. The conclusion is that buildings occupied by energy spenders could consume more than twice the energy of the energy savers. However, the most interesting part is the actual methodology that starts from the variation of the inputs and ends with a range of results, in contrast with the traditional approach, as seen in Figure 2.1.

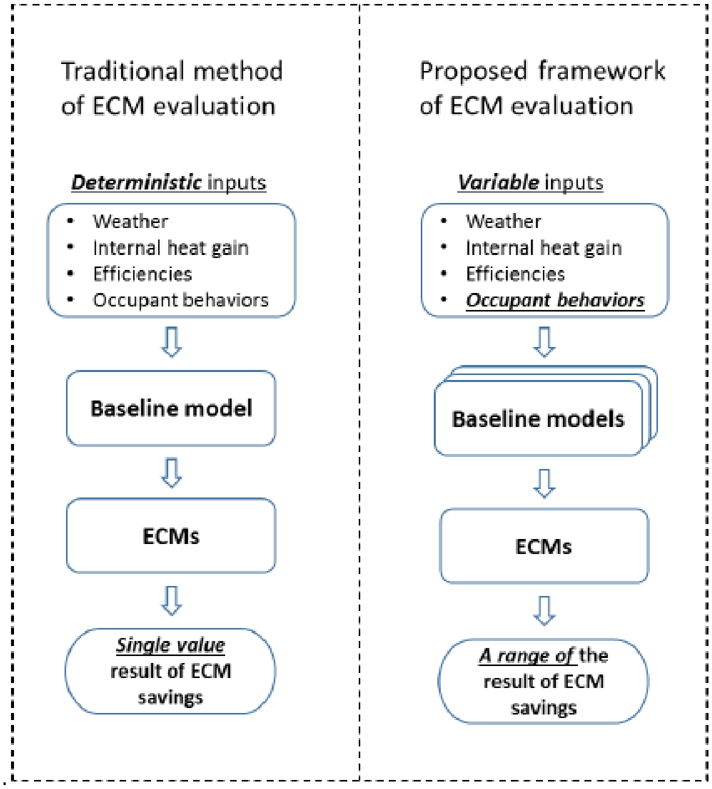


Figure 2.1: Traditional and proposed approaches to energy conservation measures evaluation in building energy modelling. From ref. [26]

Azar et al. [17] proposed a framework for Building Performance Simulation (BPS) and Agent-Based Modelling (ABM) using a regression surrogate model. Such methodology tries to overcome the limitations of BPS in modelling human behaviour. In this study, the impact of uncertainty in human actions on energy performance is quantified. The conclusion is that the level of control given to occupants and/or facility managers, and the way in which these uncertainties are taken into account can influence the energy performance of buildings and can highly change the range of the possible actual results (Figure 2.2).

Also, the study of Gaetani et al. [33], shows how uncertainties can influence the building performance predictions. In their opinion, it is crucial to include the modelling of uncertainties within BPS models. The final aim of the study is a step-by-step strategy – the fit-for-purpose OB modelling (FFP-OBm) strategy – to select the appropriate OB modelling complexity. However, simulations are run also to assess the influence of thermophysical properties uncertainty on energy demand for cooling and heating and to assess the effects of increasing modelling complexity for light and blinds operation. The result is a series of ranges and still not a single value that tries to predict "precisely" the actual energy demand of the considered building office.

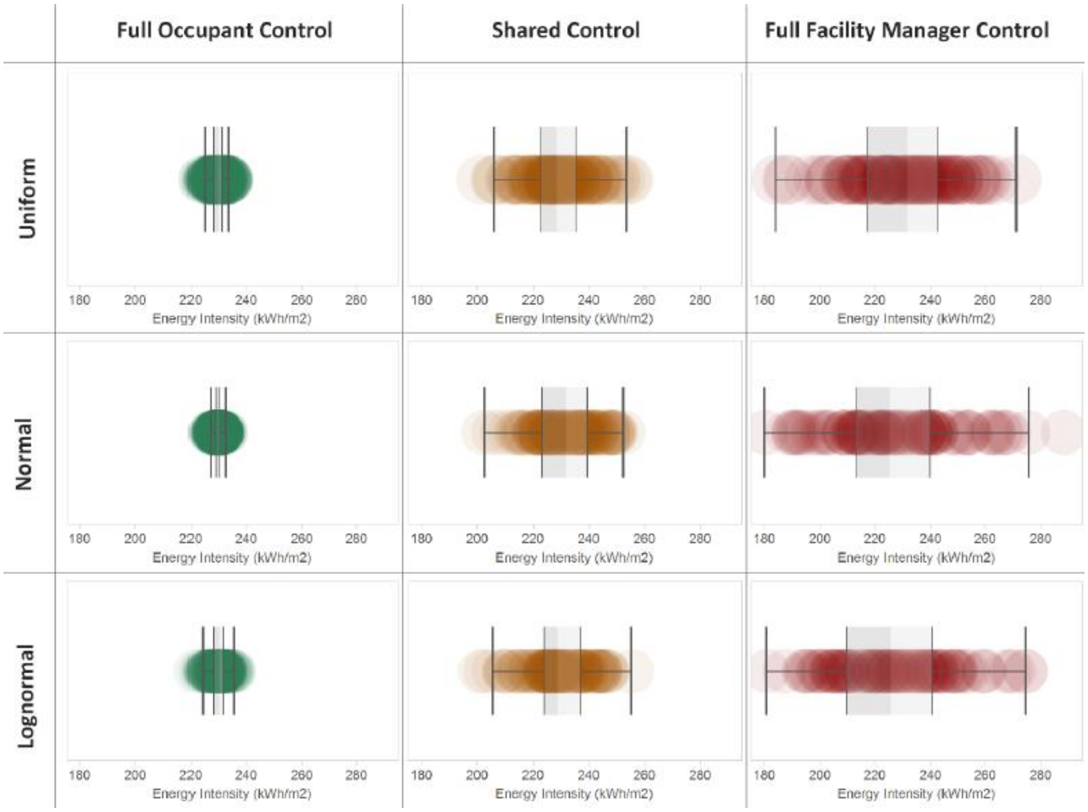


Figure 2.2: Ranges of energy results based on different level of controls and different uncertainties analyses. From ref. [17]

2.2. Electric load clustering

This thesis tries to find a stochastic methodology, in contrast with the standard deterministic approaches, to deal with the occupants' behaviour in a real case study. To achieve this goal, clusters with different scenarios of internal gains due to occupants should be developed starting from the registered electric data. Regarding this issue, the recent studies are mainly centred on the deeper understanding of customers' daily load patterns for electric suppliers. However, they are mainly on large scale and/or on not-residential buildings.

Chicco et al. [34] studied Load Pattern-Based Classification of Electricity Customers with the aim to gain accurate knowledge of the customers' consumption patterns for electricity providers in competitive electricity markets. In their study, two methods were implemented to achieve the result: a modified follow-the-leader algorithm and a self-organizing map. The conclusion is that both the two methods can effectively assist the electricity providers in performing customer classification. The different, but to some extent complementary, characteristics of the two methods, suggest using them in a way depending on the objectives. This paper shows a case study not related to residential buildings though. All surveyed customers were industrial, services, and small-business activity buildings. Moreover, a sorting on the original data was performed since the measured load patterns refer only to weekdays. Also in the following work of Gianfranco Chicco [35], the case study consists of 400 load patterns related to non-residential costumers in a representative weekday of the intermediate season only.

Tsekouras et al. [36] developed a two-stage methodology for the classification of electricity customers of the Greek power system. It is based on unsupervised pattern recognition methods, like k-means, Kohonen adaptive vector quantization, fuzzy k-means, and hierarchical clustering. In the first stage, representative load curves of various customers are deducted with the help of pattern recognition methods. In the second stage, a classification of the customers is carried out with the same methods of the first stage. The contributions of this research are the formation of typical daily load curves for each customer, the optimization of the unsupervised pattern recognition methods' features and the comparison of the performance of the clustering algorithms. However, the study considered only the industrial customers of the company, being of more interest in that specific case.

Hernández et al. [37] developed a well-structured methodology composed by a cascade application of a Self-Organizing Map (SOM) and the clustering k-means algorithm to understand the energy consumption patterns. They decreased the environments of study from regions and nations to smaller ones. The results show that the system adequately finds different behaviour patterns without supervision and without any prior knowledge about the data. The study gives very good results, but it concerns an industrial park in Spain, in which the differences between weekend and weekdays are quite clear, and it is not tested in a residential case study. The research pointed out how the cascade method of SOM and k-means is helpful in noise reduction in confused databases, referring to the work of Vesanto et al. [38] focused on Clustering of Self-Organizing Map.

Deshani et al. [39] proposed an accurate prediction of electricity demand through improved artificial intelligent approaches. This research shows how a cluster analysis performed to group similar day types, could contribute towards selecting a better set of neuro-forecasters in neural networks. Daily total electricity demands for five years were considered for the analysis and each date was assigned to one of the thirteen day-types. Three different clusters were found using Silhouette plots, and thus three neuro-forecasters were used for predictions. This paper illustrates the proposed modified neural network procedure using electricity demand data for all Sri Lanka, analysing the daily total electricity demands and not focusing on daily load patterns nor residential buildings.

Panapakidis et al. [40] developed a methodology for the investigation of the electrical behaviour of buildings, using clustering techniques, exploiting the incorporation of smart grid technologies in the building sector. Utilizing a university campus as a case study, the proposed methodology is applied to the load curves of different buildings leading to the determination of an optimum clustering procedure. In fact, the spread of the smart grid technologies enables automatic collection of information of the customer's behaviour along with the building's performance. However, the large quantity of recorded data requires efficient processing and interpretation. The methodology presented in this research is well described, however, the residential sector, in which a lot of variables and uncertainties play important roles, is still not studied.

Also in the study of Grzegorz Dudek [41], to solve the issue of clustering and forecasting, the importance of Machine Learning techniques is stressed. In this work, several methods based on neural networks are proposed and compared, such as multilayer perceptron, radial basis function neural network, generalized regression neural network, fuzzy counter propagation neural networks, and self-organizing maps. The study nevertheless, concerns data of all Poland on a national scale and the method is not implemented on the single residential building.

Finally, the new work of Capozzoli et al. [42] proposed a general framework on load profiles characterisation in buildings, based on the recent scientific literature. The methodology concerns a combination of different pattern recognition and classification algorithms. The novelty of the research is the multi-level and multi-scale implementation of the methods, from sub-system to whole building and from a single building to the stock of buildings. Anyhow, the case study does not concern a residential building, in which the daily load patterns are noisy and confused.

Some other researchers focused their attention on the residential sector. However, the methodologies always imply very big data samples, or they are accompanied by the detailed survey.

Rhodes et al. [43] studied the measured electricity use data from 103 homes in Austin, Texas to determine the shape of demand profiles, to optimise the number of normalized representative profiles and to draw correlations based on survey data from occupants. The k-means algorithm was implemented to cluster the electricity patterns and a regression method was used to determine if homeowner survey responses could serve as predictors for the clustering results. This analysis found that Austin homes fall into one of two seasonal groups with some homes using more expensive electricity than others. The regression results indicate that variables concerning the characteristics of the family (work, hours of television watched per week, and education levels) have significant correlations with average profile shape.

Also, McLoughlin et al. [44] proposed a clustering methodology in the residential sector for Ireland starting from electricity smart metering data. They used the method of data mining, that allows for the data to be segmented before aggregation processes are applied. Moreover, segmentation allows for dimension reduction thus enabling easier manipulation of the data. The study implemented three of the most widely used unsupervised clustering methods: k-means, k-medoid and Self Organising Maps (SOM). The best performing technique is then evaluated in order to segment individual households into clusters based on their pattern of electricity use across the day. After this process, each class load profile is linked to household characteristics by applying a multi-nominal logistic regression to the data. As a result, households and the manner with which they use electricity in the home can be characterised based on individual customer attributes.

Both the studies of Viegas et al. [45,46] include surveys to better understand the differences between the tenants of residential buildings. In the first work, they proposed a methodology predicting the typical daily load profile of electricity usage based on static data obtained from surveys, with the intent to determine consumer segments based on the metering data using the k-means clustering algorithm, to correlate survey data to the segments, and to develop statistical and machine learning classification models to predict the demand profile of the consumers. Whilst in the second work, they proposed a methodology for predicting the typical daily load profile of electricity usage based on static data using fuzzy clustering and modelling.

The new work of Ali et al. [47] started from a data set similar to the one available for the study reported in the present thesis, although with a far wider data sample of 400 houses. They proposed a study on data mining techniques (including exploratory data analysis and pre-processing, frequent patterns mining and associations, classification /characterization, clustering and outlier deduction) to explain and evaluate which techniques is useful for the better understanding of electricity load profile consumption data to improve the new power system by understanding large-scale load profile data.

2.3. Occupants' presence deduction

The second task of this thesis is the determination of occupancy profiles. This topic represents one of the main issues in commercial and residential buildings. For example, heating, cooling, ventilation and lighting systems depend on the estimation of occupancy to work correctly. Even if the potential applications of occupancy detection are numerous, it is still complex and expensive. The analyses with sensors, like in the work of Jorissen et al. [48] or Kim et al. [49], are the common method to register very big data of occupancy. The prevailing detectors are passive infrared (PIR) sensors, cameras or magnetic reed switches. Nevertheless, they are expensive and complex, they must be purchased, installed, calibrated, powered and maintained. Moreover, the privacy issues inhibit the implementation of such methods in the residential buildings.

Another quite simple method to gain data on occupancy is Time-User Surveys (TUS), as reported in the study of Aerts et al. [50]. Their methodology uses data from Belgium Time Use Survey of 2005. The novelty of the model is the introduction of "typical occupancy patterns". Hierarchical clustering techniques on individual occupancy profiles are implemented and then, probabilistic occupancy profiles are obtained applying the probability to transit from a certain state to another and the duration probability, which both are time-dependent. Also, the methodology proposed by Buttitta et al. [22]

introduced a new occupancy model from Time Use Survey data, using data mining clustering techniques. The methodology is divided into two steps: identification and grouping of households with similar daily occupancy profiles, and then, the creation of probabilistic occupancy profiles. However, these relatively simple methods can be only used in residential dwelling energy modelling that use occupancy time-series as inputs, not yet available in this case study.

The papers of Kleiminger et al. [51,52] are in the intersection between two main areas of study:

- sensing deployments to detect occupancy and improve energy efficiency in residential and commercial buildings;
- analysis of electricity consumption data to observe and influence users' electricity consumption behaviour.

The presented methodology, firstly proposed by Kleiminger in his work of 2013 [51]and then improved in his work of 2015 [52] exploited the electricity meters as occupancy sensors. They showed that supervised machine learning algorithms can extract occupancy information with an accuracy between 83% and 94%. They use a feature set of 10 and 35 characteristics of the registered electric load that are related somehow to the activation state of appliances, hence to the presence of occupants.

2.4. Conclusion

This thesis work is focused specifically on residential buildings provided with a relatively small data sample and no surveys. This represents a typical work condition for professional though it has not yet been widely studied in the literature. Moreover, the proposed methodology is a comprehensive process that can be followed to improve the energy modelling results from the point of view of occupants' behaviour.

Specifically, the work of Ali et al. [47], is relevant for the overall process of data mining techniques that can be followed. Adaptation of the studies of Hernández et al. [37] and Vesanto et al. [38] can be implemented and optimized to achieve the first task of the problem. Finally, to accomplish the task of presence detection, the papers of Kleiminger et al. [51,52] are adjusted and refitted to the characteristics of the available data.

From the nature of the available data and its size, data mining and unsupervised machine learning emerged as promising techniques in this case study, for both tasks. These methods are indeed useful for noise reduction [53] and for pattern recognition in a wide variety of data samples (confused, large-scale, small, etc.).

3. Machine Learning and Artificial Neural Networks

Due to the characteristics of the database available for the case study reported in this thesis, common statistical methods are not sufficient to extract an adequate insight. Machine Learning techniques were therefore implemented. These complex topics are usually outside of the building engineering educational programs; thus, this chapter is intended to provide a knowledge background and a brief summary of the theory of Machine Learning and Artificial Neural Network.

The theory explained in this chapter is extracted from the following source books: Introduction to Machine Learning by Ethem Alpaydin [54], Neural Network and Machine Learning by Simon Haykin [55], Machine Learning by Tom Mitchell [56], A Brief Introduction to Neural Networks by David Kriesel [57], Business Intelligence by Carlo Vercellis [58], Fundamentals of Neural Networks by Laurene Fausett [59].

3.1. Machine learning

The sequence of instructions to reach an output starting from an input is called algorithm. Sometimes, the instruction is easy and available, but, in complex problems, this is not always true. Analyses such as pattern recognition, regression or classification in a large dataset can no longer be done through manual processes. The idea is to transfer the task to find the best configuration of an algorithm to solve a problem to the machine. Moreover, people who are able to perform such analyses are rare and manual analysis is time expensive. For these reasons, there is a growing interest in techniques that can be run automatically by machine that can analyse data and automatically extract information from them, learning from them. Obviously, in most cases, a base knowledge has to be provided to the machine, to learn from it. Machine learning means programming computers to find the best solution to a problem using past experiences or example data. The assumption of the machine is that the near future will be quite similar to the time in which the example data was collected, and the prediction also is expected to be right. The model may be predictive, to make predictions in the future, or descriptive, to gain knowledge from data, or both. The machine will be able to detect regularities and patterns in the close future learning from the past, even if the process may be difficult to be identified completely, result with a good and useful approximation is possible. Psychology, cognitive science or neuroscience aim to understand the process underlying the learning process in animals and humans. On the contrary, in engineering, the aim of machine learning is to build useful systems to fit models to data and reproduce the process of induction. Scientists collect data making experiments and observations, then they try to extract knowledge finding a simple model that explains what the observed in the data. This is what machine learning does: it extracts general rules from a set of examples to generalize the outputs to new similar cases.

3.1.1. Learning processes

Based on the presence or not of a target feature, the learning process can be subdivided into three main categories: supervised, unsupervised and reinforcement learning.

In supervised learning, a target feature is set for each record: it represents the ground truth, the sure result, and can be a class or a continuous quantity. In unsupervised learning, no target feature is expressed, and the aim is to find similarities and differences in the data. The group of reinforcement learning methods is characterized by the presence of a target, but not an absolute one as in supervised learning, and it depends on a sequence of targets.

Supervised learning

The supervised learning can be thought as learning with a "teacher". The "teacher" has knowledge of the environment and teaches to the machine a target attribute with a set of input-output examples without explaining the environment itself. The first example is submitted to the machine and the teacher; so the teacher provides the machine the desired response for that example. The machine can adjust its features to reach the teacher's response as close as possible. The adjustments are performed iteratively, step-by-step, till the smaller error between the teacher's output (the desired response) and the machine's output (the actual response) is reached. After a series of examples with the teacher, the machine can be trusted to be left by itself with new examples. The supervised learning processes are oriented towards prediction and interpretation of the data based on a target attribute, that has to be provided.

To summarize, the supervised learning procedures correspond to the following steps:

1. entering the input pattern,
2. forward propagation of the input and generation of the output,
3. comparing the output with the desired output (teaching input) provides an error vector,
4. corrections of the features are calculated based on the error vector,
5. corrections are applied to the features of the model.

Unsupervised learning

In unsupervised learning, there is no external teacher to oversee the process. Hence, there are not labelled examples from which the machine can learn. The aim, in this case, is to find regularities in the input. The data are structured such that certain patterns occur more often than others and the machine itself finds the way to group the similarities.

The unsupervised learning procedures correspond to the following steps:

1. initializing the regularity groups,
2. dividing all the data into these groups,
3. updating the group's features based on the data,
4. if the grouping is unchanged, stop or, return to step 2.

One of the drawbacks of this procedure is that is not always possible to understand how the machine learnt to deal with the data. Some features, which are important for a scientist, might not be important for the machine and vice-versa. Anyway, unsupervised learning is a powerful and reliable technique that can be helpful when nothing can be set for sure in the dataset.

Reinforcement learning

The reinforcement learning is linked with sequences of output. In some cases, a single output is not important, but a series of output is significant. It is not possible to define the best output in an intermediate state, but an output is good if it is part of a good sequence. Example of good series of output is given to the machine as the desired response, but it is the machine itself that finds the relation between the single action and the good series. Reinforcement learning is difficult to perform for two basic reasons:

- there is not a teacher that provides the desired output step by step,
- the learning machine must be able to assign credit and blame individually to each input in the sequence of time steps that led to the outcome.

3.1.2. Applications

Every typology of the learning process has its own practical applications (Figure 3.1): the supervised methods are implemented to classifications and regression analyses; the unsupervised learning processes are used to clustering analyses and dimensionality reduction in a big dataset; the reinforcement methods are used to learn to react to a series of data.

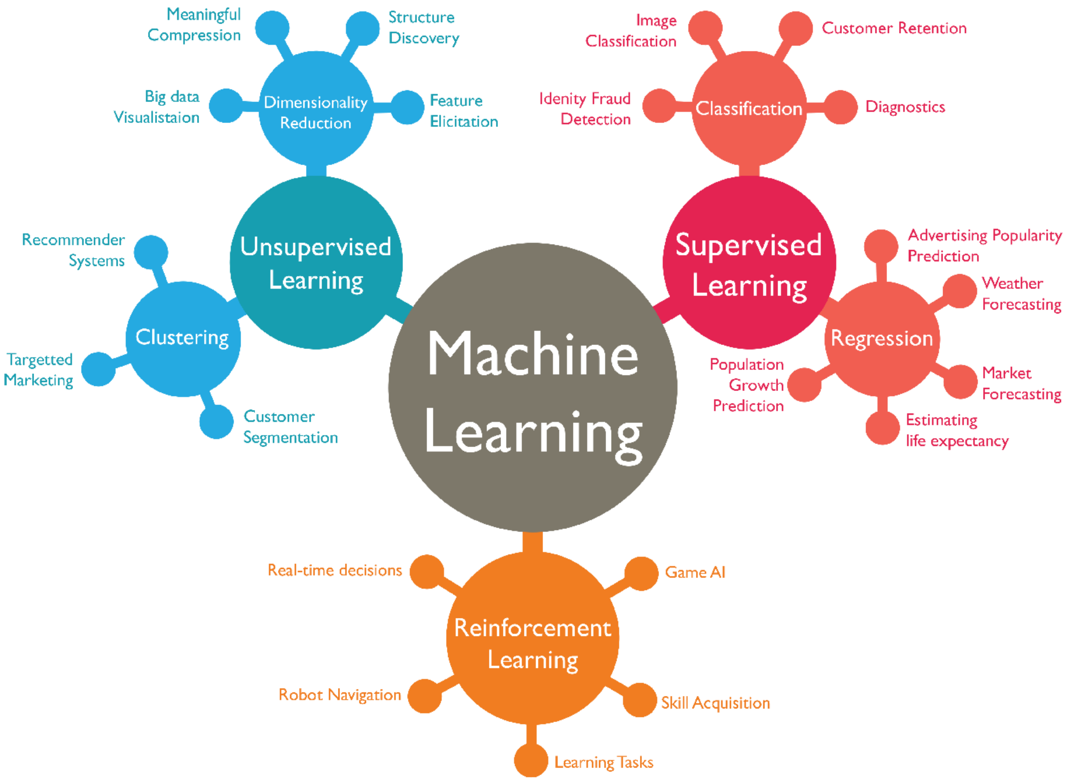


Figure 3.1: Applications of machine learning and of learning processes. From ref. [60]

Six basic machine learning applications can be identified:

- classification,
- regression,

- time-series analysis,
- clustering,
- association rules,
- description and visualization.

The first three tasks use supervised processes since a target sample exists that must be explained, based on available features or throughout its evolution in time. The other three tasks use unsupervised methods and their goal is to express the interrelationships among the available features.

Classification

Classification problems are the simplest example of the supervised process. Usually, each record of a dataset, whose target class is known, is accessible. The algorithm is able to predict the target class of future observations learning from the available observations of the past. In classification, the target attribute is a categorical variable. Therefore, the algorithm output is part of a finite and usually small number of targets. In most applications, the output can also be represented by a simple binary variable. It is the categorical nature of the output that distinguishes the classification from the regression analyses.

Regression

The process is the same of classification, but in this case, the target is not discrete but continuous. Again, looking at past examples, the algorithm can predict a target value for each new observation. Sometimes, a classification problem may be turned into a regression problem and vice versa.

Time series

In this typology a temporal dynamic is present. The algorithm is used to predict the value of the target variable for one or more future periods.

Clustering

Clustering means grouping a data set into N number of clusters C_i, i =1, 2, ..., N. Usually, it implies that this partitioning is unsupervised. Clustering algorithms are able to divide the data into a number of groups trying to minimize some criterion or error functions. A cluster is a homogeneous subgroup existing within a population. Clustering processes are able to divide the dataset into a given number of groups sharing similar features so that the data in different clusters have distinctive characteristics. Clustering methods are unsupervised procedures because there are no predefined classes or reference examples indicating the target class, so the machine itself finds the useful features for the subdivision.

Association rules

Association rules are implemented to discover recurring associations between groups in a dataset. They are different from clustering because they do not divide the dataset into groups, but they only underline frequent relations in the features of the dataset.

Description and visualization

Sometimes, the purpose is to provide a simple and concise representation of the information stored in a large dataset. Differently, from clustering and association rules, the descriptive analysis does not subdivide the dataset. These techniques are used to reach a concise description and representation of a complex dataset.

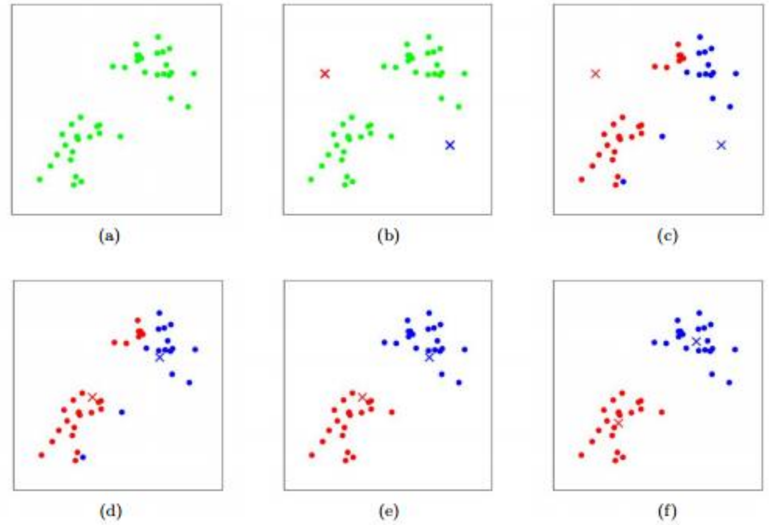
3.1.3. The algorithms

In this paragraph, the algorithms used in this thesis are explained more in detail.

The k-means

The *k*-means algorithm is one of the simplest and most commonly used unsupervised learning algorithms. It solves the problem of clustering given a fixed number (*k*) of centroids, one for each cluster. The algorithm takes each input of the dataset and associates it to the nearest centroid, in this way the early grouping is done. Next, it recalculates the *k* centroids as barycentre of the clusters resulting from the previous step. The inputs are again associate to the centroids and a new complete iteration is computed. The calculation ends when the centroids change their location of a meaningful distance and the inputs associated with a specific centroid become the cluster. For this clustering technique, the input can move from cluster to cluster at each timestep, during the analysis.

An example is given in Figures 3.2 a-f. The training examples are shown as dots (Figure 3.2 a) and the aim of the analysis is to divide them into two clusters. The first two centroids are set randomly (Figure 3.2 b). The clusters' centroids are shown as crosses. The initialization is done classifying the closest dots to the first crosses (Figure 3.2 c) and calculating the new centroids (Figure 3.2 d). Then these two steps are repeated iteratively (Figure 3.2 e) till the last grouping is performed (Figure 3.2 f). In this simple example, the last Figure is obviously the best result, but some real problems are not so simple, and they need thousands of iterations to finish. For these complex cases, usually, a maximum number of iteration is fixed also to decrease the computation time.

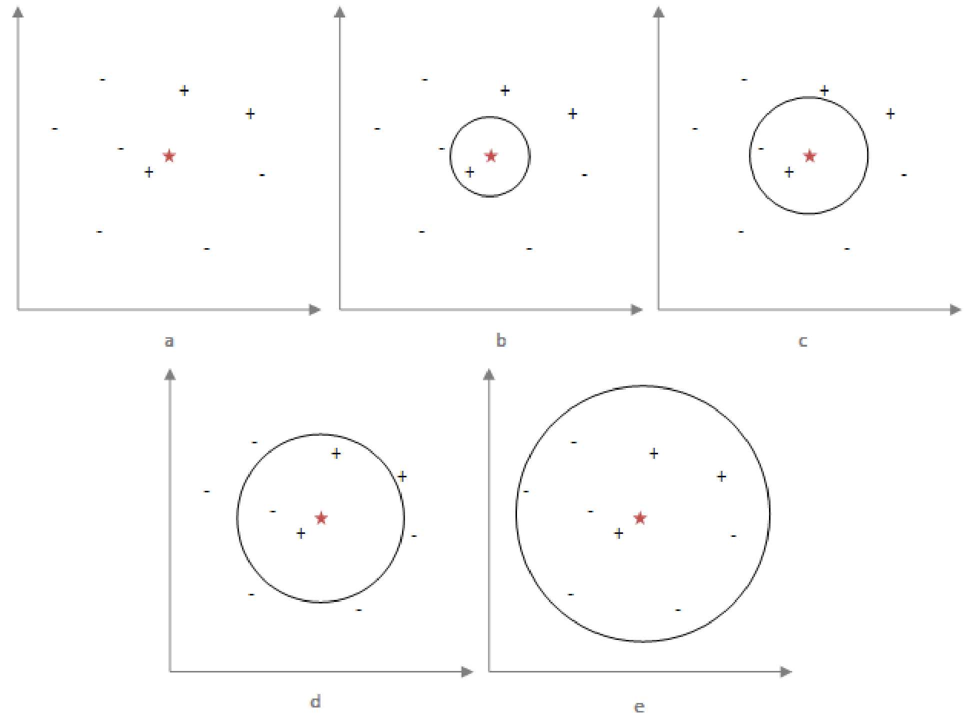


Figures 3.2: Representation of the functioning of the k-means algorithm: Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) nearest initial cluster centroids. (c-f) Illustration of running two iterations of k-means. From ref. [61]

K-Nearest-Neighbours

The *k*-Nearest-Neighbours (KNN) algorithm is used as supervised or unsupervised learning algorithms for clustering a dataset. It connects each data point to the *k* closest neighbours. Then, these groups are defined as clusters. The advantage of this procedure is that the number of clusters occurs by itself. It simply chooses the output looking at the k-nearest examples.

A simple example is represented in Figure 3.3a-e. The starting space is composed of three plus and six minus symbols and the aim is to understand to which group the red star belongs (Figure 3.3 a). Imposing k=1 the star will be in the plus group (Figure 3.3 b), increasing k=2 the star cannot be classified, going to k=3 the star is again in the plus group. Going further, till k=9 (that represent all the known sample) the star will be classified as minus because in the total group the minus is present in a larger amount (Figure 3.3e). Even if the star is very close to the three plus present in the sample, it can be classified as plus or minus depending on the number of neighbours considered.



Figures 3.3: Example of KNN the problem (b) clustering through KNN with k=1 and the output will be “plus”, (c) k=2 and the output will be unknown, (d) k=3 and the output will be “plus”, (e) k=9 and the output will be “minus”

3.2. Artificial Neural Networks

Artificial Neural Networks (ANN), consist of computational algorithms that try to simulate the behaviour of a biologic brain and its neurons. They are programmed to be capable of machine learning and pattern recognition and consequently to solve many types of problems, such as mapping, clustering or constrained optimization.

3.2.1. Biological motivations and components

The study of the ANN started around 1950 and its first aim was to understand the brain and to emulate its capacities. The brain is characterized by the capability to learn and change. Moreover, the notions are stored in a distributed way, and errors and faults affect in a minor measure its memory. On the other hand, computers are able to perform very complex calculations in a very short time, but they are passive: the largest part is only memory and data storage.

The characteristics, taken from Biology, are:

- self-organization and learning capability,

- generalization capability,
- fault tolerance.

Clearly, the starting point was the knowledge of the brain and how it works. Hence, the ANN were developed starting from the following assumptions:

- information processing occurs in many simple elements called neurons,
- signals are passed between neuron through connections links called synapses,
- each connection link has an associated weight, which, in a typical neural net, multiplies the signal transmitted,
- each neuron applies an activation function (usually nonlinear) to its net input (sum of weighted input signals) to determine its output signal.

3.2.2. The Neuron

To deeply understand how an ANN works, the starting point is studying its basic, the neuron. It is composed of three elements (Figure 3.4.):

- a group of connecting links, also called synapses, each of which is characterized by a weight or strength. Specifically, a signal x_j at the input of synapse j connected to neuron k is multiplied by the synaptic weight w_{kj}. The first subscript in w_{kj} refers to the neuron in question, and the second subscript refers to the input end of the synapse to which the weight refers. The synaptic weight of an artificial neuron may lie in a range that includes negative as well as positive values,
- an adder for summing the input signals, weighted by the respective synaptic strengths of the neuron; the operations described here constitute a linear combiner.
- an activation function to limit the amplitude of the output of a neuron. The activation function limits the permissible amplitude range of the output signal to some finite value.

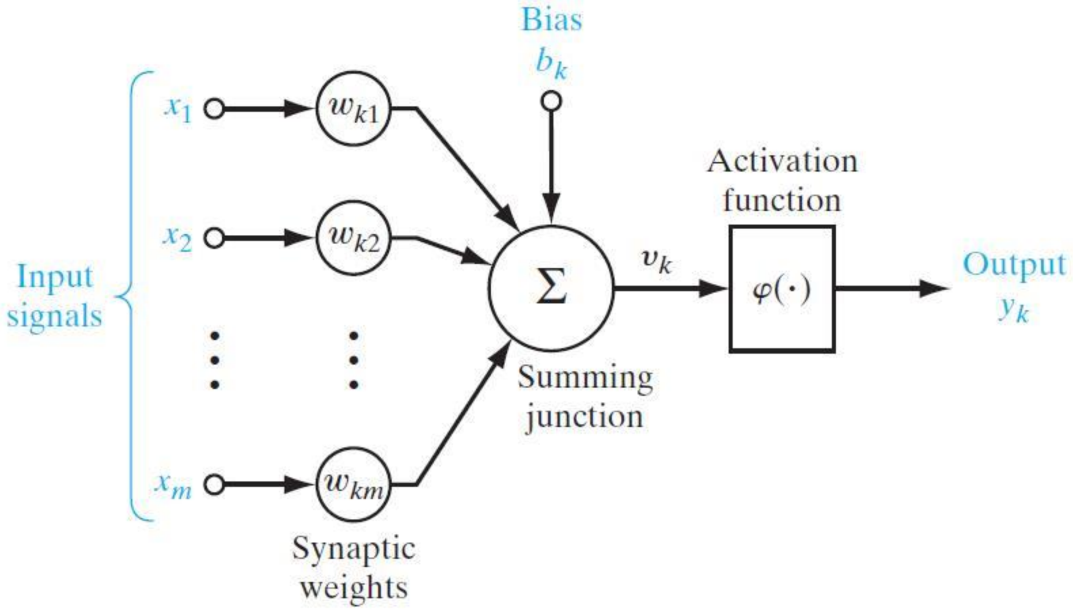


Figure 3.4: Model of a neuron, labelled k. From ref. [55]

The externally applied bias, denoted by b_k, can be present to increase or lower the net input of the activation function.

In mathematical terms the neuron *k* can be described by the equations:

$$v_k = \sum_{j=0}^{m} \omega_{kj} \cdot x_j$$

$$y_k = \varphi(v_k)$$

In which

$$v_k = u_k + b_k$$

where x_1, x_2, ..., x_m are the input signals; w_{k1}, w_{k2}, ..., w_{km} are the respective synaptic weights of neuron k; u_k (not shown in Figure 2.2.) is the linear combiner output due to the input signals; b_k is the bias; φ is the activation function; and y_k is the output signal of the neuron. The use of bias b_k has the effect of applying an affine transformation to the output u_k of the linear combiner in the model of Figure 3.2.

The activation function, defined as φ(v), defines the final output of a neuron with an induced local field v. Two main types of activation functions are described: the threshold function and the sigmoid function.

The threshold function, also called Heaviside function (Figure 3.5.a), is defined as:

$$\varphi(v) = \begin{cases} 1 & if\ v \geq 0 \\ 0 & if\ v\ < 0 \end{cases}$$

Hence, the output of the neuron *k* will be:

$$y_k = \begin{cases} 1 & if\ v_k \geq 0 \\ 0 & if\ v_k\ < 0 \end{cases}$$

In which

$$v_k = \sum_{j=0}^{m} \omega_{kj} \cdot x_j + b_k$$

The sigmoid function (Figure 3.5.b) is a strictly increasing function that is characterized by a graceful balance between linear and nonlinear behaviour. An example is a logistic function defined by:

$$\varphi(v) = \frac{1}{1 + \exp(-av)}$$

In which a is the slope parameter of the function. With the variation of the parameter a, the sigmoid functions can have different slopes, for a approaching infinite the sigmoid function becomes a threshold function. This type of function assumes a continuous range of values from 0 to 1 and is differentiable.

In non-mathematical terms, the input signals reach the neuron and are weighted in the links, and the sum of all these signals is transferred to the activation function. The function gives back the output referred to the range in which the sum of the inputs lays. After one cycle, it is understood if the neuron is activated or not.

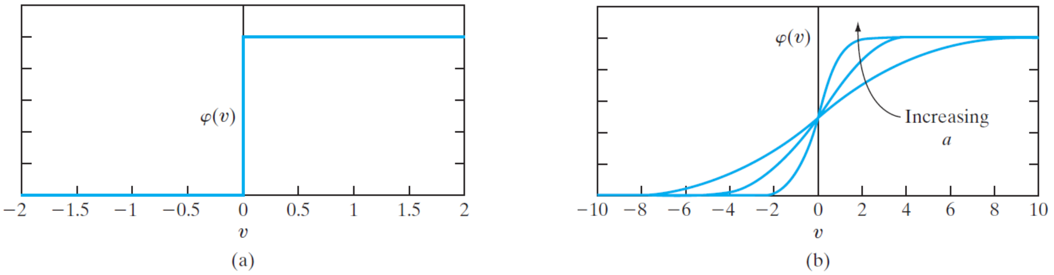


Figure 3.5: (a) threshold function, (b) sigmoid function for varying a. From ref. [55]

3.2.3. Network architectures

Neurons are usually organized in layers, how these layers are organized, and their general structure is called architecture. Three main classes of network architectures (further organization of the layers) can be identified.

The simplest form of layered network is characterized by input layers that projects into an output layer of neurons (computation nodes), but not vice versa. This typology is called Single-Layer Feedforward Networks (Figure 3.6.a). The input layer is not counted because no computation is performed there.

The second class of the feedforward neural networks is characterized by one or more hidden layers. The term "hidden" refers to the fact that these neurons are not seen directly from the input or the output layers, but they intervene between them. In this case, the computation layers are more than one, and for this reason, are called Multilayer Feedforward Networks (Figure 3.6. b). The input layer is connected to the first hidden layer, and the output signal is transferred to the third layer that can be the output itself or another hidden layer, and so on. Typically, the neurons in each layer of the network have as their inputs the output signals of the preceding layer only. A feedforward network with m source nodes, h_1 neurons in the first hidden layer, h_2 neurons in the second hidden layer, and q neurons in the output layer is referred to as an x–h_1–h_2–y network (e.g. the network in Figure 3.6. b, is called 10-4-2 network). This type of network can be fully connected, when every node in each layer is connected to every node in the next layer, or partially connected when some of the links are missing.

The third main class, of less interest in this case study, is composed of the recurrent neural networks. In this typology, there is at least one feedback loop. These loops can be self-feedback (within the same neuron), or between neurons in the same layers. The presence of feedback loops affects the learning capability of the network and its performance but increases the complexity and the running time of a network.

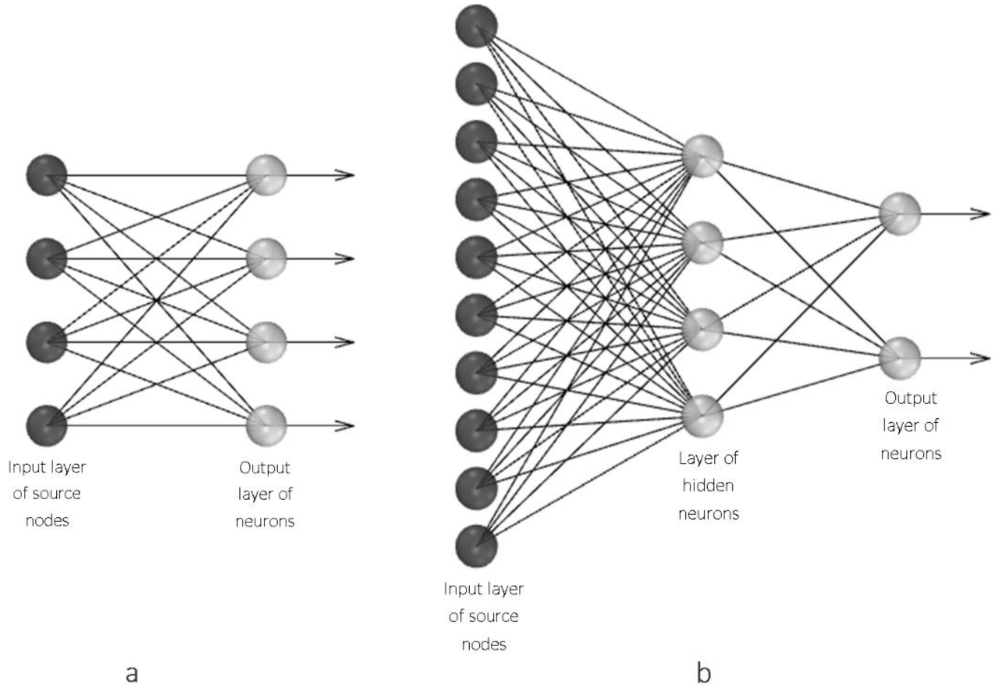


Figure 3.6: (a) Single-Layer Feedforward Network, (b) Multilayer Feedforward Networks (10-4-2)

3.2.4. Learning procedures and the training

The most interesting capability of the networks is to learn with problems by means of training and, after sufficient training, to be able to generalize. This means that the network will be able to solve unknown problems of the same typology. There are different ways in which a network can learn, and these are:

- developing new connections,
- deleting existing connections,
- changing connecting weights,
- changing the threshold values of neurons,
- varying one or more of the three neuron functions (activation, propagation and output function),

- developing new neurons,
- deleting existing neurons.

The change of the connecting weights is the most common method. It implies that there is always a rule according to the modification of the weights, and it can be described as an algorithm.

A training set (P) is a set of training patterns, which are used to train the neural net. Different typologies can be described based on the paradigm of learning, that it can be unsupervised, reinforcement and supervised learning. In the first case, the training set consists only of input patterns, and the net tries to find similarities and classes by itself. In the second case, the network, after a sequence, receives a value indicating whether the result is right or wrong and, possibly, how much it is far from the correct answer. In the last case, the training set consists of the input patterns and correct desired outputs so that the network receives a precise error vector.

Let's suppose to have a single layer neuron with randomly set weights with the aim to teach a function by means of training sample. The output of this neuron is compared with the ground truth and if there is a mismatch the weights are changed and checked again till the output is very close to the ground truth value, thus the error is zero or very small. Generally, the aim is to tweak the weights in order to minimize the error.

The learning target is:

$$y \approx \hat{y} \qquad or \qquad Err \approx 0$$

In which y is the output of the neuron, $\hat{y}$ is the ground truth value and Err is the error. The error can be calculated with several error functions, some of which are:

$$Err = y - \hat{y}$$

$$Err = |y - \hat{y}|$$

$$Err = \frac{1}{2}|y - \hat{y}|^2$$

$$Err = \sqrt{(y - \hat{y})^2}$$

The error function collects all the possible values of output due to all the combination of weights of a neuron.

3.2.5. Gradient descent procedure

In a gradient descent procedure, a function is minimized or maximized. The gradient is a vector g defined for any differentiable point of a function that points the steepest ascent. The gradient is expressed with the nabla operator ∇ and the overall notation is:

$$g(x_1, x_2, \dots, x_n) = \nabla f(x_1, x_2, \dots, x_n)$$

The gradient descent procedure consists of going downhill step by step, from any starting point to minimum toward g. In the case of neural networks, the gradient is a function of the weights, and the function is the error-one that is intended to be minimized. In Figure 3.7 an example of a two-dimension problem is described; in Figure 3.7.a three-dimensional representation of an error function is drawn with the values of the two weights on x- and y- axies and the error value on the z-axis. In Figure 3.7.b a top view of the function is presented where the steps are visible as the related gradient. Obviously, this is just an example and in neural networks the functions are usually multi-dimensional. Errors may occur during a gradient descent (Figure 3.8), like detecting a local minimum (a), quasi-standstill with small gradient (b), oscillation in canyons (c), or leaving good minima (d).

