A data-driven procedure to model occupancy and occupant-related electric load profiles in residential buildings for energy simulation

Francesco Causone[1], Salvatore Carlucci[2], Martina Ferrando[1], Alla Marchenko[2], Silvia Erba[1].

[1]Department of Energy, Politecnico di Milano, Via Lambruschini 4, Milano, Italy

[2]Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Høgskoleringen 7a, Trondheim, Norway

Abstract

Improving the reliability of energy simulation outputs is becoming a pressing task to reduce the performance gap between the design and the operation of buildings. Occupant behaviour modelling is one of the most relevant sources of uncertainty in building energy modelling and is typically modelled via *a priori* choices made by modellers. Thus, an improvement in the description of occupant behaviour is needed. To this regard, the availability of smart meter recordings might help to generate more reliable input data for building energy models. This paper discusses a novel data-driven procedure that enables to create yearly occupancy and occupant-related electric load profiles to inform building energy modelling, using a typical uneven database made available by energy operators. The procedure is subdivided into three main tasks. The first has the intent to detect representative occupant-related electric load profiles from smart meters readings. The second task aims to generate yearly occupancy profiles from the same database. The last task assesses the impact of the generated occupancy and occupant-related electric load profiles on building energy simulation outputs. The procedure is applied to the case study of a multi-residential building in Milan, Italy and is meant to show the possibility to overcome deterministic inputs that might have little relation with the actual building operation. It showed a substantial improvement in the reliability of building energy simulation and that occupant related load profiles may account for about 8 % of the building's energy need for space heating.

Keywords: Occupant behaviour; Building performance simulation; Energy modelling; Machine learning; Clustering; Classification; Residential buildings

1. Introduction

All over the world, the building sector is one of the major users of energy and materials [1,2]; therefore, in the last decades, the interest in reducing its impact has greatly increased [3,4]. At the end of 2018, the European Union launched new directives [5–8] to increase the energy efficiency of existing and new buildings and to enhance the use of renewable energy. According to the objectives expressed by these directives, building energy simulation is becoming progressively more important to support new constructions and renovation projects.

Occupant Behaviour (OB) represents one of the biggest uncertainties in building energy modelling [9,10] and has a significant impact on energy usage in buildings [11–18], especially in the residential sector [2,19,20]. Commercial software tools, nevertheless, usually lack the capacity to describe individual actions of building users and assume generic user schedules, which are not able to directly reproduce the unpredictability of OB over time [4,12,21,22]. This is becoming a relevant problem, since, in new high-performance buildings, the share of energy demand affected by OB is constantly rising [2,23,24]; thus, the interest on OB models is gaining momentum [4]. It is important not only to model more realistic OB scenarios but also to quantify the impact of them on buildings' energy performance. Employing stochastic and more accurate occupancy profiles in Building Performance Simulations (BPS) can increase the reliability of the results, or at least, provide an indication of related uncertainty. For this purpose, the International Energy Agency (IEA) approved, first, the Energy in Buildings and Communities (EBC) Annex 66 [25], whose aim is to "bridge the gap" between the built environment and OB and currently is operating the EBC Annex 79 that is addressing the issue of occupant-centric building design and operation. Among several contributions, Annex 66 provided an ontology of occupant-related phenomena that relates influencing factors with occupancy (presence, movement) and occupant behaviour (actions) [26,27] and Annex 79 is investigating the use of data-driven methods to support the modelling of occupant presence and actions (OPA) [28].

The traditional approach to energy modelling is to create a numerical model of a building with a set of deterministic input data and to calculate the model's energy performance. However, this procedure is not able to consider both epistemic and aleatory uncertainties and variability in the input data. Sun et al. [13], proposed a methodology to estimate the performance of energy conservation measures that are influenced by uncertainties. Azar et al. [12] proposed a framework for BPS and Agent-Based Modelling using a regression surrogate model. Such methodology tries to overcome the limitations of BPS in modelling human behaviour. The conclusion is that the way in which these uncertainties are considered can influence the energy performance of buildings and highly change the range of the actual results. Also, the study of Gaetani et al. [29] showed how uncertainties can influence building performance predictions. In their opinion, it is crucial to include the modelling of uncertainties among BPS models.

Data analysis applied to real dataset might help in the definition of better input to model OB. In the literature, numerous works dealt with the analysis of electric energy use datasets. Chicco et al. [30], for example, studied the 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 (SOM). The conclusion was that both can effectively assist the electricity providers in performing customer classification. Tsekouras et al. [31] developed a two-stage methodology for the classification of electricity customers of the Greek power system. It was 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 were deducted with the help of pattern recognition procedures. In the second stage, a classification of the customers was carried out with the same methods of the first stage. Hernández et al. [32] developed a well-structured methodology composed of a cascade application of a SOM and the clustering *k*-means algorithm to identify energy consumption patterns. The results showed that the system could adequately find different behaviour patterns without supervision and without any prior knowledge about the data. Deshani et al. [33] proposed an accurate prediction of electrical energy use through improved artificial intelligent approaches. This research showed how a cluster analysis performed to group similar day types, could contribute towards selecting a better set of neuro-forecasters in neural networks. The daily total electric energy use for five years was 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. Panapakidis et al. [34] developed a methodology for the investigation of the electric 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 was 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 the automatic collection of information about the customer's behaviour along with the building's performance. Also in the study of Grzegorz Dudek [35], several methods based on neural networks were proposed and compared, such as multilayer perceptron, radial basis function neural network, generalized regression neural network, fuzzy counter propagation neural networks, and SOM. Capozzoli et al. [36,37] proposed frameworks on load profiles characterisation in buildings, based on the recent scientific literature.

Other researchers tried to find a relation between electric energy use and the presence of occupants in the building. The common method to register very big data of occupancy is through sensors, like in the work of Jorissen et al. [38] or Kim et al. [39] or Khalil et al. [40]; nevertheless, privacy issues inhibit the implementation of such methods in the residential buildings. Also, Time-User Surveys (TUS) are commonly used. The methodology of Aerts et al. [41], used data from Belgium TUS of 2005. Hierarchical clustering techniques on individual occupancy profiles were implemented and then, probabilistic occupancy profiles were obtained by applying the probability to transit from a certain state to another and the duration probability, which are both time-dependent. Also, the methodology proposed by Buttitta et al. [14] introduced a new occupancy model from TUS data, using data mining clustering techniques. The methodology was 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 buildings energy models that use TUS as inputs. The works of Kleiminger et al. [42,43] exploited the electricity meters as occupancy sensors. They showed that supervised machine learning algorithms could extract occupancy

information with an accuracy between 83 % and 94 %. They used 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. In particular, they used the *k*-Nearest Neighbour (*k*-NN) algorithm [42,43], that is used as supervised learning algorithms for clustering the dataset [44]. However, adequate attention must be provided in the phase of data cleaning and processing [45,46].

Some researchers focused their attention specifically on the residential sector. Rhodes et al. [47] studied the measured electric energy 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 whether homeowner survey responses could serve as predictors for the clustering results. Also, McLoughlin et al. [48] 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. The study implemented three of the most widely used unsupervised clustering methods: *k*-means, *k*-medoid and SOM. Viegas et al. [49,50] 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. Ali et al. [46] starting from a dataset from 400 houses, proposed a study on data mining techniques to explain and evaluate which techniques are useful for better understanding large-scale use profile to improve the power system management and design.

The work presented in this paper focuses specifically on residential buildings provided with a yearly sample of smart meter readings with a time resolution of 15-minute and no surveys; a typical challenging condition for energy modellers. The research tries to provide a methodology able to overcome the limits of actual datasets, usually characterized by a lack of data, errors and noise. Following this approach, the presented work reports a procedure to analyse real monitored data to create yearly schedules for internal heat gains due to appliances and occupancy. Three types of occupants' attitudes (low, medium and high electric energy usage) are defined to represent different levels of energy consciousness in terms of control of lights and plug-in appliances. Differently from the most used deterministic approaches [51], hereby a stochastic methodology able to include the intrinsic variability of occupants is adopted. From the reported literature review, data mining and unsupervised machine learning emerged as promising techniques for this purpose. These methods are indeed useful for noise reduction [52] and for pattern recognition in a wide variety of data samples [37], automatically extracting information from the dataset [53]. For this reason, they are adopted in this case study. The work is intended for energy modellers interested in setting more reliable input for their simulations and dealing with uneven "real-world" datasets. It uses unsupervised machine learning techniques via a structured and reliable procedure that may be adopted, repeated and assumed as a tool to support a more accurate building energy simulation.

2. Case study

2.1. The building and its numerical model

The case study is a residential estate in Milan, which is composed of two blocks with a total gross floor area of 4500 m^2 with about 70 apartments for an estimated population of 200 people. The buildings were built in the '80s and have never been retrofitted so far. Based on the existing documentation, including an energy audit and in-situ inspections carried out by expert and independent engineers, a dynamic energy model is created. EnergyPlus 8.5.0 is the energy simulation engine used for the modelling and simulation tasks. Each released version of EnergyPlus undergoes two major types of verification tests [54]: analytical tests according to ASHRAE Research Projects 865 and 1052, and comparative tests according to ANSI/ASHRAE 140 [55] and IEA SHC Task34/Annex43 BESTest method.

