CrossMark

ORIGINAL ARTICLE

Influencing factors in energy use of housing blocks: a new methodology, based on clustering and energy simulations, for decision making in energy refurbishment projects

X. Cipriano · A. Vellido · J. Cipriano ·
J. Martí-Herrero · S. Danov

Abstract In recent years, big efforts have been dedicated to identify which are the factors with highest influence in the energy consumption of residential buildings. These factors include aspects such as weather dependence, user behaviour, socio-economic situation, type of the energy installations and typology of buildings. The high number of factors increases the complexity of analysis and leads to a lack of confidence in the results of the energy simulation analysis. This fact grows when we move one step up and perform global analysis of blocks of buildings. The aim of this study is to report a new methodology for the assessment of the energy performance of large groups of buildings when considering the real use of energy. We combine two clustering methods, Generative Topographic Mapping and k-means, to obtain reference dwellings that can be considered as representative of the different energy patterns and energy systems of the neighbourhood. Then, simulation of energy demand and indoor temperature against the monitored comfort conditions in a short period is performed to obtain end use load disaggregation. This methodology was applied in a district at Terrassa City (Spain), and six reference dwellings were selected. Results showed that the method was able to identify the main patterns and provide occupants with feasible recommendations so that they can make required decisions at neighbourhood level. Moreover, given that the proposed method is based on the comparison with

X. Cipriano (✉) · J. Martí-Herrero · S. Danov
Centre Internacional de Mètodes Numèrics en Enginyeria
(CIMNE), Building Energy and Environment Group, Edifici
GAIA (TR14), C/Rambla Sant Nebridi 22, 08222 Terrassa,
Barcelona, Spain
e-mail: xciprian@cimne.upc.edu

J. Martí-Herrero
e-mail: jaimemarti@cimne.upc.edu

S. Danov
e-mail: sdanov@cimne.upc.edu

A. Vellido
Computer Science, Universitat Politècnica de Catalunya (UPC
Barcelona Tech), Campus Nord UPC, 08034 Barcelona, Spain
e-mail: avellido@lsi.upc.edu

J. Cipriano
Centre Internacional de Mètodes Numèrics en Enginyeria
(CIMNE), Building Energy and Environment Group,
CIMNE-UdL Classroom, Pere de Cabrera s/n. Building, CREA.
O-ce 1.15, 25001 Lleida, Spain
e-mail: cipriano@cimne.upc.edu

J. Martí-Herrero
PROMETEO Researcher, Instituto Nacional de Eficiencia
Energética y Energías Renovables (INER), 6 de Diciembre
N33-32, Quito, Ecuador

similar buildings, it could motivate building occupants to implement community improvement actions, as well as to modify their behaviour.

Abbreviations

IEA-	International Energy Agency–Energy in
EBC	Building and Communities
EPBD	European Union Energy Performance of Buildings Directive
EUI	Energy use intensity

Highlights

- Energy audits, tenant's surveys and empirical tests in households are performed
- Normalisation and selection of relevant variables with respect to energy use are obtained
- Clustering of variables is carried out to characterise the different groups of dwellings
- Refinement of energy simulation of representative dwellings with monitoring data is presented
- Energy disaggregation and stock aggregation to the whole district are calculated
- Results serve to evaluate the energy current situation and related socio-economic impacts
- Potential impacts of energy-saving measures are finally presented

Introduction

In recent years, there has been a growing interest in understanding and analysing the real energy performance of buildings. This interest has been mainly driven by the evidence of the high variability in the energy consumption of buildings with very similar characteristics (IEA-CBCS Annex 33 2010; IEA-CBCS Annex 53 2013). This high variability, along with the lack of confidence in the estimation of real energy use, is at its most relevant, and sometimes a critical factor, in projects oriented at the neighbourhood level (IEA-CBCS Annex 33). At such level, the improvement in energy

efficiency is not only determined by urban and architectural aspects but also by the upgrading of installations and/or by changes in users' behaviour.