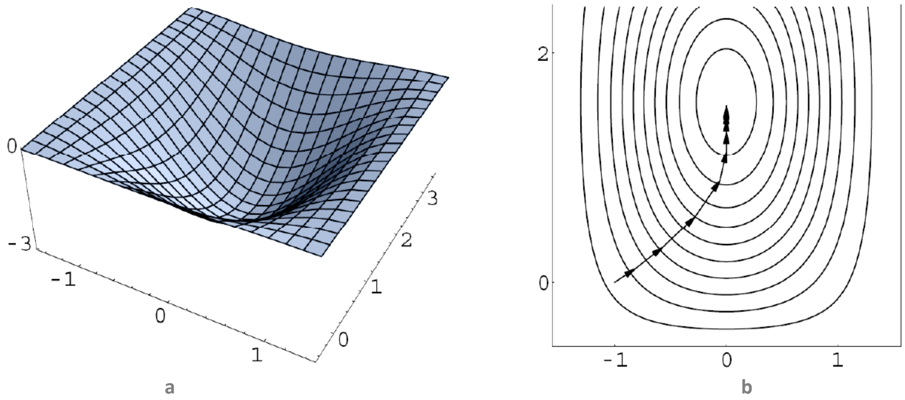


Figure 3.7: (a) 3D view of an error function, (b) gradient descent procedure step by step. From ref. [57]

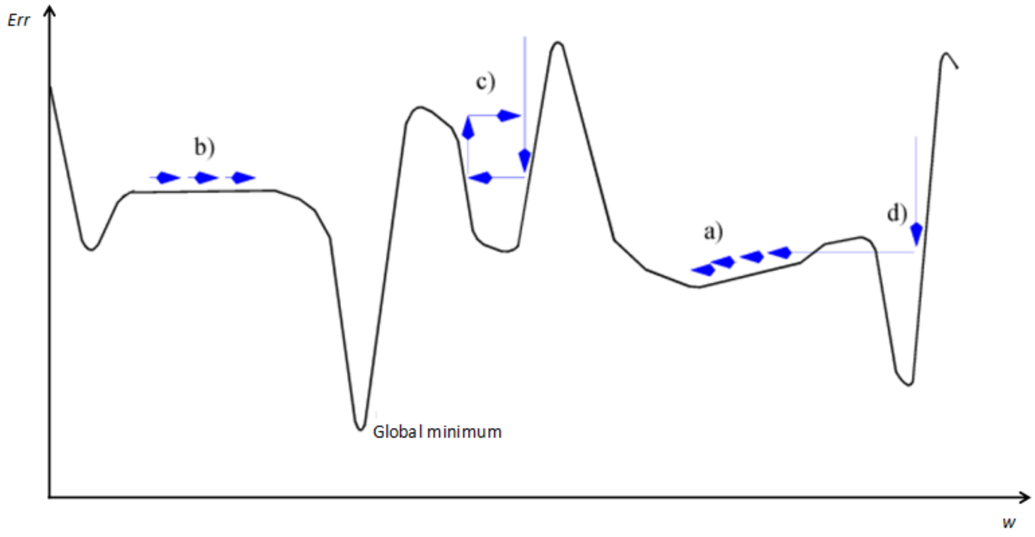


Figure 3.8: Possible errors in the gradient descent procedure. From ref. [57]

The training stops when the error is small enough (minimum or a value close enough to the minimum) or when the maximum number of iterations, which is specified in the configuration of the algorithm, is reached. In fact, in complex cases or in unsupervised networks, where the ground truth is not available, is not easy, or is even impossible, to reach a small error.

Finally, a training is computed when all the inputs are submitted to the network. An epoch is the group of all the training submitted in a procedure.

3.2.6. The Self-Organizing Map

The Self-Organizing Feature Map (shortly, Self-Organizing Map, SOM) is an unsupervised neural network method. It was firstly described by Teuvo Kohonen [62], and for this reason, the SOM is also known as Kohonen Map. This technique is able to classify data into clusters. Moreover, the SOM can display multidimensional data in a low-dimensional grid, and it represents, indeed, also a powerful visualization tool. However, SOMs do not describe what the neurons calculate but only which neuron is active at the moment. The interest is not in the exact output of the neuron but in which neuron provides the output.

The Map can be mono-dimensional, in which the neurons are like pearls on a string (Figure 3.9 a). In this case, every neuron has two neighbours, except for the two end neurons. Another type of map is the two-dimensional one. The simplest is represented by a square array of neurons, each with four neighbours except the one on the vertex which have only teo neighbours (Figure 3.9 b). A very common typology is represented by a sort of honeycomb grid, and the neighbourhood rises to 6 for each neuron (Figure 3.9 c). Irregular two-, three-, or more-dimensional grids are possible but less used.

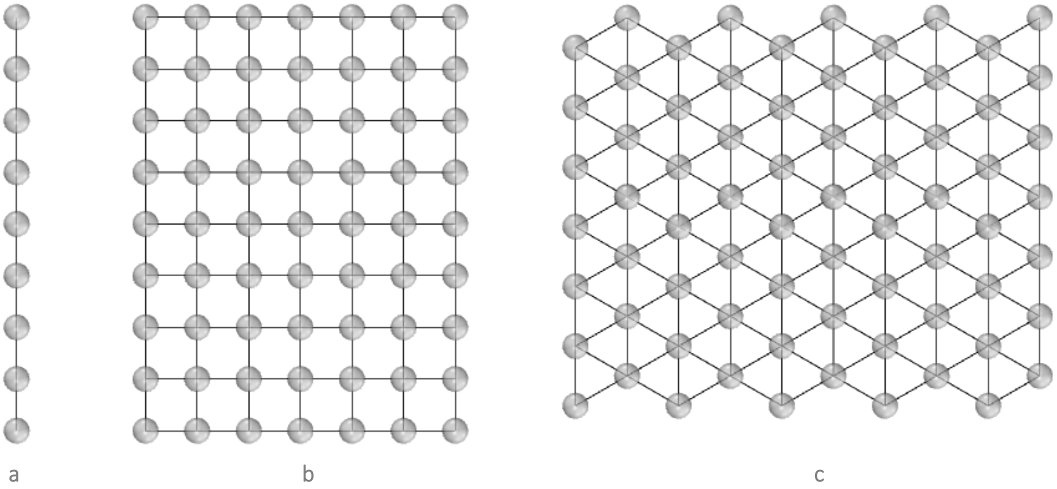


Figure 3.9: (a) mono-dimensional grid map, (b) two-dimensional square-array grid map, (c) two-dimensional honeycomb grid map

The architecture

A SOM neuron *k* does not occupy a fixed position c_k (a centre) in the input space but it moves with the training. A self-organizing map is a set K of these neurons. If an input vector is entered, the neuron closest to the input pattern is activated, in the input space. The dimension of the input space is referred to as N. The neurons are interconnected by neighbourhood relationships. These neighbourhood relationships are called topology. The training of a SOM is highly influenced by the topology. It is defined by the topology function h (i, k, t), where i is the winner neuron, k the neuron to be adapted and t the timestep. The dimension of the topology is referred to as G.

The neurons structure is a single-layer architecture (Figure 3.10): the input layer is composed by a specific number of neurons equal to the number of input variables, the actual neurons layer is a grid of $n_x \times n_y$ neurons operating in parallel. The input layer has the only role to distribute the information to the computational layer.

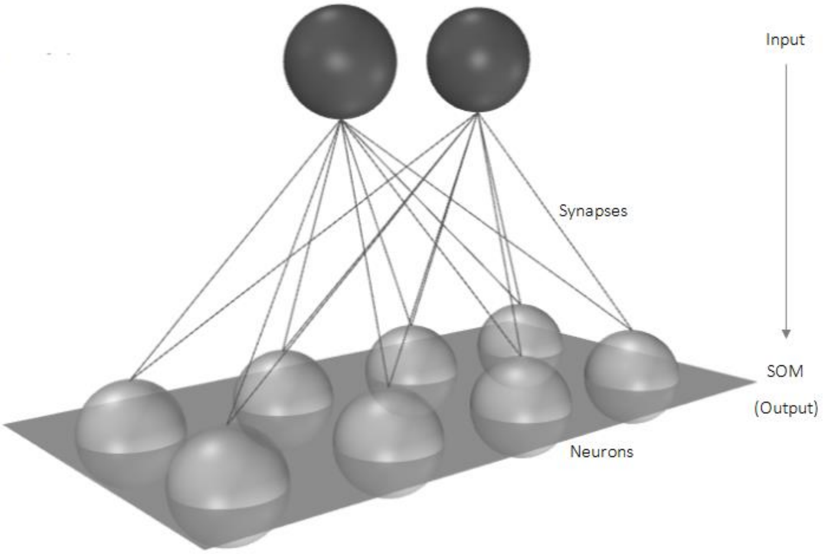
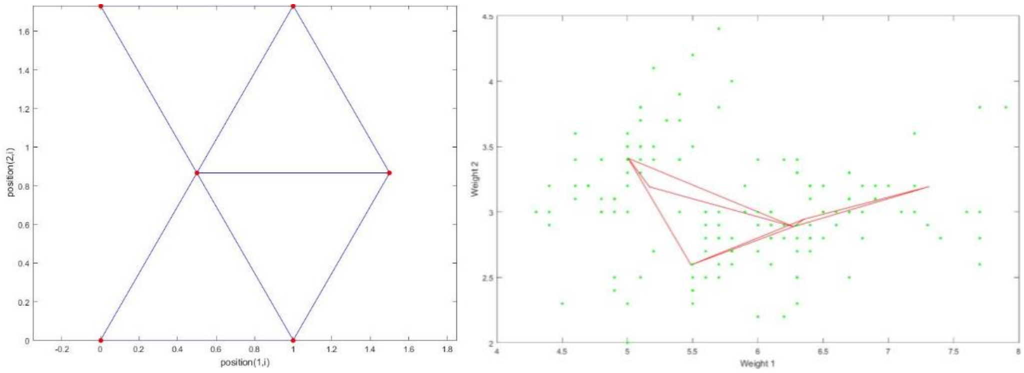


Figure 3.10: Self-Organizing Map Architecture

Competitive learning approach

The SOM uses a competitive learning approach: when an input vector is presented to the network, the similarity with each neuron's synaptic weight is computed and the weight of the neuron more similar to the input vector is the winner. The important feature of the SOM is that not only the weight of the winner is modified to be closer to the input vector, but also that the weights of all neurons within a certain neighbourhood of the winning one are updated. This means that the neurons, that at the beginning are organized according to a topology function, can move during the iterations to best fit with the inputs (Figure 3.11 shows the initial position of a hexagonal grid and the final position after 200 iterations).



Figures 3.11: Initial and Final position of the neurons' weights

The training

The training of the SOM is performed in the following steps:

1. the network is initialized with random neuron centres $c_k \in R^N$ from the input space,
2. a stimulus, i.e. a point p, is selected from the input space R^N and now this stimulus is entered into the network,
3. the distance $||p - c_k||$ is measured for every neuron *k* in the network,
4. the *winner neuron i* is determined, which is the neuron that has the smallest distance to *p*, i.e. which fulfils the condition
$$\left||p - c_i|\right| \leq \left||p - c_k|\right| \qquad \forall\, k \neq i$$
5. the neuron centres are moved within the input space according to the rule
$$\Delta c_k = \eta(t) \cdot h(i, k, t) \cdot (p - c_k)$$
where the values c_k are simply added to the existing centers. The last factor shows that the change in position of the neurons k is proportional to the distance to the input pattern *p* and, as usual, to a time-dependent learning rate (t). The above-mentioned network topology exerts its influence by means of the function h(i, k, t).

η is the learning rate. Its function is to avoid that the later training phases forcefully pull the entire map towards a new pattern. For this reason, the SOMs often work with temporally monotonically decreasing learning rates and neighbourhood sizes. This means that at the beginning of the learning process a moving neuron "pulls along” many neurons in its vicinity, in this way the randomly initialized network can unfold fast and properly. At the end of the process, the network is “stiff” and only a few neurons are moved.

The topology function

The topology function $h(i, k, t)$ is not defined on the input space but on the grid, and it represents the neighbourhood relationships between the neurons. It can be time-dependent. The parameter k is the index running through all neurons, and the parameter i is the index of the winner neuron. The topology function must be unimodal, meaning that it must have exactly one maximum. This maximum must be next to the winner neuron i, for which the distance to itself certainly is zero.

A simple example

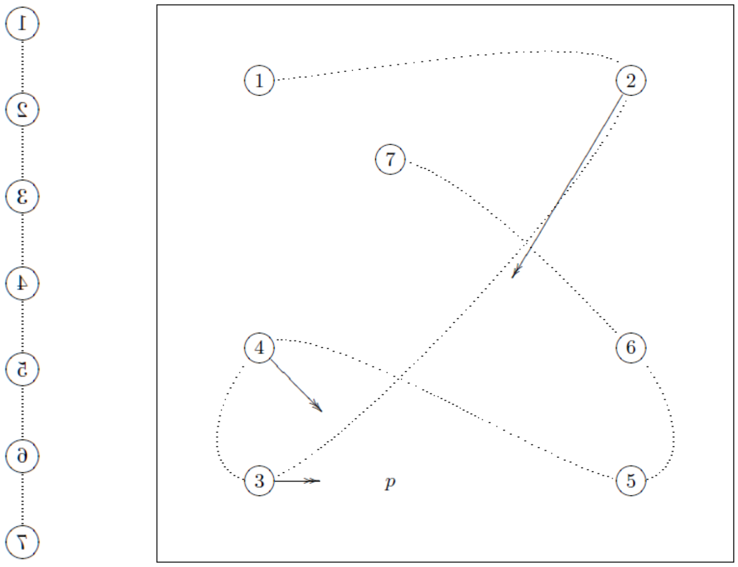
To better understand the SOMs a simple example from [57] is explained step-by-step.

The SOM taken into account is a two-dimensional space, N = 2, while the grid structure is one-dimensional (G = 1). The architecture is composed of a layer of 7 neurons and a learning rate η = 0,5 (Figure 2.12). The neighbour function is

$$h(i,k,t) = \begin{cases} 1\ if\ k\ direct\ neighbour\ of\ i, \\ 1\ if\ k = 1 \\ 0\ otherwise. \end{cases}$$

The first step is the random initialization of the centres c$_k$ (Figure 3.12) and a training sample p is submitted. Obviously, the closest neuron to p is represented by the number 3, hence this is the winning neuron. The position of the winning neuron and its neighbours can be recalculated according to the learning rule:

$$\Delta c_k = \eta(t) \cdot h(i,k,t) \cdot (p - c_k)$$



Figures 3.12: SOM example, on the right the one-dimensional topology space and on the left the two-dimensional input space. The dotted line represents the one-dimensional topology in the two-dimensional input space. The neuron 3 is the winning neuron since is the closest to p. The neurons 2 and 4 are moved because are in the neighbourhood of 3 in the topology. The arrows represent the movement of the neurons towards the training sample p.

In which: $\eta(t)$ is fixed to 0,5 for simplicity and it is not a function of time; $h(i,k,t)$ indicates that only the winner neuron and its two closest neighbours (here: 2 and 4) are allowed to learn by returning 0 for all other neurons; and, the factor $(p - c_k)$ indicates the vector of the neuron k to the pattern p. After the adaptation of the neurons 2, 3 and 4 the next pattern is applied, and so on.

It is important to notice that, even if the neuron 7 seen from the input space is considerably closer to the number 3 compared to the neuron 2, neuron 2 is learning while the number 7 is not. This is helpful to remind that is the network topology that specifies which neuron will learn and which not, and not the position in the input space.

4. Case study

The case study is a retrofit project of a residential building in the South-East area of Milan.

Several studies have confirmed the wide potential of retrofits in the residential sector. For example, Ballarini et al. [63], conducted a research for IEE TABULA project on residential buildings in Europe. They studied the effect of two types of retrofits: a standard one, concerning the common measures used within the country and an advanced one, concerning the application of the best technologies. The study was conducted on four countries representative of the North, Middle, South and East areas of Europe (respectively Denmark, Germany, Italy and the Czech Republic). They concluded that an energy saving of over 45 % can be reached, even with the standard refurbishment. Moreover, the variation between the actual situation and the standard retrofit is bigger than the variation between standard and advanced renovation. This means that being the current built environment poor in efficiency, even with a basic retrofit the reduction of consumption and the related emission can be considerable. This study only concerns the renovation of the physical part of the buildings without considering the occupancy behaviour and other specific issues. However, it clearly states that there is a big potential in the retrofit of residential buildings. In Europe, this is particularly high because the built environment is old and outdated, mainly built before energy standards. The entire sector is characterized by low efficiency, especially the buildings constructed between 1945 and 1980 [64].

Therefore the building industry is now facing a critical period in which it is trying to aggressively reduce the energy use from one side and it is coping with a lot of uncertainties on the other side, such as climate change, building operations, government policy change and human behaviour change [26]. World's building energy consumption has steadily increased since 2008, and this trend can be caused by population growth, increase in building services and comfort requirements, but also by the rise of the hours spent inside buildings [65]. To face these problems, the Energy Efficiency Directive [4], the Energy Performance of Buildings Directive [11] and the Annex 66 [28] can be recalled.

The project taken into exam is a retrofit project. The building is a property of the Municipality of Milan and thus it is part of the ownership of the central government that has to be renovated to fulfil the goal of the energy efficiency directives [5, 6]. The project aims to become an example for further renovations, with improvements and modifications and can become a source of information for future policies in Milan [64].

The proposed methodology does not deal with the retrofitting itself, but it takes inspiration from it. The final aim is to improve generally the energy modelling of the building. The idea is to use the pre-retrofit model (with registered data) to increase the knowledge of occupants' behaviour. The final results might be implemented in the post-retrofit model to assess a more realistic energy consumption.

4.1. Project description

The building is composed of two blocks, named B1 and B2 (Figure 4.1), with a total gross surface area of 4633 m^2. It hosts 66 residential units for an estimated population of 210 people. The building was constructed during the 1980's with an envelope in prefabricated concrete elements, with almost no thermal insulation and low-performance windows. The building is made of basements and garages, four stories and an attic. B1 has two staircases and a total number of twenty-four flats, whilst B2 has three staircases and forty-two flats. The building is surrounded by a garden included in the property, in which up to 20 meters height trees are present.

The heating system is outdated, and it consists of a centralized system that uses fuel oil as the energy carrier. Then, in each apartment, a local gas boiler is exploited for domestic hot water (DHW) generation. All the other energy uses depend on electric energy from the national grid, except for cookers that use natural gas.

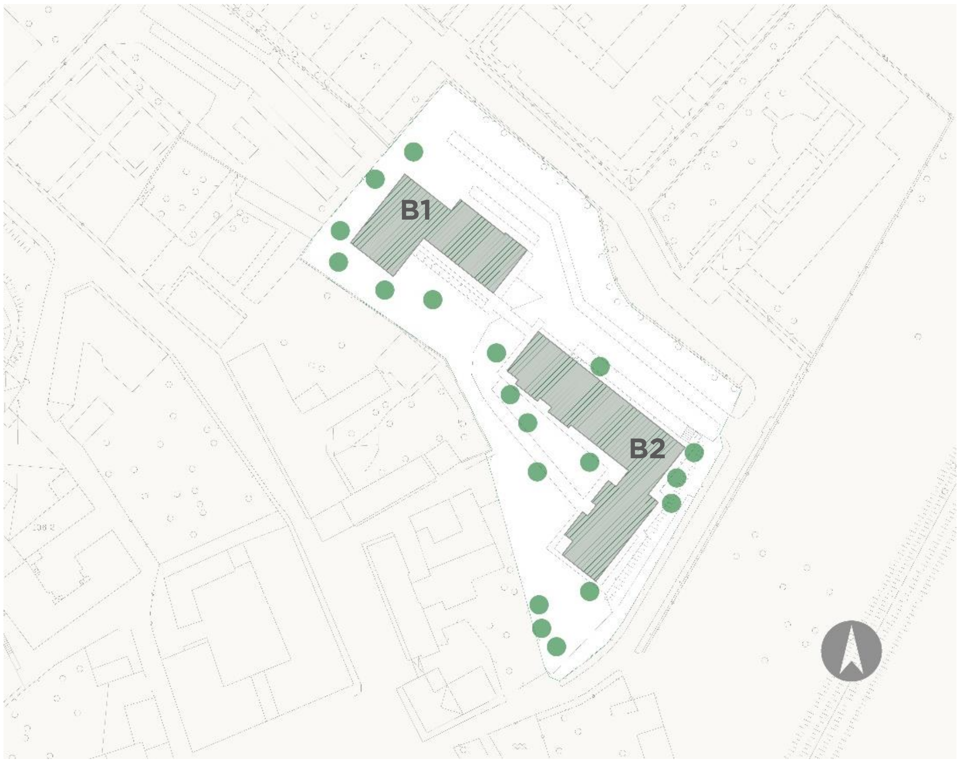


Figure 4.1: Masterplan view of the case study

The façade renovation, both of the opaque and transparent part, is one of the key points of the energy retrofit. An external thermal insulation composite system with rendering will be used, as it is common in retrofit projects. The external insulation, in fact, will lead to a strong reduction of the thermal transmittance and thermal bridges effects and a consequent increase of internal thermal comfort. The windows will be renovated with modern high-efficiency double glazed units. The shading devices (roller blinds with low-insulated cases) will be replaced with Venetian blinds. This type of shading is packable and does not need a case, therefore the voids will be filled with insulation decreasing, even more, the thermal bridges in the external envelope. In order to control heat loss due to ventilation, allowing for an adequate level of indoor air quality (IAQ), a mechanical ventilation system, equipped with heat recovery and by-pass, will be installed. A high performance centralized heating system will be placed with heat pumps as generation system. On the roof, photovoltaic and solar thermal systems will be installed, including electrical storage batteries and an energy management system. To reduce as much as possible, the electric consumption, the common areas will be equipped with high-efficiency LED lamps. However, no intervention will be possible to improve both the electrical and the heating systems inside the flats, since the renovation must take place with tenants keeping on living inside their apartments.

4.2. The energy model

On the basis of the existing documentation and in-situ inspections, the characteristics and the technical performance of the building were assessed, and a dynamic energy model has been created. Energy Plus 8.5.0 was used to model the building. The drawing tool was SketchUp 2015, linked with the main software through OpenStudio 1.7.0. The description of the energy model in the following refers to B1, chosen as representative of the overall project.

Each flat has been modelled as one heated thermal zone, whereas ground floor and the attic were considered as unheated zones (both modelled as a unique zone). Also, the two staircases (divided in one zone each for the first floor, and one thermal zone each for the other three floors together) were considered unheated areas. The surroundings of the building have been characterized by the presence of the trees, with a height of ten or eighteen meters (Figure 4.2).

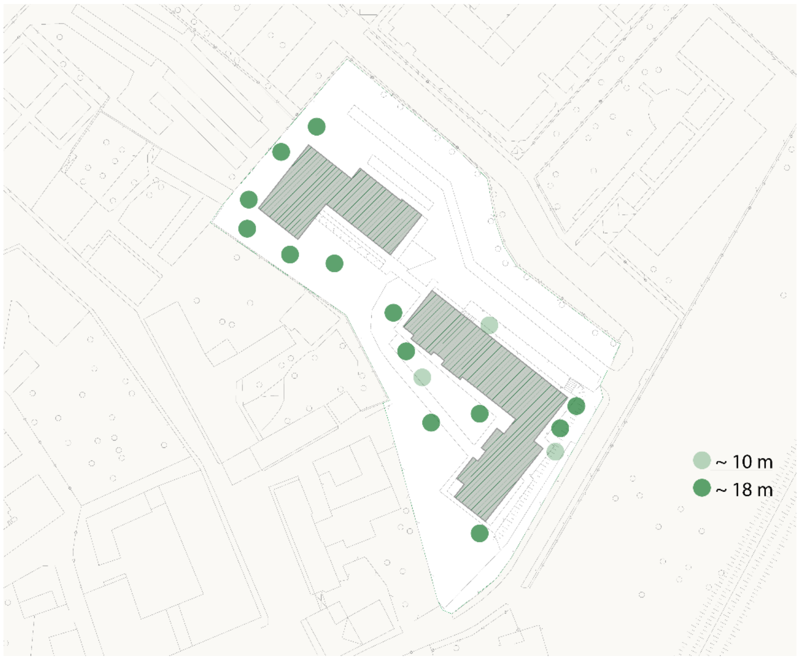


Figure 4.2: Modelled trees underlined in the site plan

Figure 4.3 shows the model of B1 while Figure 4.4 displays also the surrounding trees and B2. The complete model was used to run the analyses. Each zone was named univocally and consequently all the elements that formed it, such as walls, windows, roof, and ground (an example can be seen in Figure 4.5). Details of the used constructions and features of the model can be found in Annex I.

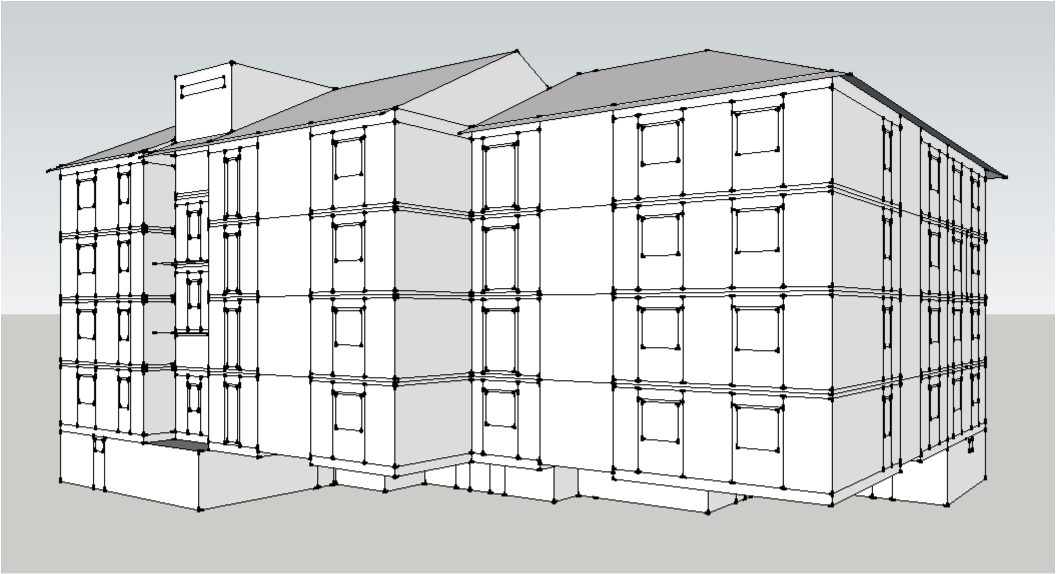


Figure 4.3: Model of B1 without context

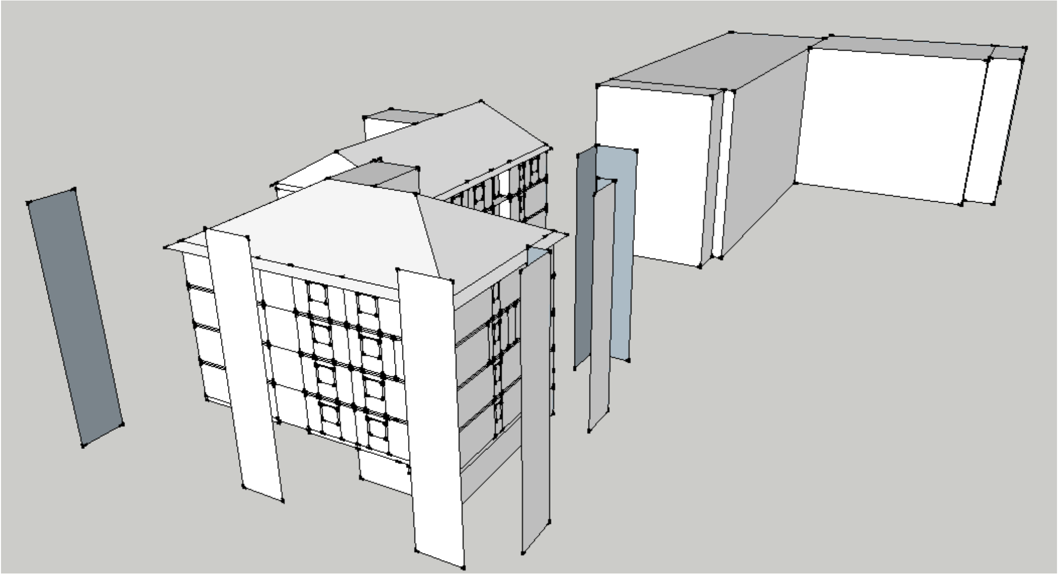
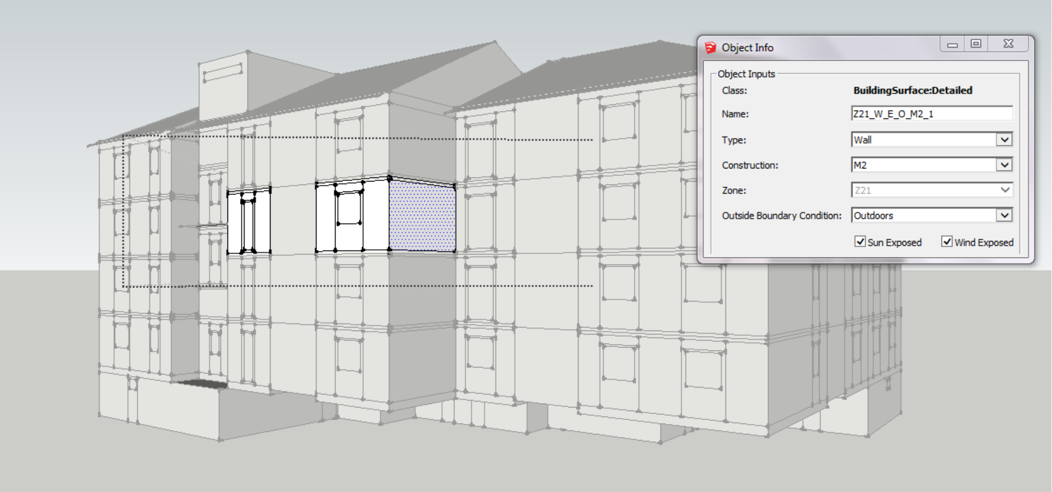


Figure 4.4: Model of B1 with context, composed of trees and B2



Class: **BuildingSurface:Detailed**

Name: Z21_W_E_O_M2_1

Type: Wall

Construction: M2

Zone: Z21

Outside Boundary Condition: Outdoors

☑ Sun Exposed ☑ Wind Exposed

Figure 4.5: Example of a univocal name used in the model
ZONE_TYPOLOGY_EXPOSURE_ORIENTATION_COSTRUCTION_PROGRESSIVE NUMBER

The simulations have been run on yearly base, with hourly time step. To have a more precise result the ground temperature has been obtained from the weather data ITA_Milano-Linate.160800_IGDG.epw as the monthly average at two meters depth.

Since the aim of this thesis does not concern the plants themselves, the focus result is the energy need for space heating to guarantee the comfort in the heated thermal zones. Therefore, the heating system was characterized in EnergyPlus by an ideal system able to maintain a temperature of 20 °C during the heating season, which is defined according to Italian national regulations from 15 October to 15 April, for the considered climatic zone.

4.3. The electric registered database

The available dataset includes registration for the year 2016 from 1st of February to 31st of August, of the building modelled in EnergyPlus. The database includes 24 households with a 15-minutes time step. The flats in the thesis work are sometimes called “zones”, this term derives from the creation of the energy model with EnergyPlus where each flat is modelled as a thermal zone. Each thermal zone is associated with a progressive number, thus a final code to indicate a specific flat can be, for example, Z4, in which Z stands for “thermal Zone” and 4 is the related progressive number.

The raw registered database is shown in Figure 4.6.

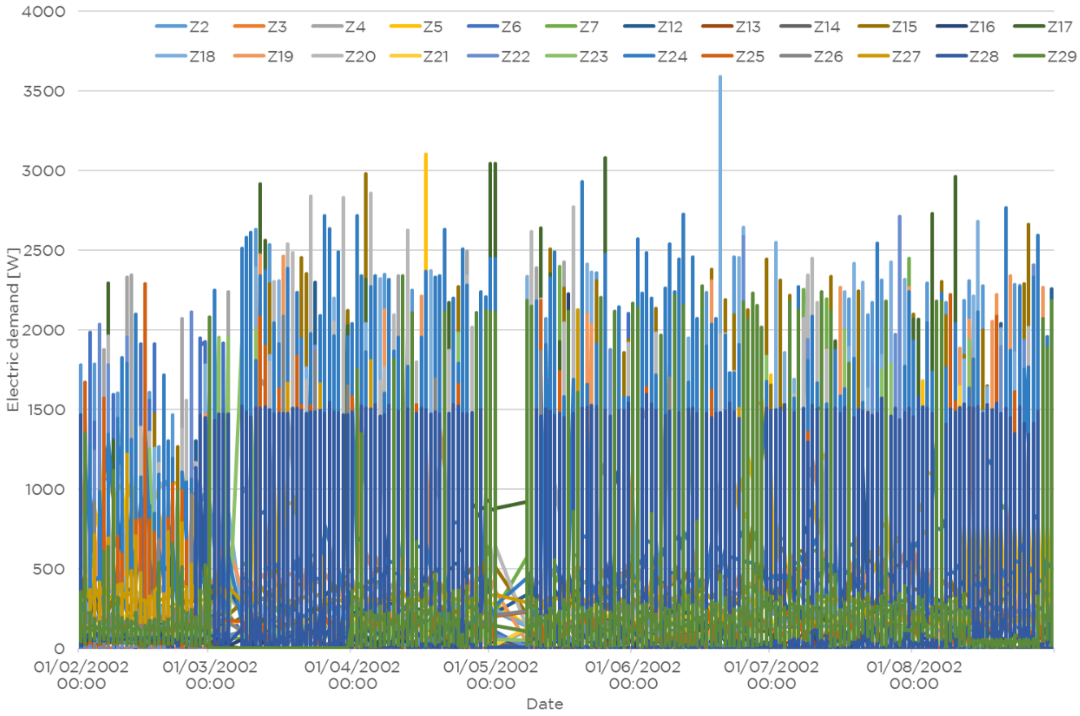


Figure 4.6: Raw registered data from 01/12/16 to 31/08/16 for the 24 households

As seen in Figure 4.6, the raw registered load of the households shows some recording errors and it cannot be easily interpreted. In fact, each household's load is the result of many different variables and characteristics. Moreover, within a household, the electric load is composed of a number of appliances, which behave differently in terms of energy consumption.

All data are completely anonymous, and no other data was available at the time of this thesis for privacy reason. This situation is similar to the one in which a modeller can find himself in a real designing project and he/she tries to improve the energy model with the available data.

4.4. The weather data

The research of Erba et al. [66] stresses the impact of different weather datasets in the energy modelling of buildings. They compared the energy needs resulting from the use of different datasets. The simulations were run with five different weather datasets available for Milan. All of them provide hourly values of the features needed to run the energy simulations: dry-bulb temperature, dew-point temperature, relative humidity, atmospheric pressure, wind direction and speed, global and diffuse horizontal radiation, direct normal radiation, and others. Some of these parameters were directly measured, some others calculated. The paper shows that the energy simulations are highly affected by the choice of the datasets. In fact, the weather represents one of the biggest uncertainties in the modelling that has to be taken into account.

The aim of this thesis work regards the assessment of the impact on the energy needs of the schedule of occupants. As validation of the model, a direct comparison with the energy needs registered in 2016 can be performed. For this reason, a weather data related to 2016 was necessary. The Agenzia Regionale per la Protezione dell'Ambiente (A.R.P.A.) provides hourly registered data for eight weather stations in the city of Milan. The available data are shown in yellow in Figure 4.7 [67]: A-Milano-Niguarda, B-Via Marche, C-Via Confalonieri, D-Via Rosellini, E-Via Brera, F-Via Feltre, G-Milano-Lambrate, H-Via Juvara. The chosen weather station is the one in Via Juvara 22 (H in the Figure) because is the closest to the building site, with a distance of about 9 km.

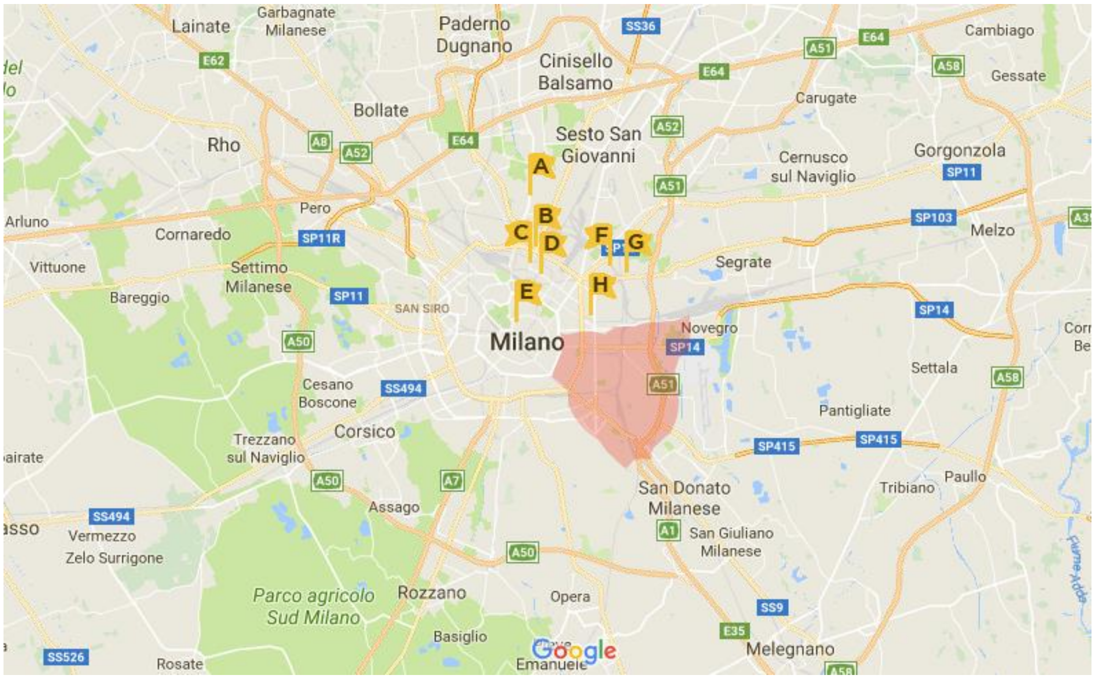


Figure 4.7: Available registered data from A.R.P.A. Lombardia and the area in which building site is located highlighted in red

The characteristics of this weather station are:

NAME:	Milano-Via-Juvara
PROVINCE:	MI
REGION:	LOMBARDIA
ALTITUDE:	122 m a.s.l. + 28 meters on the building roof
LATITUDE:	45° 28’ N
LONGITUDE:	9° 13’ E
TIME FRAME:	11 years

The existing registration for this station includes Precipitation, Dry-Bulb Temperature, Atmospheric Pressure, Relative Humidity, Global Solar Radiation, Wind Velocity and Direction. The dataset is in hourly steps; therefore the data are hourly averages within the previous hour, except Precipitation that is a cumulative value. A distinction between direct, diffuse and global radiation is missing.

4.4.1. Watanabe Method

To complete the weather data to be used in EnergyPlus, the global radiation should be divided into direct, diffuse and global solar radiation. To fill this lack of registered data the Watanabe simplified method based on the location of the weather station was used.

The Watanabe model, developed in 1983 [68], is able to separate the total global irradiance into direct and diffuse components based on the location. This method was developed for Japan and it is based on simple geographic features of the location. Japan, in terms of latitude, is not much different from Italy, and thus, the model can be assumed to be suitable for the case study. As a matter of fact, Japan’s latitude goes from 30° N to 45° N, Italy’s one goes from 36° N to 47° N. The equations that regulate the model are [69]:

$$I_b = I_0 \cdot SH \cdot K_{DS} \cdot (K_T - K_{DS})/(1 - K_{DS})$$

$$I_d = I_0 \cdot SH \cdot (K_T - K_{DS})/(1 - K_{DS})$$

In which

I = global solar irradiance in W/m^2

I_b = direct normal solar irradiance in W/m^2

I_d = diffuse solar irradiance in W/m^2

I_0 = solar constant set at 1355 W/m^2

And

$$K_T = \frac{I}{I_0} \cdot SH$$

$$K_{TC} = 0.4268 + 0.1934 \ \cdot SH$$

$$K_{DS} = K_T - (1.107 + 0.03569 \ \cdot SH + 1.687 \ \cdot SH^2) \cdot (1 - K_T)^2 \qquad when\ K_T \geq K_{TC}$$

$$K_{DS} = (3.996 - 3.862 \ \cdot SH + 1.540 \ \cdot SH^2) \cdot K_T^3 \qquad when\ K_T < K_{TC}$$

$$SH = \sin(a)$$

a is the solar altitude (Figure 4.8) and SH is called “sine of solar altitude”.

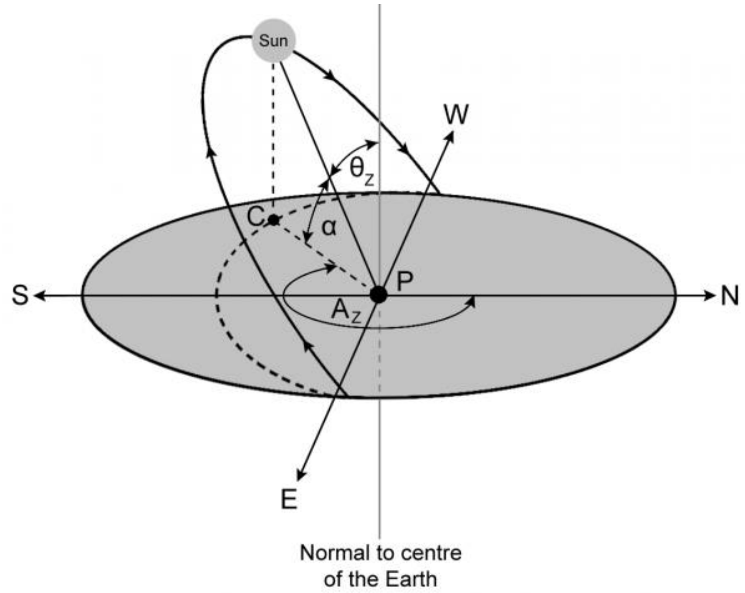


Figure 4.8: Angles and geometry of the Sun when viewed from point P. From ref. [70]

5. Methodology

5.1. Introduction

The aim of the presented methodology is to identify methods and techniques that, starting for the actual electricity readings at a building's meter, support the generation of reliable and trustful schedules that represent occupancy and the use of electric devises in dwellings to be used in BPS. Furthermore, a method is proposed to identify three types of schedules: for typical energy uses, for intense energy uses and for energy-aware uses.

Seldom the data of electric demand are full and complete, and a deep survey or expensive sensors should be used to detect the presence of people inside buildings. With the shown methodology the modeller should be able to create schedules of the presence of people and their electric uses for a full year starting from a relatively small sample of data.

This work can be divided into three main sub-tasks:

1. from the data registered by a smart meter, generation of the yearly schedule of electric uses for all the thermal zones in the energy model,
2. from the data registered by a smart meter, detection of the presence of people,
3. assessment of the impact of the electricity use and occupancy schedules on the overall energy demand for space heating of a residential building.

An overview of the methodology is given in Figure 5.1 and is explained in detail in this chapter.

The registered raw data were pre-processed to be cleaned up by errors in the data registration and storage. Next, all missing values were filled with null values. Thus, a workable dataset was created for each apartment. Finally, the original dataset was normalized to allow clustering into groups. The creation of the clusters was fundamental to identify patterns in the electricity use of each apartment and was adopted to generate three completely new yearly schedules for each thermal zone of the building's energy model (TASK 1). After this step, starting again from the workable datasets, occupancy inside each thermal zone was deducted and the related yearly schedules were created with an hourly resolution (TASK 2). With the full-year schedules, the energy model of the building was set, and some energy simulations were run to assess the impact of these three different schedules on the energy need of the entire building, and the results were compared with the actual energy need of the building (TASK 3).