Each flat is modelled as a single heated thermal zone, whereas the ground floor and the attic are modelled with two individual thermal zones and are unheated spaces. Also, the staircases are considered as unheated spaces. The surroundings of the building include the presence of trees, with a height of ten to eighteen meters. The simulations are run for a typical year with an hourly resolution. The main settings of EnergyPlus are:

- North axis set at 39 °,
- Terrain set as Suburbs,
- Solar Distribution set as Full Exterior,
- Minimum number of warmup days set at 25,
- Ground temperature set as the 2 meter-depth temperature given by the weather data IGDG of Milano-Linate,
- Heat conduction in constructions calculated with the finite difference method and with a 3-minute time step.

The main settings of the building fabric are summarized in Table 1. In the model, the thermal bridges are accounted for increasing the steady-state thermal transmittance on the envelope constructions. The window glazing has a U-value of 3,0 W/(m^2 K) with a g-value of 0,75 and a visible transmittance of 0,82, while the frame has a thermal transmittance of 5,4 W/(m^2 K). These two components are used for all the nine windows' combinations, characterised by a different number of shutters and dimensions. Roller blinds are set with a nocturnal schedule (closed from 10:00 p.m. to 6:00 a.m., otherwise open) and are characterized by a solar transmittance of 0,05, a solar reflectance of 0,5, and an infrared transmittance of 0,05. The internal gain from electric equipment is divided into a radiant and a lost fraction, respectively set as 0,3 and 0,5. The infiltration is constant and equal to 1 air change per hour, counting both for natural ventilation and infiltrations. This is a strong assumption that was made because no data on windows opening was available and to not arbitrarily influence the simulation outcome with *ad-hoc* control rules.
The building is equipped with a fuel oil centralised system for space heating and gas boilers for domestic hot water (DHW) installed in each apartment. No mechanical ventilation systems nor centralised mechanical cooling systems are installed.

Table 1: Main building fabric settings in the EnergyPlus building model

	Code	Description	Thermal transmittance W/(m^2 K)
Constructions	**M2**	Main external wall	1,22
	M7	External wall of stairs	4,00
	M7b	External wall of the attic	4,00
	MS	External wall on loggias' sides	0,55
	M31	Roller blind case	0,93
	M4	External wall on loggias	1,09
	M6	Vertical internal partition among flats	4,00
	M8	Vertical internal partition on stair cases	1,22
	MP	Vertical internal partition inside flats	1,22
	P1	External floor on the ground	1,75
	P2	External floor on loggias	2,70
	S2	Roof on the attic	4,00
	P3	Internal floor	2,78

2.2. The dataset

The building is provided with electric energy metering data for the year 2016, from the 1st February to the 31st August. Data is completely anonymous, and no additional data or survey are available. The dataset includes 24 households with a 15-minute time step registration. The raw registered data shows some recording errors that cannot be easily interpreted and required a data cleaning process. The electric profiles of each apartment are the accumulation of the electric absorptions of several unknown appliances;

therefore, the registered data includes all the electric appliances installed in each flat and could also comprehend the absorptions from small electric space heating (or cooling) devices, which are not reflected in the energy need for space heating (or cooling). This represents a typical, although uneven, dataset available for energy modellers. The aim of this work is to obtain enough information from this dataset to generate simulation input, in terms of occupancy and occupant-related load profiles, to relieve the energy modeller from the unfair and difficult choice of setting deterministic occupancy and heat gain schedules.

2.3. The weather dataset

Erba et al. [56] and Moazami et al. [57] stressed the impact of different weather datasets on the energy modelling of buildings, therefore, a weather file that could represent the actual conditions at the location was selected. The *Agenzia Regionale per la Protezione dell'Ambiente* (ARPA) provides years of registered data for the weather station in Via Juvara 22, close to the building site. Table 2 summaries the technical characteristics of the installed sensors. The existing registration for this weather station includes dry-bulb temperature, relative humidity, wind velocity and direction, precipitation, atmospheric pressure and global solar radiation. To complete the weather data to be used in EnergyPlus, the global radiation is split into direct, diffuse and global solar radiation using the Watanabe simplified method [58] based on the location of the weather station.

Table 2: Technical characteristics of the installed sensors in the ARPA weather station of Via Juvara, Milan, Italy

Measured variable	Sensor	Sensible element	Accuracy	Range	Resolution
Dry-bulb temperature	Thermohygrometer	Pt100 1/3 DIN-B	± 0,10 °C	- 50 ÷ 70 °C	0,06 °C
Relative humidity		Capacitive	± 1,5 %	0 ÷ 100 %	0,50%
Wind velocity	Tachoanemometer	Optoelectronic sensor	< 35 m/s: ± 2 % > 35 m/s: ± 3 %	0 ÷ 50 m/s	0,01 m/s
Wind direction	Gonioanemometer	Optoelectronic sensor	± 2 °	0 ÷ 360 °	1 °
Precipitation	Electric rain gauge	Collector cone and double chamber bascule	± 1%	0 ÷10 mm/min	0,2 mm
Atmospheric Pressure	Barometer	Piezoresistive sensor	± 1 hPa	800 ÷1100 hPa	0,1 hPa
Global solar radiation	Pyranometer	Thermopile	± 5 % (daily)	0 ÷ 1500 W/m^2	1 W/m^2

3. Methodology

The aim of this session is to describe the process required to create different yearly schedules for internal heat gains due to appliances and occupancy and to assess their impact on the energy performance of a multi-residential building. Seldom actual data of electric energy usage is complete, and a deep survey or expensive sensors should be used to detect the presence of people inside buildings. With the shown procedure, the modeller should be able to create schedules of occupancy and of occupant-related electric power used by appliances for a full year, starting from a relatively small sample of data, improving the reliability of the outcome of energy simulations.

The work is divided into three main sub-tasks (Figure 1): (i) generation of schedules of the standardized occupant-related electric load profiles for use in the energy model, from the registered electricity use; (ii) generation of the standardized occupancy schedules for all the flats in the energy model, from the registered electricity use; (iii) assessment of the impact of the generated schedules on the energy need for space heating of the building.

To perform the first two tasks, IBM SPSS Statistics 24 and MATLAB R2017a are exploited; whereas to perform the third task, EnergyPlus 8.5.0 is used.

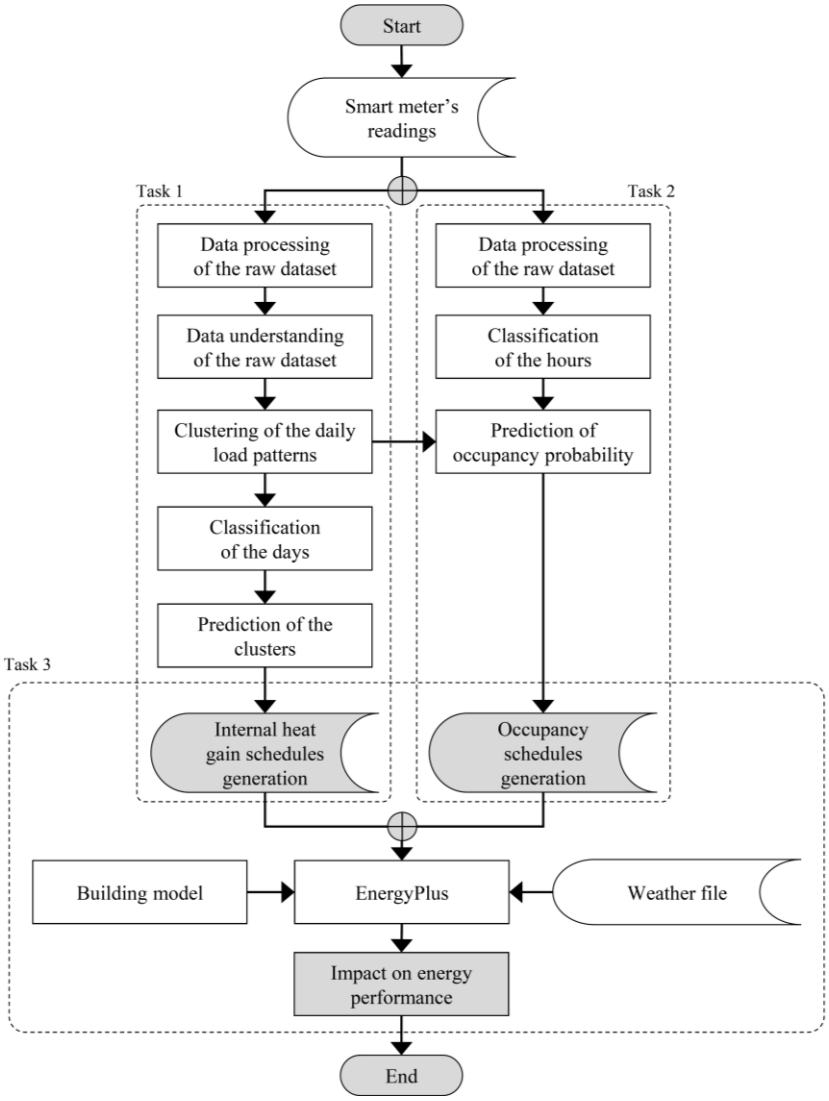

Figure 1: Flowchart of the proposed methodology

3.1. Task 1

The input of this section is the raw registered dataset of electricity use, and the output is a yearly schedule for the occupant-related electric power used by appliances. The aim of the task is to cluster the daily load curves of electricity in different meaningful groups, each represented by the first, second and third quartile. The three final scenarios can be interpreted as the electric profiles for low, normal, and high electric energy users.