This variety of factors increases the complexity of analysis of the real energy use of buildings. This complexity can be partly explained by defining the influencing factors in energy use of buildings. In the IEA-EBC Annex 53, the main influencing factors of building energy consumption were proposed to fall into six categories: (1) climate, (2) building envelope, (3) building services and energy systems, (4) building operation and maintenance, (5) occupant activities and behaviour and (6) indoor environmental quality. The three first categories are related to variables influencing building energy performance, and as defined by the European Union Energy Performance of Buildings Directive (EPBD) (Directive 2010/31/EU 2010), they are usually calculated by fixing standard conditions for the other three categories, which are specifically related to actual building functions. As a consequence, the building energy performance is calculated, assuming that all of the analysed buildings operate under the same standardised functioning conditions, as outlined in IEA-EBC Annex 33 (2010). This approach allows a coherent comparison of the calculated building energy performance, but this calculation is not strictly related to the real energy consumption (IEA-CBCS Annex 53 2013). This type of calculation allows obtaining the asset rating energy performance indicators, in contrast to the operational ratings, which are based on measured energy use, often normalised for relevant variables like climate and level of energy service (Goldstein and Eley 2014). When the focus moves to the real use of buildings, all six categories of influencing factors must be taken into account. The influencing factors could be seen as driving forces for changing energy use and are of great relevance if we extend the analysis from an isolated building to a group of buildings or to a neighbourhood level.

A first approach of this new paradigm is shown in Li et al. (2014), where a portfolio analysis and detailed case studies of the energy use of 51 high-performance buildings around the world is performed. This study concluded that no single factor determines the actual energy performance of these types of buildings and adding multiple efficient technologies does not necessarily improve building energy performance; therefore, an integrated design approach that takes account of climate, technology, occupant behaviour, and operations and maintenance practices should be implemented to

maximise energy savings in HPBs. Several models for the integration of the influence of occupants' behaviour into building energy performance calculations have recently been proposed. A number of studies (Nakagami 1996; Lopes et al. 2005; Yu et al. 2011) suggest that an optimal approach to the quantification of the global effect of occupants' behaviour should be based on knowledge extraction from monitored data and from occupants' surveys rather than on improving theoretical building energy simulation models. Moreover, neighbourhoods or large group of buildings often yield less data (and are less frequently surveyed) than individual buildings. This fact increases the uncertainty of simulations due to the broad assumptions about input data that must be often relied upon. The results of the reviewed studies (IEA-CBCS Annex 33 2010; Yu et al. 2010; Yu et al. 2011) show that a combination of statistical analysis with prediction models (both heuristic simulation and inverse models), complemented in some cases with monitoring data analysis, can be a powerful tool for the development of energy urban actions aiming at reducing the energy consumption in existing buildings. According to this approach, descriptive statistics have been used to identify the most important factors and reference members of the set, by grouping the buildings/houses according to them (Räsänen et al. 2008; McLoughlin et al. 2012). The identification of factors may help the better implementation of subsequent steps of the simulation of the current situation and of energy improvement scenarios (Yu et al. 2011; Ueno et al. 2006). In the same way, results of a framework to model personalised occupancy profiles for representing occupants' long-term presence patterns presented in Yang and Becerik-Gerber (2014) show that the personalised occupancy profiles acquired through time-series modelling, pattern recognition modelling and stochastic process modelling outperform the fixed design profiles currently used in building energy simulations. A brief description of common bottom-up modelling techniques (statistical and building physics-based) can be found in Kavgic et al. (2010) and Murray et al. (2014). An example of statistical modelling is also described in Yu et al. (2010), where a decision tree method for building energy demand characterisation was proposed and applied to historical data from a sample of Japanese residential buildings. Taking the same statistical approach, some studies about classification of buildings according to the relevant factors and the different hourly profiles of users have been carried out. The user behaviour

in these studies is usually represented as time-based profiles or patterns. An example of data mining applied to investigate the behaviour of occupants adjusting their thermostat settings and heating system operations in a 62-unit affordable housing complex can be found in Ren et al. (2015). The results from this study affirm that data-mining techniques are an effective method to analyse large data sets and extract hidden patterns to inform design and improve operations. Decision trees were used to ascertain occupant behaviour patterns. As a general rule in this approach, clustering is used to group energy consumers of similar characteristics (Chicco et al. 2003; Chicco 2012); to predict future energy demand; or to detect atypical, usually undesired, behaviours (Räsänen et al. 2008; Tsekouras et al. 2008; Yang and Becerik-Gerber 2014; Li et al. 2010). Following with this approach, a framework combining statistical analysis with two data-mining techniques, clustering and association rules, was employed to identify occupant behaviour patterns of window opening and closing in a natural ventilated office building in Germany, using detailed time interval-measured building data (Oca and Hong 2014). Results showed that there were two typical window opening office user profiles, one mainly physical environmental driven and one mainly contextual driven. Conclusions showed that, although occupant behaviour seemed to be somewhat predictive and subject to the constraints or motivating factors of thermal comfort and time management, behavioural patterns are not only statistical relevant driver-response conditioning clusters but also incorporate the motivational dimension with typical window opening habits.