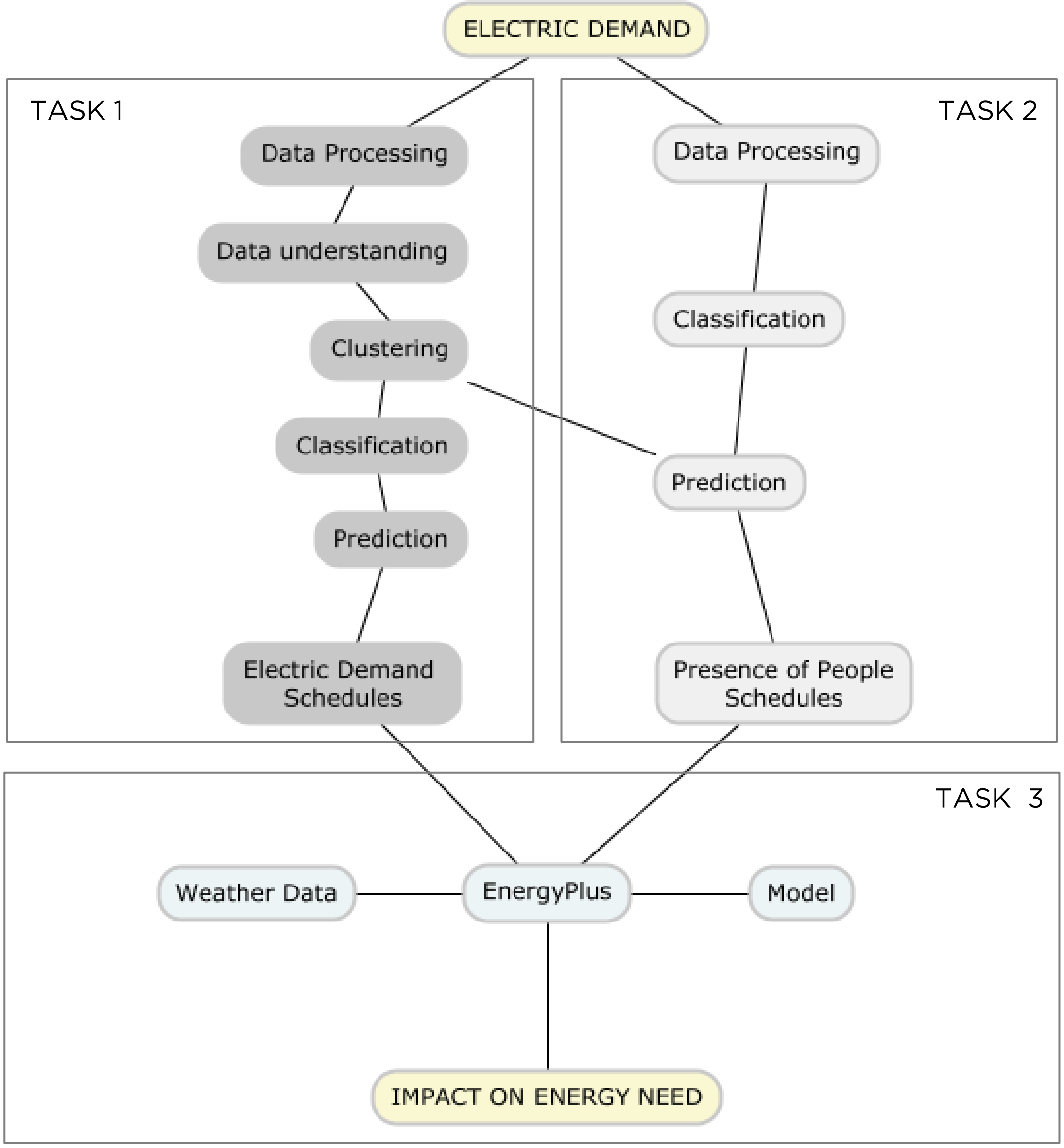


Figure 5.1: Flow chart of the methodology

Utilized software

To accomplish the first two tasks, a few statistical and machine learning techniques were exploited with IBM SPSS Statistics 24 and MATLAB R2017a.

To perform the third task, an energy model of the building was built for and simulated with the dynamic energy simulation engine Energy Plus 8.5.0.

In particular:

- SPSS was exploited for the statistical analysis of the data,
- MATLAB was used for exploiting the appropriate machine learning techniques and for managing big matrixes of data,
- EnergyPlus was exploited for the energy simulations of the building model.

5.2. Task 1

This task aims at generating a representative yearly schedule for the electric demand for each apartment, starting from the workable datasets of dwellings' electricity use. To accomplish this objective, the daily electricity load curves were clustered in several meaningful clusters. For each cluster, three load curves were drawn: the typical represented by the median, the energy intense use represented by the third quartile and the energy-aware use represented by the first quartile. The three final scenarios correspond to low, medium and high electrical consumption. A detailed flowchart of the steps adopted in this task is reported in Figure 5.2.

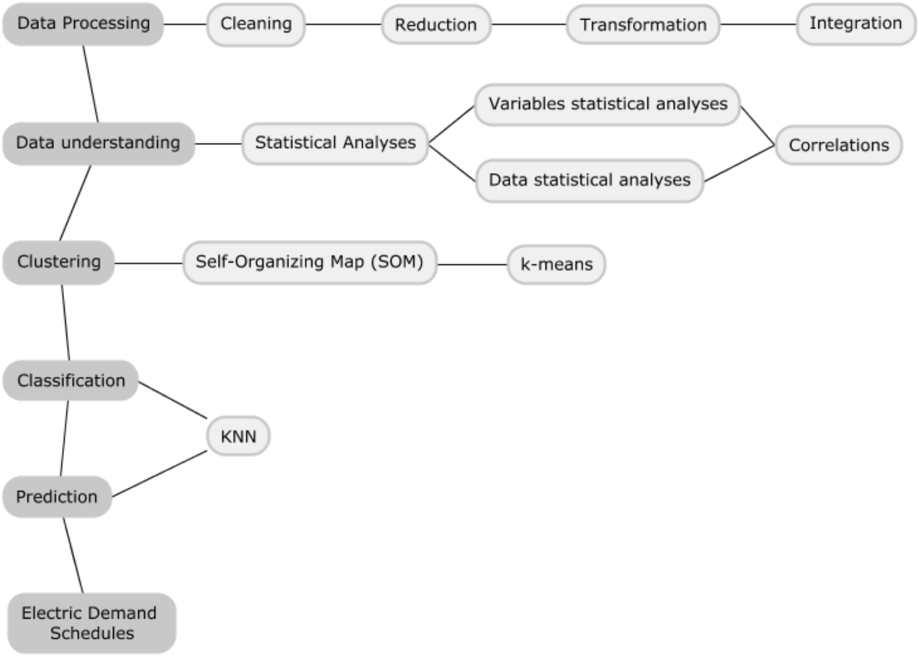


Figure 5.2: Flowchart of the steps of Task 1

The starting point of the data is the 15-minutes-step electric demand registered for 23 flats. The process is composed by a pre-processing phase in which the data is cleaned and formatted into a csv file, and by an understanding phase useful to go deeper in the study of the variables and of the registered data from a qualitative and quantitative point of view. Next, in third phase, a clustering technique is adopted to collect into groups similar daily electricity load curves. To extend the results to a full-year schedule, a classification of the known days is coupled with a prediction phase, using a machine learning technique.

5.2.1. Data processing

Data processing is an important step that can affect the final result. Usually, the real raw data is incomplete and contains errors or outliers. Moreover, to be able to perform statistical analyses is important to assign codes and to names numerical values, easily understandable by the commonly used software (SPSS in this case). The steps followed to create the actual datasets for statistical analyses (Figure 5.3) are:

- data cleaning: outliers are removed, and the inconsistency of data is resolved;
- data reduction: the representation of data is reduced but producing similar analytical results (in this case, the reduction is in terms of time step, from a 15-minutes registration the datasets is reduced at hourly timesteps);
- data transformation: the data are normalized and aggregate if needed;
- data integration: integration of multiple database and completion with attributes into a single and useable formatted file.

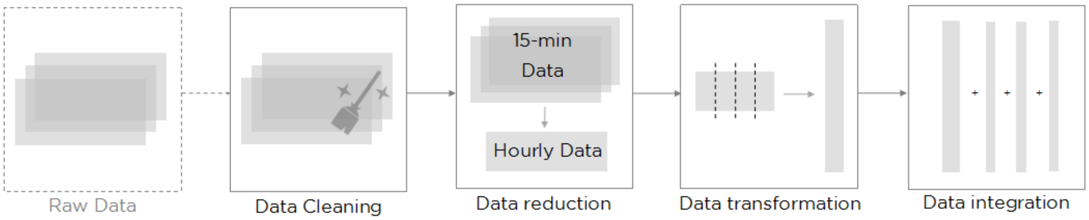


Figure 5.3: Data Pre-processing steps

Part of the Data transformation is the association between the electric demand with the characteristics of the hour, of the day or of the flat that is related to a registration. Therefore, a fundamental step is the selection of the features useable in this specific problem and that can be relevant to the work. From the literature, all the variables that can affect the registered data (electric demand) were found, and then, among them, only the exploitable ones were chosen. The selection was made mainly on two bases: availability and exploitability of the variable. In fact, some variables were not available in the datasets, by consequence, even if from literature they are indicated as highly connected with the electric use of buildings might not be exploited in this case study. On the other hand, some other variables can be available but not exploitable in this case because constant within all the datasets. As a matter of fact, the registration is coming from different flats in the same building, thus with the same architectural and locational characteristics.

The result is a clean and functional file to be used in a statistical tool.

5.2.2. Data understanding

The data understanding is performed with different statistical techniques, with the aim to develop a deep insight of the dataset and to identify relationships between several variables of the problem. The outcomes of this phase are summary tables and graphs that show the trends and main characteristics of the datasets. The followed scheme is performed in steps:

1. statistical analyses of variables that can affect the electric demand;
2. correlation analyses between variables;
3. statistical analyses of the registered data;
4. correlation analyses between the registered data and variables.

The main statistical analyses used in this step are correlation analysis, for exploring the direct relationships in the sample, T-test and ANOVA to investigate the difference between groups of data. These methods are used both to compare the variables between one another but also to check if there is a direct relationship between the variables that can affect the electric demand and the registered data.

Variables analyses

The first step is understanding the nature of the variables and data involved in the analyses. The variables are divided into continuous and categorical [71].

The continuous variables, also called quantitative variables, are the one in which the value can be any number in the range of the extremes. This category is further subdivided into interval and ratio variables. The first case groups the variables that can be measured along a continuum and they have a numerical value. The ratio variables are interval variables, but the 0 (zero) corresponds to none of that variable. An example is the temperature measured in Kelvin, the zero in this case means that there is no temperature; other examples are height, mass or distance.

The categorical variables, also called discrete or qualitative variables, can be further classified as nominal, ordinal or dichotomous. Nominal variables have two or more categories but without an order. An example of a nominal variable would be classifying where people live in the Italy by region. In this case, there will be many more levels of the nominal variable (20 in fact), but the number associated with each region is not referred to any order. Dichotomous variables are nominal variables which have only two categories or levels. Ordinal variables are variables that have two or more categories just like nominal variables only the categories can also be ordered or ranked.

Relation between variables

Correlation is used to study the strength of linear relations between two continuous variables [72]. They give the strength of the correlation and the direction (positive or negative). If a correlation is positive means that if one of the variables increases the same does the other if it is negative with the increase of the first variable a decrease in the other is registered. The resulting index can go from -1 to +1, the extreme values correspond to perfect correlations, whilst zero is a complete no correlation result, the other cases can be interpreted as in Figure 5.4, according to Deborah J. Rumsey [73]:

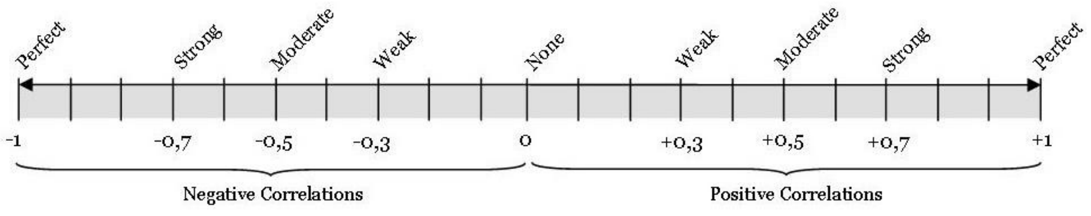


Figure 5.4: Interpretation of the correlation results.

The correlation has a direct representation with the scatterplots. They show on the two axes the variables taken into considerations and each point is positioned with the corresponding two values. The perfect negative or positive correlation means that the points are all positioned in one line, on the contrary when no correlation occurs, the points are in a noisy cloud. Scatterplots related to specific values of correlation are shown in Figure 5.5.

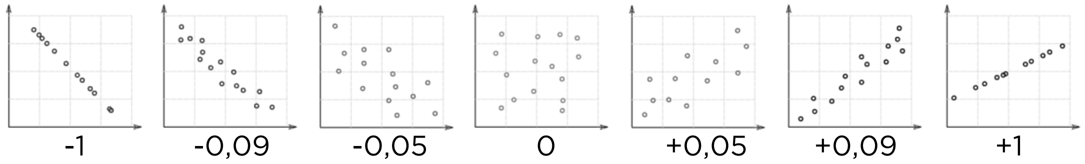


Figure 5.5: Correspondence between correlation values and scatterplots. From ref. [74]

In the performed analyses the Spearman's rho correlation was used. The more commonly used Pearson's correlation coefficient is used if the dataset is normally distributed, thus it performs a parametric analysis, and it works on the data in the sample. The Spearman's rank correlation coefficient performed a not-parametric analysis, thus when the data are not normally distributed. Moreover, it works on the ranks of the datasets, in fact, it orders the two datasets, object of the correlation, from the smallest to the biggest and it analyses the increasing trends.

Differences between groups

To assess the differences between groups T-test and ANOVA are implemented. They are both tests to assess the differences in the variance of groups but with differences in the assumptions and characteristics of the independent variables.

The T-test compares only two groups at the time, thus the independent variable must be a binary variable, it means that divides into two groups the sample. The independent-sample T-test was exploited to compare the mean scores of two different groups of conditions. SPSS gives the results of Levene's test for equality of variance, it checks if the variances of the two compared groups are approximatively equally distributed. If the Levene's test Sig. value is ≤ 0,05 (i.e. probability) means that the variances of the two groups are not the same, therefore the assumption of equal variance is violated. In this second case, SPSS is able to calculate a T-test that compensates the difference in the variances called “equal variances not assumed”. There is a significant difference between the groups if the Sig. (2-tailed) values are below 0,05 in the actual T-test ‘s results.

The ANOVA (analysis of variance) compares the mean scores of more than two groups. There is an independent variable which has different levels, corresponding to features or groups. Firstly, it tests the null hypothesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

In which μ is the group mean and k is the number of groups. If this is false it means that the alternative one is true:

$$H_1: means\ are\ not\ all\ equal$$

Then, it compares the variance between the different groups with the variability in each group and the ratio *F* of this two variability is calculated. The higher is the F, the larger is the variability between the groups compared to the one within the groups.

5.2.3. Clustering

Clustering means grouping a data set into an N number of clusters ***C****i*, I =1, 2, …, N. To solve the clustering problem two methods of machine learning were used: Self-Organizing Map and *k*-means. Chapter 3 is recalled for theory. Both are partitive clustering algorithms. They are able to divide the data into a number of groups trying to minimize some criterion or error functions. The number of clusters is predefined in both cases, and the followed steps are:

1) initialize the clusters centroids,
2) group the data,
3) update the cluster centroids,
4) if the partitioning is unchanged stop, otherwise return to step 2.

Two level approach

It this methodology, a two level-approach clustering method was used, combining the unsupervised neural network method, Self-Organizing Map (SOM), with the more classic unsupervised k-means algorithm (Figure 5.6).

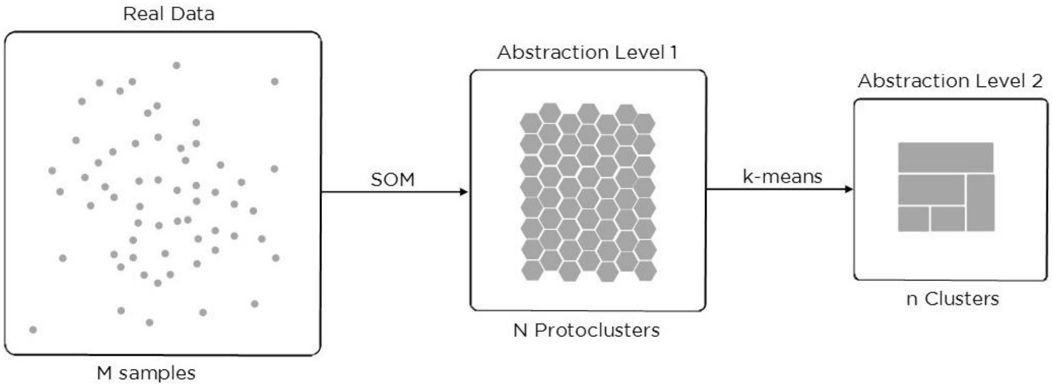


Figure 5.6: Two-Levels Clustering approach

The Self-Organizing Map is an unsupervised neural network method and it is able to classify data into clusters. The SOM is a powerful visualization tool because it can display multidimensional data in a low-dimensional grid.

The *k*-means algorithm is one of the simplest and most commonly used unsupervised learning algorithms. It solves the problem of clustering given a fixed number (*k*) of centroids, one for each cluster. The algorithm takes each input of the data set and associates it to the nearest centroid. As next step, it recalculates the *k* centroids as barycentre of the clusters resulting from the previous step. The inputs are again associated with the centroids and a new complete iteration is computed.

The SOM algorithm is used to create proto-clusters that are further grouped with a k-means algorithm to find the final clusters. As shown by Vesanto et al. [38] and Hernàndez et al. [37], this two-levels approach gives better results than directly clustering of dataset. The two main benefits, are the minimization of the computational cost and the noise reduction. The proto-clusters are local averages of the original samples and, for this reason, less sensitive to single high or low cases in the data sample.

SOM's features

The first step is to define the number of proto-clusters and cluster from which the features of the SOM and k-means will depend. To achieve a good result it is fundamental the choice of a number of clusters as outputs of the SOM. The SOM Toolbox for Matlab Report [53] was followed to set this parameter. The final choice is: a 2-dimensional map with hexagonal lattice and the size is given by the heuristic formula

$$m = 5\sqrt{n}$$

In which m is the final number of proto-clusters, n is the number of data sample given as input. Finally, the ratio of the side-lengths of the lattice would be the ratio between the two biggest eigenvalues of the covariance matrix of the given data, and the actual side-lengths are then set in such a way that their product is as close as possible to the desire m.

To improve the results of the SOM, on the real value of the hourly electric use, a normalization on the maximum value reached in the day was performed. In this way, the SOM can recognize the actual peaks and the shape of the daily electric pattern without being disturbed by the absolute value of it.

k-means algorithm's features

The number of final clusters useful to describe the data sample was set using the Davies-Bouldin Index (DBI). This index was introduced in 1979 [75] and it is able to evaluate a clustering algorithm. It is an internal evaluation method, this means that the validation is made using features and quantities of the dataset itself. After several preliminary analyses, the number k of the final clusters in the k-means is fixed at 5. This value gives a good result based on the description of data, but at the same time do not create small clusters. It is able to assess the difference between the clusters on the base that each cluster should be different from the others. A low value of the DBI means that the clustering is better.

An increase in the number of clusters leads to better results but will increment the complexity of the following analysis and result less robust to generalizations.

The result is shown in Figure 5.7:

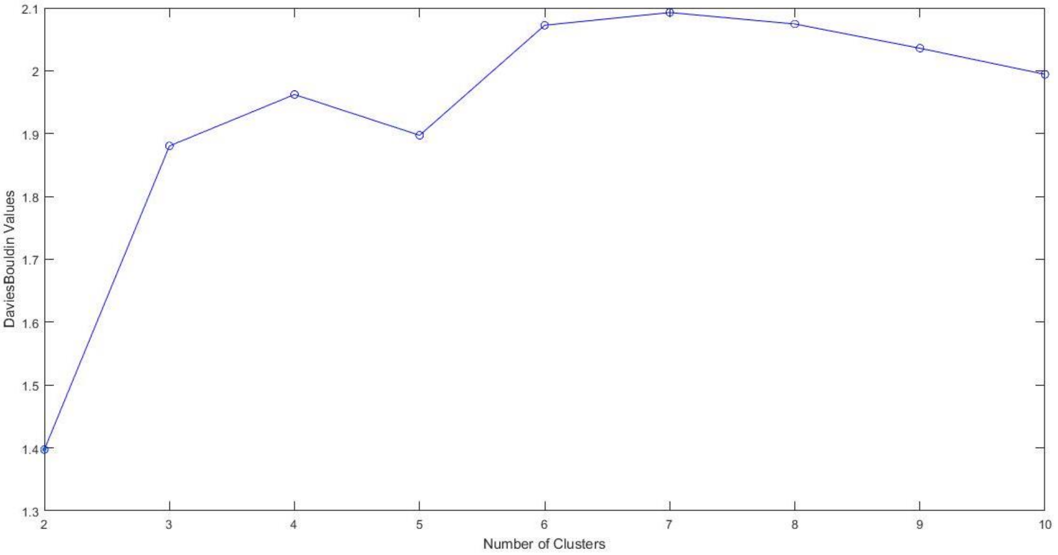


Figure 5.7: Results of DBI analyses on the dataset

5.2.4. Classification

Classification is also known as pattern recognition or discrimination. There are a lot of Classification methods, such as decision trees and rule induction, machine learning techniques such as k-nearest neighbours (KNN) and density estimation. In this methodology, a K-Nearest-Neighbours algorithm from machine learning is used to solve the classification problem because it is commonly used because it is easy to interpret, and it has a low calculation time.

The algorithm, being a supervised learning procedure, needs a training set and an application set. In this case, the training set is the part of the year in which the registration is available and on which the previous tasks were performed. Each day, in this set, is characterized by the association with a final cluster. The application set, instead, is the part of the year in which no registration of electric consumption was available, thus it could not be associated with any cluster. To run these analyses, the application of MATLAB Classification Learner was used. A cross-validation of 5 folds is implemented to avoid overfitting problems. The key issue is to define the predictors and the responses. The algorithm will learn automatically how to link a response to one or more features and it will be able to predict responses to a new set of features. The response, in this case, is the cluster in which the day taken into exam is. The difference between Working day or Not Working day, the difference between Heating and Cooling season was used as predictors.

Moreover, to give a time sequence to the days of the year without using the months and day (of not interested in this specific case) the average daily external temperature and its variation were added as predictors. The KNN is able then, to assign to each day of the year the cluster related to its specific predictors.

The result of this task is the full year schedule for all the flats in the building. Each day of the year is related to one cluster, with corresponding first, second and third quartiles of the average electric consumption of all the days that are in the specific cluster. These three daily loads can be seen as three typologies of families corresponding to different scenarios: energy-aware user, standard user and energy-intensive user.

5.3. Task 2

The input of this task is again the measured electric use, whilst the goal is to detect the occupancy presence in the flats. The used methodology is an adaptation to this case study from the work of Kleiminger et al. [51,52]. They identify, in the electrical load curve, features that may be indicative of the presence of occupants in the household. For example, clear indicators for occupancy are switching events in the load curve that require the interaction of occupants. On the other hand, electric consumption related to appliances such as fridges, freezers and standby devices are not directly related to the occupancy state. The idea is that some numerical features of the electric consumption within an hour can be seen as indicative of the presence of occupants. The average electric consumption within an hour, its standard deviation, its minimum and the maximum values and its sum of the absolute differences (SAD) are all quantities related to the presence of people that are using and changing the electric consumption. In particular: a high average value, a high minimum and high maximum values usually corresponds to the use of electric equipment that can increase the consumption from the base electric use; a high standard deviation and a high SAD corresponds both to high changes in the electric consumption within the hour that can be associated with turning on/off of devices and usually it’s related with the presence of people in the flat.

To perform this analysis some steps were implemented to prepare the data set and run the analyses. The process is shown in Figure 5.8.

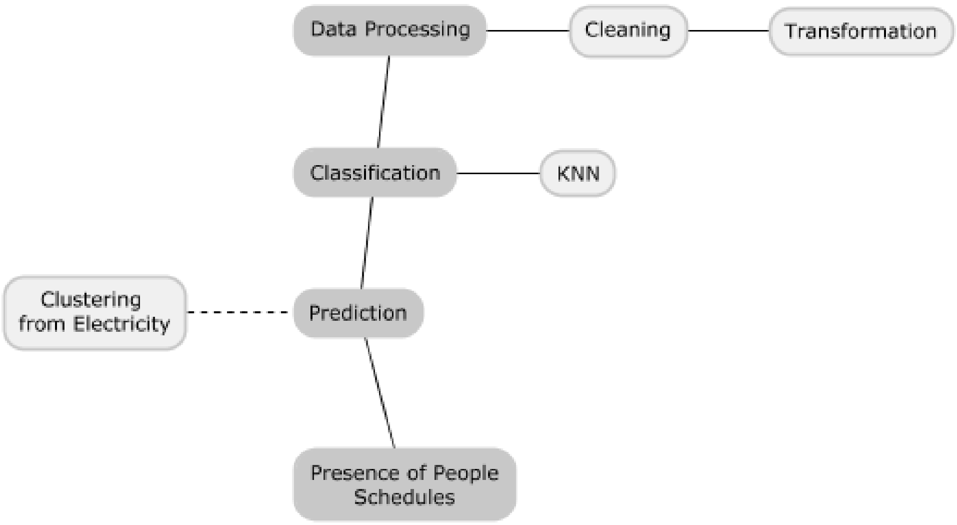


Figure 5.8: Flowchart of task 2

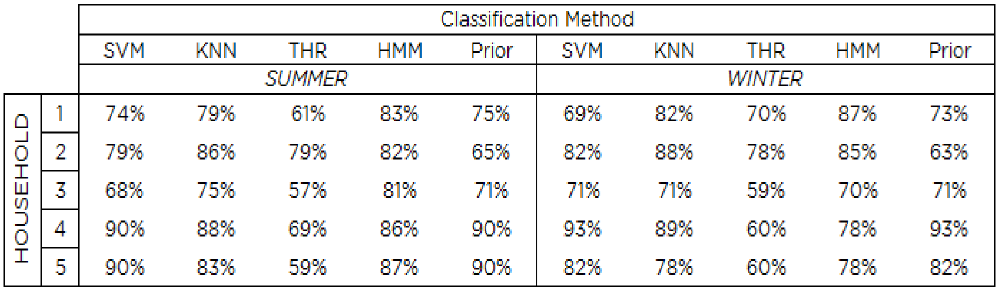
5.3.1. Data processing

The data processing of this task is composed of cleaning and transformation. These two steps have to be performed with the specific scope that the data has to be useable for the exampled methodology by Kleiminger. This means that all the hours in which the data is not sufficiently accurate to detect the 15-minutes step registration might not be part of the sample. The transformation, in this case, is the association to each hour the listed features: average, minimum and maximum, standard deviation and SAD.

5.3.2. Classification

The presented methodology is not related to the assessment of the accuracy of different Machine Learning classification methods. For this reason, according to Kleiminger et al. [7], the *k*-Nearest Neighbours algorithm was used, because it is the one with the highest average accuracy within all the households and seasons. The paper shows the accuracy for different classification methods, such as Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Thresholding (THR) and Hidden Markov Model (HMM), the results are shown in Table 5.1 and Figure 5.9.

The classification is based on the learning procedure explained in the previous chapter, and to simplify the problem a simple heuristic unsupervised occupancy detection was implemented, by comparing the current electricity consumption to the mean of the night-time consumption and proposing possible ground truth occupancy schedules to the user, as suggested in [52]. The ground truth is a dichotomous variable, in which 1 is related to the presence and 0 corresponds to the absence of people.



		Classification Method									
		SVM	KNN	THR	HMM	Prior	SVM	KNN	THR	HMM	Prior
		SUMMER					*WINTER*				
HOUSEHOLD	1	74%	79%	61%	83%	75%	69%	82%	70%	87%	73%
	2	79%	86%	79%	82%	65%	82%	88%	78%	85%	63%
	3	68%	75%	57%	81%	71%	71%	71%	59%	70%	71%
	4	90%	88%	69%	86%	90%	93%	89%	60%	78%	93%
	5	90%	83%	59%	87%	90%	82%	78%	60%	78%	82%

Table 5.1: Results of the accuracy of different classification methods. From ref. [51]

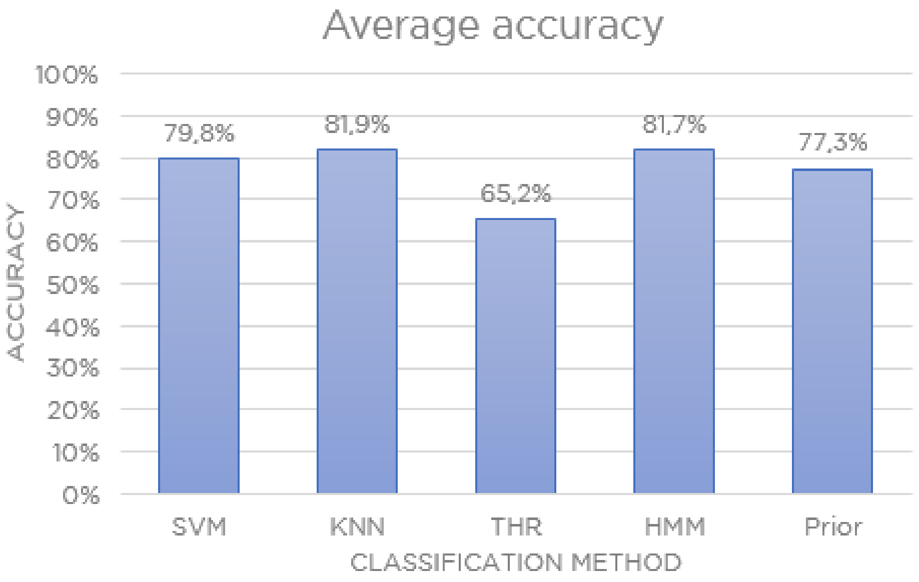


Figure 5.9: Average accuracy of the different classification methods taken into exam. From ref. [51]

5.3.3. Prediction

These final step of the process is called prediction, in fact, no clustering analyses were performed. The clusters of the previous task were taken and applied to the occupancy daily patterns of the classification outcome. The result is a continuous variable (average value), between 0 and 1, corresponding to the probability of occupancy in a specific hour of the day for each cluster.

For the presence of people, one scenario is proposed, linked to the probability of presence. When the average is below 0,33 the corresponding probability will be 0% when is between 0,33 and 0,66 will be 50% and, when above 0,66 will correspond to 100%. This is shown in Figure 5.10.

The final probability is then multiplied by the number of people that are supposed to live in the apartment on the base of the net floor area of the bedrooms. If the area is above 12 m^2 is considered suitable for 2 people, if above 8 m^2 but below 12 m^2 1 person is considered according to the Regolamento Edilizio del Comune di Milano [76].

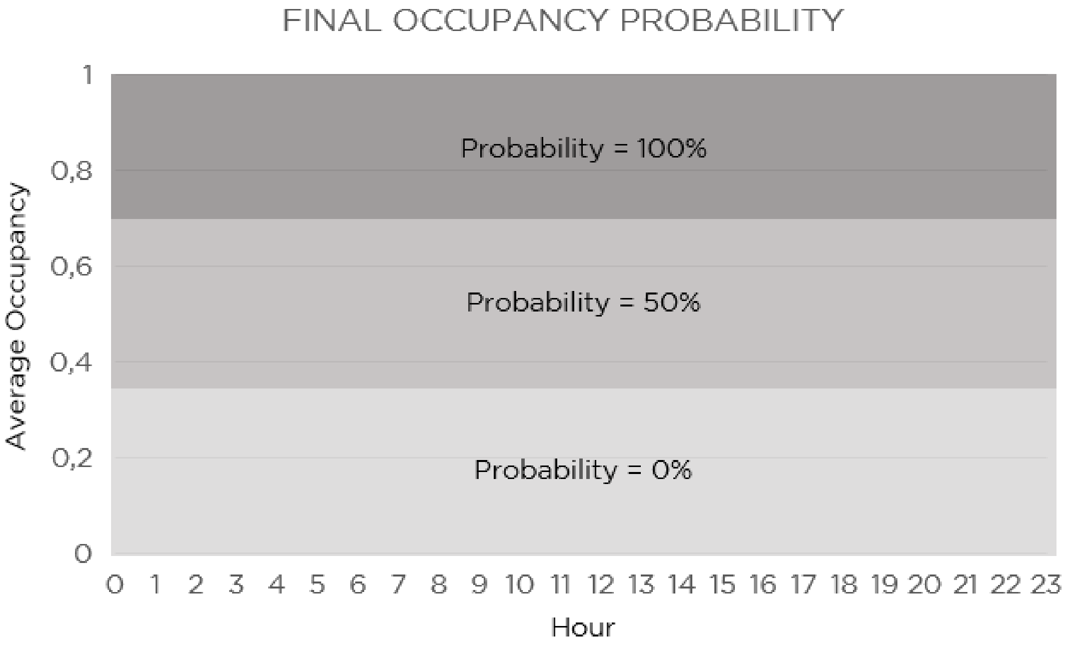


Figure 5.10: Relation between the average occupancy value and the final probability of occupancy used in the schedule

5.4. Task 3

Task 3 concerns the energy simulation of the building with the schedules created in task 1 and 2. These schedules affect the internal gains due to the people in the building and thus, they affect the energy use of the building itself. To run the final simulations EnergyPlus was deployed.

EnergyPlus is exploited running three scenarios, related to the average case and the two extremes. The schedules are used to create combinations of different inputs in the EnergyPlus model and it gives as result a range of energy needs (Figure 5.11). Modelling the extremes cases in terms of internal gains due to presence of people and electric appliances, the result are the two extreme cases also in terms of energy need.

The focus will be on the span in the energy needs due to the different combinations of the occupancy schedules used. The final result is not anymore a single value as in the traditional energy modelling, but range of results in which the real case should be.

As final validation of the methodology, these results are compared with the real registered data of 2016 of the building block. The supply company provided the delivered energy to the building site for 2016, whilst, the energy simulation will result with the energy need for the building. Thus, the energy need for space heating has been calculated applying an estimated global seasonal energy efficiency of 0,7.

In these simulations, the weather data of 2016 was used. It is created from the registration of the weather station of Milano-Via Juvara provided by A.R.P.A. Lombardia. For details, read the paragraph 4.4.

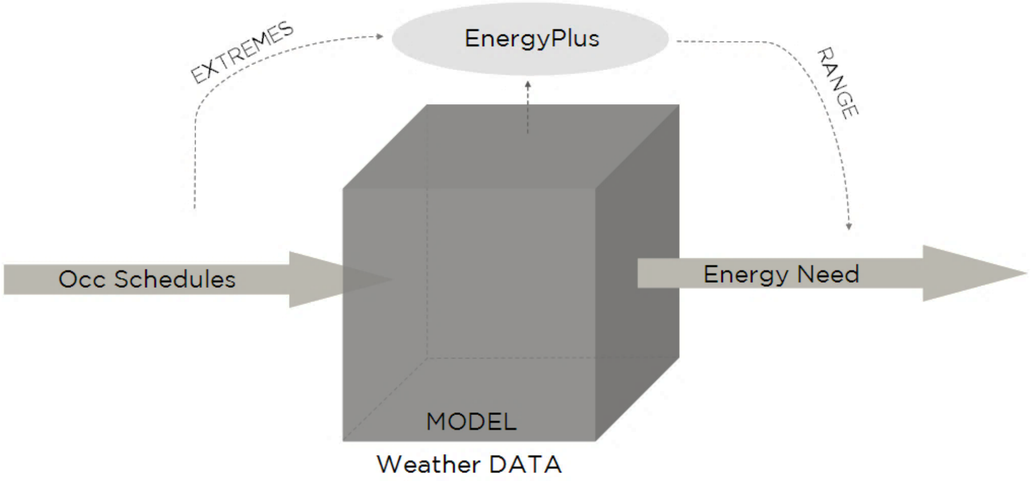


Figure 5.11: Task 3 scheme

6. Results and Discussion

It is reminded that the flats in this thesis are sometimes called "zones", this term derives from the creation of the energy model in EnergyPlus where each flat is modelled as a thermal zone and defined with a progressive number. Thus, a final code to indicate a specific flat can be Z4, in which Z stands for thermal Zone and 4 is the related progressive number.

6.1. Task 1

6.1.1. Data processing

Cleaning

Some periods of the original dataset are characterised by errors or they are totally missing the registration. To correct the inconsistent data, these periods are deleted from the dataset. In fact, a substitution can alter the dataset and modify the result. The dataset includes the electric load with a 15-minutes time-step from the 1st of February 2016 to the 31st of August 2016, for 24 flats of the B1. Some days in all the dataset are missing, they are 29/02, 6-7/03, 3-8/05.

Moreover, other periods of time are characterized by a lack of data, in particular:

- Completely Z3 and Z13, because, respectively, closed contract and empty,
- Z4 from 12/06 to 31/08,
- Z5 from 14/08 to 31/08,
- Z14 from 19/06 to 29/08,
- Z29 from 02/03 to 31/03.

Reduction

The first actual alteration of the data consists in the reduction. This step has the aim to reduce the representation of the data in volume, obtaining same or similar results. In this case, the data reduction is performed on the time-step base. The available dataset is registered every 15 minutes as electric power in Watt, to obtain an hourly value the average within the hour is performed. In this way, possible eluded outliers are reduced, and they affect less the overall results. An example of the data reduction from 15-minutes time step to hourly values can be seen in Figure 6.1. The red line represents the registered original data for the zone number 4 in the 1st of February, whilst the blue line represents the hourly values on which the further data processes are made. This reduction is applied to all the data set, decreasing the noises and the complexity of the problem, being the final aim an hourly schedule.

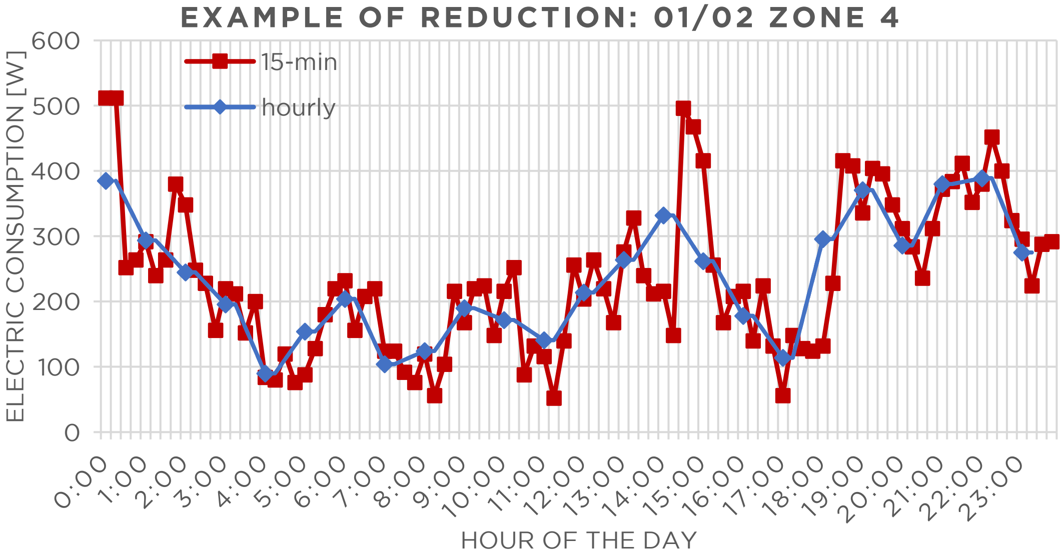


Figure 6.1: Example of data reduction from 15-min to hourly values for 01/02/2016 of Z4

Transformation

The aim of the first statistical analyses is to understand if the electric load can be easily predicted looking at some features, on different scales (hourly, daily or flat-scale). As a first approximation, the electric load can be thought to be influenced by some variables, such as the installed electrical equipment, or the number of people living in a household. These influencers are numerous; therefore, a detailed literature review is performed. In Table 6.1, all the variables that can affect the electric load are listed with one or more references in which each one is explicitly related to the electric use.

The first column of Table 6.1 categorizes the “family” of the variable. Four main groups of influencers are found to be related to:

- the location and the weather,
- the features of the household,
- the indoor conditions,
- the characteristics of the family living in the flat.

The second column lists the variables in different colours. The colour of the variable corresponds to the final exploitability, thus:

- red is for the variables that are not exploitable in this case study,
- green is for the variables that are exploitable and mark a difference within hours,
- blue is for the variables that are exploitable and mark a difference within flats.

The third column lists how the variable is affecting the load. In fact, the registered electric demand is composed by the sum of the electric demand of all the installed electric appliances inside a household. To simplify the understanding of which variable can affect the load and how, four main electric spenders can be highlighted: lighting, equipment (e.g. kettle, microwave, phone etc), small electric heating or cooling system (e.g. fans, stoves, etc), and leisure appliances (e.g. computers, television, etc). Even if the heating system of the building block does not rely on the electric load, in the energy spenders is important to add the small electric heating and cooling systems because are quite popular in the Italian context. Movable or ceiling fans, small heaters, or added conditioners are common equipment to improve the internal comfort during very hot or cold days, especially in old buildings in which local discomfort can be registered. In Table 6.1, in the column called “Affecting”, these four main categories are listed as follows:

- Lt = lighting,
- S = small electric heating/cooling systems,
- E = equipment,
- Ls = leisure.

The last column lists the works in which can be found the variable. Some of them are indicated in many papers; however, just a few of them are inserted in the table as a reference.

Family	Variable	Availability	Affecting	Reference
Location /Weather	External radiation	Yes	Lt	Mardaljevic 2009 [77]
	External temperature	Yes	S	Sandels 2015 [78]
	Workdays / Holidays	Yes	All	Paatero 2006 [79]
	Day of the week	Yes	All	Butitta 2017 [22]
	Precipitation	Yes	All	-
	Cost of electricity	Yes	E + S + Ls	-
	Hour of the day	Yes	Lt + E + Ls	Paatero, 2006 [79]
	Heating/Cooling Season	Yes	All	Paatero 2006 [79], Sandels 2015 [78]
	Renewables on site	Yes but Constant	All	Galvin 2016 [21]
	House demand limit	Yes but Constant	Sum	Capasso 1994 [80]
Flat characteristics	Orientation	Yes	Lt + S	Mardaljevic 2009 [77]
	Floor	Yes	Lt + S	Menezes 2012 [81]
	n° rooms	Yes	E + S + Ls	Yohanis 2007 [82]
	Floor Area	Yes	E + S + Ls	Yohanis 2007 [82]
	Window/Wall ratio	Yes	Lt	Bokel 2007 [83]
	Insulation	Yes but Constant	S	Sandels 2015 [78]
	g-value	Yes but Constant	Lt	Mardaljevic 2009 [77]
	Shading type	Yes but Constant	Lt	Tzempelikos 2005 [84]
	Typology	Yes but Constant	All	Yohanis 2007 [82]
Indoor	Indoor Air Temperature	No	S	Sandels 2015 [78]
	Internal Illuminance	No	Lt	Mardaljevic 2009 [77]
Family type	n° people	No	All	Capasso 1994 [80], Yohanis 2007 [82]
	Sex	No	All	Capasso 1994 [80], Yohanis 2007 [82]
	Age	No	All	Shimoda 2004 [85]
	Income	No	All	Capasso 1994 [80], Yohanis 2007 [82]
	Nationality	No	All	-
	Occupation	No	All	Capasso 1994 [80], Yohanis 2007 [82]
	Shading operation	No	Lt	Tzempelikos 2005 [84]
	Efficiency	No	All	Menezes 2012 [81]
	Electric Car	No	E	Clement-Nyns 2010 [86]
	Installed equipment	No	All	Capasso 1994 [80], Menezes 2012 [81]

Table 6.1: List of variables that can affect the electric demand with related references

The variables without reference are just hypothesis that can (or might not) be verified in this case study. Precipitation, at first look not related to the electric demand, could be a variable in the non-working day on the decision to stay at home or go out, especially during the summer period. The nationality of the family is added to the variables, assuming that different habits and comfort perception can differ due to the culture and the origin, even if not available in the dataset.

Integration

The integration step is fundamental to get the data in a usable format to run the analyses. The variables are chosen on the base of their exploitability and added in a spreadsheet with the registered data. To implement the dataset in IBM SPSS Statistics, the characteristics of each variable have to be set. The Table 6.2 summaries the list of the selected variables with their features. In the table, the code used in SPSS Statistics is listed, with the unit of measurement if available and the domain.