3.1.1. Data processing

Data processing is an important step that can spoil the final quality of results. Usually, the actual raw data is incomplete and contains errors. The steps followed to create the dataset for statistical analyses are: (i) data pre-processing (cleaning), in which the outliers are identified and removed, and the inconsistency of data is resolved; (ii) data dimensionality reduction/discretization, in which the representation of data is reduced but producing similar analytical results; (iii) data transformation, in which the data is normalized and aggregated if needed; (iv) data integration, in which integration of multiple dataset and completion with attributes are set into a single and useable format. The result is a clean and functional file suitable for statistical analyses.

3.1.2. Data understanding

The data understanding is performed with different statistical techniques and basic summaries, with the aim to have a deep insight into the dataset and the relations among the different possible influencing factors of the problem. A distinction is made between the registered data and the possible influencing factors. The followed steps are: (i) statistical analyses of the possible influencing factors; (ii) relation analyses among influencing factors; (iii) statistical analyses of the registered data; (iv) relation analyses between the

registered data and the possible influencing factors. The main statistical analysis used in this step is the correlation test, to explore the direct relationships in the sample, T-test and analysis of variance (ANOVA), to perform the difference among groups analyses. These methods are used both to compare the influencing factors between one another and to check whether there is a direct relationship between the influencing factors and the registered data.

3.1.3. Clustering

Clustering means grouping a dataset into an N number of clusters $\boldsymbol{C}_i$, where i = {1, 2, …, N}. To solve the clustering problem two methods of machine learning are used (Figure 2): SOM and *k*-means. They are able to allocate the data into a number of groups trying to minimize some criterion or error functions. The number of clusters is predefined in both cases. Clustering requires the following steps: (i) initialize the clusters' centroids; (ii) group the data; (iii) update the cluster centroids; (iv) if the partitioning is unchanged stop, otherwise return to step ii.

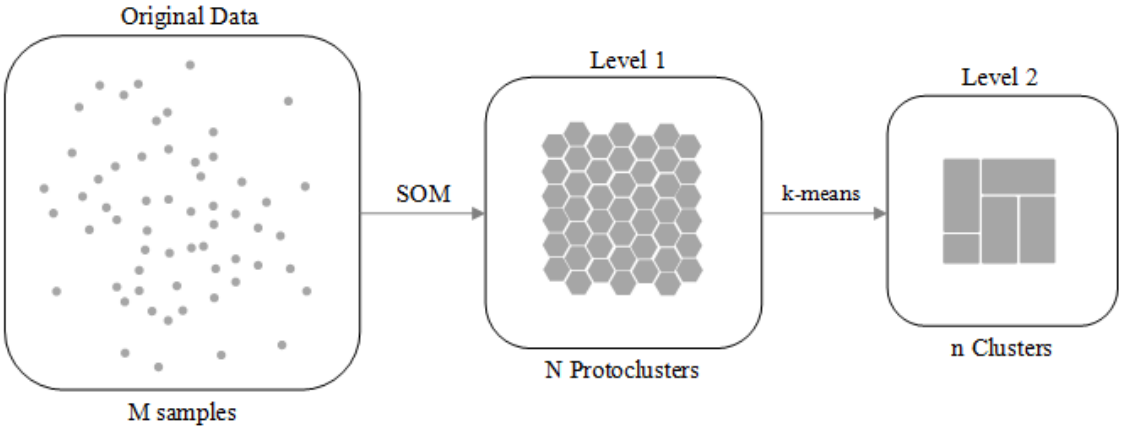

Figure 2: Two-levels clustering approach. From Ref. [59]

The SOM algorithm is used to create protoclusters that are further grouped with a *k*-means algorithm to find the final clusters. As shown by Vesanto et al. [59] and Hernàndez et al. [32], this two-levels approach gives better results than directly clustering the data. The two main benefits are the minimization of computational cost and noise reduction. The protoclusters are local averages of the original samples and so, less sensitive to single high or low cases in the data sample. The SOM algorithm classifies unlabelled data into clusters [60], it can display multidimensional data in a low-dimensional grid and is also a powerful visualization tool. A SOM neuron k does not occupy a fixed position c_k in the input space, but it moves due to weight adjustments during the training process. If an input vector is entered, the neuron closest to the input pattern is activated in the input space. The neurons structure is a single-layer architecture: 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. The important feature of the SOM is that through it, not only the weight of the winner is modified to be closer to the input vector, but also the weights of all neurons, in a certain neighbour of the winning one, are updated. This means that the neurons, which at the beginning are organized according to a topology function, can move during the iterations to best fit with the inputs. The *k*-means algorithm, instead, is one of the simplest and most commonly used unsupervised learning algorithms [61]. It solves the problem of clustering given a fixed number (k) of centroids, one for each cluster. For this clustering technique, the input can move from cluster to cluster during the analysis.

To achieve a good result, it is fundamental the choice of a proper number of clusters as outputs of the SOM. The SOM Toolbox for MATLAB Report [52] is 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:

$$N = 5\sqrt{M} \qquad (1)$$

in which N is the final number of protoclusters, M 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. The actual side-lengths are then set so that their product is as close as possible to the desired N. To improve the results of the SOM, a normalization

on the maximum value reached in the day is performed. The number of final clusters (n) useful to describe the data sample is set using the Davies-Bouldin Index (DBI). After several analyses, the number n 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 does not create negligible not-representative clusters.

3.1.4. Classification

In this methodology, a *k*-NN algorithm is used to solve the classification problem. The recorded data covers only seven months; thus, a classification algorithm is necessary to complete the yearly schedule that can be used as input in the building simulation software. At this step, the days covered by the registration data are characterised by a cluster and some influencing predictors. These predictors are the difference between *Working day* and *Not-working day* and the difference between the *Heating season* and the *Cooling season* are used as predictors. Moreover, to give a time sequence to the days of the year without using the months and days, the average daily external temperature and its variation are added as predictors. For the rest of the year, not covered by the registration, just these influencing predictors are available. Thus, in this section, a *k*-NN algorithm is exploited to assign one of the five clusters obtained in the previous steps to the remaining days of the year on the base of the predictors. The *k*-NN algorithm, being a supervised learning procedure, needs a training set and a testing set. In this case, the training set is the known part of the year in which each day is characterized by the association with a final cluster. The testing set is the part of the year in which no registration of electric consumption is available. To run these analyses, the application of MATLAB Classification Learner is used. Cross-validation of five folds is implemented to avoid overfitting problems. The response, in this case, is the cluster in which the considered day falls in. The *k*-NN is able then, to assign to each day of the year the cluster related to its specific predictors.

The result of this task is an hourly full year schedule (8760 values) for all the flats in the building. Each day of the year is related to one cluster. Then, the curves corresponding to the first, second and third quartiles of the electric use are extracted, for all the days in the specific cluster. These three daily load profiles might be interpreted as three typologies of potential user: low, normal, and high electric energy user.

3.2. Task 2

The input of this task is again the measured electric energy use, whilst the goal is to generate reliable occupancy profiles from it, in order to associate, in the building simulation, heat gains due to occupants to a meaningful profile of occurrence. The used procedure is an adaptation of the work of Kleiminger et al. [42,43]. The idea is that some numerical features of the electric use within an hour can be indicative of the presence of occupants. The average electric use within an hour, its standard deviation, its minimum and maximum values and the sum of the absolute differences (SAD) are all quantities related to the presence of people that are using and changing the electric energy use. For example, a high standard deviation and a high SAD corresponds both to high changes in the electric use within the hour that can be associated with turning on/off the devices and it is usually related with the presence of people in the flat [42,43] since no building automation system is installed.

3.2.1. Data processing

The data processing of this task is composed of cleaning and transformation. The parts of the database in which the 15-minute step registration is not available (e.g., due to gaps or errors) are erased from the sample. The transformation, in this case, is the association with the hourly listed features: average, minimum and maximum, standard deviation and SAD.

3.2.2. Classification

A simple heuristic unsupervised occupancy detection is used to simplify the problem by comparing the current electric energy use to the mean of the night-time use and proposing possible ground truth occupancy profiles to the user, as suggested by Kleiminger et al. [43]. The *k*-NN algorithm is used for this activity. The outcome of the classification is a binary variable (e.g., it can take either 0 or 1) corresponding to the absence or presence of occupants for each hour of each day.

3.2.3. Prediction

Occupancy profiles expressed by a binary variable are too sharp for describing the occupancy in apartments used by several members. It is preferable to provide the probability of occupancy, which can better describe the real occupancy patterns in the

apartments. Thus, from the previous task, each day of the year is clustered in one of the five identified clusters. Then, for each cluster, the first, second and third quartiles are calculated from all the daily binary profiles of occupancy derived from the classification. The results are three continuous variables (first, second, third quartiles) ranging from 0 to 1, corresponding to the probability of occupancy in a specific hour of the day for each cluster. Since this probability should be associated with the heat gains due to the building occupants, it is expressed with integer numbers, to avoid setting the heat gain due to fractions of a physical person. It is assumed that when the value of the continuous variable is below 0,33, it will be rounded to 0 % probability, when the value of the variable is between 0,33 and 0,66, the considered probability will be 50 % and, when it is higher than 0,66, it corresponds to 100 % probability of occupancy. For each cluster, proper daily schedules are created from the results of the classification and transformation process. The probability is then multiplied by the nominal number of people that are supposed to live in an apartment (2 or 4), which is estimated on the basis of the type of bedroom: rooms with a net floor area lower than 14 m^2 are single bedrooms otherwise are double bedrooms, according to the prescriptions of the building regulation of the Municipality of Milan [62].