On the other hand, and considering only the building physic models at district level, the reported approaches generally include the energy calculation of a sample of houses considered to be representative of the neighbourhood/district/nation stock, as described in Swan et al. (2009). In some cases, simulation methods were used to conduct building energy consumption calculations, in order to identify the correlation between building energy consumption and different influencing factors (e.g. building relative compactness, building control strategies) (Ourghi et al. 2007).

However, simulation methods do not perform so well in simulating energy performance for occupied buildings as compared to non-occupied buildings, due to a lack of sufficient knowledge about occupant behaviour patterns, which are normally very difficult to parameterise. Moreover, the calibration of

building simulation programs against real conditions is a normally complex undertaking and the learning process is time-consuming (Yu et al. 2011). Complementary approaches which try to reduce the gap between simulation and real consumption are based in using data-mining techniques to identify patterns of use of usual occupancy schedules, or working profiles that can be used as input to current building energy modelling programs, such as Energy Plus or IDA-ICE, to investigate impact of occupant presence on design, operation and energy use in office buildings (Oca and Hong 2015). The main goal of the current study is the assessment of the energy performance of medium and large groups of buildings when aiming at eliciting common characteristics of building/dwelling typologies and the main factors influencing in their energy consumption. Improving the understanding of these influencing factors will allow us not only to improve the accuracy of prediction or classification methods but also to incorporate the socio-economic impact in decision making for urban refurbishment projects. In order to achieve this objective, the paper is structured as follows: first, the methodology is described as a combination of innovative and standard statistic methods for clustering (Generative Topographic Mapping and *k*-means), together with simulation tools that are employed to obtain realistic assumptions about user behaviour in the main representative groups of dwellings of a neighbourhood. These assumptions enable the estimation of present energy consumption at the level of individual properties as well as at the neighbourhood level. Then, this approach is implemented in a case study involving a district of the city of Terrassa (Spain), including an estimation of the potential impact of improvement measures. Finally, a discussion of the appropriateness of the approach is also provided.

Methodology

The analytical framework involves both quantitative and qualitative household information (inputs), the steps of the working process (process) and the results obtained at each step (outputs). The developed methodology is presented in Fig. 1; the quantitative data set comprises electricity and gas bills, complemented with electricity consumption and indoor temperature measurements over 15-min periods, as well as results of blowing door tests. The qualitative data set includes the household occupants' responses to surveys and interviews carried out by the researcher team.

**Fig. 1** Global view of the working process

Data acquisition and treatment

A study entitled "Diagnostic and analysis of energy improvements in low-income districts in Catalonia region: case study in Can Jofresa's neighbourhood (Terrassa)" was carried out by CIMNE from May 2008 to May 2010. The investigated neighbourhood is located in the city of Terrassa (Barcelona, population 215,517, as of 2014), in NE Spain, and consists of 12 H-shaped 15-story tower blocks (60 dwellings per tower, 720 households in total (see Fig. 2). The tower had the stairs and lifts in the centre. The apartments' distribution is always the same for the whole building.

For this project, the data was specifically collected following two phases of data acquisition. A first stage based on field surveys of energy-related data, monthly energy bills (gas and electricity) and equipment inventory (number of appliances with their nominal power), as well as an exhaustive list of the lighting systems, were carried out in those occupied dwellings; inhabitants accepted visits to fill surveys and to supply their energy bills (166 dwellings). The rationale behind these surveys was to obtain an overall vision of all the aspects of the household related to energy that can be parameterised (as indicator, numeric value or categorical value). There were defined 43 questions with predefined answers to choose. These 43 questions were classified in those sections: (i) general information about the household, (ii) economic data (incomes), (iii) energy and water consumption, (iii) data about perception of energy consumption, (iv) comfort and (v) time at home and use of appliances.