	Name	Code	Unit	Domain	Characteristics	
BETWEEN HOURS/DAYS	External radiation	Rad	W/m^2	[0 ≤ x ≤ 931,30]	Continuous	Scale
	External temperature	Temp	°C	[1,60 ≤ x ≤ 33,80]	Continuous	Scale
	Precipitation	Prec	mm	[0 ≤ x ≤ 29,60]	Continuous	Scale
	Month	Month	-	[2; 3; 4; 5; 6; 7; 8]	Categorical	Ordinal
	Day of the Month	Day	-	[1; 2; 3; 4; 5; 6; …31]	Categorical	Ordinal
	Hour of the day	Hour	-	[0; 1; 2; …; 23]	Categorical	Ordinal
	Day of the Week	WD	-	[0; 1; 2; 3; 4; 5; 6; 7]	Categorical	Ordinal
	Day or Night	DN	-	[-1; 0; 1]	Categorical	Nominal
	Workdays / Not Working	WNW	-	[-1; 1]	Categorical	Nominal
	Season (heating or cooling)	CH	-	[-1; 1]	Categorical	Nominal
	Cost of electricity	Cost	-	[-1; 1]	Categorical	Ordinal
BETWEEN ZONES	Orientation	Orien	-	[1; 2; 3; 4]	Categorical	Nominal
	Zones	Zone	-	[2; 4-7; 14-29]	Categorical	Nominal
	Floor	Floor	-	[0; 1; 2; 3]	Categorical	Nominal
	n° rooms	n°Rooms	-	[1; 2; 3]	Categorical	Ordinal
	Floor Area	Area	m^2	[37,87 ≤ x ≤ 95,28]	Continuous	Scale
	Window / Floor ratio	WlFl	-	[0,103 ≤ x ≤ 0,196]	Continuous	Scale

Table 6.2: List of selected variables and their features

The first distinction is between categorical and continuous variables. Categorical variables are subdivided into ranks, whilst the continuous ones allow all the possible values in the domain. A further distinction between the categorical ones is between nominal and ordinal. In the first case, the ranks do not correspond to a sequence, but just to mere numerical substitution of names. An example can be the distinction between working and non-working day, since there is not an order between the two categories and the associated number is helpful to distinguish them. On the other hand, a categorical ordinal variable expresses a sequence, usually temporal. For example, the number associated with the day of the week is related to the name itself but also to an order repeated from 1 to 7, and the same can be said for the months.

Variables description

The weather variables are derived from the registration of the weather station of A.R.P.A. Lombardia, located in Via Juvara [67]. The external radiation is the global radiation in Watt on square meters, calculated as the hourly average of the measured data. The external temperature is the hourly average temperature of the registration in Celsius Degrees. Precipitation is the hourly cumulative value in mm.

The month, the day of the month, the hour of the day and the day of the week variables are inserted to give a temporal distinction that can be used as the variable itself, but also to subdivide into groups the data sample in the SPSS software.

Day or night is inserted as a categorical variable. Day is indicated with 1, and it is related to the hours in which there is solar radiation in the shortest day of the year (the winter solstice), thus from 8 a.m. to 4 p.m. Whilst, night is indicated with -1, and it is related to the hours in which there is not solar radiation in the shortest night of the year (the summer solstice), thus from 10 p.m. to 4 a.m.

The variable working/non-working day is introduced as the difference between the weekdays and all the other days in which usually people do not go to work or school. The non-working days are the weekends and all the national holidays.

The difference between the heating and the cooling season is set according to Art. 9 of D.P.R. 26/08/93 [87]. Milan belongs to climatic zone E, so the heating season is from 15/10 to 15/4.

The cost of electricity is added as variable because, in bi-hourly tariff contracts the cost of the electricity decreases during the nights, weekends and national holidays, a graph of the typical week can be seen in Figure 6.2. Some families, especially in periods of economic hardship, can decide to shift the use of some electric devices (such as the washing machine or the dishwasher) in the hours in which the cost is lower.

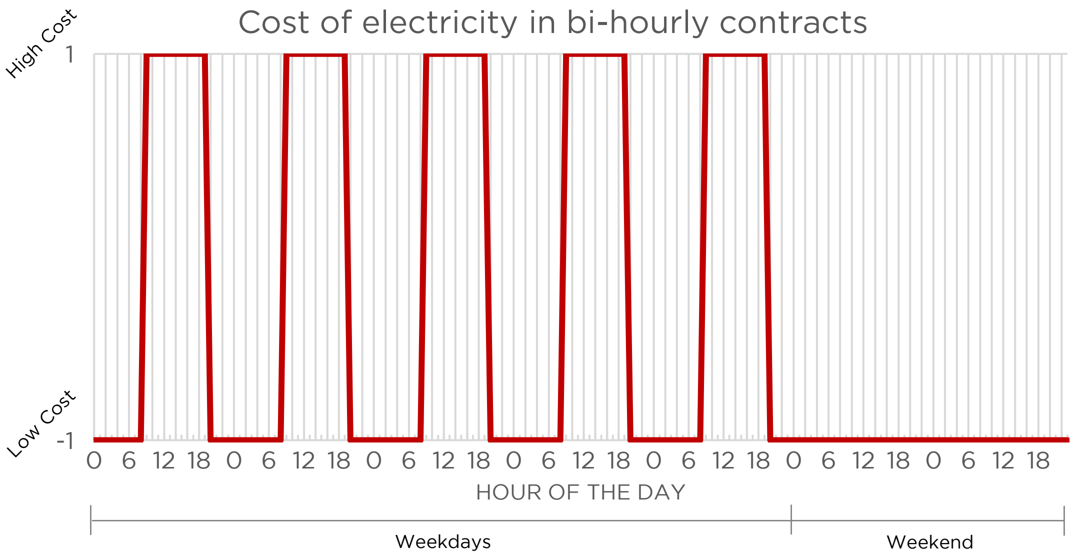


Figure 6.2: Distribution of the "electricity cost" variable during a typical week, with the distinction between high and low cost

The orientation is set according to the position of the main windows of each flat. The orientation is a categorical variable with values from 1 to 4, in which 1 is South-West, 2 is Noth-West, 3 is North-East and 4 is South-East.

The zone is simply the progressive number used to distinguish the flat. It is useful to create graphs and groups in SPSS Statistics.

The floor is related to the storey at which the flat is, it is a categorical variable with a range from 0 to 3. 0 is for the mezzanine, 1, 2, 3 are for the first, second and third floor.

The variable "n°Rooms" is set as the number of bedrooms in the flat, the area is calculated as the net useful floor area, and the window/floor ratio is the ratio between the net area of the windows and the net useful floor area.

6.1.2. Data understanding

In this section, traditional statistical methods and visualization graphs are exploited to understand the kind of variables and of database. Lastly, a correlation test is performed to understand better the links between the variables and the electric registered data.

Statistical analyses on and within variables

For the categorical variables, codebooks and descriptive statistics are enough, whilst, in the case of continuous variables, histograms and box plots are added to have a visual description. The relative tables and graphs can be found in Section I of Annex II.

The result of the correlation test is of major importance and it is shown in Table 6.3. The Spearman's rho correlation coefficient expresses the relation between the two variables taken into exam, with a maximum of 1 or -1, corresponding to perfect positive and negative correlations. In the case of nominal categorical variables, the sign is not always relevant, being not related to an order.

A positive correlation can be found between the variable day/night and the global radiation, as expected. A moderate correlation is registered between day/night and the cost of electricity. A positive correlation is registered between the cooling and heating season variable and the external temperature. This result is predictable. An expected strong correlation is registered also between working and non-working day and the day of the week. An almost moderate correlation is found between the working/non-working variable and the cost of electricity. A moderate correlation is registered between the cost of electricity and the radiation, in fact, during the night hours, the electricity is always low-cost. Finally, a very strong correlation (almost perfect) is registered between the floor qrea and the number of bedrooms in a flat. This result is predictable. Further graphs to describe the main correlations can be found in the second section of Annex II.

From these simple test, can be said, that some variables are dependent on others, and for this reason, they can be avoided to simplify the problem. Nevertheless, any variable is deleted before the correlation test with the electric demand data. For example, one variable between the number of rooms and the area could be avoided, but to understand which one, a correlation test with the registered data should be performed to identify which one affects more the electricity demand.

Correlations

		DN	CH	WNW	WD	Prec	Temp	Rad	Cost	n°Rooms	Area	WiFI
Spearman's rho Correlation Coefficient	DN	1,000	,000	,000	,000	-,004	,177**	,881**	,646**	,000	,000	,000
	CH	,000	1,000	-,029**	,005	-,046**	,731**	,173**	-,013**	,000	,000	,000
	WNW	,000	-,029**	1,000	-,727**	-,074**	-,004	,007	,451**	,000	,000	,000
	WD	,000	,005	-,727**	1,000	,052**	,005	-,011**	-,328**	,000	,000	,000
	Prec	-,004	-,046**	-,074**	,052**	1,000	-,196**	-,063**	-,051**	,000	,000	,000
	Temp	,177**	,731**	-,004	,005	-,196**	1,000	,369**	,171**	,000	,000	,000
	Rad	,881**	,173**	,007	-,011**	-,063**	,369**	1,000	,623**	,000	,000	,000
	Cost	,646**	-,013**	,451**	-,328**	-,051**	,171**	,623**	1,000	,000	,000	,000
	n°Rooms	,000	,000	,000	,000	,000	,000	,000	,000	1,000	,935**	-,078**
	Area	,000	,000	,000	,000	,000	,000	,000	,000	,935**	1,000	-,229**
	WiFI	,000	,000	,000	,000	,000	,000	,000	,000	-,078**	-,229**	1,000

**. Correlation is significant

Table 6.3: Correlation results between variables

Statistical analyses of the registered electric data

The analyses start with descriptive statistics, summarized in Table 6.4. The maximum is less than the standard maximum of 3 kW. This value shows that ideally there are no outliers that can seriously affect the result. The values of Skewness and Kurtosis indexes prove that the distribution is far from a normal distribution, as can be seen also in the frequency graph in Figure 6.3. The values are moved towards zero (as the positive Skewness value shows) and the high-values are rare compared to the low-ones.

Descriptives

		Statistic	Std. Error
Electric Demand W	Mean	193,3	,646
	Median	136,0	
	Variance	42968,1	
	Std. Deviation	207,3	
	Minimum	0	
	Maximum	2691	
	Range	2691	
	Skewness	3,1	,008
	Kurtosis	13,5	,015

Table 6.4: Descriptive statistics of the electric data

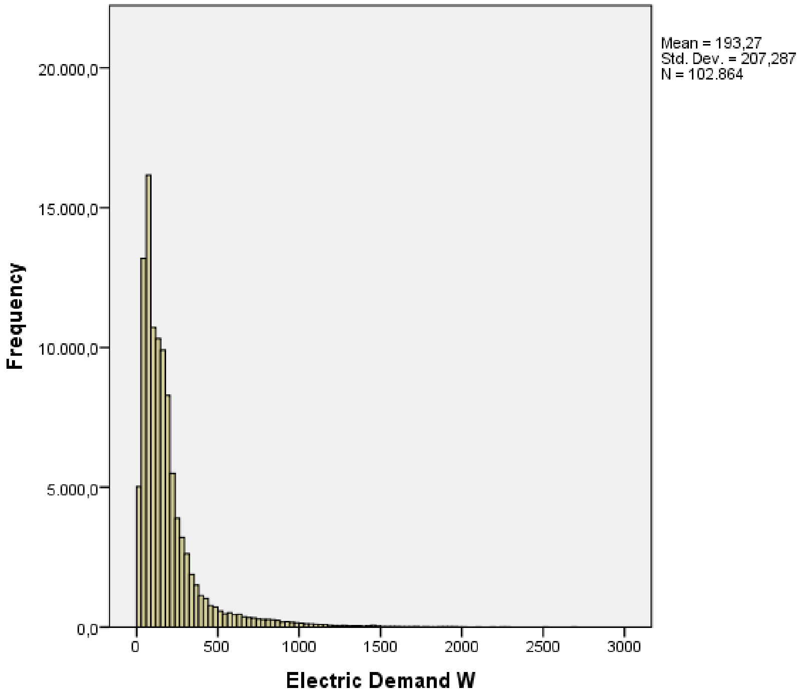


Figure 6.3: Frequency graph of the electric data

To understand better the trends of the electric demand in the building some other graphs are exploited for a visual and easy interpretation.

Figure 6.4 shows the daily sum of the electric demand of all the flats in the whole period from 01/02 to 31/08. The use of electricity in the dwellings slightly decrease along the seasons. This trend can be simply justified thinking about the decrease in the electricity spent for lighting during the summer period, in which the days are longer with a higher incidence of sunny days.

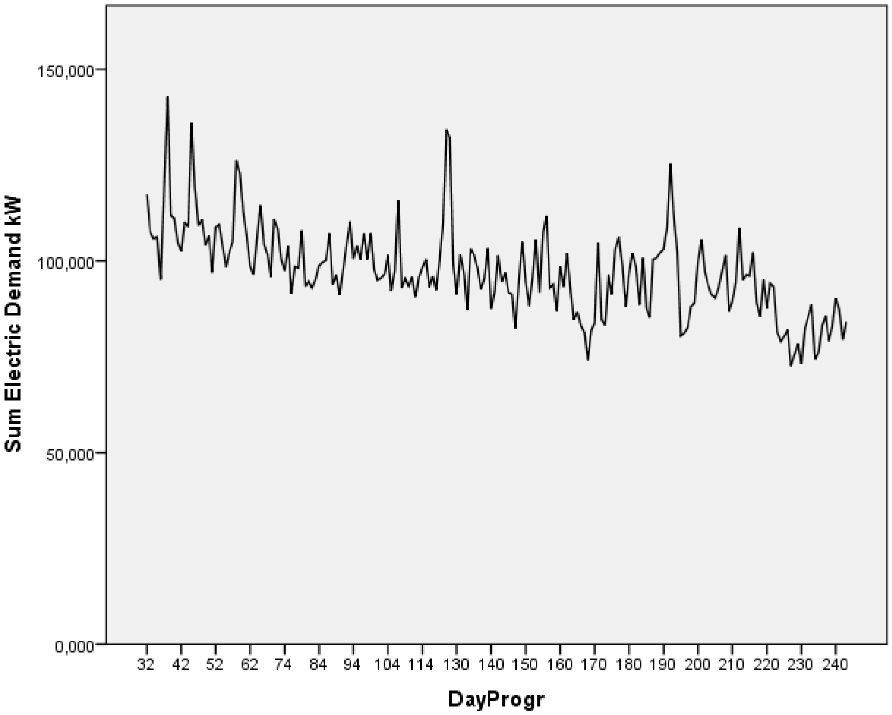


Figure 6.4: Total electric demand of the flats in the period from 01/02 to 31/08

Figure 6.5 shows the monthly average electric demand of the flats and again a negative trend can be seen going from February to August. However, July is characterized by an increase in the electric consumptions compared to the close moths, showing an average value comparable with February and March. A hypothesis to this behaviour is the increased use of small cooling electric devices in that month, one of the hottest according to the Global Climate Report of 2016 [88]. August shows a lower consumption probably due to the fact that summer holidays in Italy are usually in this month, and in the building, some families could be out for long periods. However, this is just a hypothesis that cannot be taken for granted.

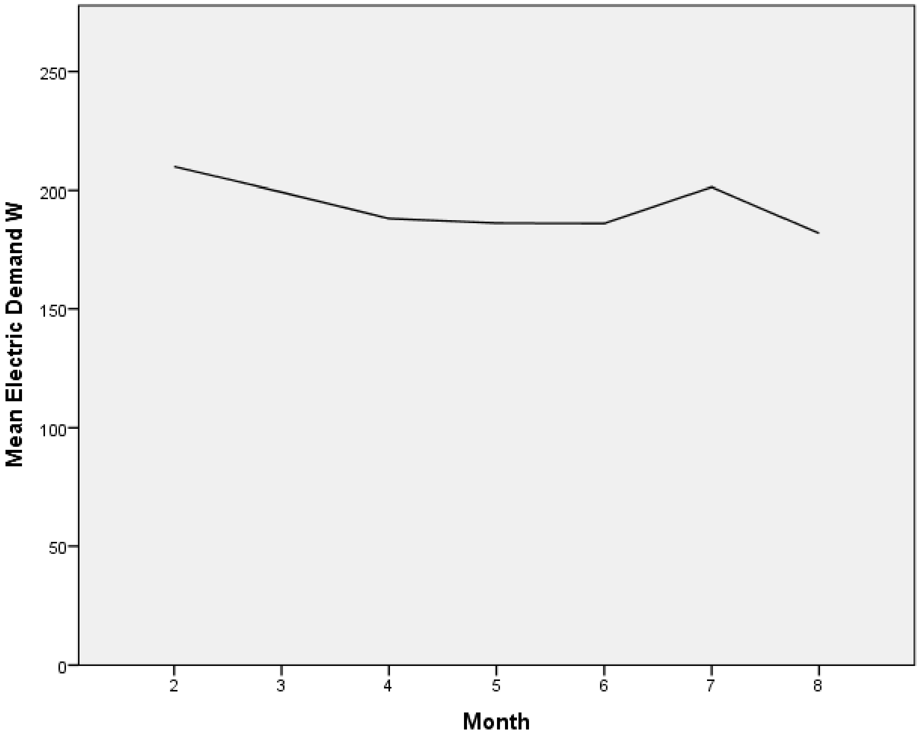


Figure 6.5: Monthly average of the electric demand

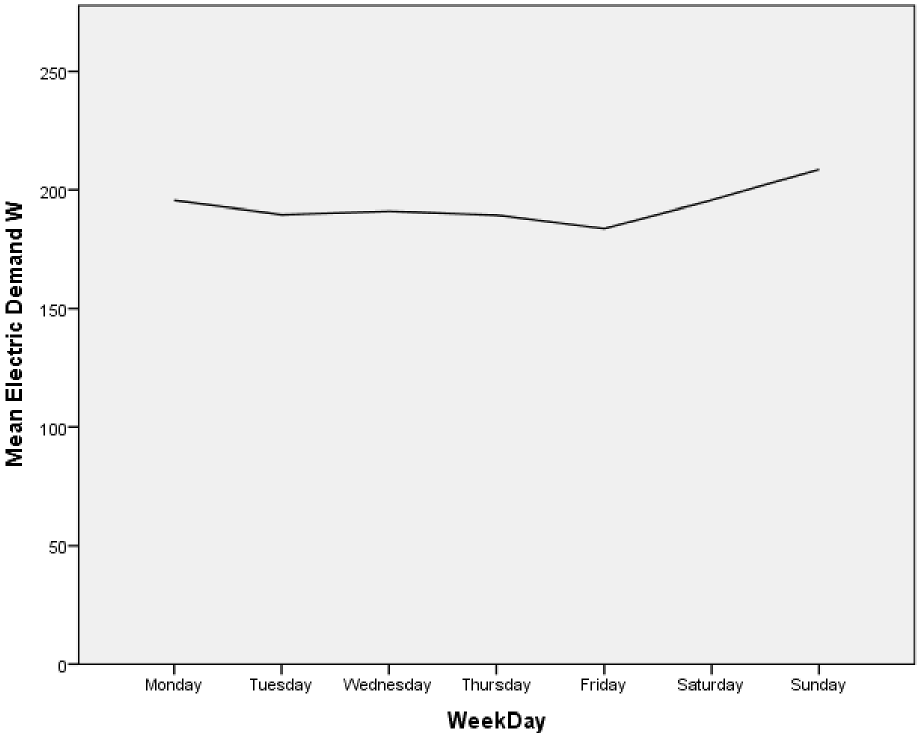


Figure 6.6: Daily average of the electric demand

Figure 6.6 shows the average electric demand of the flats on a daily base. Sunday is characterized by the highest value of average electric use. This result can be ascribed to the fact that people could stay at home more than working days, resulting in an increase of the electric demand.

Moving to the hourly base, Figure 6.7 shows the mean electric demand in the building. This daily pattern can be used to deduce some characteristics of the electric daily load. The early morning is characterized by a very low electric demand with the minimum reached around 4 a.m., then, the electricity spent increases till the lunch time, around noon. During the afternoon there is an almost constant electricity consumption and the maximum values are registered in the evening, approximately from 7 to 10 p.m. The evening period is sharply higher than the rest of the day, due to the fact that probably, almost all the tenants are at home, having dinner, using lighting and/or using leisure electric equipment such as television or personal computers.

To study deeper this pattern, Figure 6.8 shows the same average hourly values subdivided into thermal zones, thus in flats. The low-electricity hours around 4 a.m. are common to all zones at different levels. The absolute value of these hours can be proportional to the basic equipment or the standby ones, such as the fridge, the freezer or modem, etc. During the day, the different load curves differ one from another. It can be noticed that the values can change a lot, with the maximum average for zone 24 and minimum averages for zones 6, 14 or 16. Also, the trends are different, for example, zone 12 shows a low consumption during the afternoon, but two peaks during the lunch and dinner times. Whilst, zone 2 shows a peak also during the morning.

Already from these basic analyses can be seen that the trend is quite complex to be studied. The variables that can affect the result are numerous and the electric demand is unpredictable. For these reasons, a correlation test between the variable and the electric demand is run.

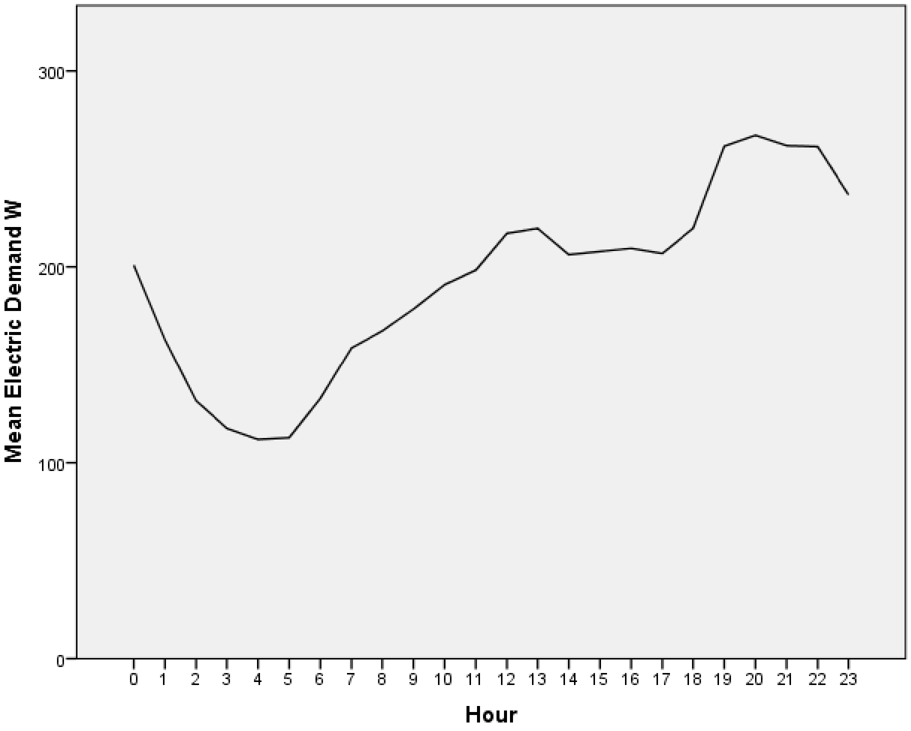


Figure 6.7: Hourly average of the electric demand

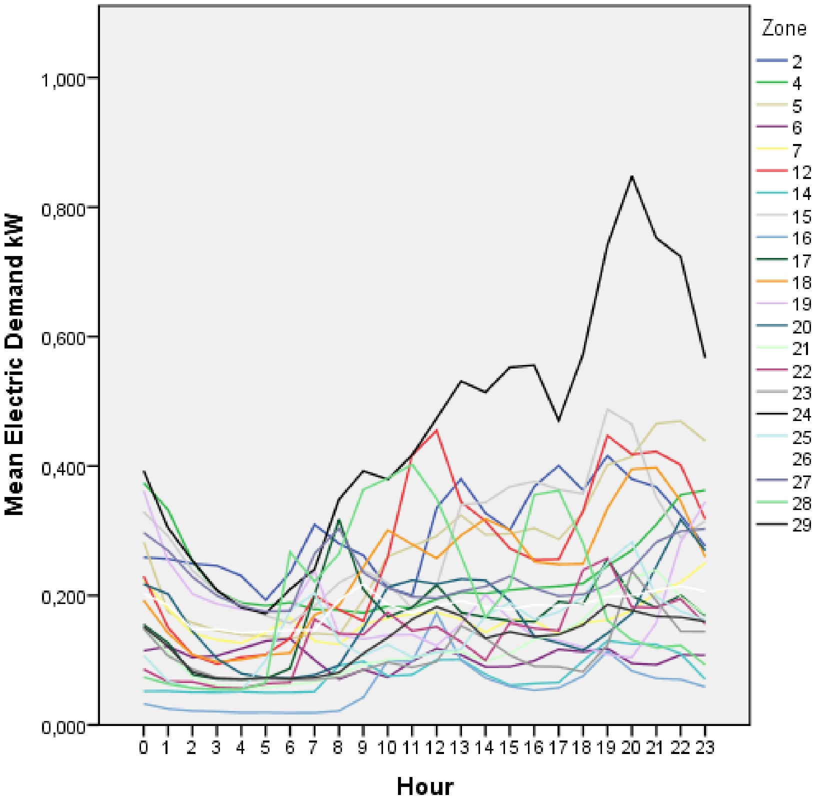
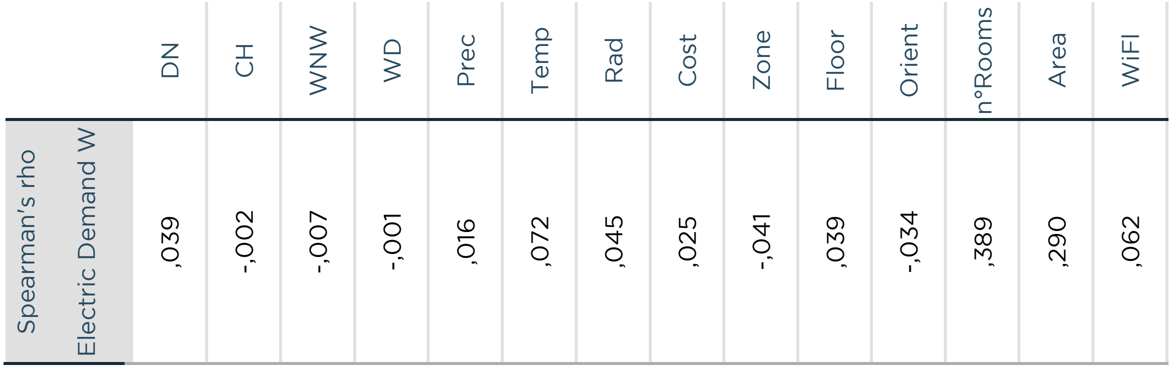


Figure 6.8: Hourly average of the electric demand divided into zones

Correlation test

Table 6.5 shows that no variable can be considered as highly or moderately correlated with the electric demand. A weak correlation is shown with the number of rooms and with the floor area. Thus, to a bigger flat corresponds a higher electric use, probably due to the higher number of electric appliances installed.

In the beginning, due to this result, the number of rooms was taken as normalization variable for the clustering step. After several analyses, the normalization against the maximum daily value was used though, simplifying the clustering step.



		DN	CH	WNW	WD	Prec	Temp	Rad	Cost	Zone	Floor	Orient	n°Rooms	Area	WiFI
Spearman's rho	Electric Demand W	,039	-,002	-,007	-,001	,016	,072	,045	,025	-,041	,039	-,034	,389	,290	,062

Table 6.5: Correlation results between variables and electric demand

6.1.3. Clustering

Self-Organizing Map

The data is normalized to the daily maximum. The final size of the Self-Organizing Map is 8 x 42. Thanks to Matlab, it is possible to create graphs to help in visualizing a multi-dimensional input space. Figure 6.9 (a) shows the topology of the used SOM, and Figure 6.9 (b) shows the connections between the neurons. This graph uses blue hexagons to represent the neurons, whilst the red lines represent the connection between neighbouring neurons. After the running of the SOM, each neuron represents a proto-cluster. Another useful graph is the SOM Sample Hits (Figure 6.9 c). It shows how many data points are associated with each neuron. The distribution is not even and some neurons group many days. Finally, Figure 6.9 (d) shows the SOM Neighbour Distances, which presents the following colour coding:

- the blue hexagons represent the neurons,

- the red lines connect neighbouring neurons,
- the colours in the regions containing the red lines indicate the distances between neurons,
- the darker colours represent larger distances,
- the lighter colours represent smaller distances.

The proto-clusters are not sharply subdivided, and any proto-cluster is isolated from the others, they are mainly linked together. Obviously, there is a direct relation between the two graphs (6.9 c and d).

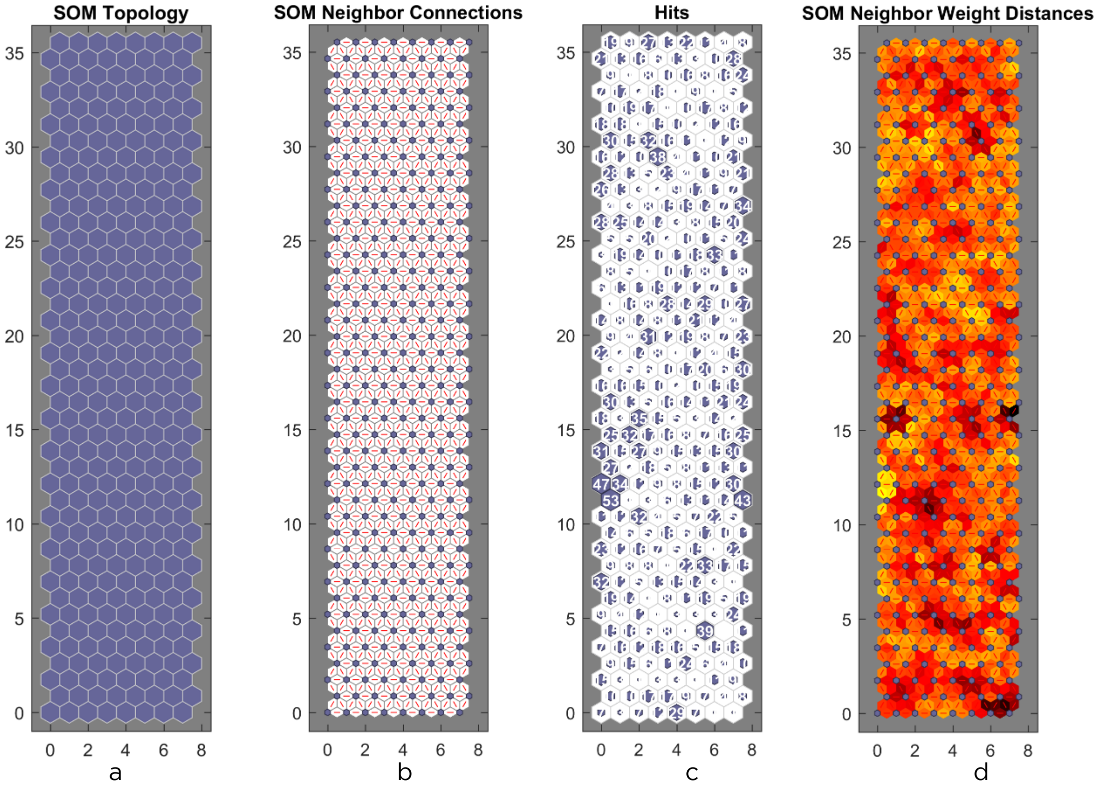


Figure 6.9: (a) Topology, (b) Connections, (c) Hits and (d) Weighted distances of the SOM

The weights themselves can be visualized with the graphs in Figure 6.10. There is a weight plane for each element of the input vector (in this case the 24 hours of the day). They are a visualization of the weights that connect each input of the neurons. The colour coding is the same used in Figure 6.9 d. These graphs express the strength of the SOM as a visualization tool. Thus, if the connection patterns of two inputs are very similar, can be assumed that the inputs are highly correlated. In this case, the inputs of the hours from 2 a.m. to 5 a.m. are very similar and at the same time, they are very different from the other inputs.

Also the inputs from 9 p.m. to midnight are very similar one to the other. This means that these groups of hours show similar values in the data sample. Especially the group from 2 a.m. to 5 a.m. corresponds to the daily minimum period underlined in Figures 6.7 and 6.8.

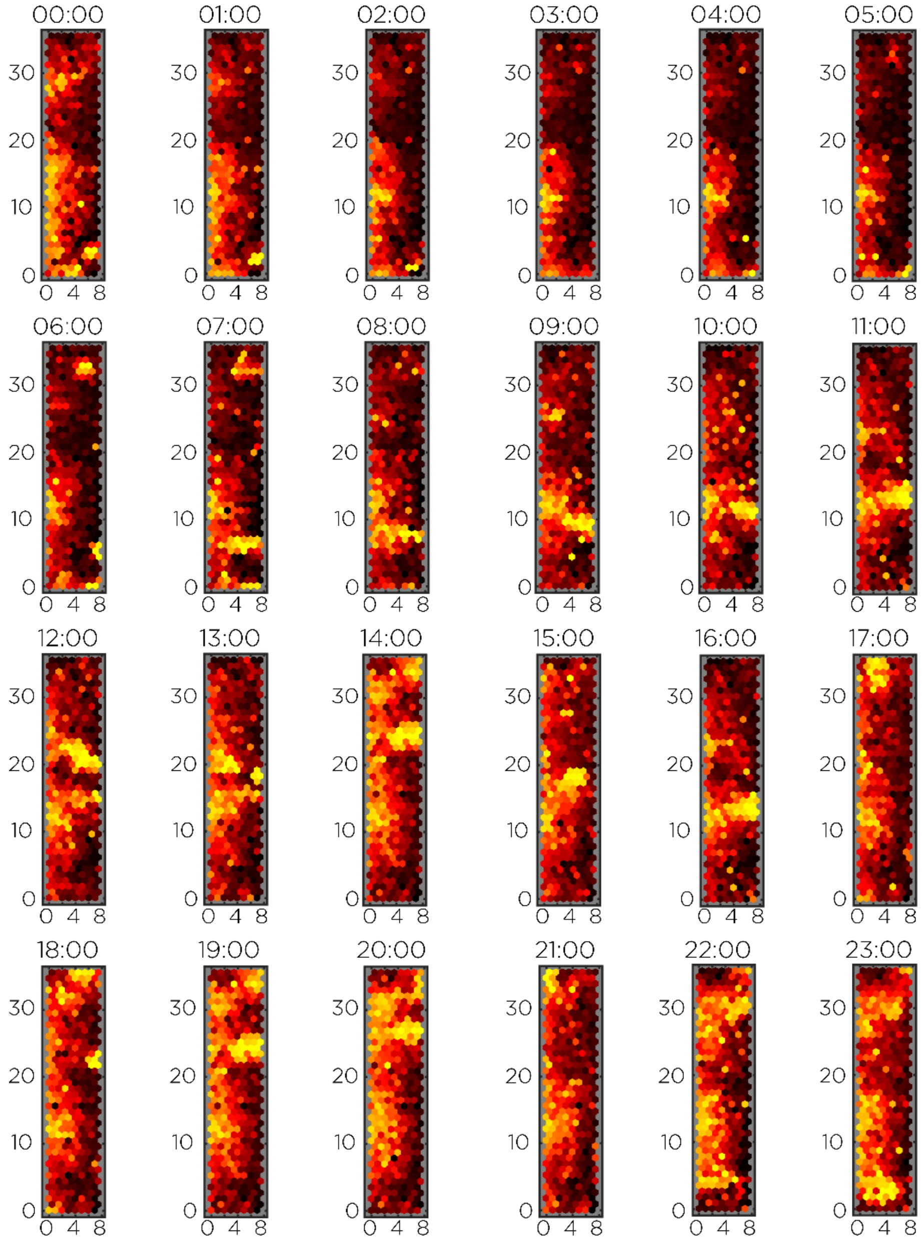


Figure 6.10: Weights planes for each input

In Figures 6.11, 6.12 and 6.13 are reported, as examples, the first three proto-clusters with the daily loads grouped. Figures 6.14, 6.15 and 6.16 show the corresponding loads not normalized.

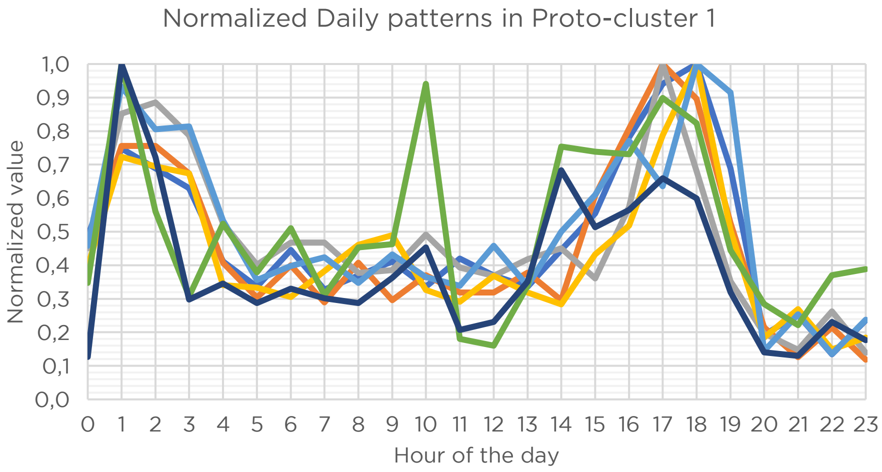


Figure 6.11: Normalized daily load curves in proto-cluster 1

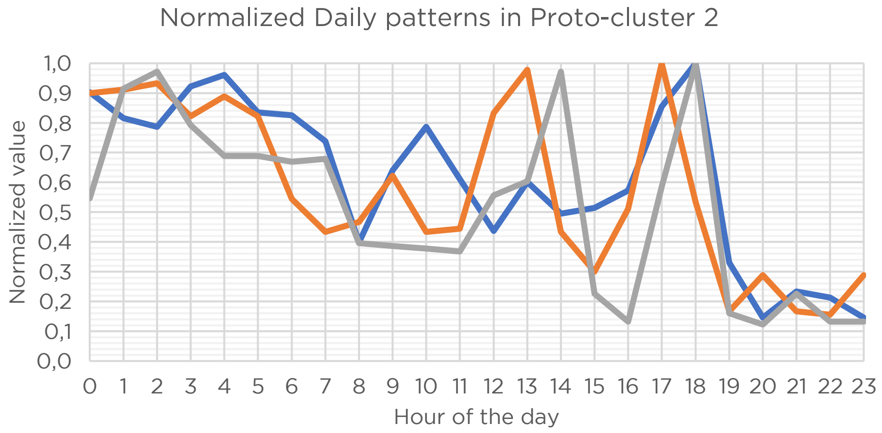


Figure 6.12: Normalized daily load curves in proto-cluster 2

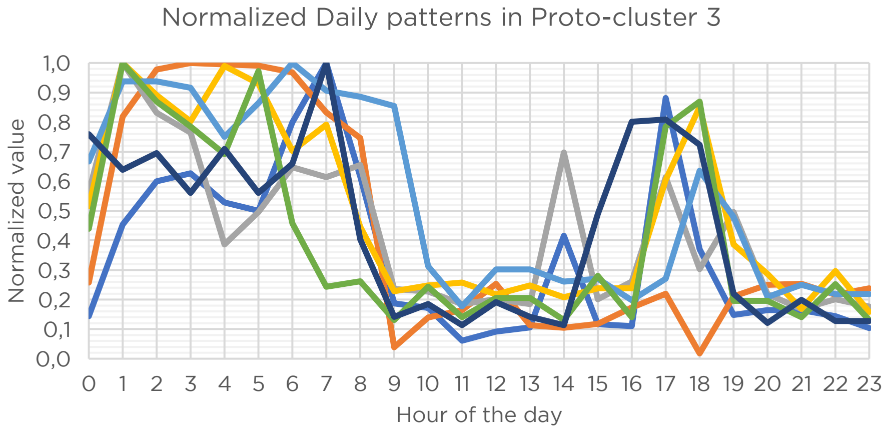


Figure 6.13: Normalized daily load curves in proto-cluster 3

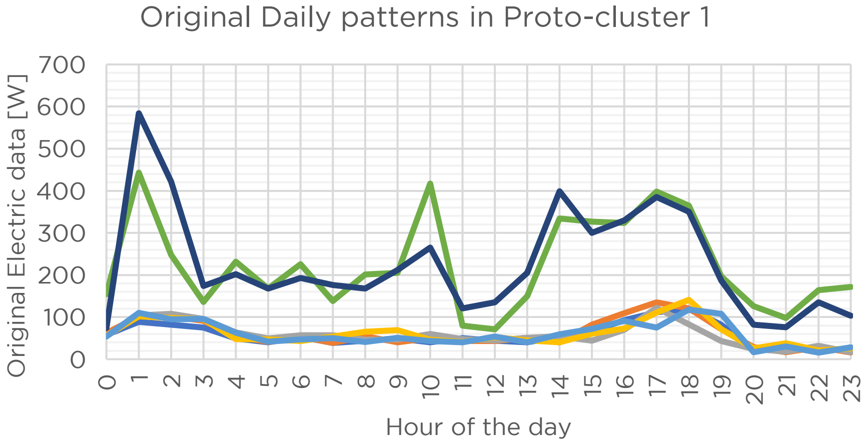


Figure 6.14: Corresponding original daily load curves in proto-cluster 1

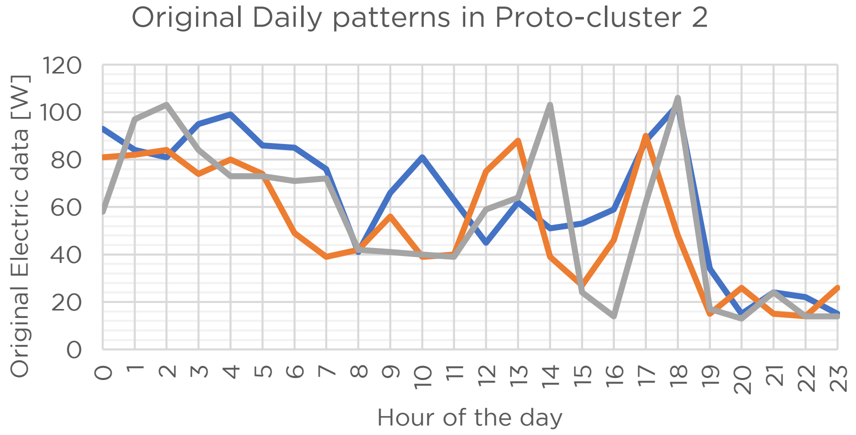


Figure 6.15: Corresponding original daily load curves in proto-cluster 2

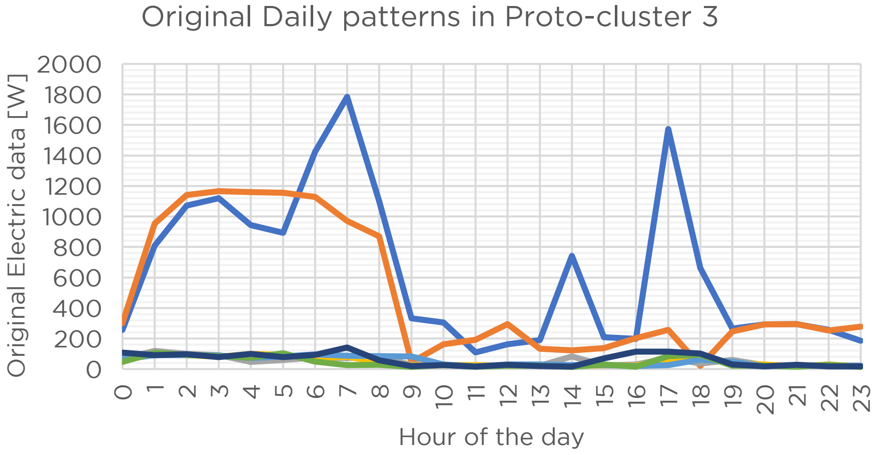


Figure 6.16: Corresponding original daily load curves in proto-cluster 3

The normalized graphs are clean and net. This is the evidence of how the SOM is working properly to reach the final aim of recognizing the different daily patterns in the big data sample. The normalized graphs show, respect to the ones with the original values, how the normalization helps to find the daily pattern. The different absolute values in the original daily profiles can confuse the SOM, bringing to a not-satisfying clustering. The average of the normalized daily load in the same proto-clusters is calculated and then submitted to the k-means algorithm.

k-means

For each cluster, two graphs are shown. The first graph explains the three different scenarios: energy-aware user, standard user and energy-intensive user, calculated as the three quartiles (first, second and third). The second graph shows the proto-clusters inside a cluster. These graphs can appear quite confused, but a trend in the overall daily loads can be found and it is expressed with the three final loads shown in the first graph. Figure 6.17 shows the representativeness of each cluster in the data sample. 28 % of the days of the original data set are grouped in Cluster 5, 25 % are in cluster 4, 20 % are in Cluster 1, 14 % are in Cluster 2, and 13 % are in cluster 3. Any cluster is far more representative than others, and thus, the 5 clusters are sorted properly. The first analyses run without normalization showed always one cluster far more numerous respect to the others. More than 50 % of the daily loads were in this cluster. These daily loads were the ones without evident peaks but also the ones with a lower average electric demand even if characterized by evident peaks if normalized.