3.3. Task 3

Task 3 concerns the energy simulation of the building with the schedules created in task 1 and 2. These schedules affect the occupant-related electric power used by appliances in the building and, beyond the electrical energy use, the energy need for space heating and cooling (in the following case study cooling will not be considered since the building is not equipped with a mechanical cooling system). Three cases are run:

1. a *low internal heat gains scenario* that considers, for each cluster, the first-quartile load curve for appliances (i.e. low electric energy users) and the low probability of presence,
2. a *medium internal heat gains scenario* characterised by the median load curve (i.e. medium electric energy users) for appliances and the medium occupancy probability,
3. a *high internal heat gains scenario* with the third-quartile load curve for appliances (i.e. high electric energy users) and the high probability of presence.

Finally, as a conclusive validation of the methodology, the energy simulation results are compared against the registered energy use for space heating of the building block for the year 2016. It is reminded that the simulation has been run with the actual weather data of 2016 from the weather station in via Juvara in Milan.

4. Results and Discussion

4.1. Task 1

4.1.1. Data processing

A first necessary step is the cleaning of the dataset from errors and data gaps. To correct the inconsistent data, these periods are erased from the dataset. In fact, a substitution can alter the dataset and modify the result. The dataset includes the electric energy use with a 15-minute time step from 1st February 2016 to 31st August 2016, for 24 flats. The data of 29th February, 6th and 7th March, and from 3rd to 8th May are totally missing. Moreover, other periods of time are characterized by a lack of data, in particular:

- Flat 3 and Flat 13 have no data because they are empty or without an active contract,
- Flat 4 from 12th June to 31st August,
- Flat 5 from 14th August to 31st August,
- Flat 14 from 19th June to 29th August,
- Flat 29 from 2nd March to 31st March.

The second fundamental step is the reduction of data. In this case, it is performed on the time step basis. The available dataset is registered every 15 minutes as electric power in watt, to obtain hourly values, the average within the hour is performed. In this way, possible eluded outliers are reduced, and so their effect on the overall results. Moreover, the hourly time step is appropriate for the creation of a yearly schedule for energy simulation software.

The aim of the first statistical analysis is to understand if the electric energy use can be easily predicted considering some features, on different scales (hourly-, daily- or flat-scale). As a first approximation, electric energy use can be thought to be affected by some influencing factors, such as the installed electrical appliances, or the number of people living in a household, etc. These influencing factors are numerous; therefore, a detailed literature review is performed. In Table 3, all the influencing factors that can affect the electric energy use are listed with one or more references in which each one is explicitly related to building electric use.

Table 3: List of influencing factors that can affect electric energy use with related references

Family	Influencing factor	Reference(s)
Location/Weather/Habits	External radiation†	[63]
	External temperature†	[64]
	Working days / Not-working days†	[65]
	Day of the week†	[14]
	Precipitation†	-
	Hour of the day†	[65]
	Heating season/Cooling season†	[64][65]
	Renewables on site◊	[66]
	House demand limit◊	[67]
Flat characteristics	Orientation*	[63]
	Floor*	[68]
	Number of rooms*	[69]
	Floor area*	[69]
	Window-to-wall ratio*	[70]
	Insulation level◊	[64]
	g-value◊	[63]
	Shading type◊	[71]
	Typology◊	[69]
Indoor	Indoor air temperature∞	[64]
	Internal illuminance∞	[63]
Family type	Number of people∞	[67][69]
	Sex∞	[67][69]
	Age∞	[72]
	Income∞	[67][69]
	Occupation∞	[67][69]
	Shading operation∞	[71]
	Appliances' efficiency∞	[68]
	Availability of electric car∞	[73]
	Installed equipment∞	[67][68]

∞ is for the influencing factors that are not available in this case study.

◊ is for the influencing factors that are not exploitable in this case study because they are constant all over the dataset.

† is for the influencing factors that are exploitable and mark a difference among hours.

* is for the influencing factors that are exploitable and mark a difference among flats.

The integration step is necessary to get the data in a usable format to run the statistical analysis. The influencing factors are chosen on the basis of their exploitability and added in a spreadsheet with the registered data. Table 4 summaries the list of the selected influencing factors.

Table 4: List of selected influencing factors and their features

	Variable name	Unit of measure	Range of variation - Continuous [Interval, step] - Categorical {discrete values}	Type of variable
AMONG HOURS/DAYS	External-radiation	W/m^2	$[0 \leq x \leq 931,3]$	Continuous
	External-temperature	°C	$[1,6 \leq x \leq 33,8]$	Continuous
	Precipitation	mm	$[0 \leq x \leq 29,6]$	Continuous
	Month	-	{2-8}	Interval
	Day-of-the-month	-	{1-31}	Interval
	Hour-of-the-day	-	{0-23}	Interval
	Day-of-the-week	-	{1-7}	Interval
	Day/Night	-	{-1; 0; 1}	Categorical
	Workdays/Not-working-days	-	{-1; 1}	Binary
	Heating/Cooling-season	-	{-1; 1}	Binary
AMONG FLATS	Orientation	-	{1-4}	Categorical
	(Flat-number)	-	{2; 4-7; 14-29}	Categorical
	Floor	-	{0-3}	Ordinal
	Number of bedrooms	-	{1; 2; 3}	Categorical
	Floor-area	m^2	$[37,9 \leq x \leq 95,3]$	Continuous
	Window-to-floor-ratio	-	{1-4}	Ordinal

Description of influencing factors

The climate influencing factors are derived from the registration of the A.R.P.A. Lombardia weather station located in Via Juvara [67]. *External-radiation* is the global radiation incident on a horizontal plane expressed in W/m^2 and calculated as the hourly average of the measured data. *External-temperature* is the hourly average air temperature in degree Celsius (˚C). *Precipitation* is the hourly cumulative value in millimetres (mm).

The *Month*, the *Day-of-the-month*, the *Hour-of-the-day* and the *Day-of-the-week* are inserted to give a temporal distinction that is used as an influencing factor and to sample data in SPSS. A categorical variable (*Day/Night*) is inserted to distinguish between day and night. *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 8:00 a.m. to 4:00 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 10:00 p.m. to 4:00 a.m. The third group, composed by the hours in between, that can change to be day or night during the year, is indicated with 0. The influencing factor *Workdays/Not-working-days* discriminates between the working days and weekends plus national holidays.

The difference between the *Heating/Cooling-season* is set according to the Art. 9 of DPR 26/08/93 [87]. Milan belongs to climatic zone E, so the heating season is from 15th October to 15th April; the rest of the year is set as cooling season, although no active mechanical cooling is available in the building.

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

Flat-number is simply the progressive number used to name the different flats. It is useful only to group data and to represent features in a graphical format, thus, the correlation results for this factor is not reported. *Floor* is the storey at which a flat is located. *Number of bedrooms* is set as the number of bedrooms in the flat, the area is calculated as the net useful floor area. *Window-to-floor-ratio* is the ratio between the net window area belonging to an apartment and the net floor area of the whole apartment.

4.1.2. Data understanding

In this section, statistical methods and graphs are exploited to understand the variables of the dataset. Then, a correlation test is performed to investigate the links between the influencing factors and the actual electric registered data. Visual techniques are used to detect possible patterns in the building's electric energy usage. The daily sum of the electric energy usage of all the flats in the entire period from 1st February to 31st August shows that the use of electricity in the dwellings slightly decreases along the seasons. In the monthly average electric energy use of the flats, again, a negative trend is registered going from February to August. Nevertheless, July is characterized by an increase in electric use compared to the closest months, showing an average value comparable with February and March. Calculating the average electric energy use of the flats on a daily basis, 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 longer than during working days, resulting in higher electric energy usage. Moving to the hourly resolution, some characteristics of the electric daily energy use can be deduced from Figure 3. The early morning is characterized by a very low electric demand with the minimum reached around 4:00 a.m.; then, the values increase till lunchtime, around noon. During the afternoon, there is almost constant electricity demand, and the maximum values are registered in the evening, from 7:00 to 10:00 p.m. The energy usage in the evening is sharply higher than the rest of the day, since almost all the tenants are at home, having dinner, using lighting and/or using leisure electric equipment such as television or personal computers. The influencing factors that can affect the electric energy usage in the residential buildings are numerous and thus, the trends are not easily predictable. For these reasons, a correlation test between the influencing factors and electric energy use is carried out.

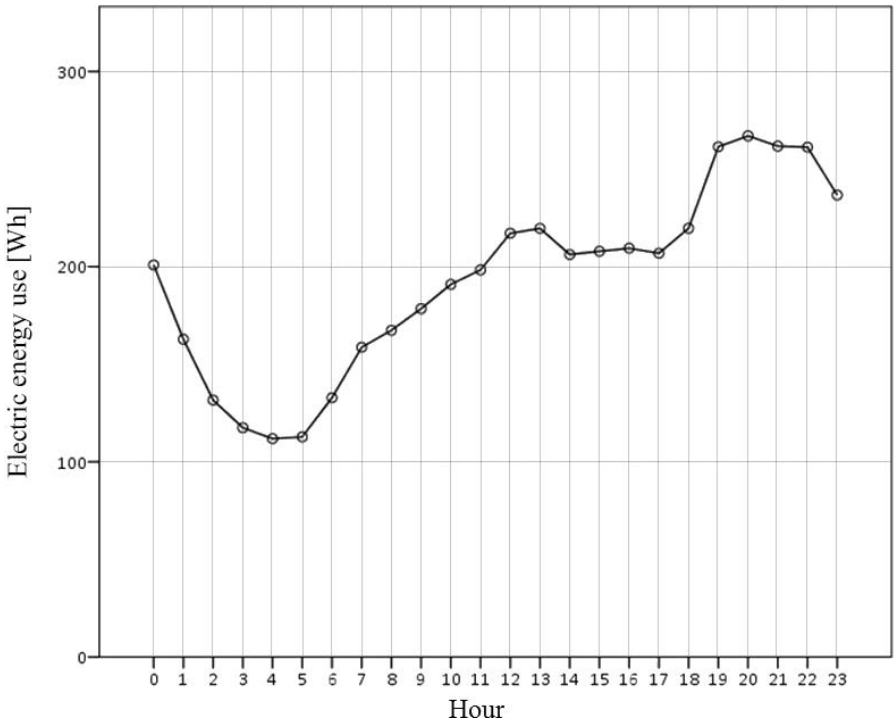

Figure 3: Hourly-averaged electric energy use of all flats

Correlation analysis

Table 5 shows that no influencing factor can be considered as highly correlated with electric energy use according to the criteria that assess the strength of the correlation proposed by Rumsey [74]. Spearman's rho correlation is exploited because the variables were not always normally distributed. Moderate and almost moderate correlations are registered between the energy usage and the number of bedrooms and the floor area. These correlations can be justified considering that a larger flat (larger floor and more bedrooms) has more electric appliances installed, which are also typically used by a higher number of occupants. If strong correlations were registered, this analysis could lead to a simplified clustering process. However, in this case study, no significant result is found, and advanced clustering processes need to be explored. Machine learning techniques emerged from the literature review as the most promising clustering techniques and are, therefore, implemented to achieve this goal. The objective is to find daily patterns that can be attributed to different family types and habits.

381 *Table 5: Correlation results between the possible influencing factors and electric energy usage*

Influencing factor	**Spearman's rho**
Day/Night	0,039
Heating/Cooling-season	-0,002
Workdays/Not-working-days	-0,007
Day-of-the-week	-0,001
Precipitation	0,016
External-temperature	0,072
External-radiation	0,045
Floor	0,039
Orientation	-0,034
Number of bedrooms	0,389
Floor-area	0,290
Window-to-floor-ratio	0,062

382
383 ***4.1.3. Clustering***

384 The data is normalized to the daily maximum. The final size of the SOM is 8 x 42. Figure 4a shows the topology of the used SOM,
385 and Figure 4b shows the connections among the neurons. This figure uses blue hexagons to represent the neurons, whilst the red
386 lines represent the connection among neighbouring neurons. After running the SOM, each neuron represents a protocluster. Another
387 useful figure is the SOM Sample Hits (Figure 4c). It shows how many data points are associated with each neuron. The distribution
388 is not even, and some neurons group many days. Finally, Figure 4d shows the SOM Weighted Neighbour Distances, which presents
389 the following colour coding:

390 – the blue hexagons represent the neurons,
391 – the red lines connect neighbouring neurons,
392 – the colours in the regions containing the red lines indicate the distances among neurons,
393 – the darker colours represent larger distances,
394 – the lighter colours represent smaller distances.

395 From Figure 4d, it is visible that the protoclusters are not sharply subdivided; they are mainly linked together. The average of the
396 normalized daily use in each protocluster is calculated and then submitted to the *k*-means algorithm.
397 After the clustering phase, the dataset is subdivided into five groups of days with similar daily patterns; the results are shown in
398 Figure 5. The graphs on the right side of Figure 5 show the protoclusters inside a cluster. On the left side, the graphs show the three
399 final scenarios: low, medium and high electric energy users, calculated as the first, second and third quartile respectively. The SOM
400 plus *k*-mean method is able to subdivide into groups days with similar patterns. 28 % of the days of the original dataset are grouped
401 in cluster 5, 25 % are in cluster 4, 20 % are in cluster 1, 14 % are in cluster 2, and 13 % are in cluster 3. No cluster is apparently
402 more representative than others.

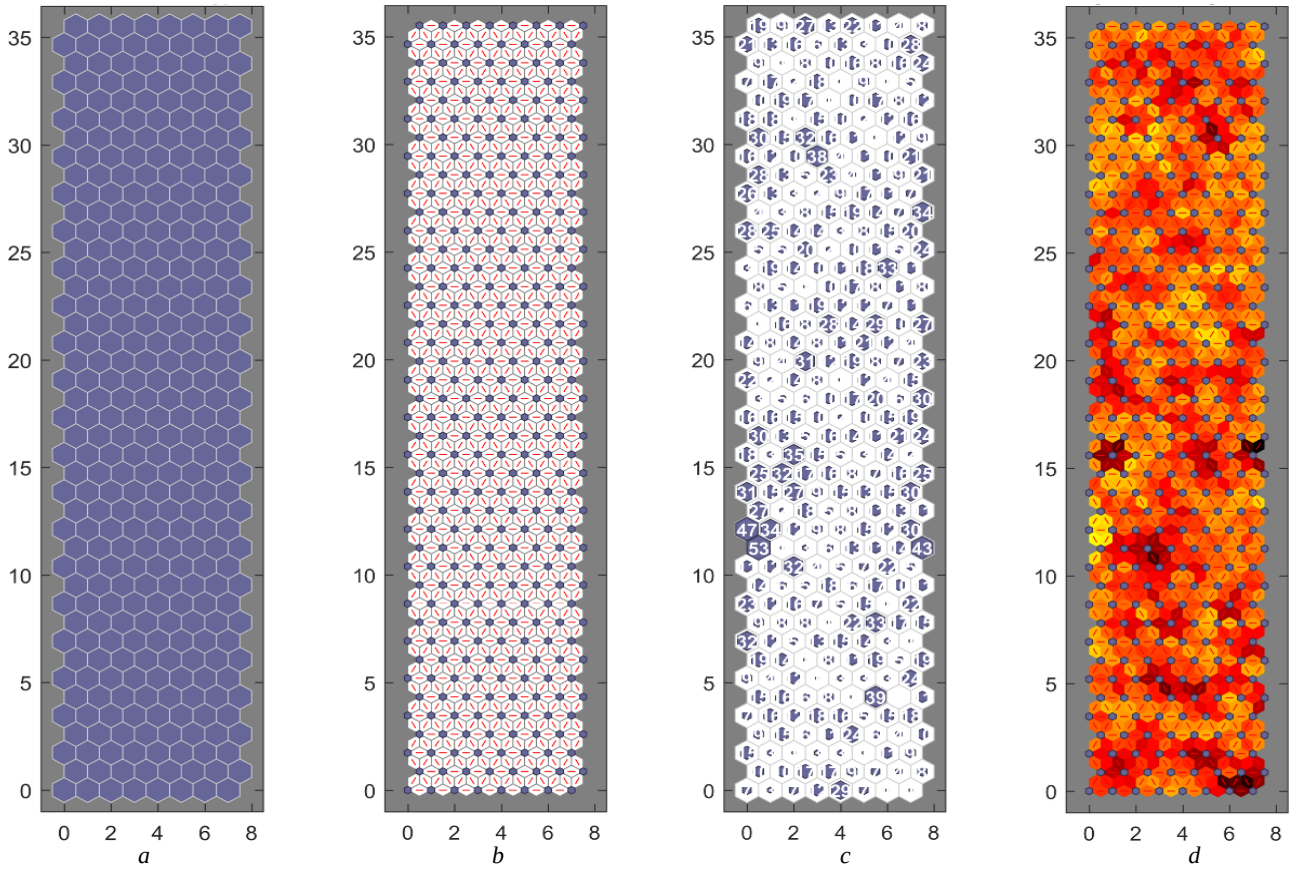

Figure 4: (a) Topology, (b) Connections, (c) Hits and (d) Weighted distances of the SOM

4.1.4. Classification

The *k*-NN is then applied with a cubic calculation of the distances with 10 neighbourhoods., *Workdays/Not-working-days*, *Heating/Cooling-season*, *External-temperature* and its variation are used as predictors to complete those months that missed data. The *k*-NN algorithm is used because it was evaluated as one of the techniques with the highest average accuracy [51][52]. The accuracy, in the current case study, is high (with a *k*-NN run with the cubic calculation of distances and with 10 neighbours), with an average of 86 % among the flats. The *k*-NN is run, and then the result is applied to the rest of the data sample. The application of the *k*-NN provides a cluster for each day of the year. To create the yearly hourly schedule, then, each daily cluster is substituted with the relative 24-hour pattern (Figure 5). Since two flats have no data registration (flat 3 and 13), to complete the yearly schedule for the whole building, the schedules of the most similar flats in the building are adopted. Similarity is estimated comparing *Floor-area*, *Number of bedrooms*, *Orientation*, *Floor*, and *Window-to-floor-ratio*. Thus, flat 6 and flat 16 are used to represent flat 3 and 13 respectively.