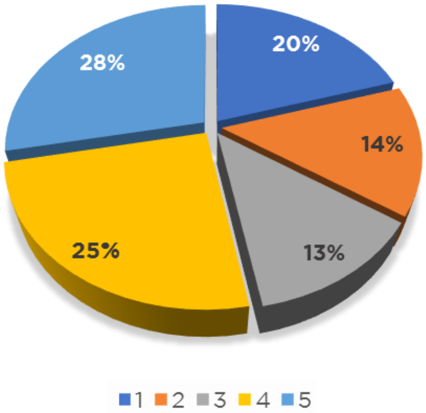


Figure 6.17: Clusters’ representativeness in the data sample

Cluster 1

Cluster 1 is characterized by all the days in which there are not evident peaks. This is visible in both the graphs that describe the cluster (Figure 6.18). The average load is around 150 W. These types of daily loads could be representative of the days in which the house is completely empty or the one in which the dwellers are at home constantly but using not so many electric appliances. For example, a couple of retired people could stay at home all day but use few electric devices. Around lunchtime and in the evening, the first graph shows in all the three quartiles a slight increase in the electric demand.

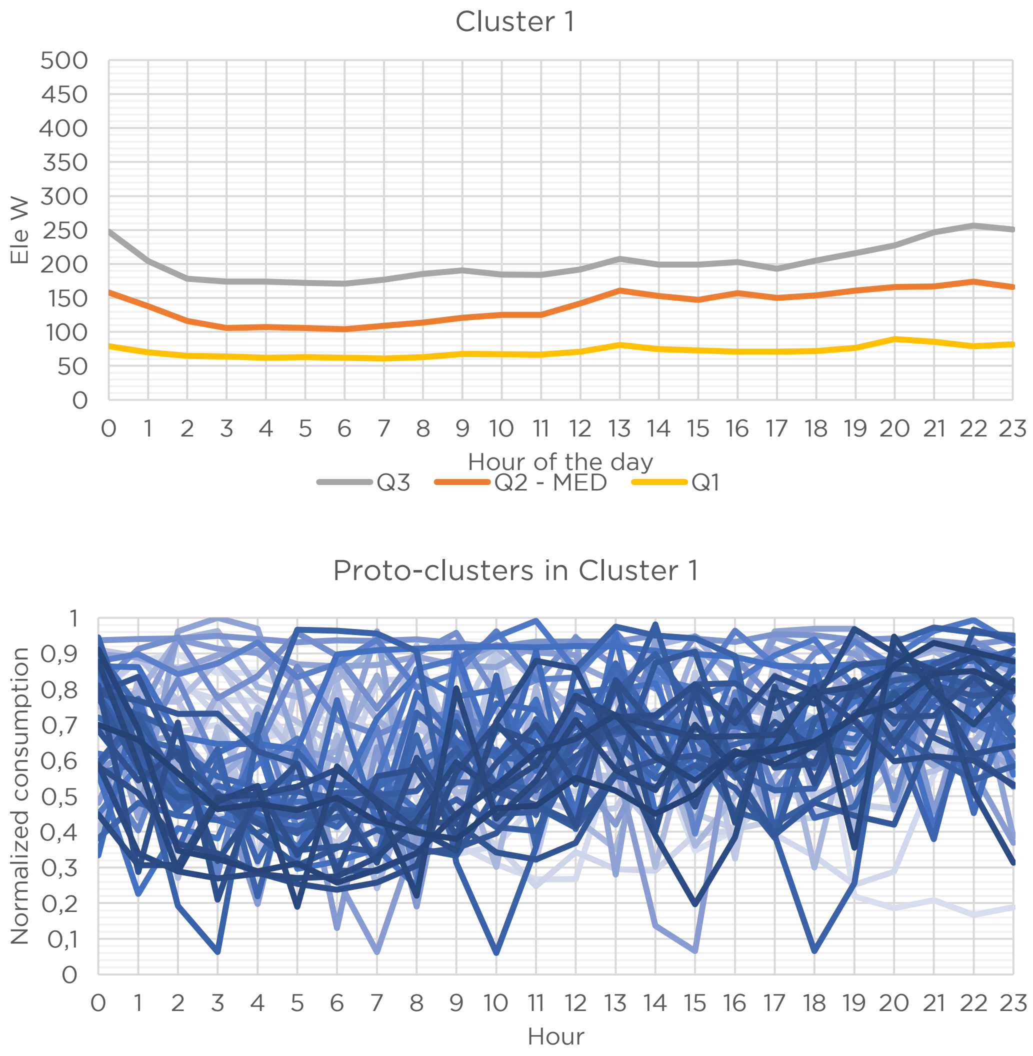


Figure 6.18: Cluster 1

Cluster 2

Cluster 2, shown in Figure 6.19, presents a two-peaks daily load. The average electric use is not far from the value of cluster 1, but the third quartile shows high values, that reach almost 500 W. The load rises around 8 a.m. and then increase progressively during the afternoon with maximum uses around 6 p.m. This cluster could be representative of days in which the occupants go out in the morning and gradually go back home during the afternoon. This could be a family with children, that go back to school in the afternoon, preparing dinner around 7 p.m. The second graph, which visually appears noisier, shows the same trend with a very low value around the lunchtime. Just a few days shows peaks in this period and in the late evening.

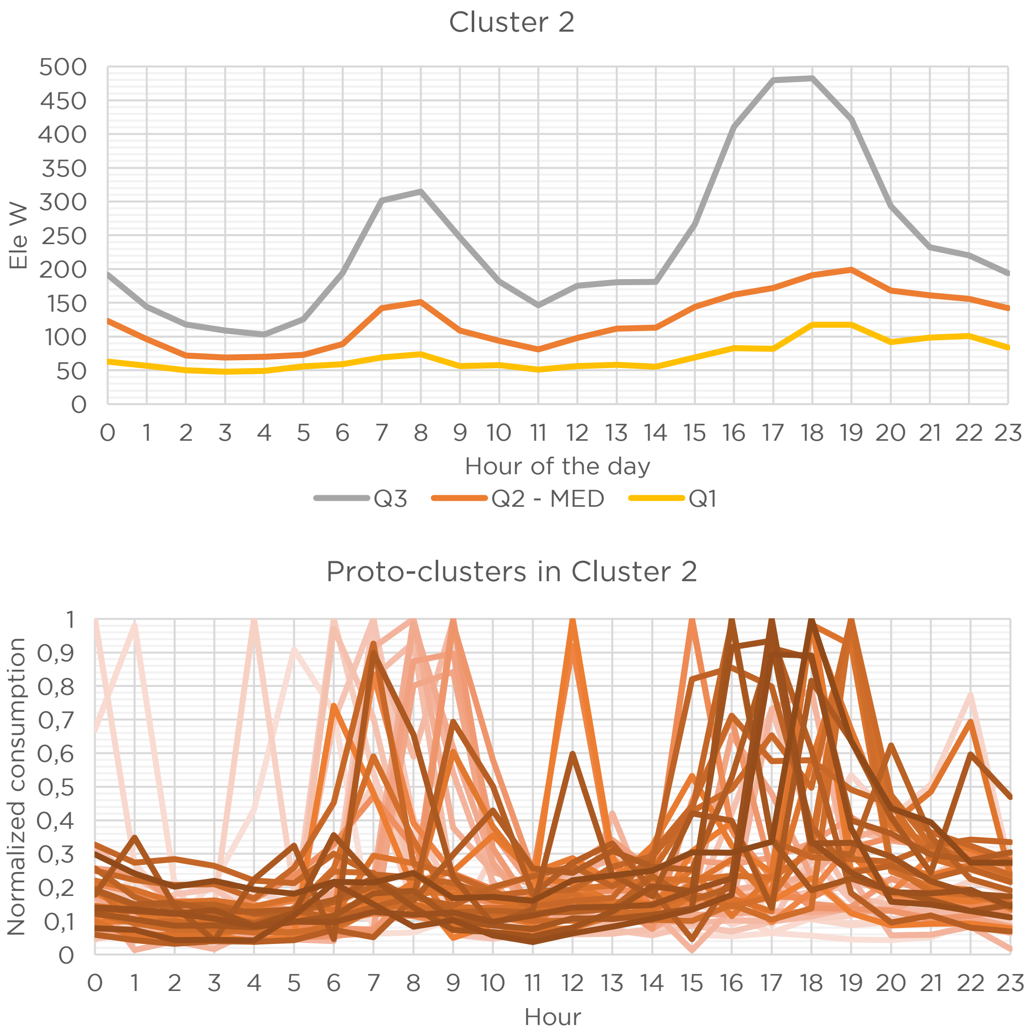


Figure 6.19: Cluster 2

Cluster 3

Cluster 3, described in Figure 6.20, shows a daily load again with an average value around 150 W, but without very high peaks. This cluster is characterized by a very low consumption during all day, with a small peak in the morning around 7 a.m., but quite high values during the night. This cluster could be representative of people that are out all day long, going out in the morning and coming home in the late evening, having dinner outside or around 8 p.m. The increase in the electric load during the night could be attributable to the use of washing machine, dishwasher and appliances used for leisure in these hours. The second graph is quite noisy but is evident the decrease in the consumption during the day, especially around noon.

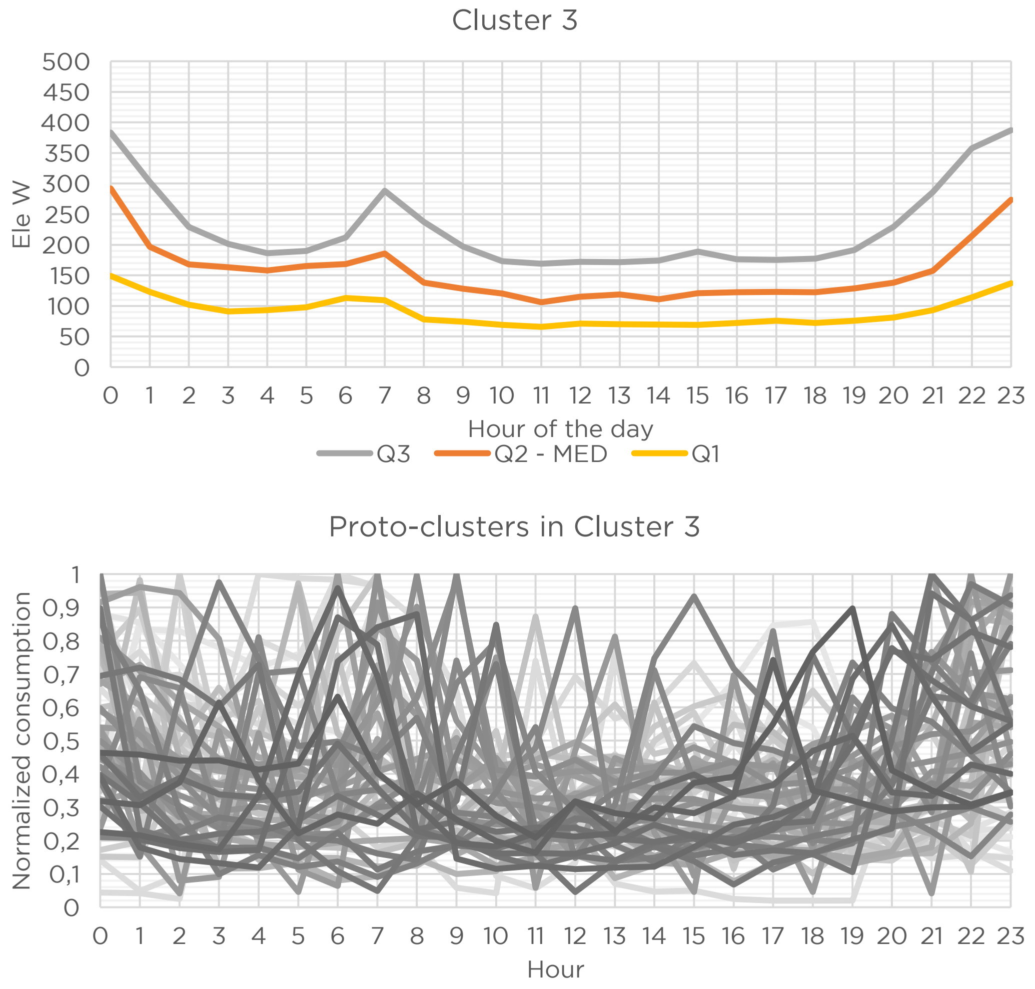


Figure 6.20: Cluster 3

Cluster 4

Cluster 4, in Figure 6.21, shows peaks during the evening, around 8 p.m. This daily load could be typical of dwellers that are out in the morning, and in the afternoon. In terms of pattern and values is not very different from Cluster 1, except from the absence of the morning load and the shifting of the evening peak from 7 p.m. to 8 p.m. This could be an indication of the different habits of the same typology of family composition. For example, Cluster 1 could be characterized by electric usage during the morning, such as for the television, radio or electric toothbrushes, razors, or kitchen tools, differently from cluster 4. In addition, the dinnertime could be different for the two cases.

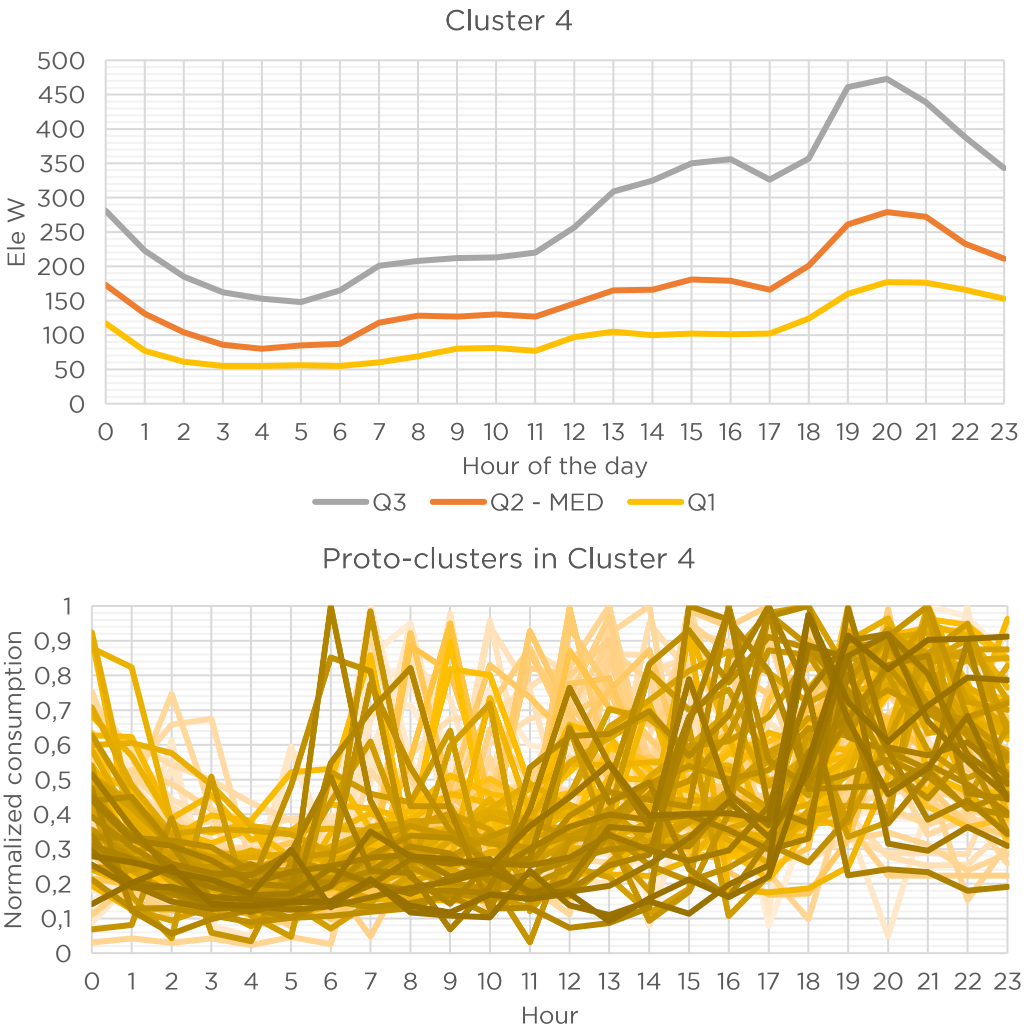


Figure 6.21: Cluster 4

Cluster 5

Cluster 5, shown in Figure 6.22, is characterized by a two-peaks load. The maximum values are reached around noon and 8 p.m. These could be respectively the lunch and dinner time for any typology of family. During the afternoon a relatively low electric use is registered, almost similar to the one registered around 4 a.m. in the morning. This could be a sign of the absence of people inside the house or of a limited use of electric appliances during these hours. The second graph of Cluster 5, shows quite clearly the pattern also with the clusterized proto-clusters.

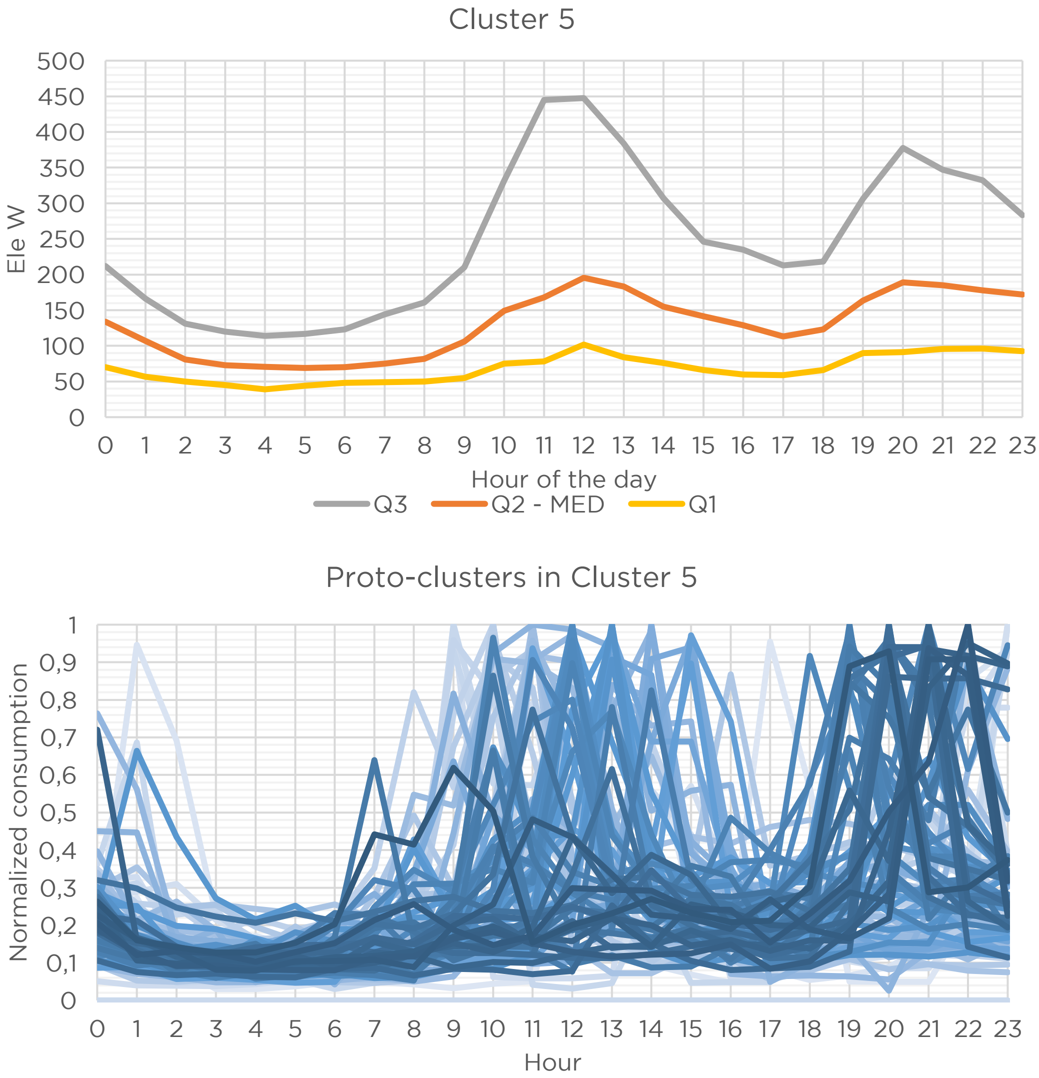


Figure 6.22: Cluster 5

6.1.4. Comparison

Before the generalization of the results of the clustering to all year, a comparison between the three scenarios and the registered data is performed. To complete the zones 3 and 13, the most similar flats in terms of characteristics (orientation, floor, area, window/floor ratio, etc) were used: respectively zone 6 and 16.

Figure 6.23 shows the sum of the electric use of the period taken into exam (from 01/02/16 to 31/08/16) in MWh. The Q2 scenario (the average) is different from the real registered data by 3 MWh. This means that probably all the three scenarios are slightly underestimating the electric demand. This difference, however, is not big and the combination of these scenarios in the final analyses will decrease more this gap.

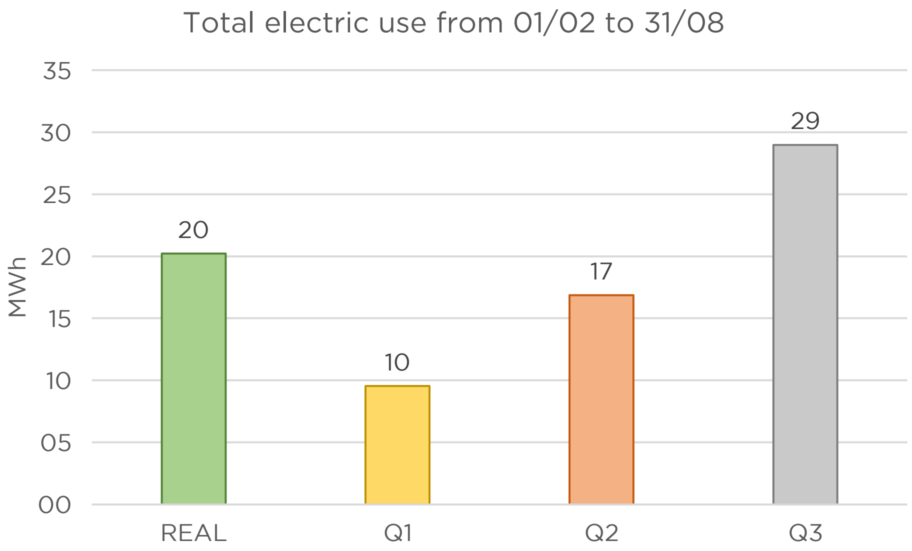


Figure 6.23: Comparison between the total sum of electric use from 01/02 to 31/08 in the building for the three scenarios and the real registered data

Figure 6.24 represents the total electric use for each flat. Clearly, the three modelled scenarios decrease the differences between the flats. It is quite evident for the zone 24, characterized by a very high electric use. However, on average they are quite representative. The graph of the averages between the zones, in Figure 6.25, shows that the span of the real registered data is really high compared to the ones of the three scenarios. In the scenarios, the range of the total electric load in the different zones is almost constant.

These results are expected and related to the averages between proto-clusters in each cluster. The unique patterns (very low or high) are reduced, together with the noise that characterized the real registered data.

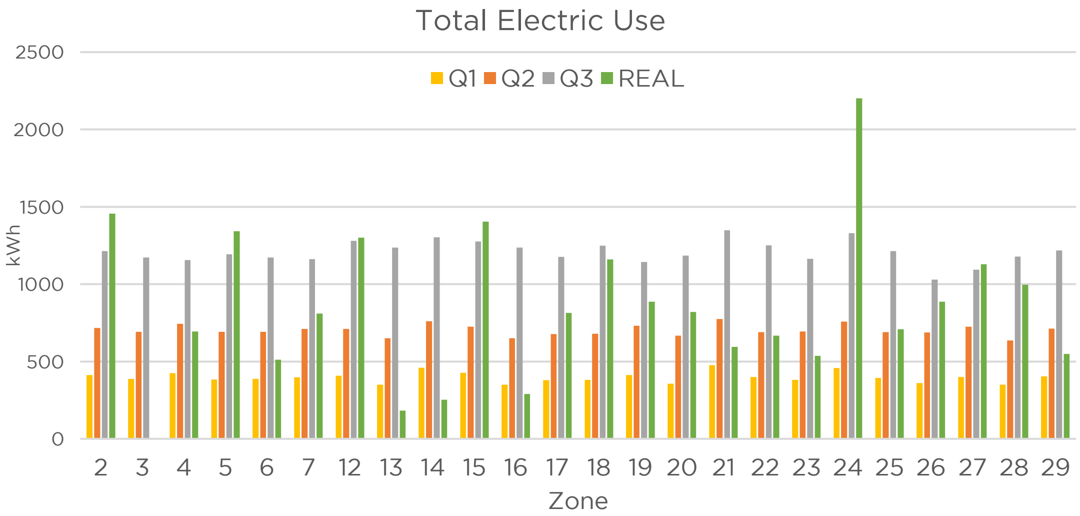


Figure 6.24: Comparison between the monthly sum of electric use for each flat for the three scenarios and the real registered data

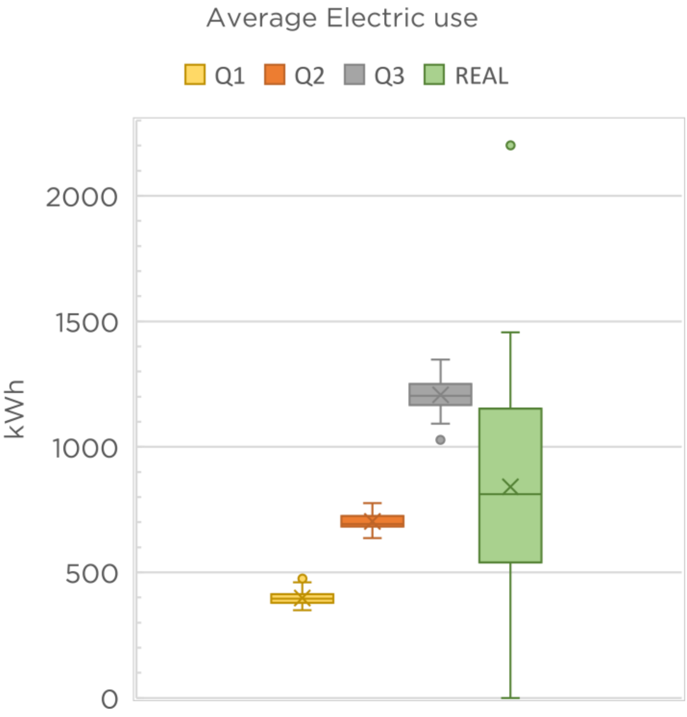


Figure 6.25: Comparison between the average electric use from 01/02 to 31/08 in the building for the three scenarios and the real registered data

6.1.5. Classification

The KNN is applied with a cubic calculation of the distances with 10 neighbourhoods. As predictors, the difference between Working day or Non-Working day, the difference between Heating and Cooling season, the average daily external temperature and its variation were used, as explained in the paragraph 5.2.4.

Application

The application of the KNN gives a cluster of each day of the year, as shown in Figure 6.26. The graph shows, as an example, the result for the zone 2. To create the yearly hourly schedule, then, each daily cluster is substituted with the relative 24 hours patter, as shown in Figure 6.27. The Figure 6.28 shows the representativeness of each cluster for the zone 2. Each zone shows different percentages, resulting in a unique yearly schedule for three different scenarios. The introduction of the predictors related to the temperatures give a temporal trend that is clearly visible in Figure 6.26, in which the winter months are represented mainly by the cluster 3 and 4, whilst the summer periods by the clusters 1, 2 and 5.

All the results for the other zones can be found in Section III of Annex II.

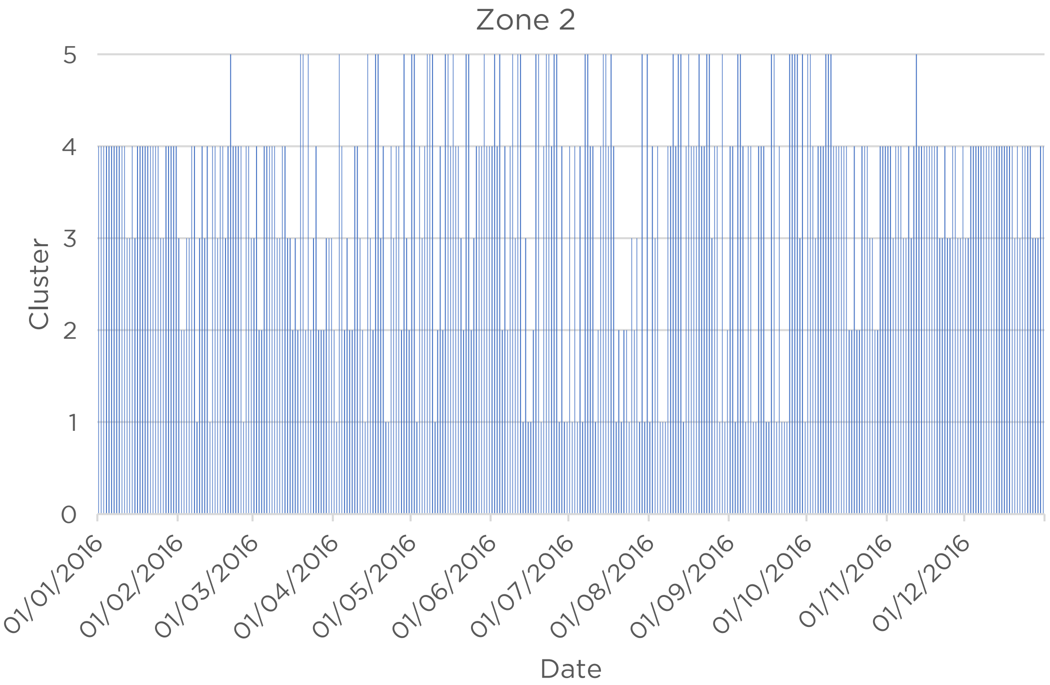


Figure 6.26: Application of the KNN algorithm to assess the yearly schedule

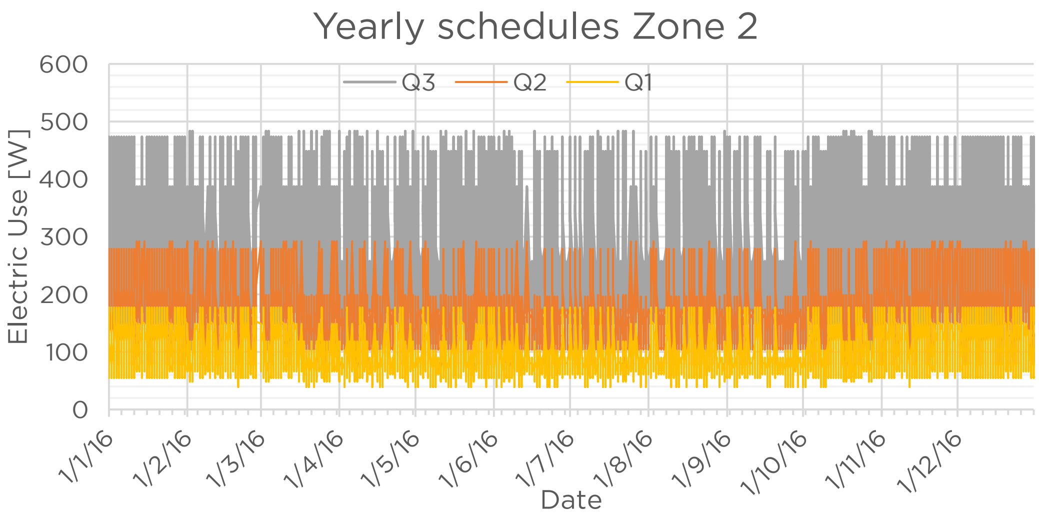


Figure 6.27: Final yearly electric schedule of scenarios Q2 of Zone 2

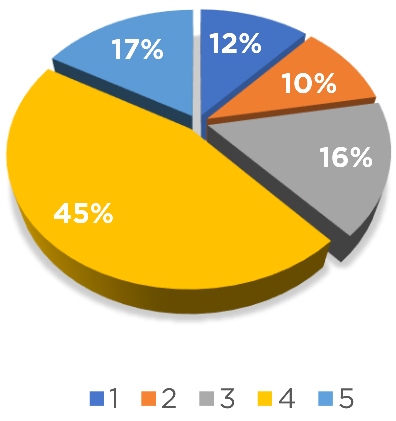


Figure 6.28: Clusters' representativeness in the Zone 2

6.2. Task 2

6.2.1. Data processing

Data processing in this task is composed just of cleaning and transformation. Thus, all the hours in which the data is not sufficiently accurate to detect the 15-minutes step registration might not be part of the sample. In almost all the flats February is registered as hourly value and for this reason it is not exploitable for this task. Two exceptions are Zone 4 and 22. The transformation is the association to each hour of the features: average, minimum and maximum, standard deviation and SAD, according to Kleiminger et al. [51].

6.2.2. Classification

The *k*-Nearest Neighbours algorithm was used because it has been evaluated as the one with the highest average accuracy in the work of 2013 of Kleiminger et al. [51] and in the update of the research of 2015 [52].

The accuracy also, in this case, is high (with a KNN run with the cubic calculation of distances and with 10 neighbours), with an average of 86 % within the zones, as can be seen in Figure 6.29.

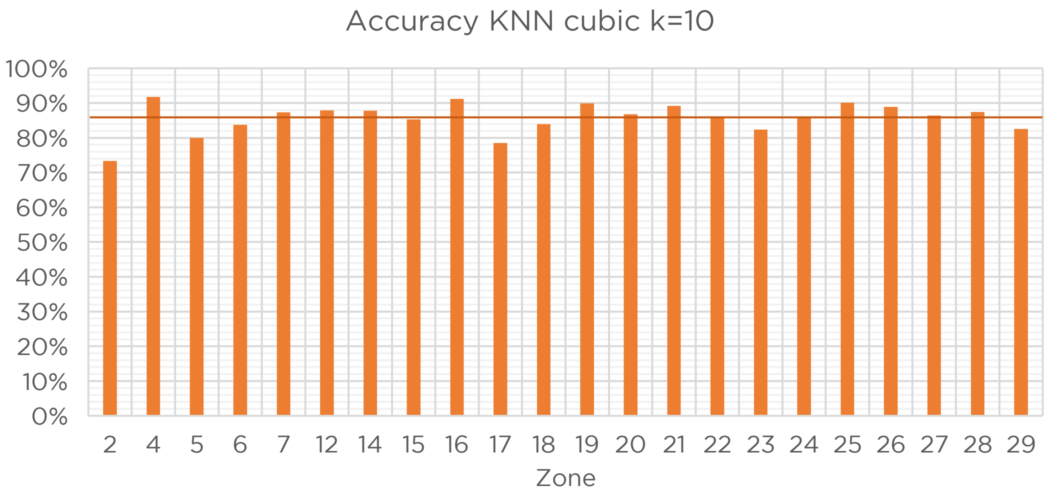


Figure 6.29: Accuracy for the used KNN algorithm for each zone

The KNN is run and then the result is applied to the rest of the data sample. The result is the assignment to each hour of the dichotomous variable 0 or 1, corresponding to 0 % or 100 % of occupancy.

6.2.3. Prediction

To actually predict the occupancy, the clusters of the previous task were taken and applied to the occupancy daily patterns of the classification outcome. The average of this outcome is performed in each cluster. The result is a continuous variable (average value), between 0 and 1, corresponding to the probability of occupancy in a specific hour of the day for each cluster. Then, when the average is below 0,33 the corresponding probability will be 0% when is between 0,33 and 0,66 will be 50% and, when above 0,66 will correspond to 100%. The results of each cluster are shown in Figures 6.30-6.34. In these graphs, the green line corresponds to the continuous variable resulting from the average of the occupancy within the same cluster, the blue line is the final discrete occupancy probability.

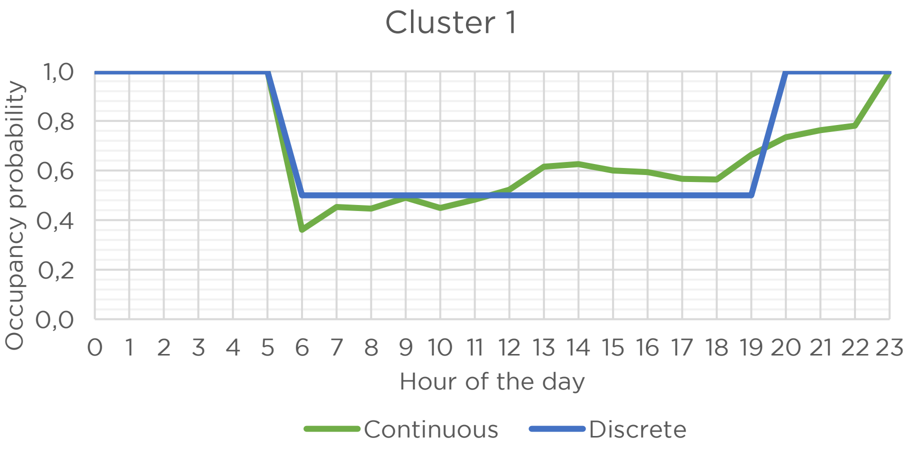


Figure 6.30: Probability of occupancy in Cluster 1

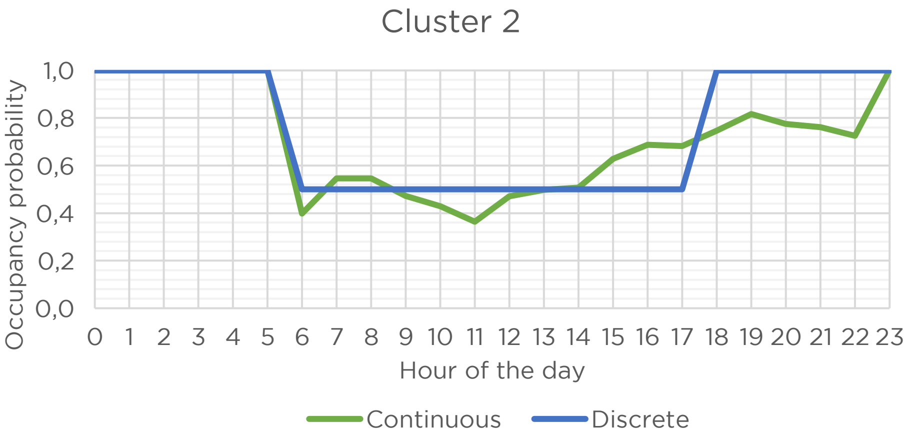


Figure 6.31: Probability of occupancy in Cluster 2

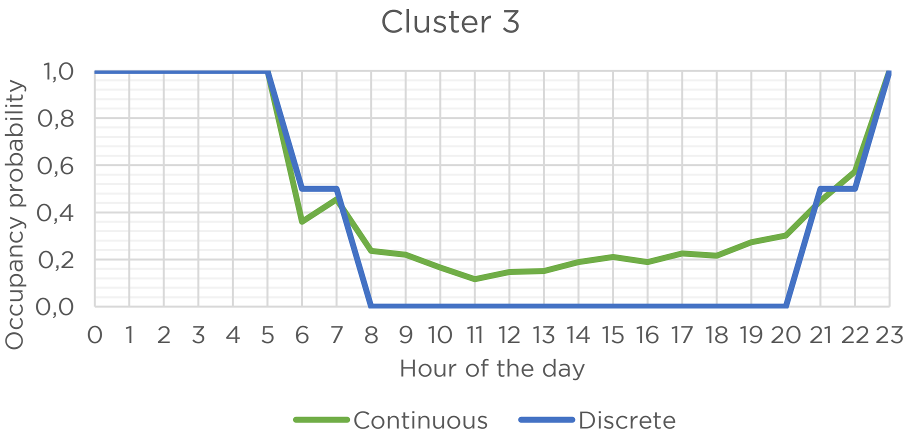


Figure 6.32: Probability of occupancy in Cluster 3

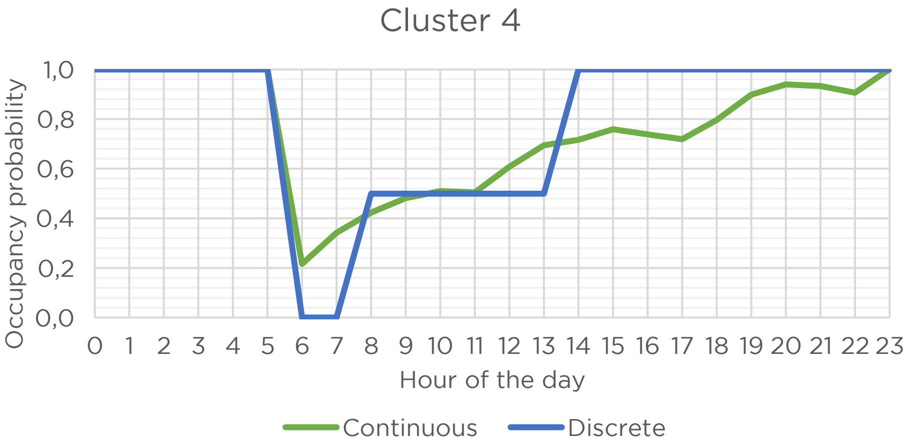


Figure 6.33: Probability of occupancy in Cluster 4

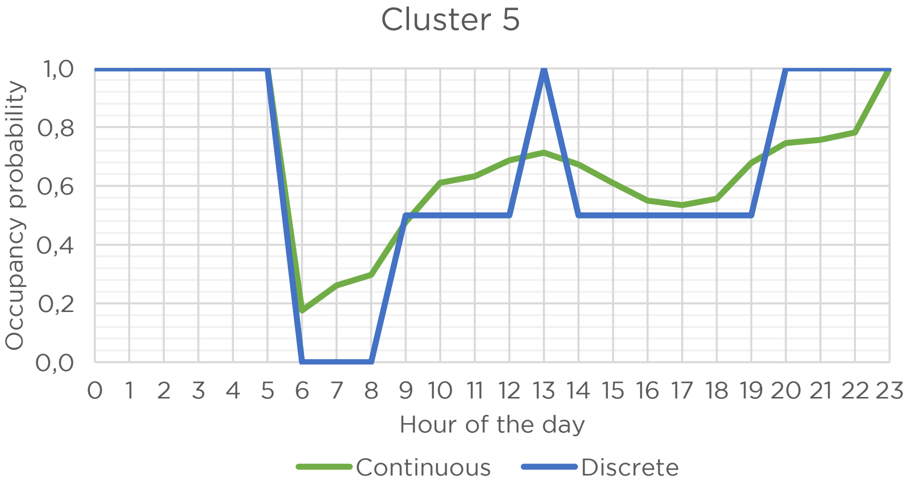


Figure 6.34: Probability of occupancy in Cluster 5

The same procedure is followed for the other two quartiles to create three scenarios of presence. The quartile 3 represents families whose components spend at home more time, whilst, the quartile 1 represents families that spend more time outside.

The result is shown in Figures 6.35-6.6.39. In these graphs the blue line represents the discrete occupancy probability given by the average, the red line represents the third quartile and the dark blue line represents the first quartile.