4.2. Task 2

This task, consisting of data processing (section 3.2.1), classification (section 3.2.2.) and prediction (section 3.2.3.), provides yearly schedules of occupancy. The procedure needs a higher granularity of the registration respect to the previous one, thus, only the parts of the dataset with an available 15-minute step registration were used. The 15-minute step registration is integrated with the selected hourly features (average, minimum and maximum, standard deviation and SAD). Then, the *k*-NN algorithm is used with a simple heuristic unsupervised occupancy detection method, to predict the final occupancy probability. The prediction method is applied to the three resulting quartiles to create three scenarios of occupancy. The result is shown in Figures 6. The quartile 3 represents families whose components spend more time at home, whilst, the quartile 1 represents families that spend more time outside. In these figures, the blue line represents the occupancy probability given by the average, the red line represents the third quartile and the dark blue line represents the first quartile. The conclusion is an occupancy probability directly related to each cluster, therefore, it can be generalized throughout the year following the results of the previous task. The final probability is then multiplied by the number of people that are supposed to live in the apartment, based on the number of bedrooms.

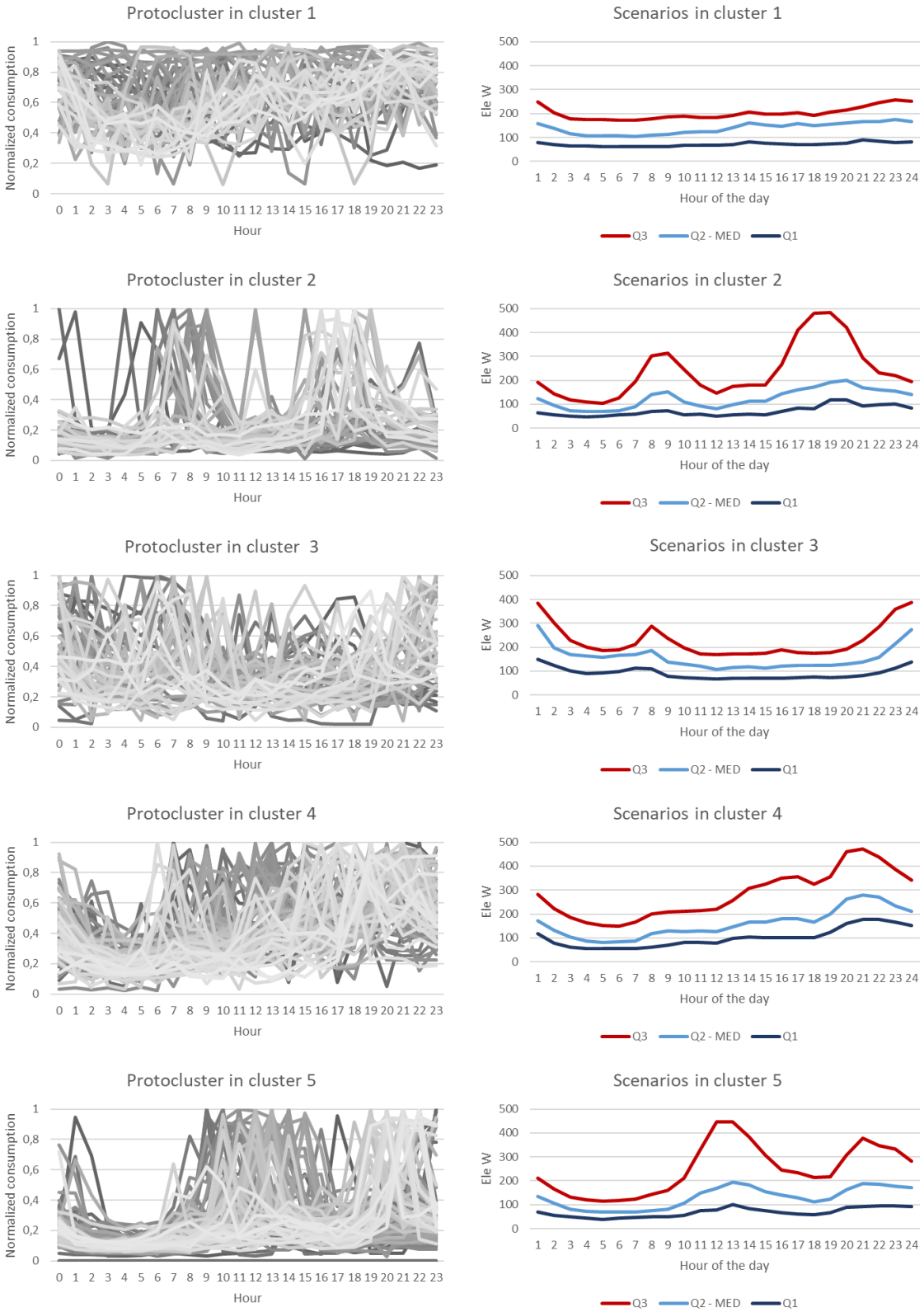

Figure 5: Protoclusters and scenarios in each cluster

4.3. Task 3

The aim of the final analyses is to assess the impact of the generated schedules on the yearly energy need for space heating. Case 1 (Figure 7) is run varying the three scenarios of electric energy use but using the medium presence profile in all the flats (second quartile). The analysis results in three different cases that can be ascribed to the internal heat gains schedule scenarios: low, medium and high. In the first case, the schedule generated with the first quartile is assigned to all the households and it corresponds to the situation of the lowest electric energy use due to lighting and appliances. For this reason, an increase in the energy need for space heating is expected. In the second case, the medium schedule is assigned. The third case corresponds to the assignment of high electric energy users' schedule to all the households in the building, resulting in an increase of the internal heat gains.

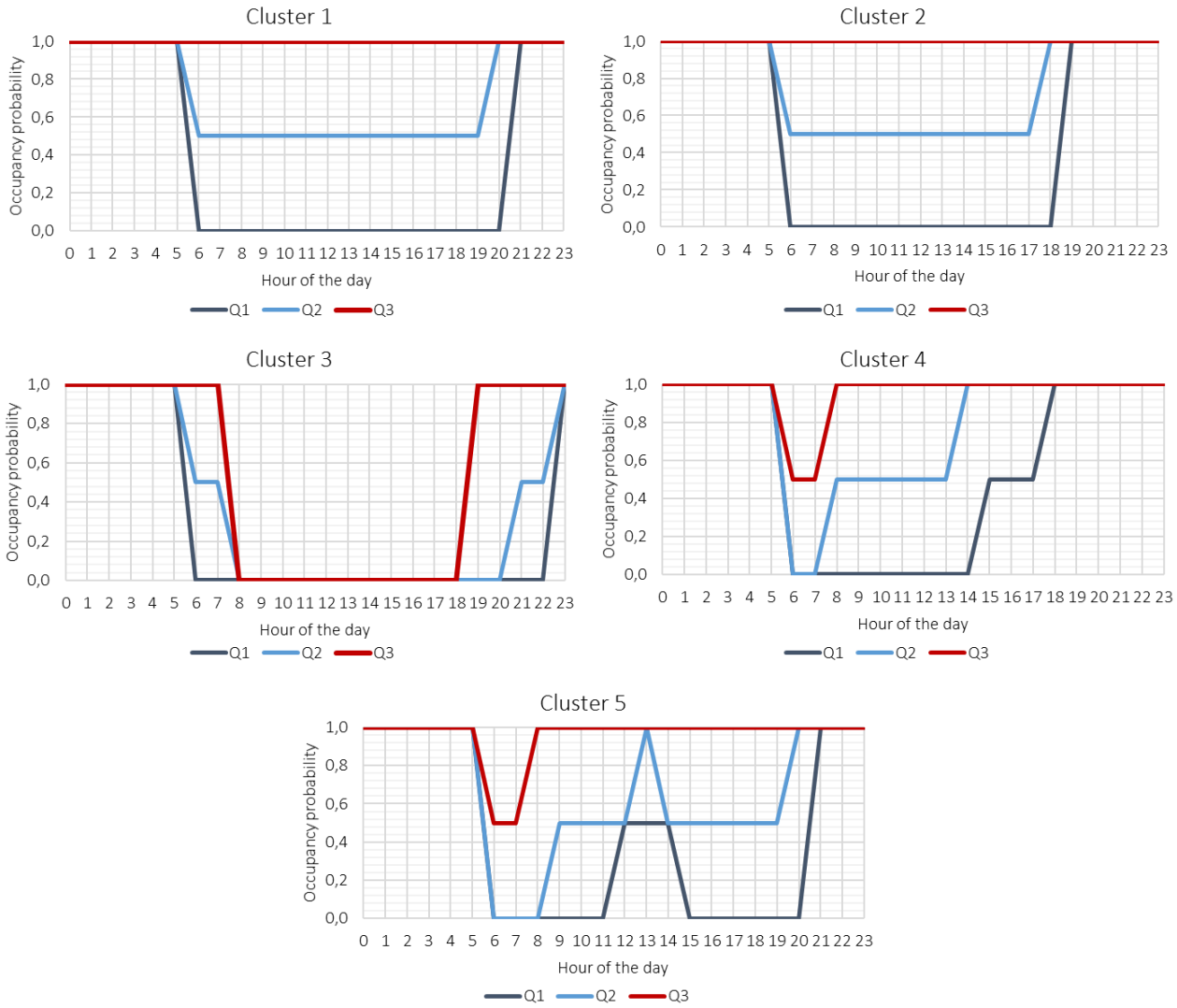

Figure 6: Three scenarios of occupancy probability in each cluster

In this last case, a decrease in the energy need to maintain thermal comfort is expected, since they refer to heating only. The average value of energy need for heating is 77,3 kWh/(m^2 a). The range given by the two extreme scenarios goes from a minimum of 74,7 kWh/(m^2 a) with the high electric energy users' schedules, to a maximum of 78,8 kWh/(m^2 a) with the low electric energy users' schedules. This variability corresponds to -2 % and +3 % of the average value. To understand the impact of the presence of the people in the flats, Case 2 (Figure 7) is run. In this case, in addition to the variation of internal heat 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 average value of energy need for space heating is the same as before, corresponding to 77,3 kWh/(m^2 a). The range given by the two extreme scenarios is increased and it goes from a minimum of 71,0 kWh/(m^2 a) with the high electric energy users' schedules and high presence, to a maximum of 82,7 kWh/(m^2 a) with the low electric energy users' schedules and low presence. This variability corresponds to -7 % and +8 % of the average value.

The aim of this task is also to compare the obtained results with the registered energy need of 2016 as validation. Even if a comparison cannot be considered as a proper validation method, it gives a good estimate of the quality of all the modelling process, thus of the presented methodology. The original reference value is expressed in terms of delivered energy for space heating. Thus, an estimated global seasonal efficiency of 0,7, due to generation, distribution, emission and regulation, is used to compute the energy need for space heating, based on the characteristics of the existing heating systems described in the energy audit and evaluated via on-site surveys. The registered data, 74,1 kWh/(m^2 a), is not far from the modelled values (Figure 7), indicating that the overall modelling of the building is able to approximate satisfactorily the result, always considering that the energy modelling implies numerous hypothesis and simplifications. The good agreement between simulated and registered energy consumptions can be attributed to both an accurate estimation of the occupant-dependent input data and a careful creation of the numerical model of the building. Specifically, epistemic uncertainties related to the specification of technical features were reduced as much as possible

thanks to the use of detailed information obtained by an energy audit and data collected in several inspections carried out by expert and independent engineers. Aleatory uncertainty due to weather variation was reduced by creating a weather file using the actual weather data collected by a meteorological station located close to the site in the same period of the energy metering.

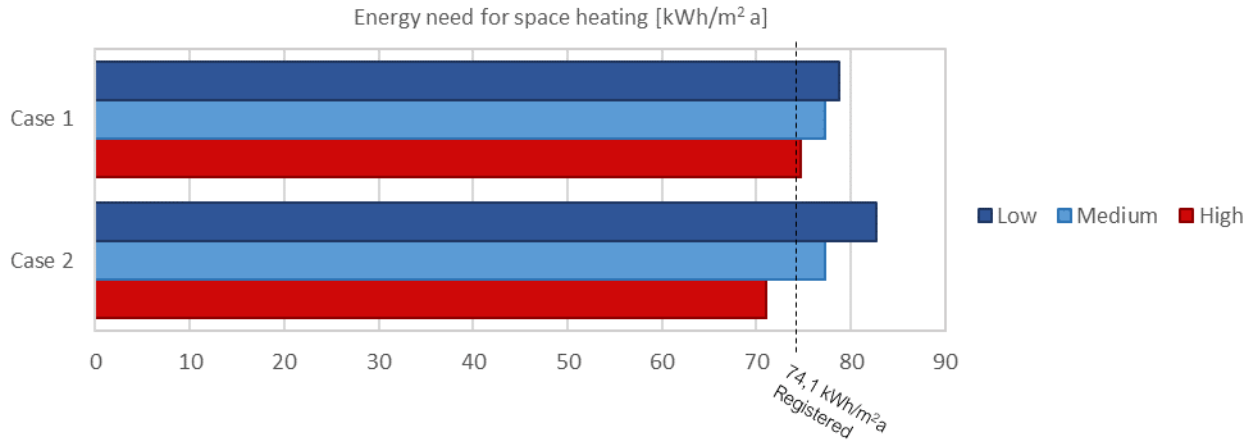

Figure 7: Result of the two cases compared to the registered data (Case 1: setting the medium schedules for occupancy, varying the electric use; Case 2: varying both electric use and occupancy schedules)

According to Figure 7, the actual behaviour of the majority of building occupants', in terms of electrical energy use and occupancy probability, seems to be medium to high.

5. Conclusions and future outlook

In this paper, a novel procedure that aims at improving the energy modelling of residential buildings is proposed. The procedure uses a few machine learning algorithms to extract information useful for generating occupant-related input schedules from the electricity recordings of smart meters. It is applied to and validated through a case study regarding a multi-residential building estate located in Milan, Italy.

The procedure is subdivided into three main phases. In the first phase, the implementation of the Self-Organizing Map (SOM) and the *k*-means algorithm was found appropriate for clustering purposes given the nature and complexity of the data sample. In particular, the two techniques are coupled for efficiently detecting representative electricity daily use profiles from actual electricity recordings. After the first run of analysis, five clusters emerged with different daily profiles that can be attributed to different types of user. Daily occupant-related load profiles were generated for each cluster. Afterwards, the *k*-Nearest Neighbour (*k*-NN) algorithm was implemented to extend the results to the whole year.

In the second phase, a classification method is proposed to estimate the occupancy in the apartments. However, occupancy estimation does not rely on data from occupancy sensors but is based on the analysis of the actual electricity recordings. The *k*-NN algorithm is used for the classification. Occupancy schedules are then generated and associated to all the apartments of the multi-residential building. The accuracy of this task, however, depends on the availability of ground truth from which the *k*-NN algorithm could have learnt. To this regard, a survey on occupancy presence would have been helpful to validate and improve the methodology. The resulted occupancy is associated with the five daily profiles assessed in the first phase that retrace the different type of user.

In the third phase, the schedules generated in the previous two steps are used to simulated the impact of the occupancy behaviour on the energy need for space heating. The result from the simulation is compared with the actual registered data, showing a range of variation, for heating, of about ±3 % changing only the internal heat gains due to electric appliances and of about ±8 % changing also occupancy schedules.

The procedure proposed in task 1, for the investigation of the electricity use profiles, may be valuable as an efficient use profile analysis in residential buildings. This type of buildings represents a peculiar case in terms of noise in the data sample, of the complexity of the variable and of privacy issues. In the residential sector, a vast amount of raw data getting available thanks to smart meters, might be processed, obtaining in-depth and useful information about electricity use behaviour. The approach used in task 2 exploits the electricity consumption registration as an occupancy sensor. Modellers, without the availability of ground truth data, can apply the proposed methodology to create occupancy schedules. The first advantage of the approach is the fact that the privacy of the tenants is respected since no invasive sensor is installed in the building and no personal information needs to be collected.

Furthermore, being able to understand how much occupants and their habits may impact on the energy need of a building is crucial for high-performance buildings. As a matter of fact, the occupancy behaviour can substantially change the result and, for this reason, the accomplishment or the failure of an energy target.

Finally, the obtained results can be helpful to different stakeholders, such as:

- modellers, who do not possess occupancy and internal heat gains schedules for their residential buildings models [10];
- tenants, who can benefit from the knowledge of good and bad behaviours to decrease their expenses and can benefit from targeted tariff plans;
- policymaker, who can optimize the energy targets considering occupants' uncertainty;
- facility managers whose aim is to develop strategies for energy savings due to the good management of resources;
- distribution system operators and transmission system operators, who can both exploit the identification of energy profiles for the management of the electricity grids and for the balance of the market;
- energy service companies involved in the building management that can exploit the information to optimize the energy savings measures.

The proposed methodology looks promising and with minor improvements could become an asset in the field. To further improve the results of the generation of standardized schedules of occupancy probability (task 2), a more detailed measurement of the electricity might be helpful. In addition, a ground truth of the presence of people can be beneficial for improving the accuracy of the *k*-NN algorithm. The procedure developed in the paper could be further extended to address other topics, such as ventilation, particularly natural ventilation (not considered in this case study for lack of data on the windows' openings).

Acknowledgements

The study was developed within the framework of the project SHAR-LLM (Sharing Cities), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 691895 and the EnergiX R&D project 269650 "*Utvikling av metoder og system for automatisk effektkontroll i bolig*". We thank our colleagues from *Sikom AS* and the *IEA EBC Annex 79 "Occupant-centric building design and operation"* who provided useful discussion points.

References

[1] Y. Peng, A. Rysanek, Z. Nagy, A. Schlüter, Using machine learning techniques for occupancy-prediction-based cooling control in office buildings, Appl. Energy. 211 (2018) 1343–1358. doi:10.1016/j.apenergy.2017.12.002.

[2] 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.

[3] X. Xu, P.J. Culligan, J.E. Taylor, Energy Saving Alignment Strategy: Achieving energy efficiency in urban buildings by matching occupant temperature preferences with a building's indoor thermal environment, Appl. Energy. 123 (2014) 209–219. doi:10.1016/j.apenergy.2014.02.039.

[4] 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.

[5] European Parliament and Council of 11 December 2018, Directive (EU) 2018/1999, 2018. doi:10.2903/j.efsa.2007.555.

[6] European Parliament and Council of 11 December 2018, Directive (EU) 2018/2002, 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018L2002&from=EN.

[7] European Parliament and Council of 11 December 2018, Directive (EU) 2018/844, 2018.

[8] European Parliament and Council of 11 December 2018, Directive (EU) 2018/2001, 2018.

[9] I. Gaetani, P.J. Hoes, J.L.M. Hensen, Estimating the influence of occupant behavior on building heating and cooling energy in one simulation run, Appl. Energy. 223 (2018) 159–171. doi:10.1016/j.apenergy.2018.03.108.

[10] W. O'Brien, I. Gaetani, S. Gilani, S. Carlucci, P.J. Hoes, J. Hensen, International survey on current occupant modelling approaches in building performance simulation, J. Build. Perform. Simul. 10 (2017) 653–671. doi:10.1080/19401493.2016.1243731.