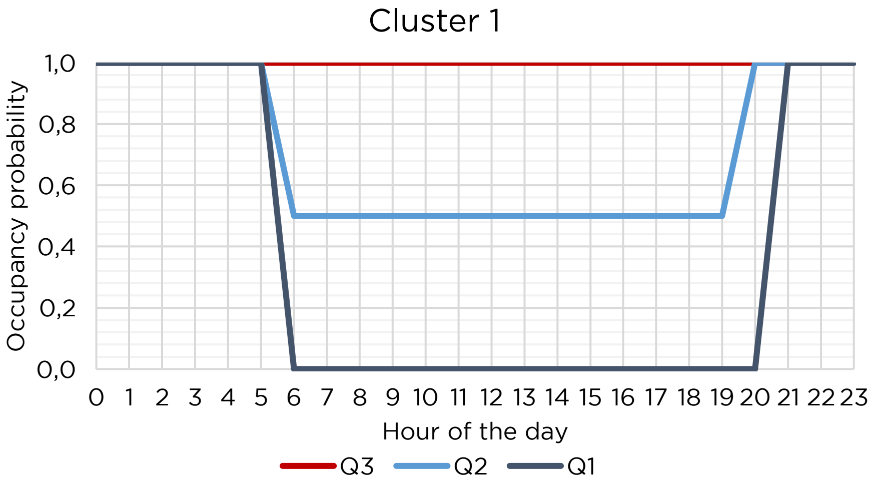


Figure 6.35: Three scenarios of the probability of occupancy in Cluster 1

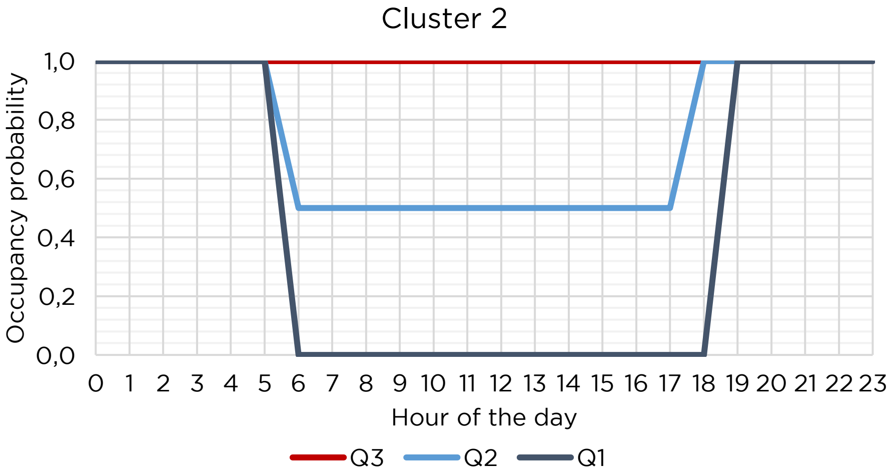


Figure 6.36: Three scenarios of the probability of occupancy in Cluster 2

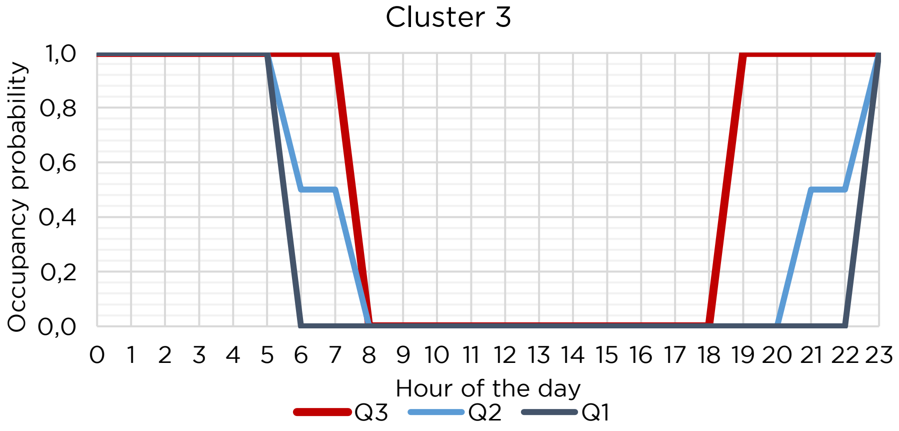


Figure 6.37: Three scenarios of the probability of occupancy in Cluster 3

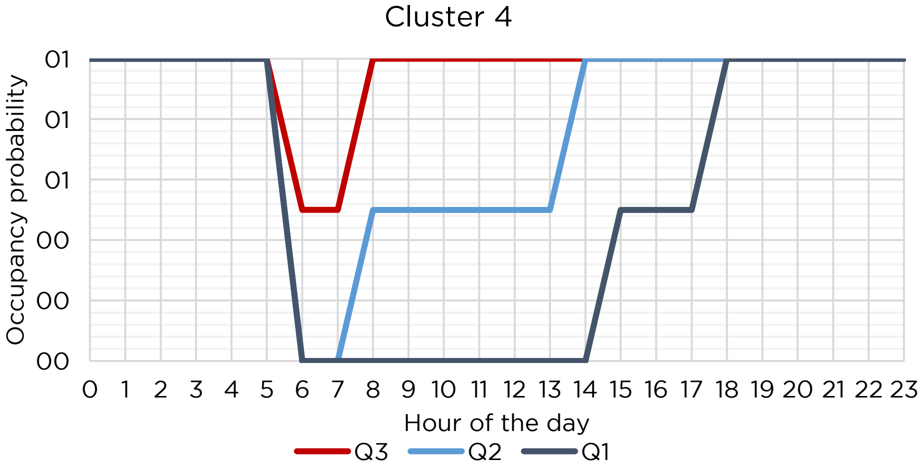


Figure 6.38: Three scenarios of the probability of occupancy in Cluster 4

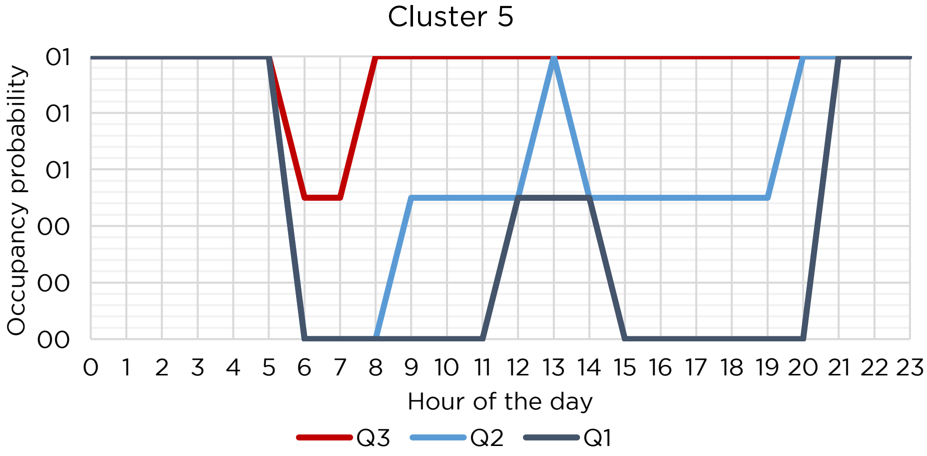


Figure 6.39: Three scenarios of the probability of occupancy in Cluster 5

6.2.4. Conclusion

The conclusion is an occupancy probability directly related with each cluster, thus it is generalized to all year following the results of the previous task shown in Figure 6.26. The three final scenarios with the probability of occupancy are summarized in figures 6.40-6.44 for each cluster. In the graphs, the daily electric use loads rely on the first axis (on the left), whilst the average occupancy probability relies on the secondary axis (on the right).

The final probability is then multiplied by the number of people that are supposed to live in the apartment on the base of the area of the bedrooms.

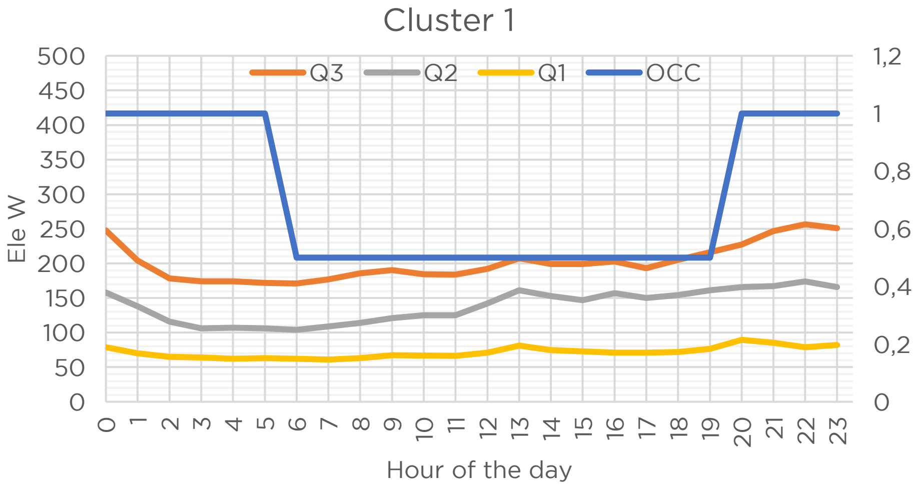


Figure 6.40: Final three scenarios and relative occupancy probability of Cluster 1

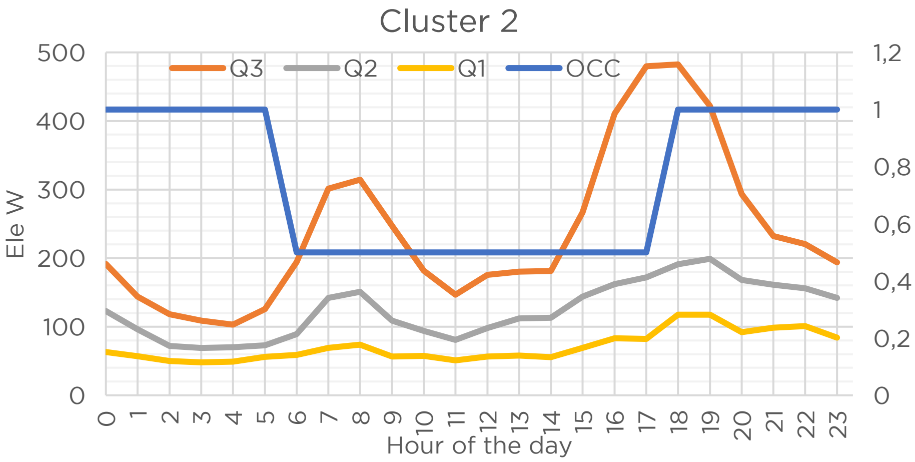


Figure 6.41: Final three scenarios and relative occupancy probability of Cluster 2

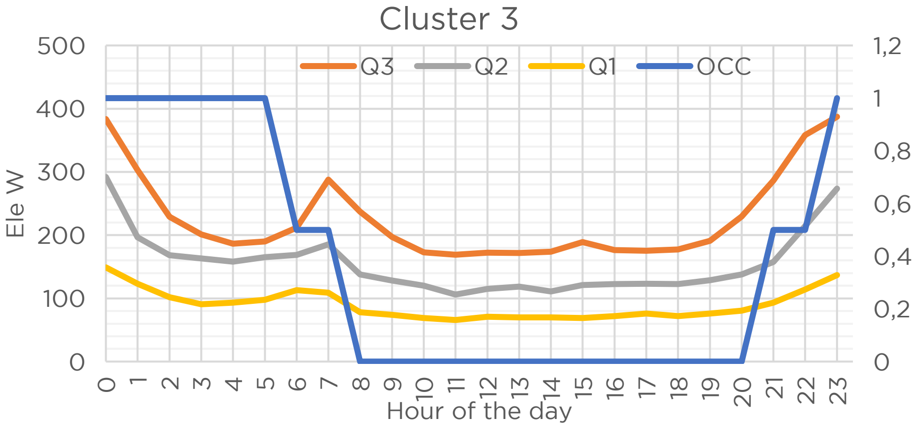


Figure 6.42: Final three scenarios and relative occupancy probability of Cluster 3

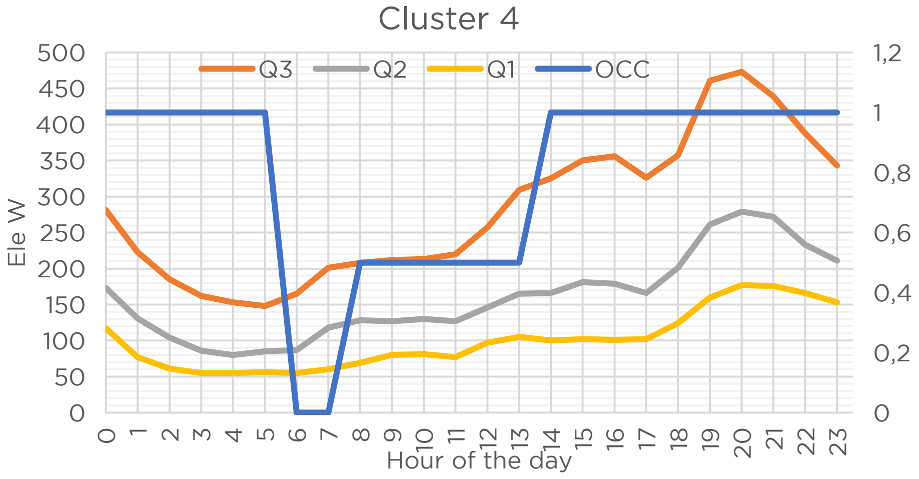


Figure 6.43: Final three scenarios and relative occupancy probability of Cluster 4

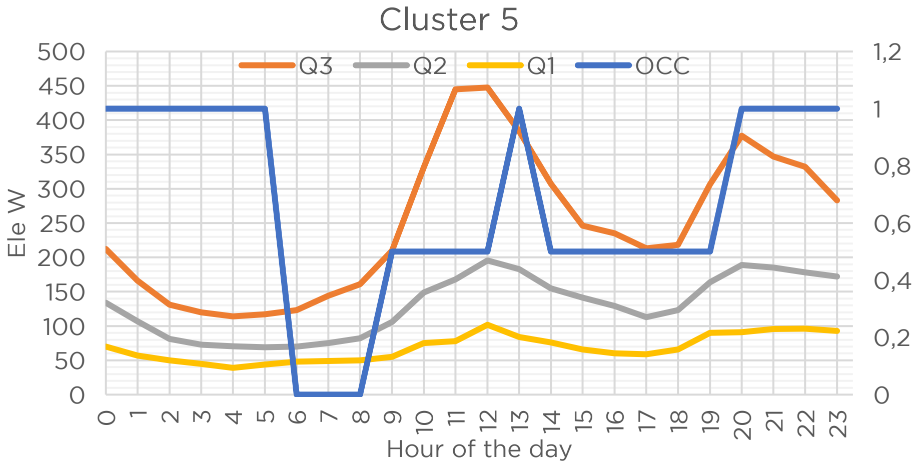


Figure 6.44: Final three scenarios and relative occupancy probability of Cluster 5

6.3. Task 3

The aim of the final analyses is to assess the impact of the generated schedules on the yearly energy need.

To achieve this goal, the first analysis is run varying the three scenarios of electric demand but using the fixed average presence profile. The analyses show 3 scenarios: energy-aware user, standard user and energy-intensive user. In the first case, the schedule generated with the first quartile is assigned to all the households and it corresponds to the situation of lowest internal gains due to lighting and appliances. For this reason, an increase of the energy need is expected. In the second case, the average schedule is assigned. The third case corresponds to the assignment of high-consumers' schedule to all the households in the building, resulting in the increase of the internal gains. In this last case, a decrease in the energy need to maintain the thermal comfort is expected.

The result of this analysis is shown in Figure 6.45.

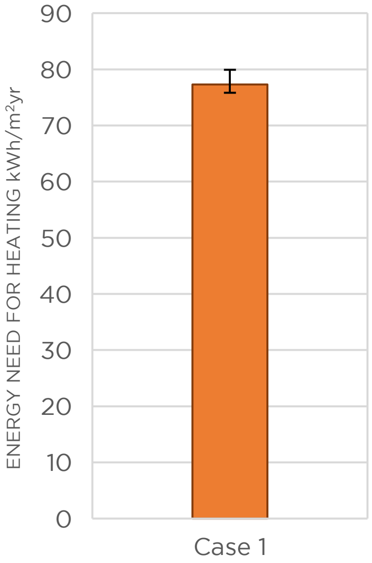


Figure 6.45: Result of the Case 1 analysis

The average value of energy need for heating is 77,30 kWh/m^{2}yr. The range given by the two extreme scenarios goes from a maximum of 78,78 kWh/m^{2}yr with the high-consumers' schedules to a minimum of 74,69 kWh/m^{2}yr with the conservative users' schedules. This variability corresponds to +3 % and -2 %

around the average. As expected the increase of the internal gains brought to a decrease in the energy need for heating and vice versa with the decrease of the internal gains. Considering different internal gains for appliances and lighting, the final energy need for heating can vary by 5 % around the average scenario.

To understand the impact of also the presence of the people in the thermal zones, a second case analysis is run. The idea is the same, run an average case with the standard users' schedule, and the two extremes cases, lowest and highest internal gains. In this case, in addition to the variation of internal gains due to appliances and lights, there is also the variation of the internal gains due to people and, thus, an increase in the variability of the results is expected. The outcome of this second case is shown in Figure 6.46.

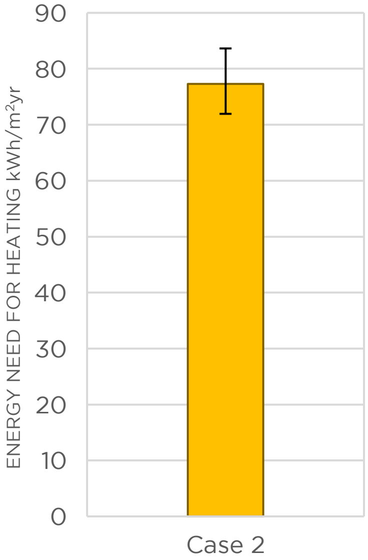


Figure 6.46: Result of the Case 2 analysis

The average value of energy need for heating is the same as before, corresponding to 77,30 kWh/m^{2}yr. The range given by the two extreme scenarios is increased and it goes from a maximum of 82,65 kWh/m^{2}yr with the high-consumers' schedules and high presence, to a minimum of 70,96 kWh/m^{2}yr with the conservative users' schedules and low presence. This variability corresponds to +8 % and -7 % of the average. As expected the further increase of the internal gains brought to a further decrease in the energy need for heating and vice versa. Considering different internal gains for appliances/lighting and presence, the final energy need for heating can vary by 15 %.

In this thesis, the natural ventilation is not modelled in detail, but it can heavily affect the results. To understand which the effects can be changing the setting of the natural ventilation, two analyses are run with the minimum value of 0,05 air change/hour as infiltration (ideally set to zero the natural ventilation). The two cases are set as the Case 1 and Case 2, thus with a variation of the internal gains due to appliances and a variation in the presence of people. The outcomes of these two analyses are shown in Figure 6.47.

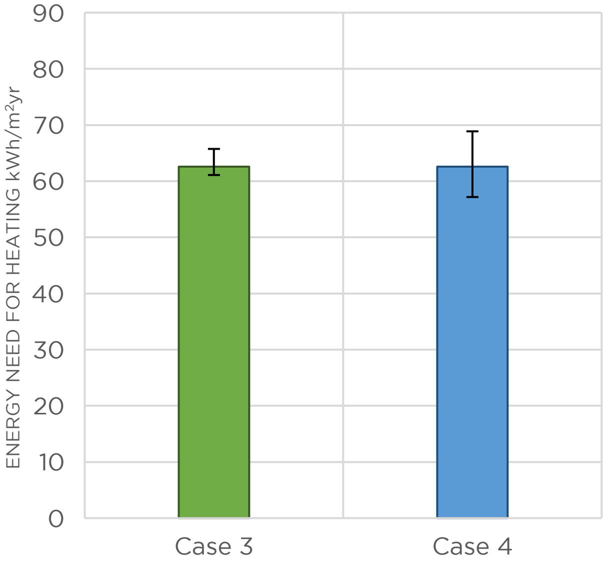


Figure 6.47: Result of the Case 3 and Case 4

The average of these analyses is decreased to 65,58 kWh/m^{2}yr, 21 % of the cases with a higher natural ventilation. The range of Case 3 (with the only variation of the electric schedules) is within 64,07 kWh/m^{2}yr and 59,42 kWh/m^{2}yr, corresponding to +5 % and -2 %. Instead, the range in Case 4 (considering also the variation of the people) goes from a maximum of 67,99 kWh/m^{2}yr to a minimum of 56,29 kWh/m^{2}yr, corresponding to +10 % and -9 %. The variations are similar to the cases with a higher natural ventilation, but a slight increase is registered on the percentage variation in both cases.

The aim of this task was also to compare the obtained result with the registered energy need of 2016. The supply company provided the delivered energy for heating, thus an estimated global seasonal efficiency of 0,7 was applied to calculate the energy need for heating.

The registered data is not far from the modelled value (Figure 6.48), indicating that the overall modelling of the building is able to approximate satisfactorily the result, always considering that the energy modelling implies numerous variables. However, the followed methodology looks promising and with implementation and improvements can really become an asset in the field.

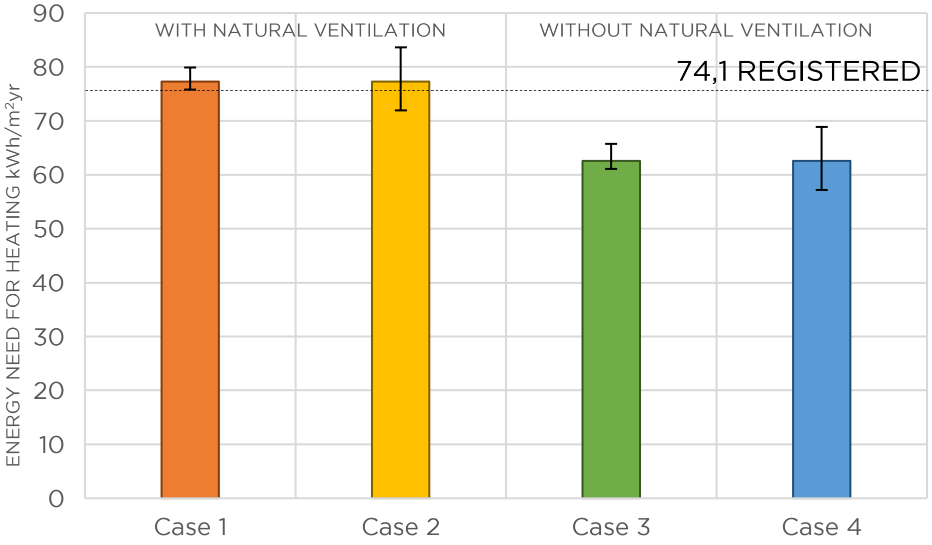


Figure 6.48: Result of four Cases compared to the registered data

7. Conclusion and Future outlooks

In this thesis, a comprehensive data-driven approach for the assessment of the impact of occupant behaviour is proposed. The methodology is focused on the improvement of the generation of schedules related to occupancy behaviour in energy modelling of residential buildings. The presented methods are applied and validated through a case study regarding a residential building block in Milan, Italy.

In the first phase, a clustering methodology for creating five representative electricity daily load profiles for the residential sector in Italy is presented. The implementation of machine learning techniques was found to be appropriate for the nature of the data sample and its complexity. In particular, the Self Organizing Map, a Neural Network technique, is coupled with the k-means algorithm, a classic machine learning method. Moreover, to extend the results to the whole year, the k-Nearest Neighbours algorithm is implemented after a validation through comparison with the real registered data. Five clusters emerged with different daily profiles, that can be ascribed to different types of families and habits.

In the second phase, a detection method is proposed to estimate the presence of occupants in the household. The technique does not rely on sensors, but it is based on the analysis of the electricity consumption data. To achieve the goal, also in this phase, machine learning techniques were implemented. In particular, the k-Nearest Neighbours algorithm for the detection of the occupancy was used. The extension to the whole year relies on the clusters obtained in the previous task. This phase, however, depends on an approximation for the creation of the ground truth from which the KNN learnt. For this reason, an at least partial survey on occupancy presence, could be helpful to validate or improve the methodology. The resulted presences are associated with the five daily profiles assessed in the previous step and they retrace the different habits of the families.

In the third phase, the schedule generated in the previous steps are used to assess the impact of the occupancy behaviour on the heating energy need during a year. The modelled result is compared with the real registered data. The range of results for heating can vary around 7 % changing the internal gains due to electric appliances and around 15 % changing also the presence of people. The uncertainty of 20 % due to different natural ventilation schedules implemented in the analyses is also of great interest.

7.1. Applications

The methodology may be a useful tool for several applications.

The method used in task 1, for the investigation of the electricity load profiles, may be valuable as an efficient load profiling analysis in residential buildings, a peculiar case in terms of noise in the data sample, complexity of the variable and privacy issues. In the residential sector a vast amount of raw data available thanks to the smart meters, will need to be processed, to obtain in-depth and useful information of the electricity behaviour. This knowledge can be exploited by different actors, such as:

- managers whose aim is to develop strategies for energy savings due to good management of renewable systems,
- distribution system operators and transmission system operators can both exploit the identification of energy profiles for the management of the grid and of the markets,
- modellers who do not possess electricity loads for their residential buildings models,
- energy service companies involved in the building management that can exploit the information to optimize the energy savings measures,
- tenants, who can benefit from targeted tariff plans,
- policymakers who can benefit from the profiles characterization to optimize the actions.

The approach used in task 2 exploits the electricity consumption registration as an occupancy sensor. Modellers, without an available presence ground truth data, or a partial one, can apply the proposed methodology to create presence schedules. The first advantage of the approach is the fact that the privacy of the tenants is respected. Moreover, with building-size databases, this method is a simple and rapid implementation to obtain occupancy schedules.

The assessment of the occupants in building energy modelling, performed in task 3, is of major importance in the last years' researchers. Understand deeply how much the occupants and their habits can impact on the energy need of a building is crucial for high-performance buildings. As a matter of fact, the occupancy behaviour can change the result and, for this reason, the accomplishment or the failure of an energy target. Knowledge of the impact of occupants can be helpful for different operators, such as:

- policymaker, who can optimize the energy targets considering this uncertainty in a defined way,
- managers whose aim is to develop strategies for energy savings due to good management of resources,
- tenants, who can benefit from the knowledge of good and bad behaviours to decrease their expenses,
- modellers, who can evaluate better the impact of variables in their models.

7.2. Future works

In future, to improve the results of the second task, a more detailed measurement of the electricity might be helpful. A time step of 1 second could be optimal. In addition, a ground truth of the presence of people, at least partial, can be beneficial for the accuracy of the KNN algorithm.

From the results of the energy analyses, it is clear that the natural ventilation is a big uncertainty that can have a huge impact on the final energy consumption of the building. The methodology developed in the thesis for occupancy profile generation could be further extended to address natural ventilation too.

Annex I

In this annex, the main details of the EnergyPlus model are listed.

Constructions

Vertical Closure

Code **M2** Description **MAIN EXTERNAL WALL**

Thermal Resistance **0.82 m²K /W**

N	Layer	th [mm]	λ [W/mK]	R [m²K/W]	Vol Mass [kg/m³]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Lime and Gypsum Plaster	5.00	0.70	0.007	1400	1.00	11
2	Reinforced Concrete (1% steel)	105.00	2.30	0.046	2300	1.00	130
3	EPS 1985	40.00	0.07	0.571	25	1.00	100
4	Reinforced Concrete (1% steel)	55.00	2.30	0.024	2300	1.00	130
5	Lime and concrete Mortar	5.00	0.90	0.006	1800	1.00	23
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.1: Layers description of construction M2

Code **M7** Description **EXTERNAL WALL OF STAIRS**

Thermal Resistance **0.25 m²K /W**

N	Layer	th [mm]	λ [W/mK]	R [m²K/W]	Vol Mass [kg/m³]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Lime and Gypsum Plaster	5.00	0.70	0.007	1400	1.00	11
2	Reinforced Concrete (1% steel)	150.00	2.30	0.065	2300	1.00	130
3	Lime and concrete Mortar	5.00	0.90	0.006	1800	1.00	22
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.2: Layers description of construction M7

Code **M7b** Description **EXTERNAL WALL OF THE ATTIC**

Thermal Resistance **0.25 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Lime and Gypsum Plaster	15.00	0.70	0.021	1400	1.00	11
2	Reinforced Concrete (1% steel)	150.00	2.30	0.065	2300	1.00	130
3	Lime and concrete Mortar	15.00	0.90	0.017	1800	1.00	22
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.3: Layers description of construction M7b

Code **MS** Description **EXTERNAL WALL ON LOGGIA SIDES**

Thermal Resistance **1.81 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Lime and Gypsum Plaster	5.00	0.70	0.007	1400	1.00	11
2	Reinforced Concrete (1% steel)	150.00	2.30	0.065	2300	1.00	130
3	Lime and Cement Plaster	5.00	0.90	0.006	1800	1.00	22
4	Sintered expanded polystyrene	40.00	0.04	1.000	30	1.45	60
5	External Plastic Plaster	10.00	0.300	0.033	1300	0.84	30
6	Sintered expanded polystyrene	20.00	0.04	0.500	30	1.45	60
7	External Plastic Plaster	10.00	0.3	0.033	1300	0.84	30
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.4: Layers description of construction MS

Code **M31** Description **ROLLER BLIND CASE**

Thermal Resistance **1.07 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Lime and Gypsum Plaster	5.00	0.70	0.007	1400	1.00	10
2	Reinforced Concrete (1% steel)	55.00	2.30	0.024	2300	1.00	130
3	EPS 1985	40.00	0.07	0.571	25	1.00	100
4	Reinforced Concrete (1% steel)	55.00	2.30	0.024	2300	1.00	130
6	Internal Gap of not ventilated air	200.00	1.11	0.180	-	-	-
8	Expanded poliurethane	30.00	0.035	0.857	70	1.03	1
9	Glued wood Panels	13.00	0.14	0.093	500	1.70	30
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.5: Layers description of construction M31

Code **M130/M4** Description **EXTERNAL WALL ON LOGGIAS**

Thermal Resistance **0.92 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Lime and Gypsum Plaster	5.00	0.70	0.007	1400	1.00	11
2	Hollow Bricks	80.00	0.29	0.280	825	0.84	9
3	Not-ventilated Air gap Av<500 mm^2/m	50.00	0.28	0.180	-	-	-
4	Hollow Bricks	80.00	0.29	0.280	825	0.84	9
5	Cement and Lime Plaster	5.00	0.90	0.006	1800	1.00	22
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.6: Layers description of constructions M130 and M4

Vertical Partitions

Code **M6** Description **VERTICAL PARTITION BETWEEN FLATS**

Thermal Resistance **0.25 m²K /W**

N	Layer	th [mm]	λ [W/mK]	R [m²K/W]	Vol Mass [kg/m³]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Lime and Gypsum Plaster	5.00	0.70	0.007	1400	1.00	11
2	Reinforced Concrete (1% steel)	150.00	2.30	0.065	2300	1.00	130
3	Lime and concrete Mortar	5.00	0.90	0.006	1800	1.00	22
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.7: Layers description of construction M6

Code **M8** Description **INTERNAL PARTITION ON STAIR CASES**

Thermal Resistance **0.82 m²K /W**

N	Layer	th [mm]	λ [W/mK]	R [m²K/W]	Vol Mass [kg/m³]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Lime and Gypsum Plaster	5.00	0.70	0.007	1400	1.00	10
2	Reinforced Concrete (1% steel)	105.00	2.30	0.046	2300	1.00	130
3	EPS 1985	40.00	0.07	0.571	25	1.00	100
4	Reinforced Concrete (1% steel)	55.00	2.30	0.024	2300	1.00	130
5	Lime and concrete Mortar	5.00	0.90	0.006	1800	1.00	23
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.8: Layers description of construction M8

Code **MP** Description **INTERNAL PARTITION INSIDE FLATS**

Thermal Resistance **0.82 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Lime and concrete Mortar	15.00	0.90	0.017	1400	1.00	23
2	Reinforced Concrete (1% steel)	55.00	2.30	0.024	2300	1.00	130
3	EPS 1985	40.00	0.07	0.571	25	1.00	100
4	Reinforced Concrete (1% steel)	55.00	2.30	0.024	2300	1.00	130
5	Lime and concrete Mortar	15.00	0.90	0.017	1800	1.00	23
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.9: Layers description of construction M8

Horizontal Closures

Code **P1** Description **EXTERNAL FLOOR ON GROUND**

Thermal Resistance **0.57 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.170	-	-	-
1	Ceramic Tiles	10.00	1.30	0.008	2300	0.84	9999999
2	Screed in Concrete	50.00	1.49	0.034	2200	0.88	70
3	Deck in bricks	220.00	0.72	0.306	1800	0.84	9
4	Lime and concrete Mortar	10.00	0.90	0.011	1800	1.00	23
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.10: Layers description of construction P1

Code **P2** Description **EXTERNAL FLOOR ON LOGGIAS**

Thermal Resistance **0.37 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.170	-	-	-
1	Ceramic Tiles	20.00	1.30	0.015	2300	0.84	9999999
2	Screed in Lean Concrete	50.00	0.90	0.056	1800	0.88	30
3	Screed in Reinforced Concrete	180.00	2.15	0.084	2400	0.88	100
4	Lime and concrete Mortar	5.00	0.90	0.006	1800	1.00	22
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.11: Layers description of construction P2

Code **S2** Description **ROOF ON THE ATTIC**

Thermal Resistance **0.25 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	External Superficial Resistance	-	-	0.040	-	-	-
1	Screed in Reinforced Cocrete	180.00	2.15	0.084	2400	0.88	100
2	Gypsum and Sand Plaster	20.00	0.80	0.025	1600	1.00	10
-	Internal Superficial Resistance	-	-	0.100	-	-	-

Table AI.12: Layers description of construction S2

Horizontal Partitions

Code **P3** Description **INTERNAL FLOOR**

Thermal Resistance **0.36 m²K /W**

N	Layer	th [mm]	λ [W/mK]	R [m²K/W]	Vol Mass [kg/m³]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.170	-	-	-
1	Ceramic Tiles	20.00	1.30	0.015	2300	0.84	9999999
2	Screed in Lean Concrete	50.00	0.90	0.056	1800	0.88	30
3	Screed in Reinforced Concrete	180.00	2.15	0.084	2400	0.88	100
-	External Superficial Resistance	-	-	0.040	-	-	-

Table AI.13: Layers description of construction P3

Thermal Bridges

Code	Thermal Resistance m²K/W	Thermal Transmittance W/m²K	Description
PT1	0.650	1.54	Main external Thermal Bridge applied on the façades between floors (M2, M7)
PT2	0.410	2.44	External Thermal Bridge applied on the loggias (MS, M130)
PT3	0.296	3.38	Internal Thermal Bridge with the stairs applied between floors (M6, M8, M8S)
PT4	0.650	1.54	External Thermal Bridge applied on the façade at the last floor
PT5	0.650	1.54	External Thermal Bridge applied on the loggias at the last floor
PT6	0.830	1.20	External Thermal Bridge applied on the stairs at the last floor

Table AI.14: Thermal bridges description

Glazed Units

Glass 4+4.2/16/3+3.2	U-Factor W/m^{2}K	3
	SHGC	0.75
	Visible Transmittance	0.82
Frame and Divider	Conductance W/m^{2}K	60
	Width m	0.075
	Solar Absorptance	0.7
	Divider Width m	0.15
	Divider Cond W/m^{2}K	60

Table AI.15: Glass, Frame and divider description

Code	General Description	Number of Sashes	Number of Dividers	Width m	Height m	A_w total m^2	A_g glass m^2	A_f frame m^2
W1	Window used in the bedrooms	2	1	1.45	1.35	1.96	1.37	0.59
W2	Window used rarely instead of W1	3	2	2.10	1.35	2.84	1.76	1.08
W3	Window used in the bathrooms	1	0	0.80	1.35	1.08	0.78	0.30
W5	Window used in the kitchens	1	0	0.80	2.30	1.84	1.40	0.44
W5n	Window used on the stair cases	1	0	0.80	1.35	1.08	0.78	0.30
W6	French window used on loggias	2	1	1.45	2.30	3.34	2.60	0.73
WD	Entrance door at ground level	1	0	~ 2.00	~ 2.6	~ 5.20	~ 4.50	~ 0.70
WI	Window used at the ground floor	1	0	0.80	0.65	0.52	0.33	0.20
WL	Window used in the attic	1	0	0.57	3.15	1.80	1.26	0.54

Table AI.16: Windows description

Shading

Schedule	Nocturnal 10 p.m.-6 a.m.
Solar Transmittance	0,05
Solar Reflectance	0,5
Visible Transmittanve	0,05
Visible Reflectance	0,5
Infrared Hemispherical Emissivity	0,9
Infrared Transmittance	0,05
Thickness	0,015 m
Conductivity	0,1 W/mK
Shade to Glass Distance	0,34 m
Airflow Permeability	0,05

Table AI.17: Shading Setting

Doors

Code **M30** Description **DOOR**

Thermal Resistance **0.55 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	HD Wood Fibers Panels	9.00	0.160	0.056	900	1.70	72
2	Internal Gap of not ventilated air	15.70	-	0.180	-	-	-
3	HD Wood Fibers Panels	9.00	0.160	0.056	900	1.70	72
-	Internal Superficial Resistance	-	-	0.130	-	-	-

Table AI.18: Layers description of construction M30

Code **M29** Description **DOOR AT GROUND LEVEL**

Thermal Resistance **1.51 m^{2}K /W**

N	Layer	th [mm]	λ [W/mK]	R [m^{2}K/W]	Vol Mass [kg/m^3]	c [kJ/kgK]	Vap Res [-]
-	Internal Superficial Resistance	-	-	0.130	-	-	-
1	Metallic Coating	-	-	0.001	-	-	-
2	Mineral Wool	50.00	0.040	1.250	165	1.03	1
3	Metallic Coating	-	-	0.001	-	-	-

-	Internal Superficial Resistance	-	-	0.130	-	-	-

Table AI.19: Layers description of construction M29

General Settings

North axis	39°
Terrain	Suburbs
Solar Distribution	Full Exterior
Minimum number of warup days	25

The ground temperature on the building location is set considering the 2 m-depth temperature given by the weather data IGDG of Milano-Linate.

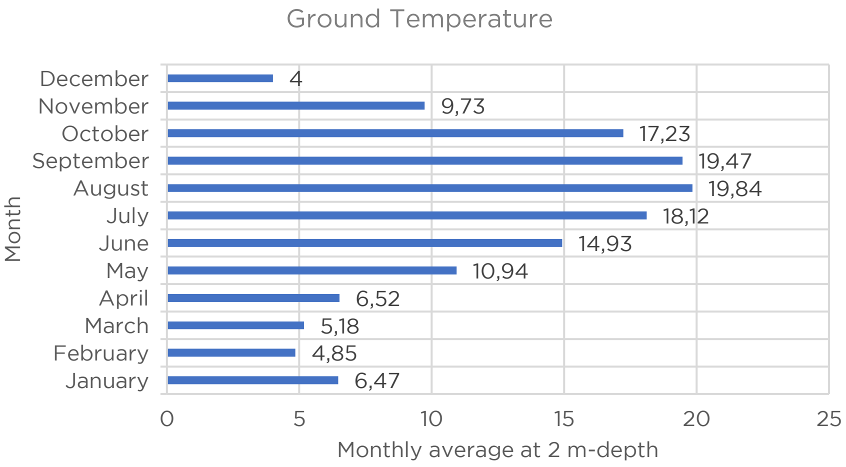


Figure AI.1: Set ground temperature in the EnergyPlus model

Internal gain from Electric Equipment

Fraction radiat	0,3
Fraction lost	0,5

Annex II

Section I

Descriptive statistics of the continuous variables

Cumulative Precipitation

Descriptives

		Statistic	Std. Error
Hourly Cumulative Precipitation mm	Mean	,1357	,00302
	Median	,0000	
	Variance	1,024	
	Std. Deviation	1,01187	
	Minimum	,00	
	Maximum	29,60	
	Range	29,60	
	Skewness	15,933	,007
	Kurtosis	337,462	,015

Table AII.1: Descriptive Statistics of hourly cumulative precipitation variable

Hourly average temperature

Descriptives

		Statistic	Std. Error
Hourly Average Temperature °C	Mean	18,0938	,02205
	Median	18,0000	
	Variance	54,421	
	Std. Deviation	7,37706	
	Minimum	1,60	
	Maximum	33,80	
	Range	32,20	
	Skewness	-,038	,007
	Kurtosis	-,975	,015

Table AII.2: Descriptive Statistics of hourly average temperature variable

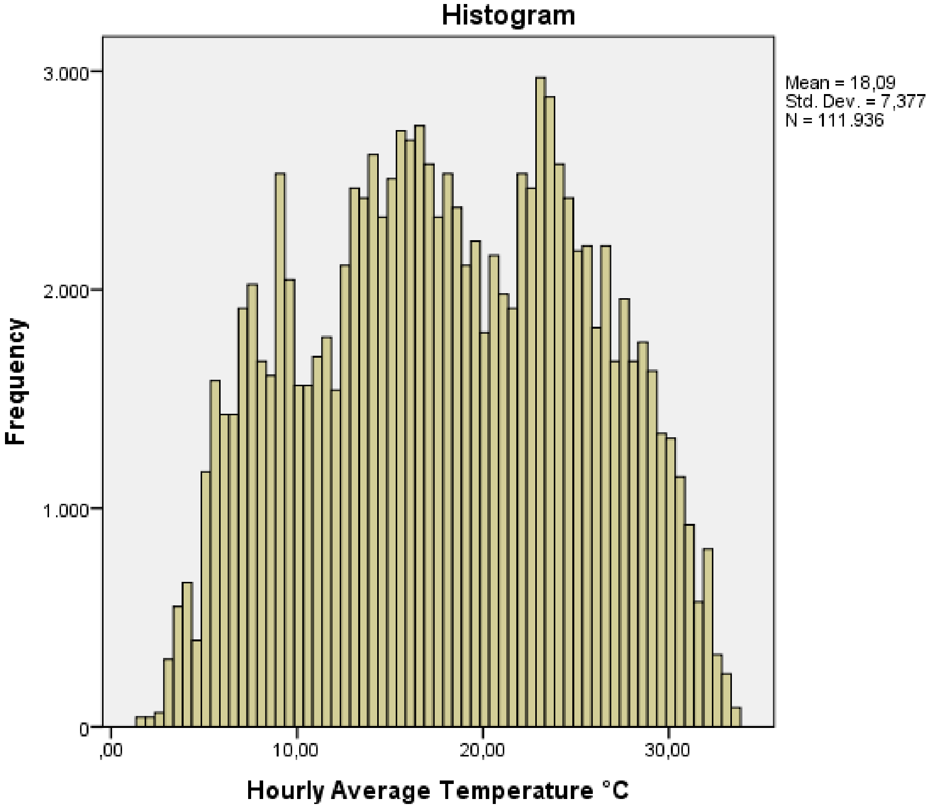


Figure AII.1: Histogram of the hourly average temperature variable

Global radiation

Descriptives

		Statistic	Std. Error
Global Radiation W/m2	Mean	198,4626	,80836
	Median	29,8500	
	Variance	73144,391	
	Std. Deviation	270,45220	
	Minimum	,00	
	Maximum	931,30	
	Range	931,30	
	Skewness	1,147	,007
	Kurtosis	-,111	,015

Table AII.3: Descriptive Statistics of global radiation variable

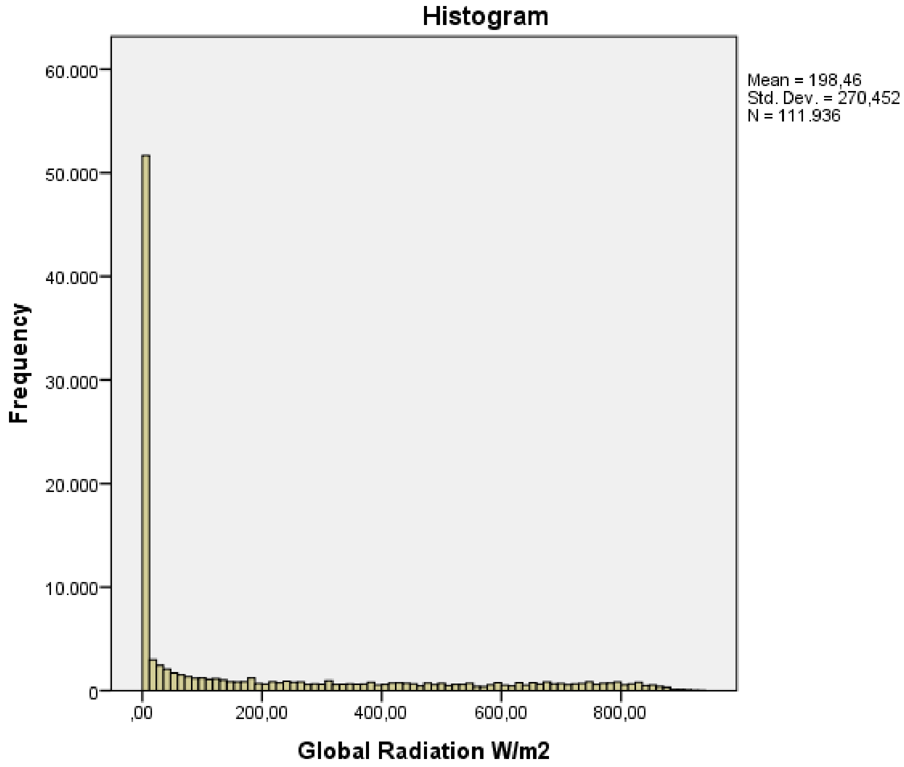


Figure AII.2: Histogram of the global radiation variable

Floor Area

Descriptives

		Statistic	Std. Error
Area	Mean	67,7968	,06229
	Median	71,3700	
	Variance	434,250	
	Std. Deviation	20,83865	
	Minimum	37,87	
	Maximum	95,28	
	Range	57,41	
	Skewness	-,168	,007
	Kurtosis	-1,209	,015

Table AII.4: Descriptive Statistics of floor area variable

Window/Floor ratio

Descriptives

		Statistic	Std. Error
Window/Floor Area	Mean	,142170	,0000688
	Median	,138511	
	Variance	,001	
	Std. Deviation	,0230334	
	Minimum	,1026	
	Maximum	,1956	
	Range	,0930	
	Skewness	,334	,007
	Kurtosis	-,584	,015

Table AII.5: Descriptive Statistics of window/floor ratio variable

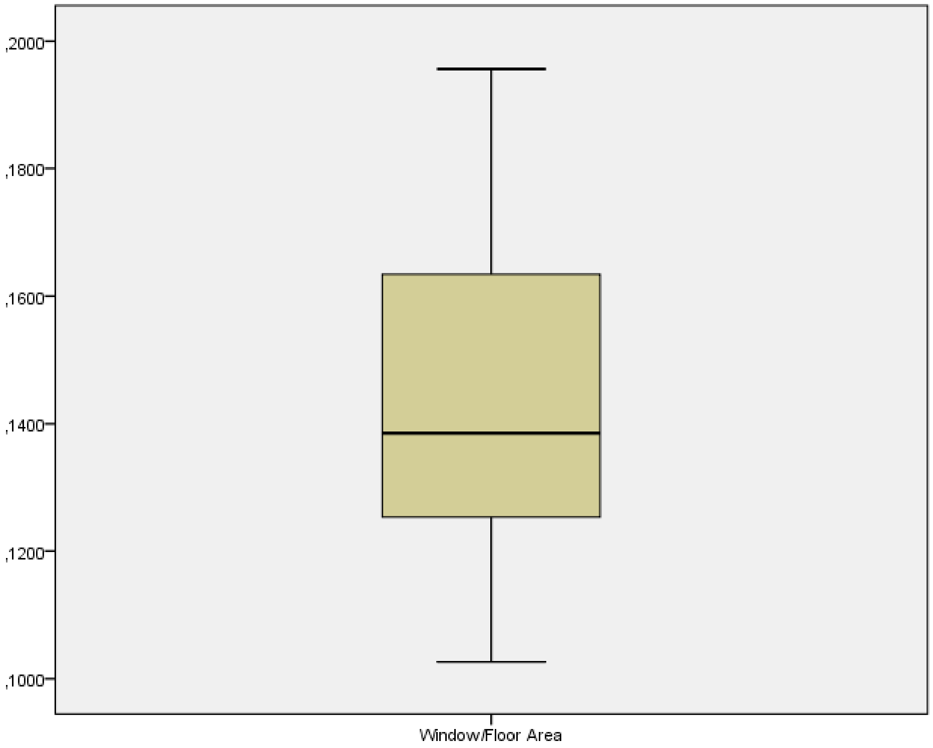


Figure AII.3: Box plot of the window/floor ratio variable

Section II

Boxplots between variables

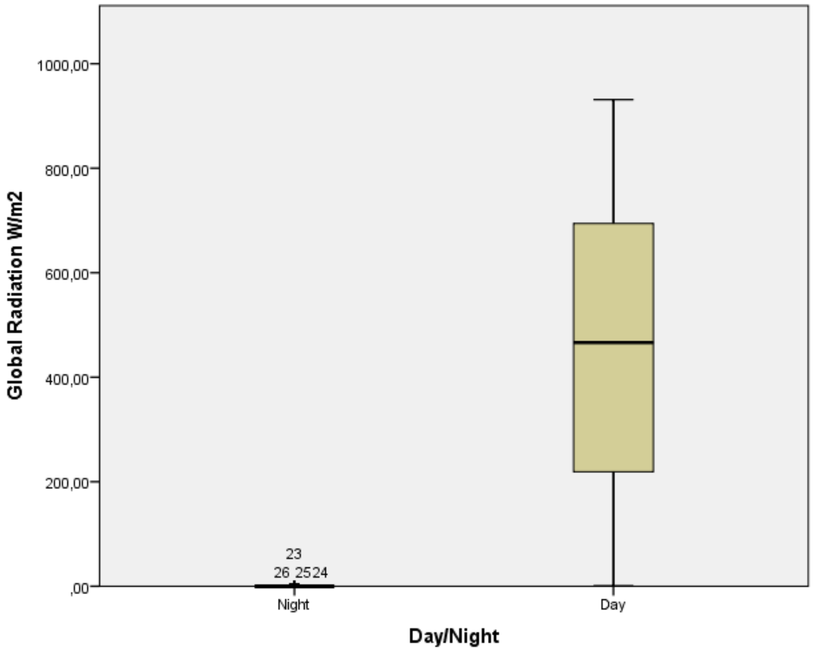


Figure AII.4: Box plot between Day/Night variable and the Global radiation

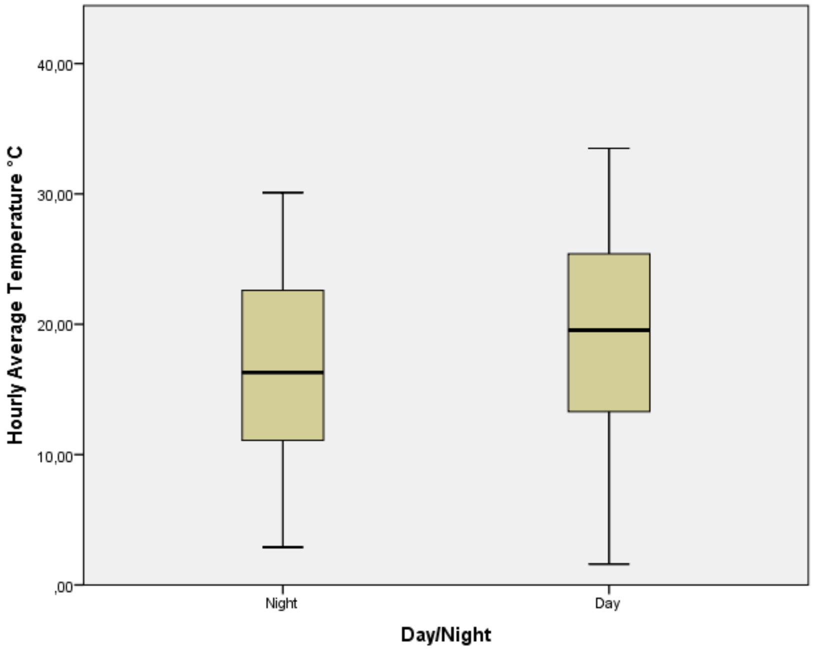


Figure AII.5: Box plot between Day/Night variable and the Hourly average temperature

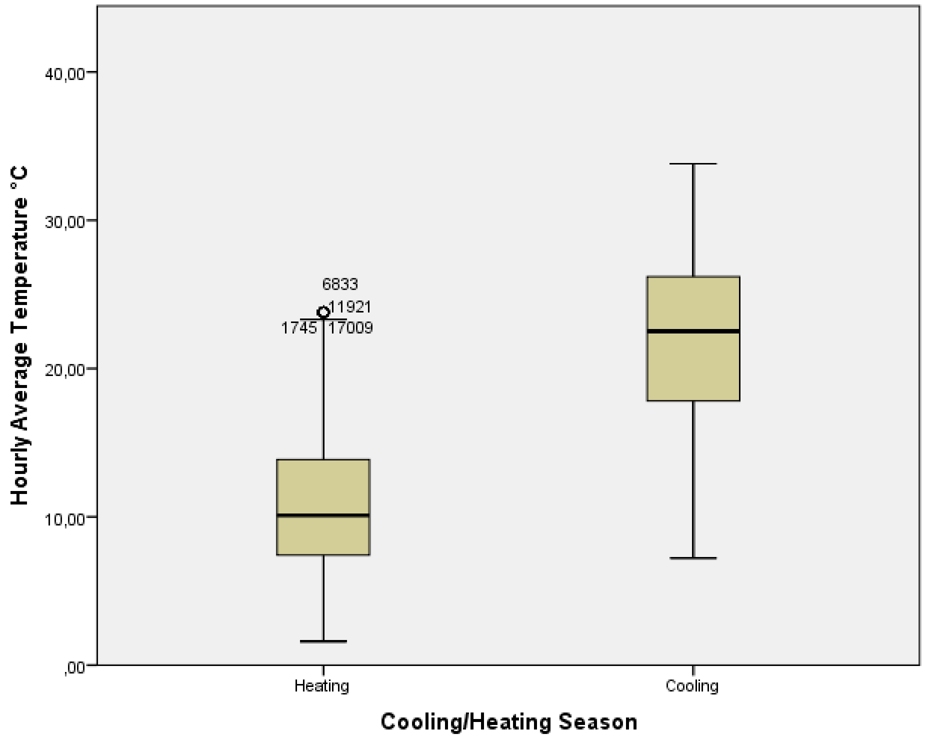


Figure AII.6: Box plot between Cooling/Heating season variable and the Average external temperature

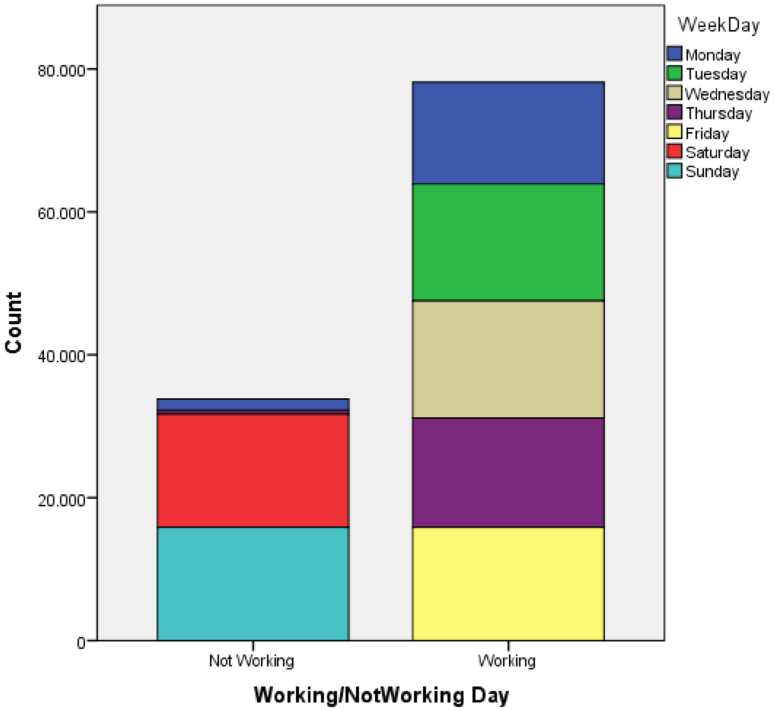


Figure AII.7: Frequency of the Weekdays within the Working or Not Working variable

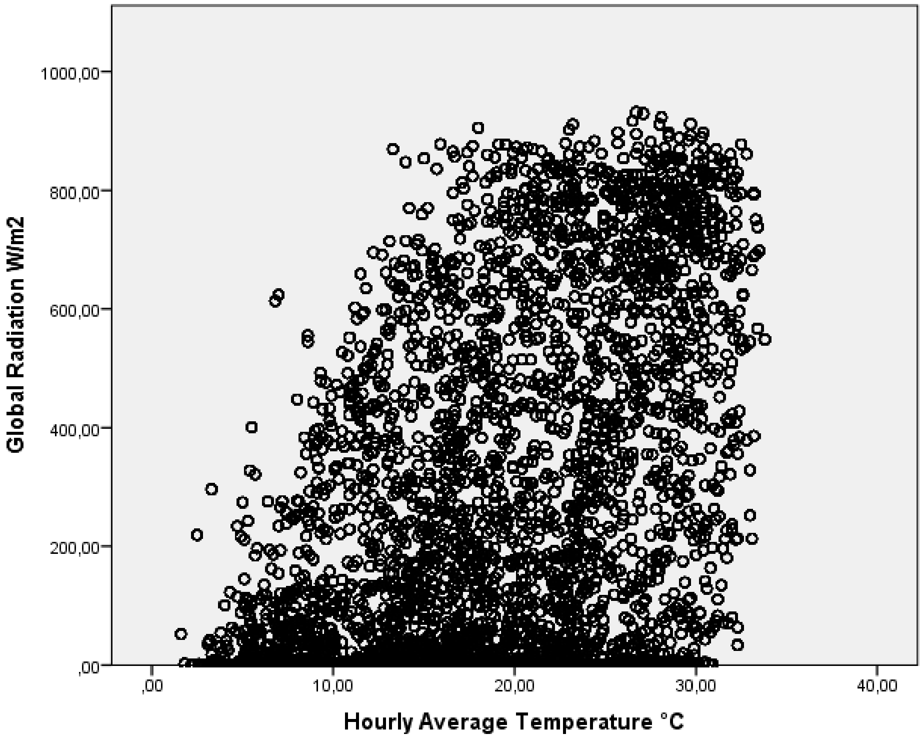


Figure AII.8: Scatter plot between Hourly average temperature and Global Radiation

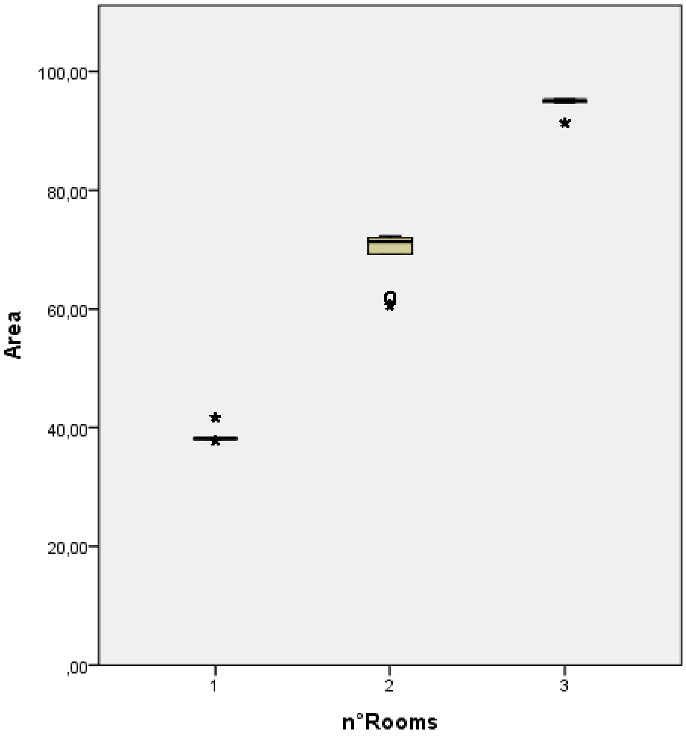


Figure AII.9: Box plot between the number of rooms and area of a flat.