[11] 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.

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

[13] 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.

[14] 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).

[15] 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.

[16] A. Wagner, W. O'Brien, B. Dong, Exploring Occupant Behavior in Buildings: Methods and Challenges, 2017.

[17] J.A. Díaz, M.J. Jiménez, Experimental assessment of room occupancy patterns in an office building. Comparison of different approaches based on CO2concentrations and computer power consumption, Appl. Energy. 199 (2017) 121–141. doi:10.1016/j.apenergy.2017.04.082.

[18] L. Pagliano, S. Carlucci, F. Causone, A. Moazami, G. Cattarin, Energy retrofit for a climate resilient child care centre, Energy Build. 127 (2016) 1117–1132. doi:10.1016/j.enbuild.2016.05.092.

[19] G.Y. Yun, K. Steemers, Behavioural, physical and socio-economic factors in household cooling energy consumption, Appl. Energy. 88 (2011) 2191–2200.

doi:10.1016/j.apenergy.2011.01.010.
[20] V.M. Barthelmes, C. Becchio, S.P. Corgnati, Occupant behavior lifestyles in a residential nearly zero energy building: Effect on energy use and thermal comfort, Sci. Technol. Built Environ. 22 (2016) 960–975. doi:10.1080/23744731.2016.1197758.
[21] K.U. Ahn, D.W. Kim, C.S. Park, P. de Wilde, Predictability of occupant presence and performance gap in building energy simulation, Appl. Energy. 208 (2017) 1639–1652. doi:10.1016/j.apenergy.2017.04.083.
[22] W. O'Brien, I. Gaetani, S. Carlucci, P.J. Hoes, J.L.M. Hensen, On occupant-centric building performance metrics, Build. Environ. 122 (2017) 373–385. doi:10.1016/j.buildenv.2017.06.028.
[23] 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.
[24] F. Causone, A. Tatti, M. Pietrobon, F. Zanghirella, L. Pagliano, Yearly operational performance of a nZEB in the Mediterranean climate, Energy Build. (2019). doi:10.1016/J.ENBUILD.2019.05.062.
[25] D. Yan, T. Hong, B. Dong, A. Mahdavi, S. D'Oca, I. Gaetani, X. Feng, 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.
[26] M. Schweiker, S. Carlucci, R.K. Andersen, B. Dong, W. O'Brien, Occupancy and Occupants' Actions, in: A. Wagner, W. O'Brien, B. Dong (Eds.), Explor. Occupant Behav. Build. Methods Challenges, Springer International Publishing, Cham, 2018: pp. 7–38. doi:10.1007/978-3-319-61464-9_2.
[27] A. Mahdavi, M. Taheri, An ontology for building monitoring, J. Build. Perform. Simul. 10 (2017) 499–508. doi:10.1080/19401493.2016.1243730.
[28] M.B. Kjærgaard, B. Dong, S. Carlucci, F.D. Salim, J. Yang, C.J. Andrews, O. Ardakanian, Data-driven Occupant Modeling Strategies and Digital Tools Enabled by IEA EBC Annex 79: Poster Abstract, in: Proc. 5th Conf. Syst. Built Environ., ACM, New York, NY, USA, 2018: pp. 188–189. doi:10.1145/3276774.3281015.
[29] I. Gaetani, P.-J. Hoes, J.L.M. Hensen, Introducing and testing a strategy for fit-for-purpose occupant behavior modeling in a simulation-aided building design process, IBPSA Build. Simul. Conf. (2017) 761–768.
[30] 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.
[31] 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.
[32] 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.
[33] K.A.D. Deshani, L.L. Hansen, M.D.T. Attygalle, A. Karunaratne, Improved Neural Network Prediction Performances of Electricity Demand: Modifying Inputs through Clustering, Second Int. Conf. Comput. Sci. Eng. (2014) 137–147. doi:10.5121/csit.2014.4412.
[34] 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.
[35] 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.
[36] A. Capozzoli, M. Savino, S. Brandi, D. Grassi, G. Chicco, Automated load patterns learning and diagnosis for enhancing energy management in smart buildings, Energy. 157 (2018) 336–352. doi:10.1016/j.energy.2018.05.127.
[37] 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.
[38] F. Jorissen, W. Boydens, L. Helsen, Simulation-based occupancy estimation in office buildings using CO 2 sensors, IBPSA Build. Simul. 2017. 2 (2017).
[39] S.H. Kim, H.J. Moon, Y.R. Yoon, Improved occupancy detection accuracy using PIR and door sensors for a smart thermostat, (2016) 2753–2758.
[40] N. Khalil, D. Benhaddou, O. Gnawali, J. Subhlok, Nonintrusive ultrasonic-based occupant identification for energy efficient smart building applications, Appl. Energy. 220 (2018) 814–828. doi:10.1016/j.apenergy.2018.03.018.
[41] 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.
[42] W. Kleiminger, C. Beckel, T. Staake, S. Santini, Occupancy Detection from Electricity Consumption Data, Proc. 5th ACM Work. Embed. Syst. Energy-Efficient Build. - BuildSys'13. (2013) 1–8. doi:10.1145/2528282.2528295.
[43] W. Kleiminger, C. Beckel, S. Santini, Household occupancy monitoring using electricity meters, Proc. 2015 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. - UbiComp '15. (2015) 975–986. doi:10.1145/2750858.2807538.
[44] E. Alpaydin, Introduction to machine learning, Third edit, 2014.
[45] F. Ferracuti, A. Fonti, L. Ciabattoni, S. Pizzuti, A. Arteconi, L. Helsen, G. Comodi, Data-driven models for short-term thermal behaviour prediction in real buildings, Appl. Energy. 204 (2017) 1375–1387. doi:10.1016/j.apenergy.2017.05.015.
[46] U. Ali, C. Buccella, C. Cecati, Households Electricity Consumption Analysis with Data Mining Techniques, Ind. Electron. Soc. , IECON 2016 - 42nd Annu. Conf. IEEE. (2016) 3966–3971. doi:10.1109/IECON.2016.7793118.
[47] 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.
[48] 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.
[49] J.L. Viegas, S.M. Vieira, J.M.C. Sousa, R. Melício, V.M.F. Mendes, Electricity demand profile prediction based on household characteristics, Int. Conf. Eur. Energy Mark. EEM. 2015-Augus (2015) 0–4. doi:10.1109/EEM.2015.7216746.
[50] J.L. Viegas, S.M. Vieira, J.M.C. Sousa, Fuzzy clustering and prediction of electricity demand based on household characteristics, Proc. 2015 Conf. Int. Fuzzy Syst. Assoc. Eur. Soc. Fuzzy Log. Technol. (2015). doi:10.2991/ifsa-eusflat-15.2015.147.
[51] H. Polinder, M. Schweiker, A. Van Der Aa, K. Schakib-Ekbatan, V. Fabi, R. Andersen, N. Morishita, C. Wang, S. Corgnati, P. Heiselberg, D. Yan, B. Olesen, T. Bednar, A. Wagner, Final Report Annex 53 - Occupant behavior and modeling (Separate Document Volume II), (2013) 153. http://www.iea-ebc.org/fileadmin/user_upload/images/Pictures/EBC_Annex_53_Appendix_Volume_2.pdf.
[52] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, SOM Toolbox for Matlab 5, 2000.
[53] S. Haykin, Neural Networks and Learning Machines, 2008. doi:978-0131471399.
[54] US-Doe, "Testing and validation" EnergyPlus Energy SImulation Software, (2014). http://apps1.eere.energy.gov/buildings/energyplus/energyplus_testing.cfm.
[55] R. and A.-C.E. Atlanta (GA), USA, American Society of Heating, ANSI/ASHRAE 140 - Standard Method of Test for the Evaluation of Building Energy Analysis Computer Programs, (2011).
[56] 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.
[57] A. Moazami, V.M. Nik, S. Carlucci, S. Geving, Impacts of future weather data typology on building energy performance – Investigating long-term patterns of climate change and extreme weather conditions, Appl. Energy. 238 (2019) 696–720. doi:10.1016/j.apenergy.2019.01.085.
[58] 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).
[59] J. Vesanto, E. Alhoniemi, Clustering of the self-organizing map, IEEE Trans. Neural Networks. 11 (2000) 586–600. doi:10.1109/72.846731.
[60] M. Tom, Chapter 4: Artificial Neural Networks, in: Mach. Learn., 1997.
[61] Piech Chris, K Means, (2013). http://stanford.edu/~cpiech/cs221/handouts/kmeans.html (accessed July 1, 2018).
[62] Comune di Milano, Regolamento edilizio, 2014.
[63] J. Mardaljevic, L. Heschong, E. Lee, Daylight metrics and energy savings, Light. Res. Technol. 41 (2009) 261–283. doi:10.1177/1477153509339703.
[64] 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.
[65] 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.

[66] R. Galvin, The Rebound Effect in Home Heating, (2016) 162.
[67] A. Capasso, W. Grattieri, R. Lamedica, A. Prudenzi, Bottom-up approach to residential load modeling, IEEE Trans. Power Syst. 9 (1994) 957–964. doi:10.1109/59.317650.
[68] 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.
[69] 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.
[70] 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.
[71] 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.
[72] 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.
[73] 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.
[74] D.J. Rumsey, Statistics For Dummies, 2nd Edition, 2016.