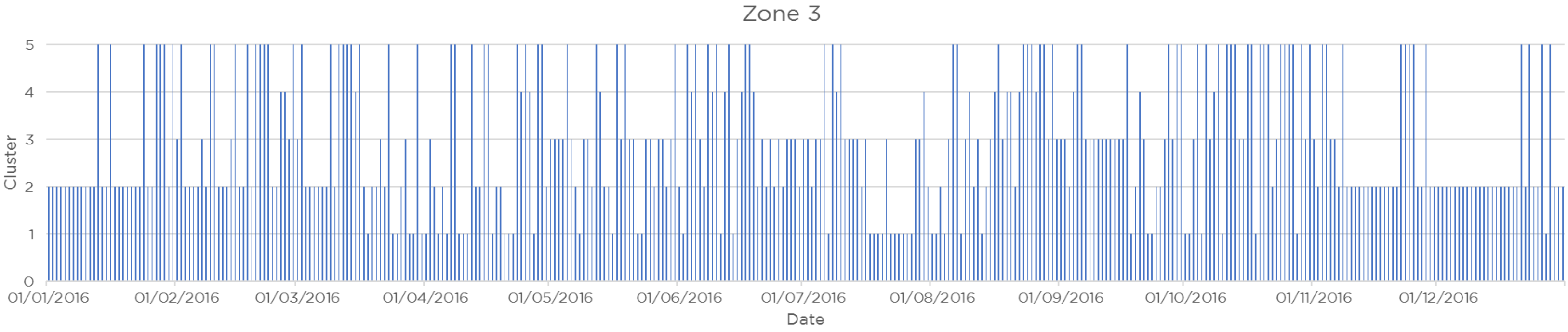


Figure AII.10: Final cluster assessment of Zone 3

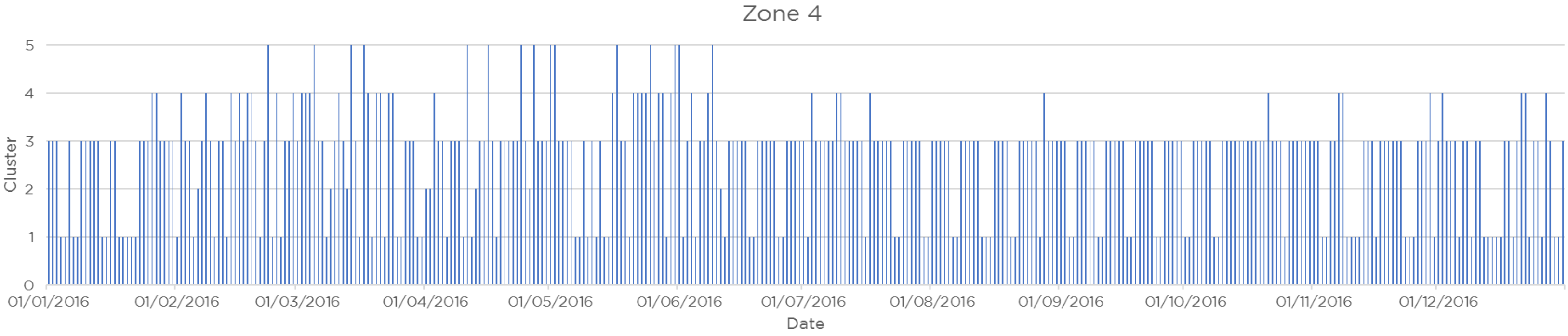


Figure AII.11: Final cluster assessment of Zone 4

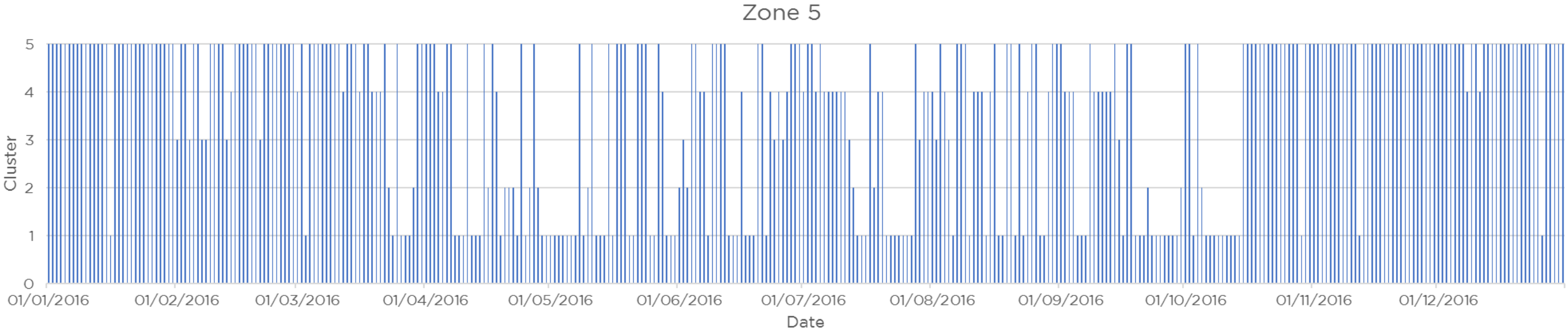


Figure AII.12: Final cluster assessment of Zone 5

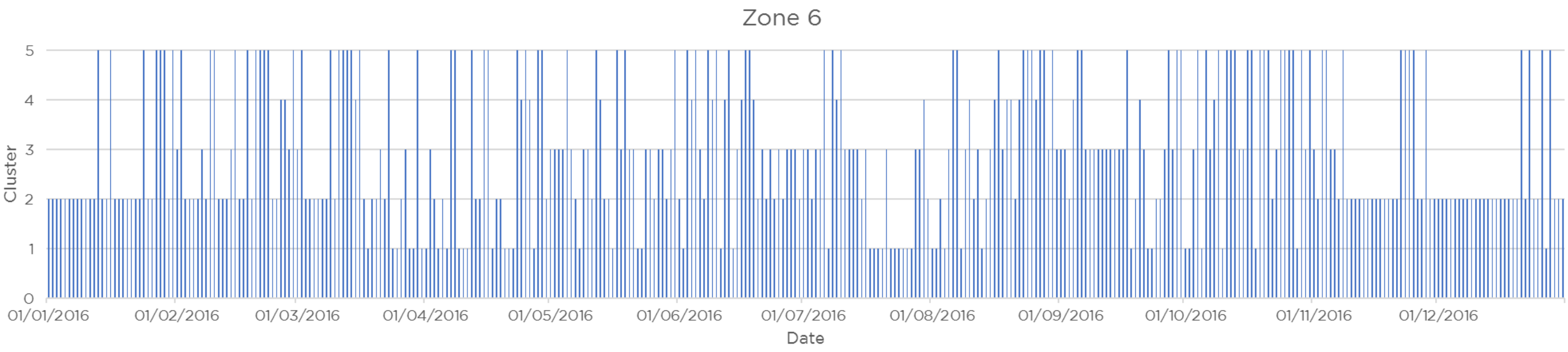


Figure AII.13: Final cluster assessment of Zone 6

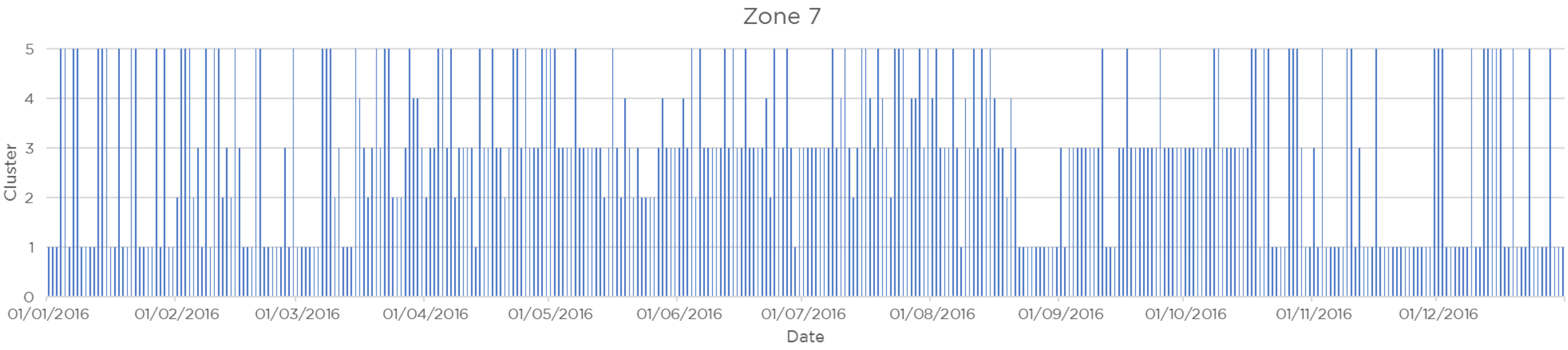


Figure AII.14: Final cluster assessment of Zone 7

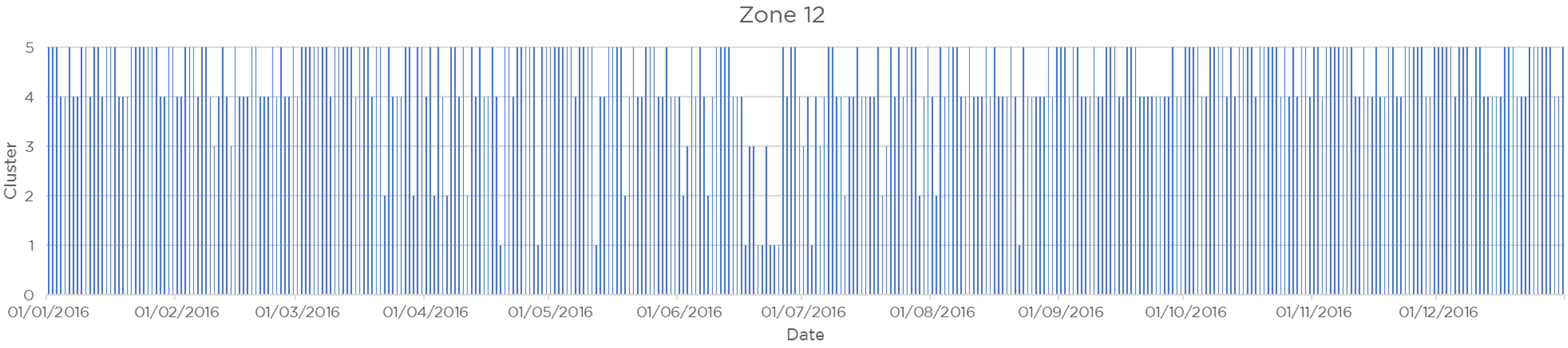


Figure AII.15: Final cluster assessment of Zone 12

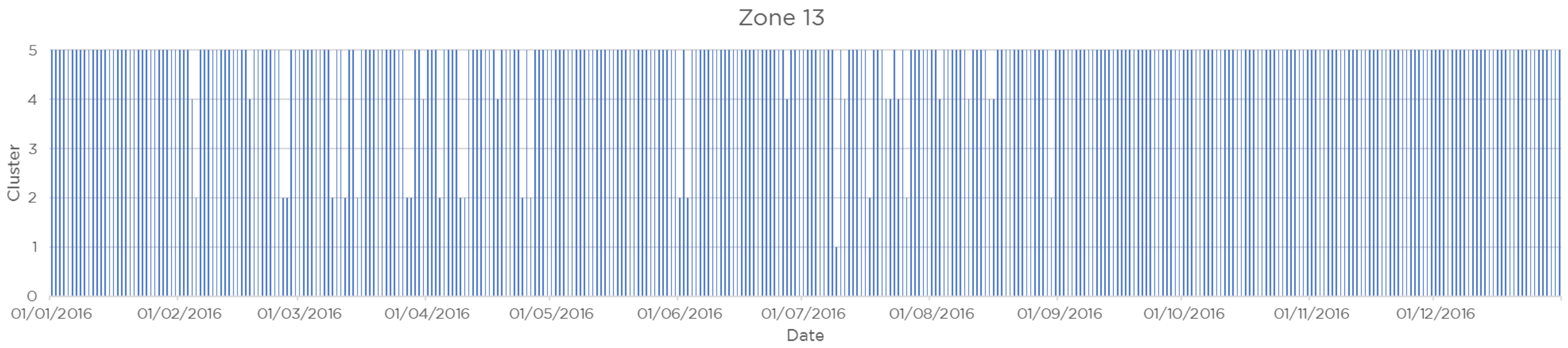


Figure AII.16: Final cluster assessment of Zone 13

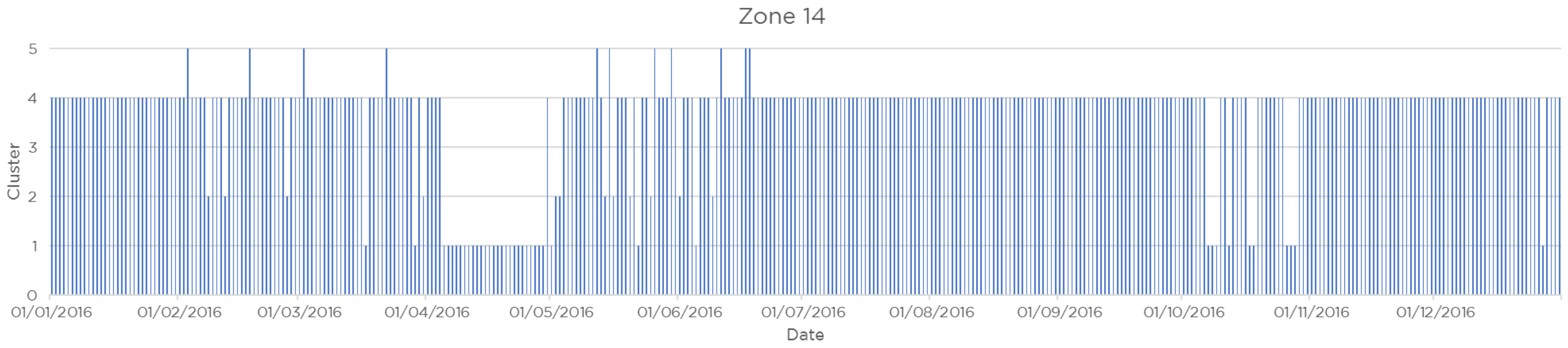


Figure AII.17: Final cluster assessment of Zone 14

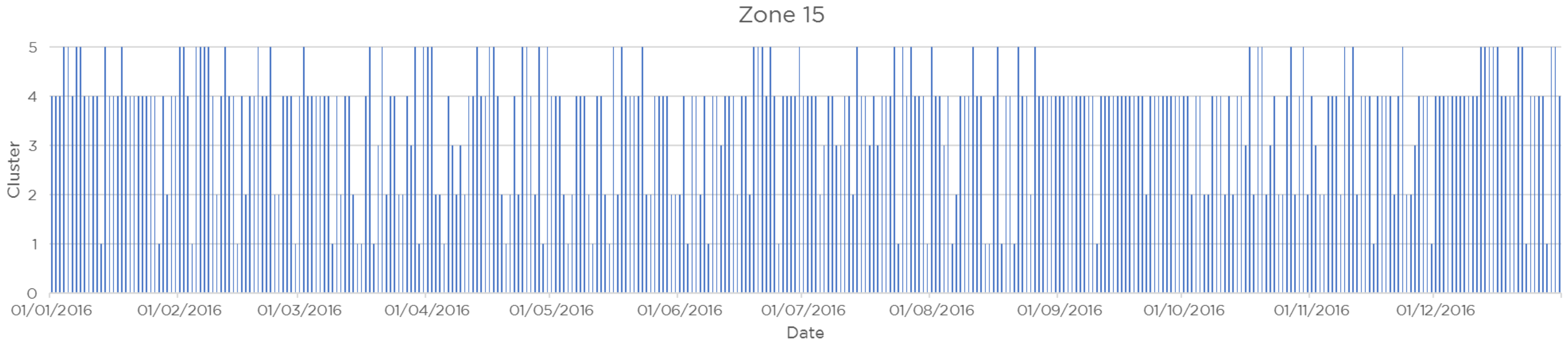


Figure AII.18: Final cluster assessment of Zone 15

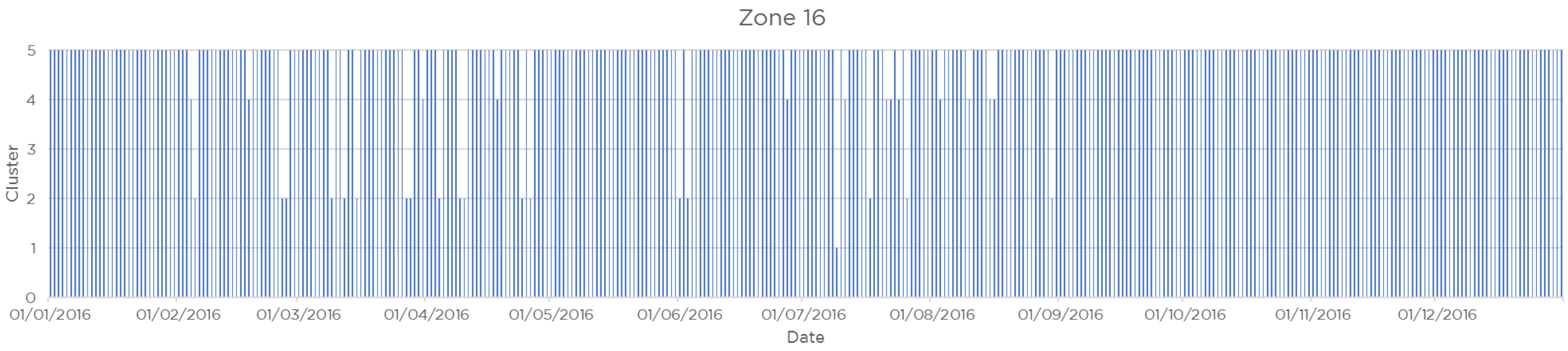


Figure AII.19: Final cluster assessment of Zone 16

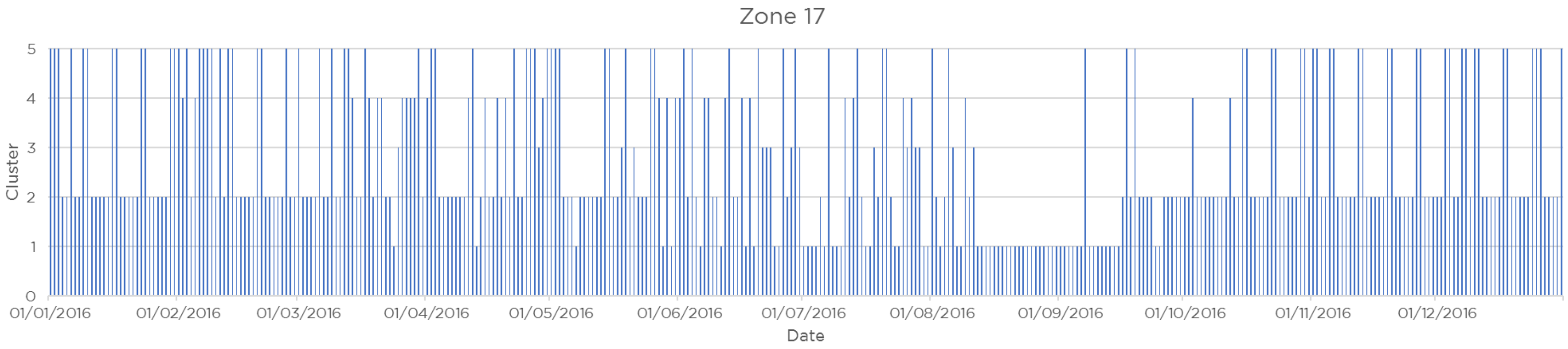


Figure AII.20: Final cluster assessment of Zone 17

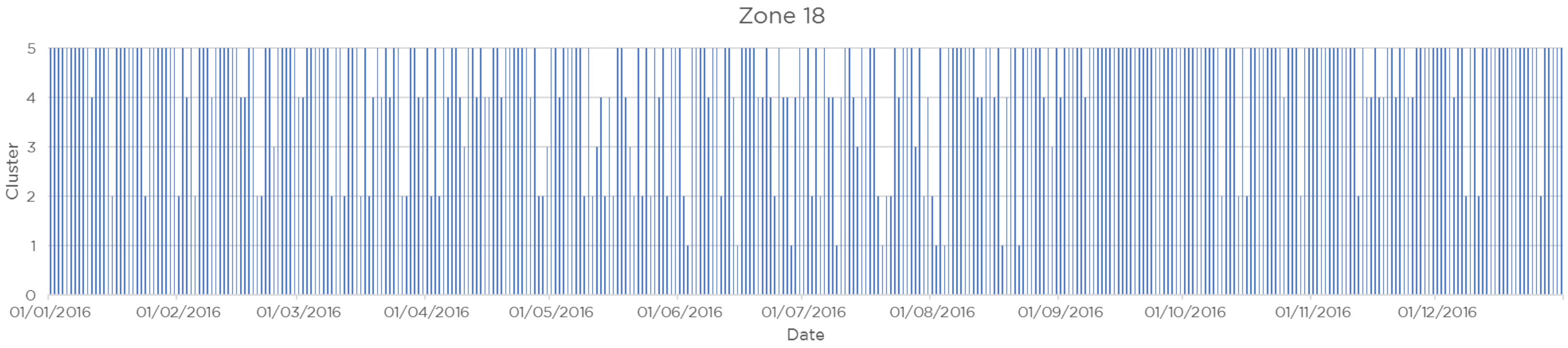


Figure AII.21: Final cluster assessment of Zone 18

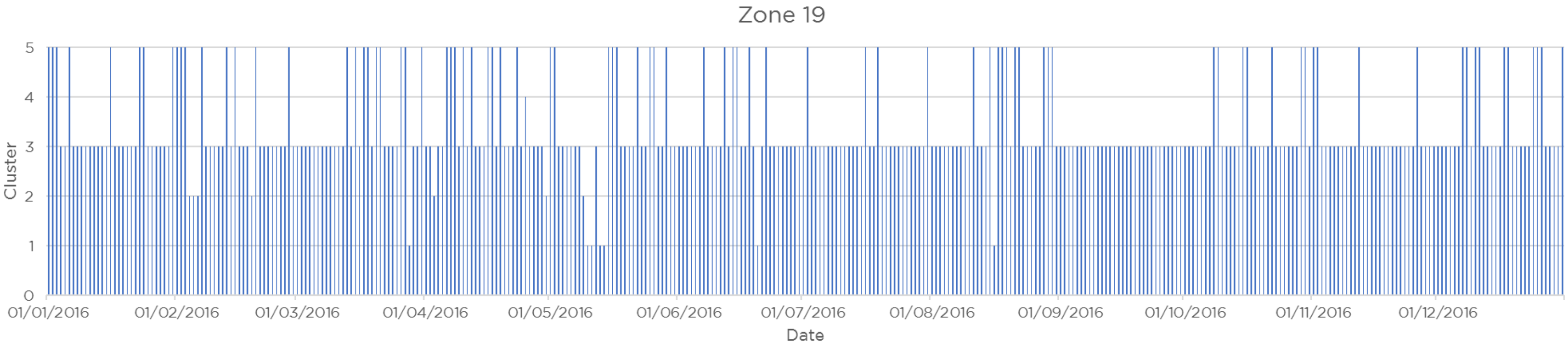


Figure AII.22: Final cluster assessment of Zone 19

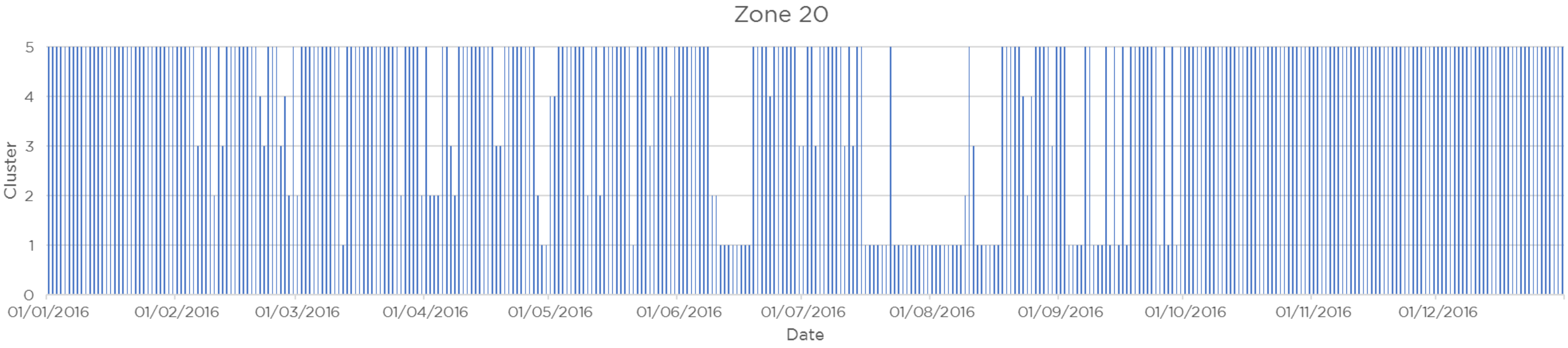


Figure AII.23: Final cluster assessment of Zone 20

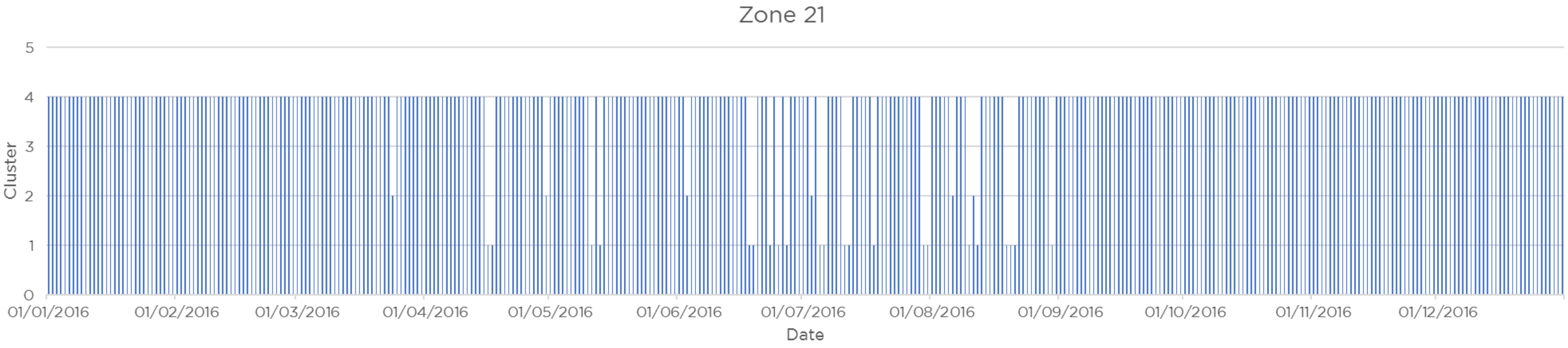


Figure AII.24: Final cluster assessment of Zone 21

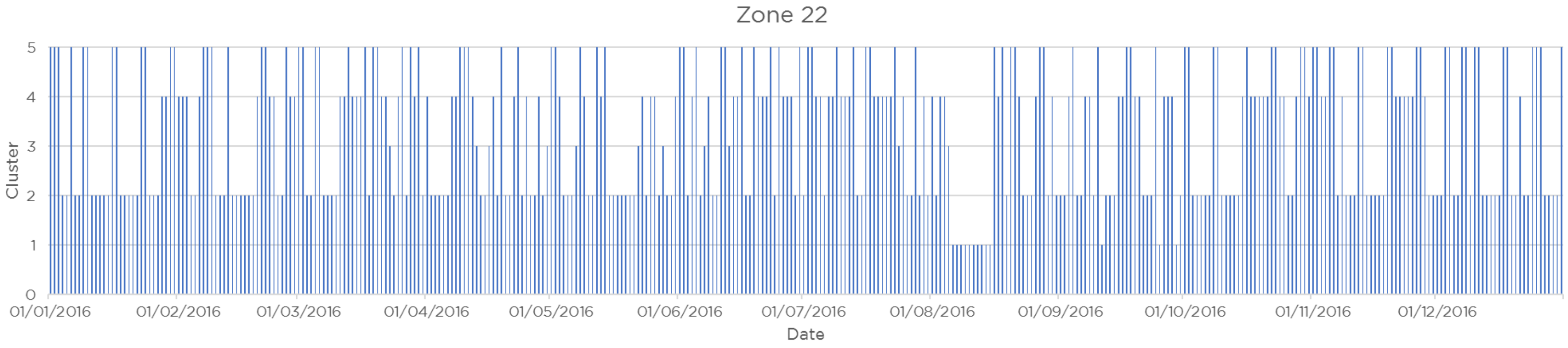


Figure AII.25: Final cluster assessment of Zone 22

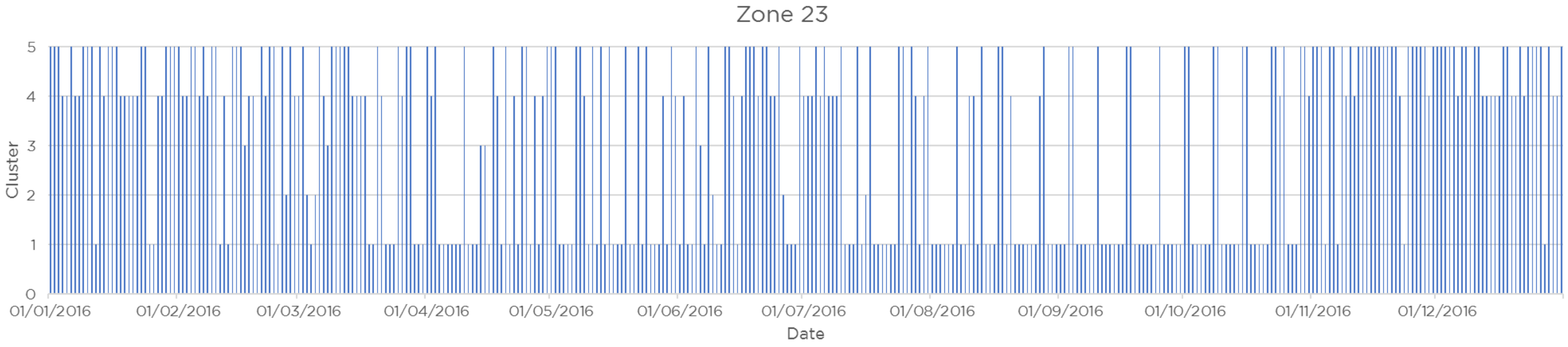


Figure AII.26: Final cluster assessment of Zone 23

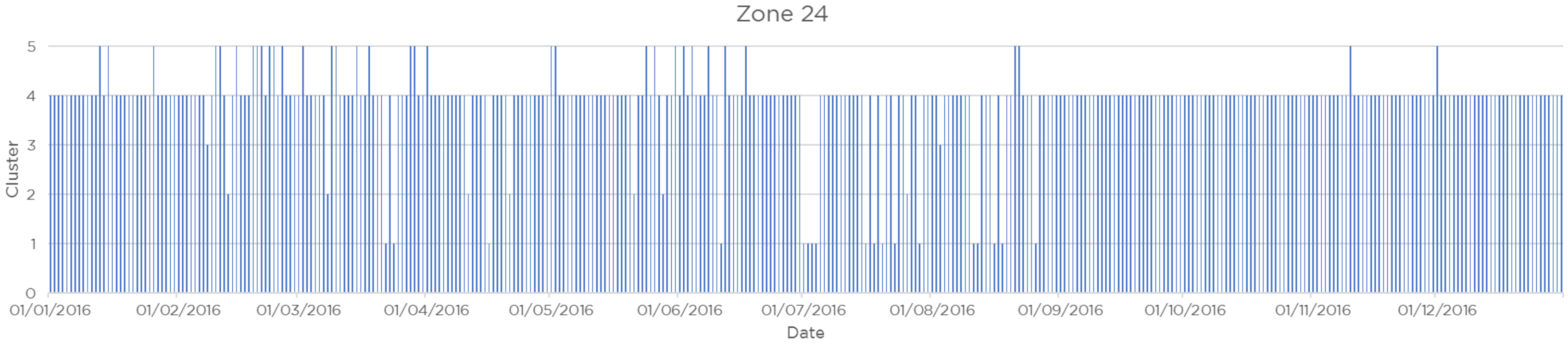


Figure AII.27: Final cluster assessment of Zone 24

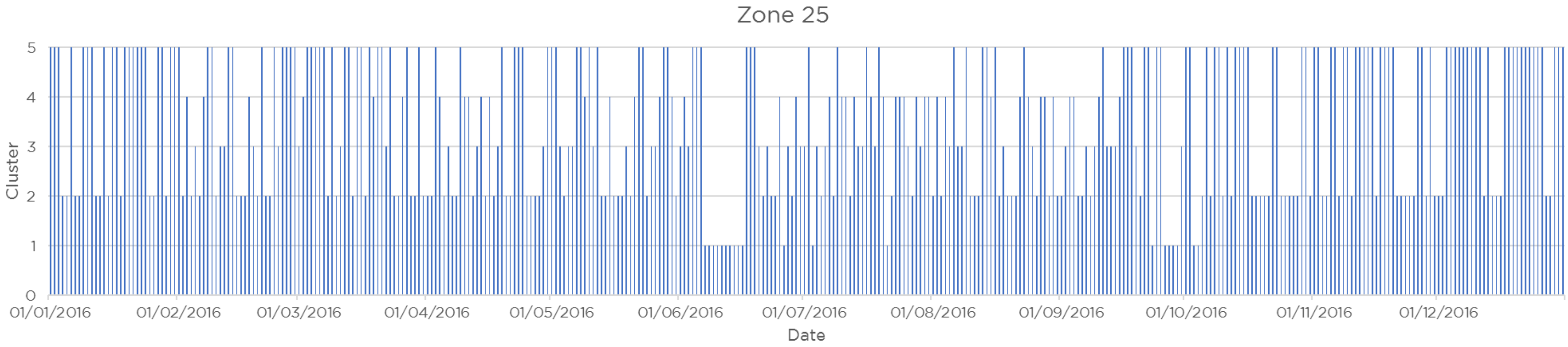


Figure AII.28: Final cluster assessment of Zone 25

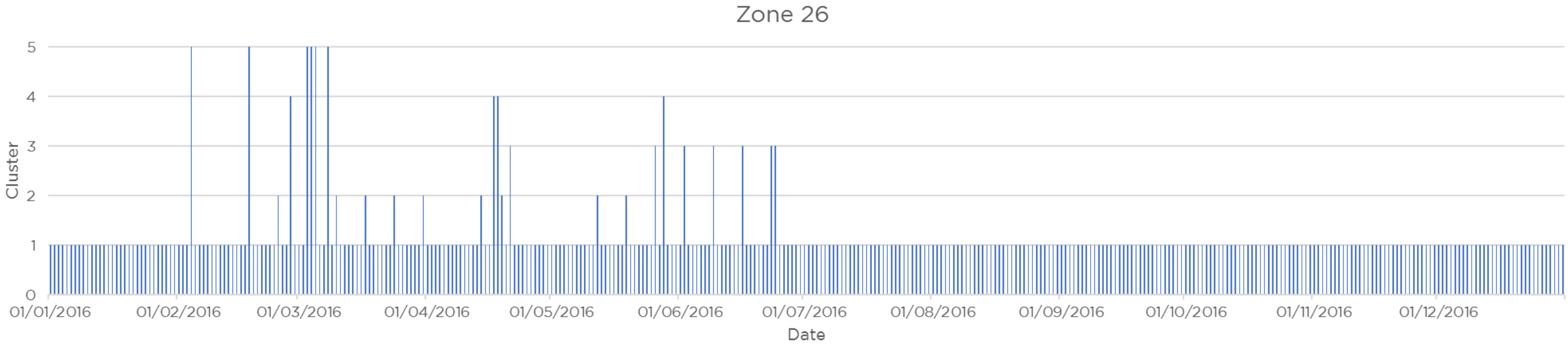


Figure AII.29: Final cluster assessment of Zone 26

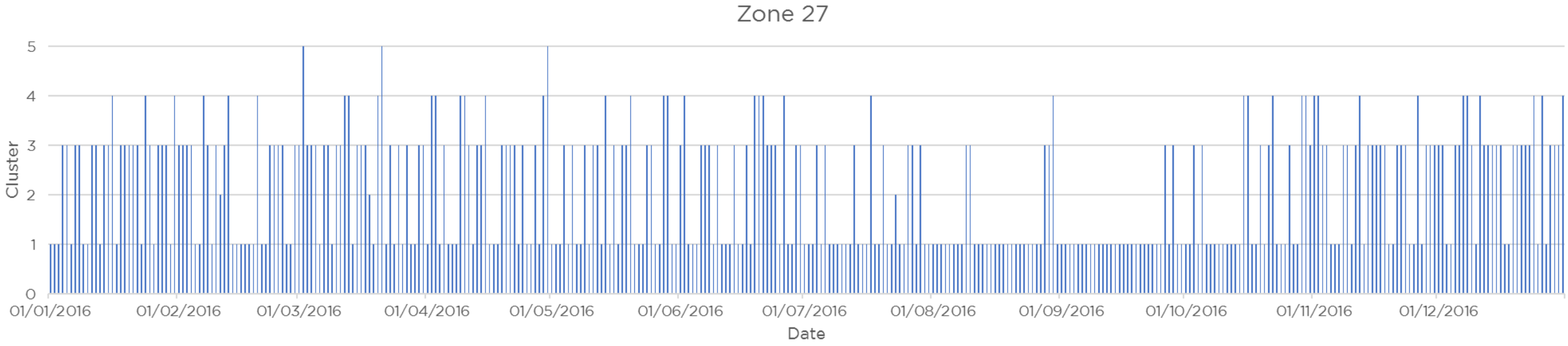


Figure AII.30: Final cluster assessment of Zone 27

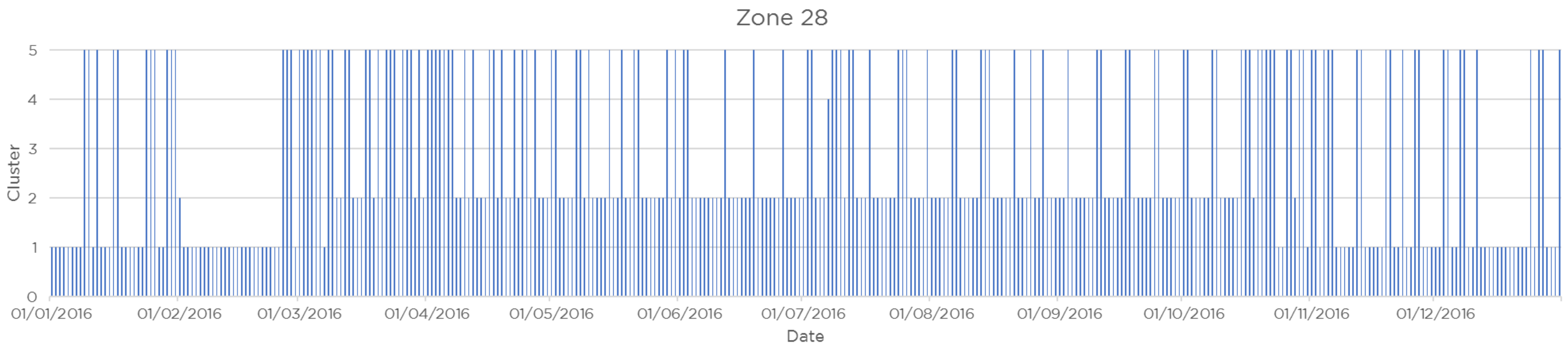


Figure AII.31: Final cluster assessment of Zone 28

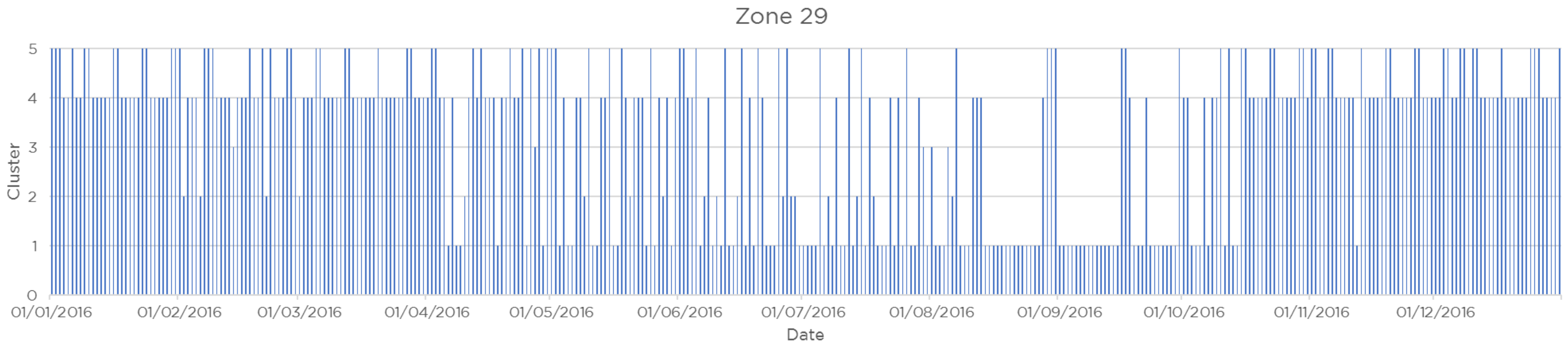


Figure AII.32: Final cluster assessment of Zone 29

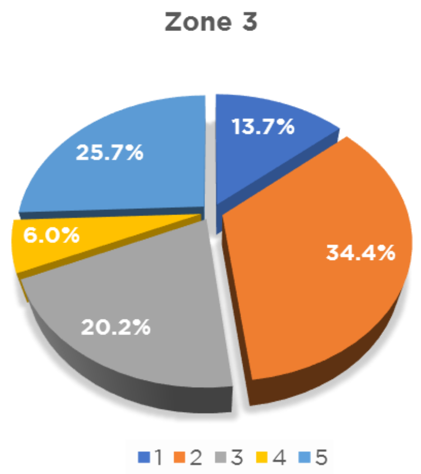


Figure AII.33: Clusters' representativeness in the Zone 3

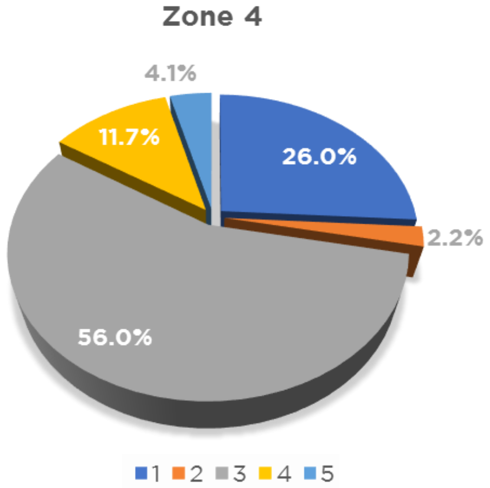


Figure AII.34: Clusters' representativeness in the Zone 4

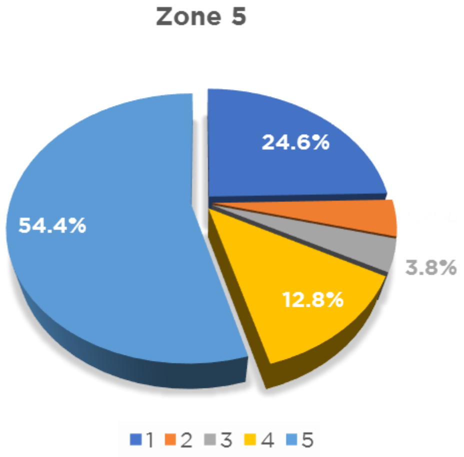


Figure AII.35: Clusters' representativeness in the Zone 5

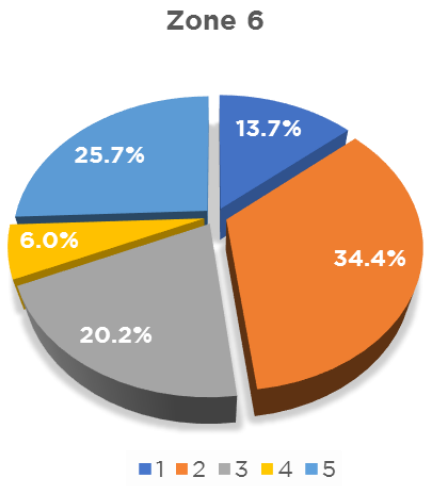


Figure AII.36: Clusters' representativeness in the Zone 6

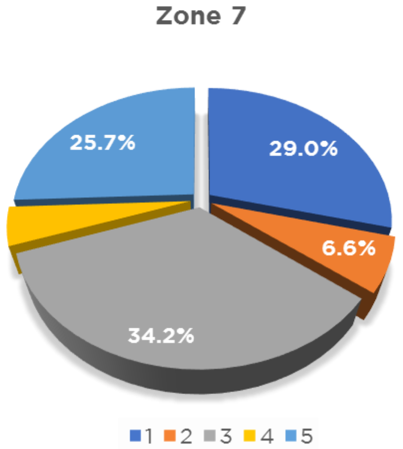


Figure AII.37: Clusters' representativeness in the Zone 7

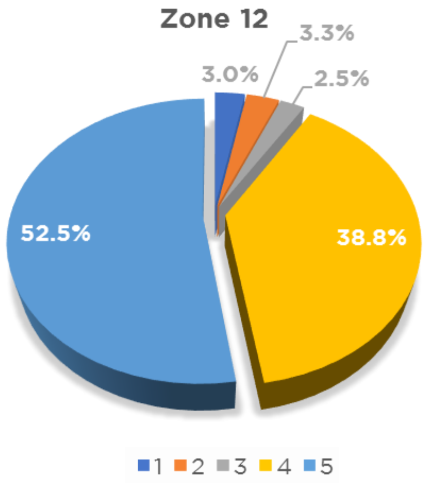


Figure AII.38: Clusters' representativeness in the Zone 12

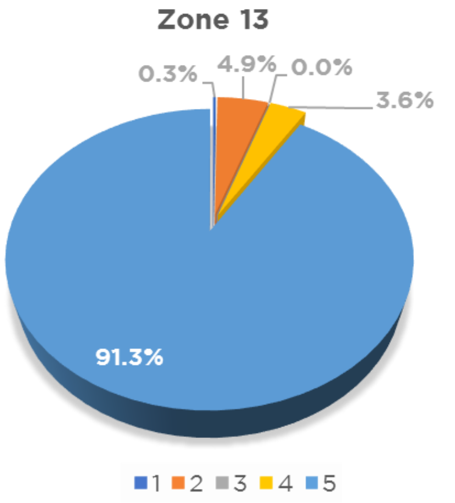


Figure AII.39: Clusters' representativeness in the Zone 13

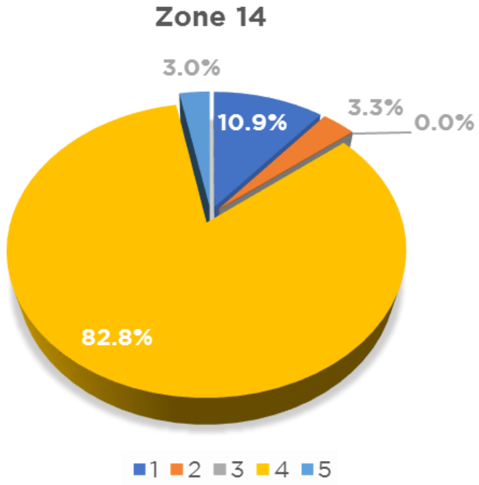


Figure AII.40: Clusters' representativeness in the Zone 14

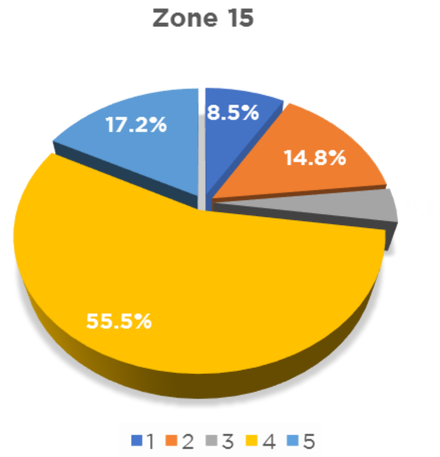


Figure AII.41: Clusters' representativeness in the Zone 15

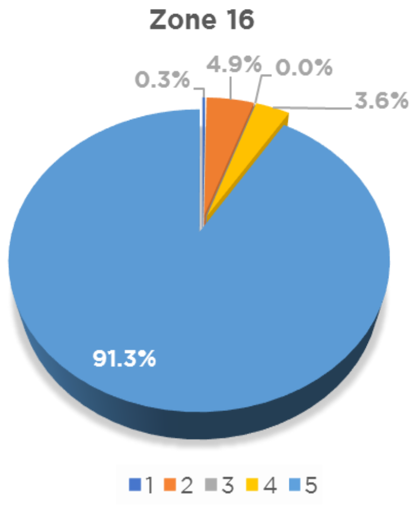


Figure AII.42: Clusters' representativeness in the Zone 16

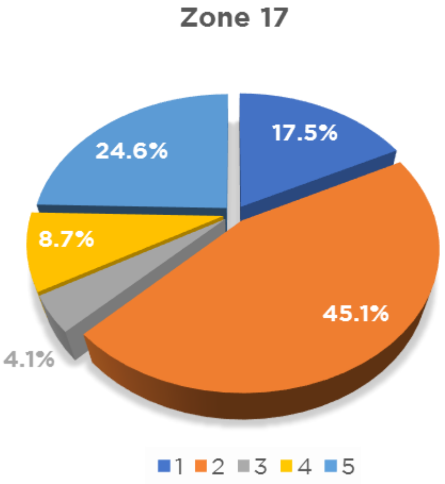


Figure AII.43: Clusters' representativeness in the Zone 17

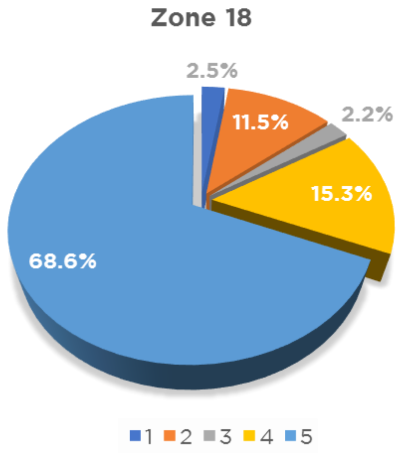


Figure AII.44: Clusters' representativeness in the Zone 18

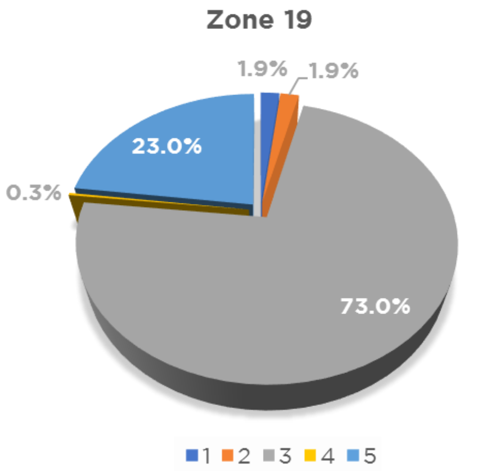


Figure AII.45: Clusters' representativeness in the Zone 19

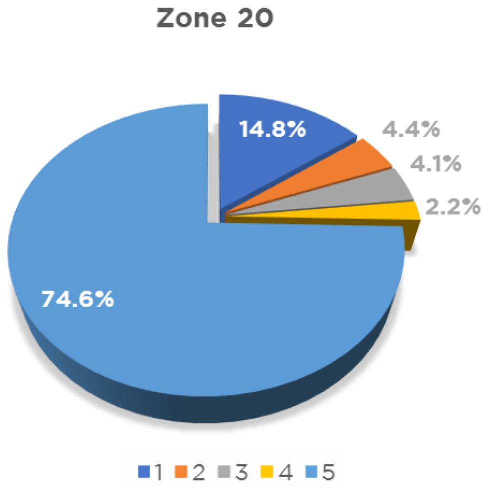


Figure AII.46: Clusters' representativeness in the Zone 20

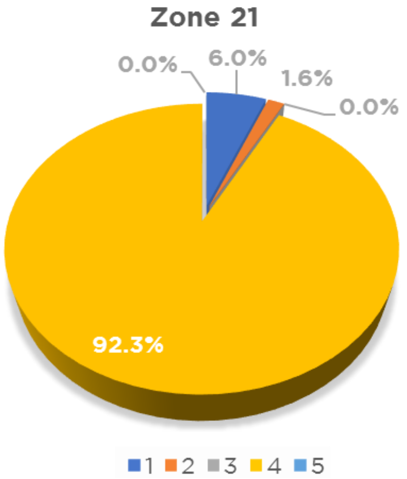


Figure AII.47: Clusters' representativeness in the Zone 21

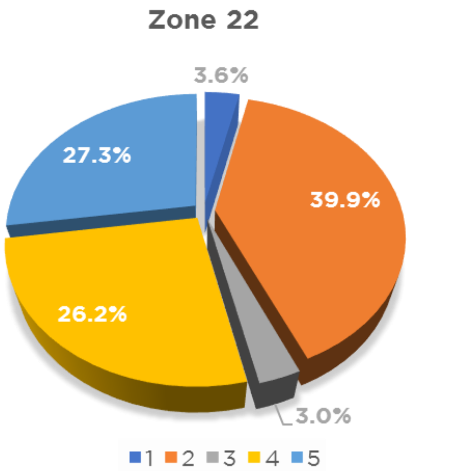


Figure AII.48: Clusters' representativeness in the Zone 22

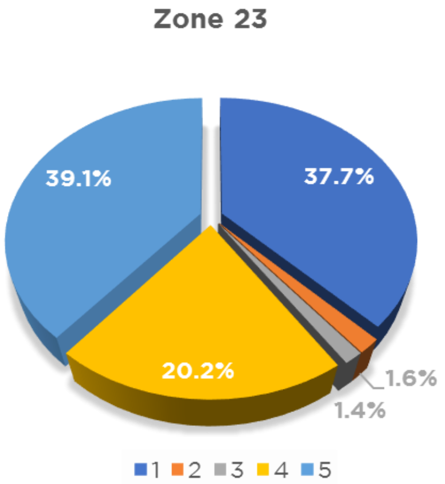


Figure AII.49: Clusters' representativeness in the Zone 23

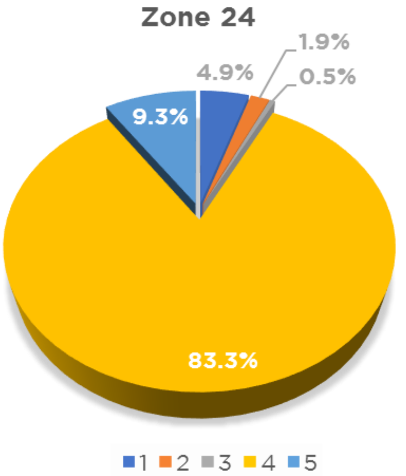


Figure AII.50: Clusters' representativeness in the Zone 24

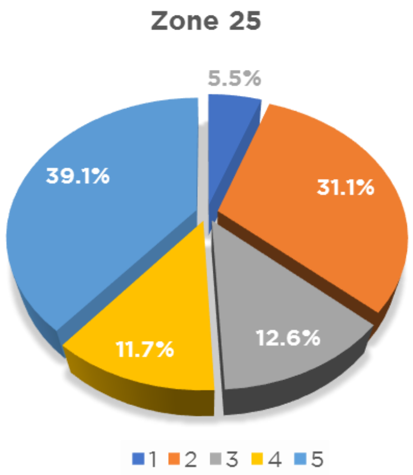


Figure AII.51: Clusters' representativeness in the Zone 25

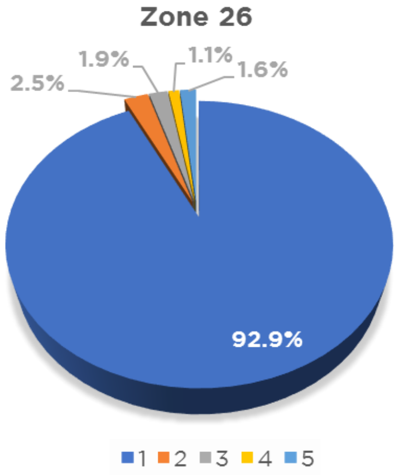


Figure AII.52: Clusters' representativeness in the Zone 26

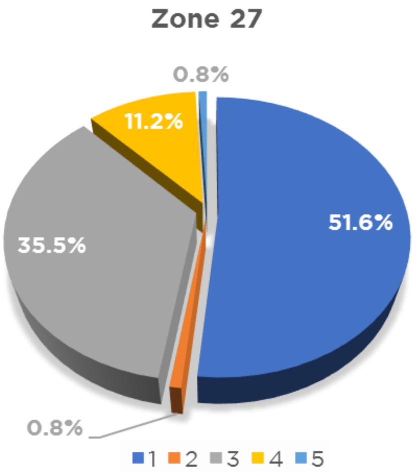


Figure AII.53: Clusters' representativeness in the Zone 27

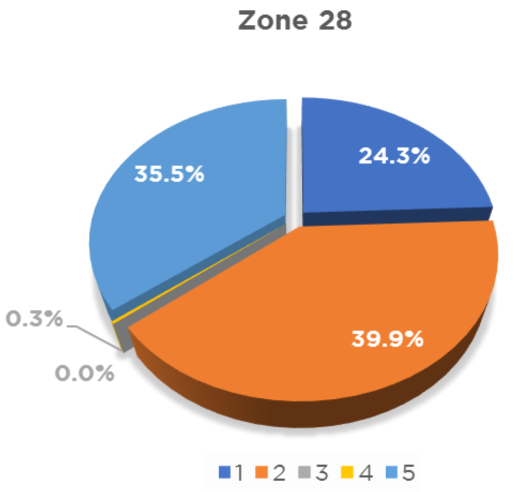


Figure AII.54: Clusters' representativeness in the Zone 28

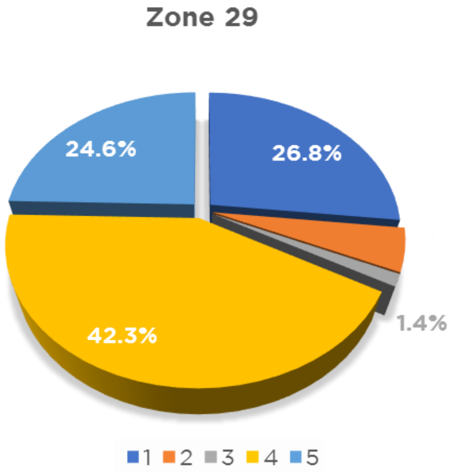


Figure AII.55: Clusters' representativeness in the Zone 29

References

[1] UNI EN 15603, Consumo energetico globale e definizione dei metodi di valutazione energetica Overall energy use and definition of energy ratings, 2011.

[2] US Department of Energy, EnergyPlus Engineering Reference: The Reference to EnergyPlus Calculations, 2010. doi:citeulike-article-id:10579266.

[3] O. Guerra Santin, L. Itard, H. Visscher, The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock, Energy Build. 41 (2009) 1223–1232. doi:10.1016/j.enbuild.2009.07.002.

[4] European Commission, Database - Eurostat, 2017. (2017). http://ec.europa.eu/eurostat/data/database (accessed January 9, 2017).

[5] Ministero dello Sviluppo Economico, Relazione annuale sull'efficienza energetica, 2017.

[6] S. Carlucci, G. Lobaccaro, Y. Li, E. Catto Lucchino, R. Ramaci, The effect of spatial and temporal randomness of stochastically generated occupancy schedules on the energy performance of a multiresidential building, Energy Build. 127 (2016) 279–300. doi:10.1016/j.enbuild.2016.05.023.

[7] F. Yousefi, Y. Gholipour, W. Yan, A study of the impact of occupant behaviors on energy performance of building envelopes using occupants' data, Energy Build. 148 (2017) 182–198. doi:10.1016/j.enbuild.2017.04.085.

[8] M. Schweiker, M. Shukuya, Comparative effects of building envelope improvements and occupant behavioural changes on the exergy consumption for heating and cooling, Energy Policy. 38 (2010) 2976–2986. doi:10.1016/j.enpol.2010.01.035.

[9] V. Fabi, V.M. Barthelmes, Y. Heo, S.P. Corgnati, Monitoring and stimulating energy behavioural change in university buildings towards post carbon cities, in: IBPSA Build. Simul. 2017, 2017: pp. 423–429.

[10] M.P. Deuble, R.J. de Dear, Green occupants for green buildings: The missing link?, Build. Environ. 56 (2012) 21–27. doi:10.1016/j.buildenv.2012.02.029.

[11] European Commission, Buildings - European Commission. https://ec.europa.eu/energy/en/topics/energy-efficiency/buildings (accessed January 13, 2017).

[12] Energy Performance of Buildings Directive (EPBD) - EuroACE. https://euroace.org/euroace-positions/energy-performance-buildings-directive-epbd/ (accessed January 13, 2017).

[13] EU Building Stock Observatory - European Commission. https://ec.europa.eu/energy/en/eubuildings (accessed December 18, 2017).

[14] A. Alfakara, Ben Corxford, Towards Better Buildings Performance Estimations ? A Framework for Integrating Dynamic Occupant Behaviour in Dynamic Buildings Simulation Tools, in: IBPSA Build. Simul. 2017, 2017.

[15] P. Hoes, J.L.M. Hensen, M.G.L.C. Loomans, B. de Vries, D. Bourgeois, User behavior in whole building simulation, Energy Build. 41 (2009) 295–302. doi:10.1016/j.enbuild.2008.09.008.

[16] S. Carlucci, F. Causone, L. Pagliano, M. Pietrobon, Zero-Energy Living Lab, in: J. Littlewood, C. Spataru, R.J. Howlett, L.C. Jain (Eds.), Smart Energy Control Syst. Sustain. Build., Smart Inno, Springer International Publishing, Cham, 2017.

[17] A. Elie, P. Sokratis, Human Behavior and Energy Consumption in Buildings : An Integrated Agent-Based Modeling and Building Performance Simulation Framework, in: IBPSA Build. Simul. 2017, 2017.

[18] S. Wei, R. Jones, P. De Wilde, Driving factors for occupant-controlled space heating in residential buildings, Energy Build. 70 (2014) 36–44. doi:10.1016/j.enbuild.2013.11.001.

[19] V. Corrado, I. Ballarini, S. Paduos, E. Primo, F. Madonna, Application of Dynamic Numerical Simulation to Investigate the Effects of Occupant Behaviour Changes in Retrofitted Buildings, in: IBPSA Build. Simul. 2017, 2017.

[20] R. Galvin, Making the “rebound effect” more useful for performance evaluation of thermal retrofits of existing homes: Defining the “energy savings deficit” and the “energy performance gap,” Energy Build. (2014).

[21] R. Galvin, The Rebound Effect in Home Heating, 1st Editio, 2016.

[22] G. Buttitta, O. Neu, W. Turner, D. Finn, Modelling Household Occupancy Profiles using Data Mining Clustering Techniques on Time Use Data, IBPSA Build. Simul. 2017. 2 (2017).

[23] S. D’Oca, V. Fabi, S.P. Corgnati, R.K. Andersen, Effect of thermostat and window opening occupant behavior models on energy use in homes, Build. Simul. 7 (2014) 683–694. doi:10.1007/s12273-014-0191-6.

[24] K. Schakib-Ekbatan, F.Z. Çakici, M. Schweiker, A. Wagner, Does the occupant behavior match the energy concept of the building? - Analysis of a German naturally ventilated office building, Build. Environ. 84 (2015) 142–150. doi:10.1016/j.buildenv.2014.10.018.

[25] H. Polinder, M. Schweiker, A. Van Der Aa, K. Schakib-Ekbatan, V. Fabi, R. Andersen, et al., Final Report Annex 53 - Occupant behavior and modeling (Separate Document Volume II), 2013.

[26] K. Sun, T. Hong, A framework for quantifying the impact of occupant behavior on energy savings of energy conservation measures, Energy Build. 146 (2017) 383–396. doi:10.1016/j.enbuild.2017.04.065.

[27] G. Ren, M. Sunikka-blank, X. Zhang, The Influence of Variation in Occupancy Pattern on Domestic Energy Simulation Prediction : A Case Study in Shanghai, in: IBPSA Build. Simul. 2017, 2017.

[28] International Energy Agency, IEA-EBC Annex 66. https://www.annex66.org/ (accessed January 15, 2017).

[29] D. Yan, T. Hong, B. Dong, A. Mahdavi, S. D'Oca, I. Gaetani, et al., IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings, Energy Build. 156 (2017) 258–270. doi:10.1016/j.enbuild.2017.09.084.

[30] D. Yan, W. O'Brien, T. Hong, X. Feng, G.B. H., F. Tahmasebi, et al., Occupant behaviour modeling for building performance simulation: Current state and future challenges, Energy Build. 107 (2015) 264–278.

[31] I. Ajzen, The theory of planned behavior, Organ. Behav. Hum. Decis. Process. 50 (1991) 179–211. doi:https://doi.org/10.1016/0749-5978(91)90020-T.

[32] R.H. Fazio, Multiple Processes by which Attitudes Guide Behavior: The Mode Model as an Integrative Framework, in: M.P.B.T.-A. in E.S.P. Zanna (Ed.), Adv. Exp. Soc. Psychol., Academic Press, 1990: pp. 75–109. doi:https://doi.org/10.1016/S0065-2601(08)60318-4.

[33] I. Gaetani, P. Hoes, J.L.M. Hensen, Introducing and testing a strategy for fit-for-purpose occupant behavior modeling in a simulation-aided building design process, in: IBPSA Build. Simul. 2017, 2017.

[34] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, C. Toader, Load pattern-based classification of electricity customers, IEEE Trans. Power Syst. 19 (2004) 1232–1239. doi:10.1109/TPWRS.2004.826810.

[35] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, Energy. 42 (2012) 68–80. doi:10.1016/j.energy.2011.12.031.

[36] G.J. Tsekouras, N.D. Hatziargyriou, E.N. Dialynas, Two-stage pattern recognition of load curves for classification of electricity customers, IEEE Trans. Power Syst. 22 (2007) 1120–1128. doi:10.1109/TPWRS.2007.901287.

[37] L. Hernández, C. Baladrón, J.M. Aguiar, B. Carro, A. Sánchez-Esguevillas, Classification and clustering of electricity demand patterns in industrial parks, Energies. 5 (2012) 5215–5228. doi:10.3390/en5125215.

[38] J. Vesanto, E. Alhoniemi, Clustering of the self-organizing map, IEEE Trans. Neural Networks. 11 (2000) 586–600. doi:10.1109/72.846731.

[39] K.A.D. Deshani, L.L. Hansen, M.D.T. Attygalle, A. Karunaratne, Improved Neural Network Prediction Performances of Electicity Demand: Modifying Inputs through Clustering, in: Second Int. Conf. Comput. Sci. Eng., 2014: pp. 137–147. doi:10.5121/csit.2014.4412.

[40] I.P. Panapakidis, T.A. Papadopoulos, G.C. Christoforidis, G.K. Papagiannis, Pattern recognition algorithms for electricity load curve analysis of buildings, Energy Build. 73 (2014) 137–145. doi:10.1016/j.enbuild.2014.01.002.

[41] G. Dudek, Neural networks for pattern-based short-term load forecasting: A comparative study, Neurocomputing. 205 (2016) 64–74. doi:10.1016/j.neucom.2016.04.021.

[42] A. Capozzoli, M.S. Piscitelli, S. Brandi, Mining typical load profiles in buildings to support energy management in the smart city context, Energy Procedia. 134 (2017) 865–874. doi:10.1016/j.egypro.2017.09.545.

[43] J.D. Rhodes, W.J. Cole, C.R. Upshaw, T.F. Edgar, M.E. Webber, Clustering analysis of residential electricity demand profiles, Appl. Energy. 135 (2014) 461–471. doi:10.1016/j.apenergy.2014.08.111.

[44] F. McLoughlin, A. Duffy, M. Conlon, A clustering approach to domestic electricity load profile characterisation using smart metering data, Appl. Energy. 141 (2015) 190–199. doi:10.1016/j.apenergy.2014.12.039.

[45] J.L. Viegas, S.M. Vieira, J.M.C. Sousa, R. Melício, V.M.F. Mendes, Electricity demand profile prediction based on household characteristics, in: Int. Conf. Eur. Energy Mark. EEM, 2015: pp. 0–4. doi:10.1109/EEM.2015.7216746.

[46] J.L. Viegas, S.M. Vieira, J.M.C. Sousa, Fuzzy clustering and prediction of electricity demand based on household characteristics, in: 2015 Conf. Int. Fuzzy Syst. Assoc. Eur. Soc. Fuzzy Log. Technol., 2015. doi:10.2991/ifsa-eusflat-15.2015.147.

[47] U. Ali, C. Buccella, C. Cecati, Households Electricity Consumption Analysis with Data Mining Techniques, in: Ind. Electron. Soc. , IECON 2016 - 42nd Annu. Conf. IEEE, 2016: pp. 3966–3971. doi:10.1109/IECON.2016.7793118.

[48] F. Jorissen, W. Boydens, L. Helsen, Simulation-based occupancy estimation in

office buildings using CO 2 sensors, in: IBPSA Build. Simul. 2017, 2017.

[49] S.H. Kim, H.J. Moon, Y.R. Yoon, Improved occupancy detection accuracy using PIR and door sensors for a smart thermostat, in: IBPSA Build. Simul. 2017, 2017.

[50] D. Aerts, J. Minnen, I. Glorieux, I. Wouters, F. Descamps, A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison, Build. Environ. 75 (2014) 67–78. doi:10.1016/j.buildenv.2014.01.021.

[51] W. Kleiminger, C. Beckel, T. Staake, S. Santini, Occupancy Detection from Electricity Consumption Data, in: Proc. 5th ACM Work. Embed. Syst. Energy-Efficient Build. - BuildSys'13, 2013: pp. 1–8. doi:10.1145/2528282.2528295.

[52] W. Kleiminger, C. Beckel, S. Santini, Household occupancy monitoring using electricity meters, in: ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. - UbiComp 2015, 2015: pp. 975–986. doi:10.1145/2750858.2807538.

[53] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, SOM Toolbox for Matlab 5, 2000.

[54] E. Alpaydin, Introduction to machine learning, Third edit, 2014.

[55] S. Haykin, Neural Networks and Learning Machines, 2008. doi:978-0131471399.

[56] T. Mitchell, Chapter 4, Artificial Neural Networks, in: Mach. Learn., 1997.

[57] D. Kriesel, A Brief Introduction to Neural Networks, available at http://www.dkriesel.com, (2007).

[58] C. Vercellis, Business Intelligence, 2013. doi:10.1017/CBO9781107415324.004.

[59] L. Fausett, Fundamentals of Neural Networks, 1994.

[60] Abdul Rahid, WordStream. https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications (accessed January 9, 2018).

[61] Piech Chris, K Means. http://stanford.edu/~cpiech/cs221/handouts/kmeans.html (accessed January 7, 2017).

[62] T. Kohonen, The self-organizing map, Proc. IEEE. 78 (1990) 1464–1480. doi:10.1109/5.58325.

[63] I. Ballarini, V. Corrado, C. Becchio, S.P. Corgnati, Energy saving potential by retrofitting residential buildings in Europe, Rehva J. (2012) 34–38.

[64] C. Sousa, F. Causone, S. Cunha, A. Pina, S. Erba, Addressing the challenges of

public housing retrofits, Energy Procedia. 134 (2017) 442–451.

[65] X. Ren, D. Yan, T. Hong, Data mining of space heating system performance in affordable housing, Build. Environ. 89 (2015) 1–13. doi:10.1016/j.buildenv.2015.02.009.

[66] S. Erba, F. Causone, R. Armani, The effect of weather datasets on building energy simulation outputs, Energy Procedia. 134 (2017) 545–554. doi:10.1016/j.egypro.2017.09.561.

[67] Agenzia Regionale per la Protezione dell'Ambiente della Lombardia. http://www.arpalombardia.it/ (accessed February 9, 2018).

[68] T. Watanabe, Y. Urano, T. Hayashi, Procedures for separating direct and diffuse insolation on a horizontal surface and prediction of insolation on tilted surfaces, Transactions, no. 330, Trans. Archit. Inst. Japan. (1983).

[69] D. Seo, M. Krarti, Development of Models for Hourly Solar Radiation Prediction, Archit. Eng. 114 (2008) 1–12.

[70] ITACA, Part 1: Solar Astronomy | ITACA. http://www.itacanet.org/the-sun-as-a-source-of-energy/part-1-solar-astronomy/ (accessed February 11, 2018).

[71] Lund Research Ltd, Understanding the different types of variable in statistics, (2013). https://statistics.laerd.com/statistical-guides/types-of-variable.php (accessed March 23, 2018).

[72] J. Pallant, SPSS Survival Manual, A step by step guide to data analysis using SPSS 4th edition, 2002.

[73] D.J. Rumsey, Statistics For Dummies, 2nd Edition, 2016.

[74] Correlation. https://www.mathsisfun.com/data/correlation.html (accessed January 8, 2017).

[75] D.L. Davies, D.W. Bouldin, A Cluster Separation Measure, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1 (1979) 224–227. doi:10.1109/TPAMI.1979.4766909.

[76] Comune di Milano, Regolamento edilizio, 2014.

[77] J. Mardaljevic, L. Heschong, E. Lee, Daylight metrics and energy savings, Light. Res. Technol. 41 (2009) 261–283. doi:10.1177/1477153509339703.

[78] C. Sandels, J. Widén, L. Nordström, E. Andersson, Day-ahead predictions of electricity consumption in a Swedish office building from weather, occupancy, and temporal data, Energy Build. 108 (2015) 279–290. doi:10.1016/j.enbuild.2015.08.052.

[79] J. V. Paatero, P.D. Lund, A model for generating household electricity load profiles, Int. J. Energy Res. 30 (2006) 273–290. doi:10.1002/er.1136.

[80] A. Capasso, R. Lamedica, A. Prudenzi, W. Grattieri, Bottom-up approach to residential load modeling, IEEE Trans. Power Syst. 9 (1994) 957–964. doi:10.1109/59.317650.

[81] A.C. Menezes, A. Cripps, D. Bouchlaghem, R. Buswell, Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap, Appl. Energy. 97 (2012) 355–364. doi:10.1016/j.apenergy.2011.11.075.

[82] Y.G. Yohanis, J.D. Mondol, A. Wright, B. Norton, Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use, Energy Build. 40 (2008) 1053–1059. doi:10.1016/j.enbuild.2007.09.001.

[83] R.M.J. Bokel, The effect of window position and window size on the energy demand for heating, cooling and electric lighting, Build. Simul. (2007) 117–121.

[84] A. Tzempelikos, A.K. Athienitis, The impact of shading design and control on building cooling and lighting demand, Sol. Energy. 81 (2007) 369–382. doi:10.1016/j.solener.2006.06.015.

[85] Y. Shimoda, T. Fujii, T. Morikawa, M. Mizuno, Residential end-use energy simulation at city scale, Build. Environ. 39 (2004) 959–967. doi:10.1016/J.BUILDENV.2004.01.020.

[86] K. Clement-Nyns, E. Haesen, J. Driesen, The impact of Charging plug-in hybrid electric vehicles on a residential distribution grid, IEEE Trans. Power Syst. 25 (2010) 371–380. doi:10.1109/TPWRS.2009.2036481.

[87] Articolo 9, Decreto del Presidente della Repubblica 26 agosto 1993, n. 412.

[88] Global Climate Report - Annual 2016 | State of the Climate | National Centers for Environmental Information (NCEI). https://www.ncdc.noaa.gov/sotc/global/201613 (accessed March 1, 2018).