Fig. 2 General site view (*left*) and pilot tower detail (*right*)
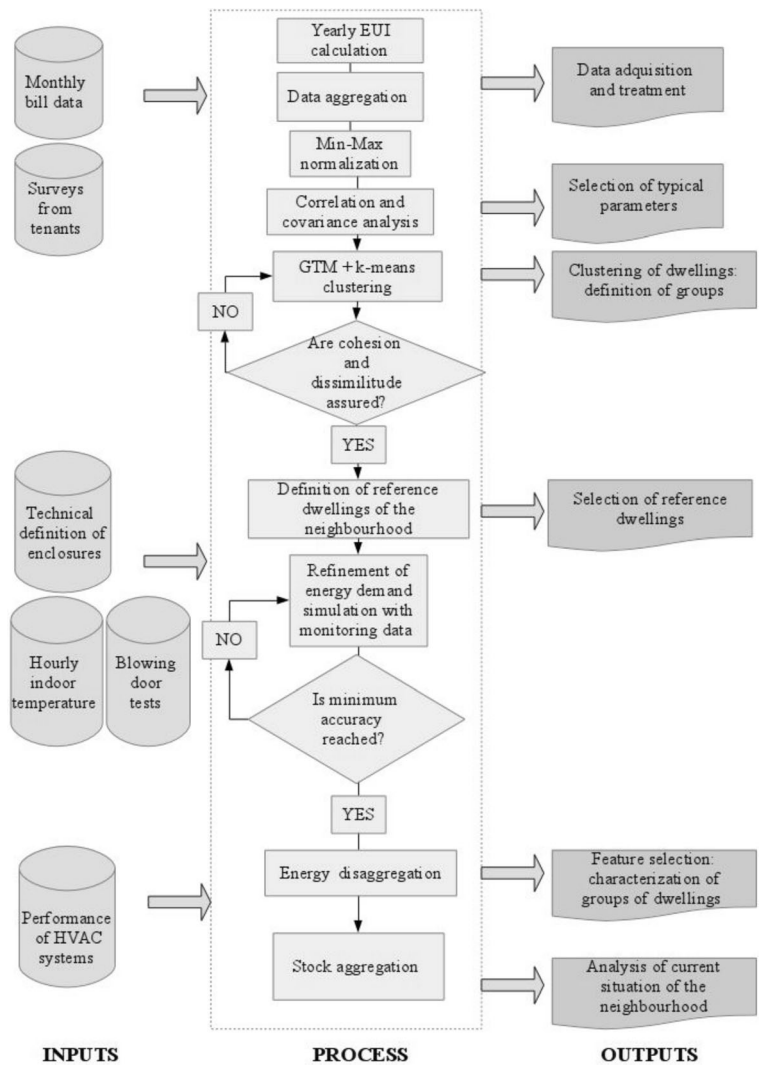
In parallel, some field measurements to estimate U values of walls, windows, thermal bridges and doors were obtained from ad hoc tests and measures performed in one of the tower blocks. Different procedures were carried out depending on the construction element type: (a) for external walls (façades), the construction was analysed by means of drilling a hole in an unoccupied ground-floor apartment; (b) regarding the internal floors, ground floors and roof, it was assumed that they consisted of the typical construction used in the 70s for this kind of social housing project; (c) regarding the internal walls and partitions, and internal and external doors, a specific construction was assumed after an external visual checking; (d) infiltration rates of the whole house were measured by means of a blowing door test, in four occupied dwellings with the different types of windows; and (e) the location and surface of thermal bridges were identified by means of an infrared camera (and high U value was assigned to this surface of the wall).

A second step of data acquisition was performed into the six reference dwellings. In this second phase, the monitoring of indoor temperature (measured in one bedroom and in the living room), electricity half hourly consumption (monitoring equipment installed next to the meter) and outdoor temperature (meteorological station in the roof of

the building) was carried for two periods of the year (3 months both in summer and in winter). Table 1 shows the survey items and corresponding investigation methods.

Data reduction and aggregation were then performed to obtain a more parsimonious representation of the original data. Normalisation of the yearly energy consumption per unit of surface (kWh/m^2 · yr) was applied (called energy use intensity (EUI)). An aggregation was carried out in some of the surveyed items for a more clear understanding of the variables under analysis. For instance, questions related to the type of window frames, type of glass and degree of windows tightness were grouped in a categorisation of the quality of windows (1 = very poor, 2 = poor, 3 = good, 4 = very good). This process of related question grouping was also carried out for the categorical answers, in order to have a more understandable classification. The categories of variables in Table 2 were selected according to their relationship with the energy consumption. For instance, the type of heating is defined by their energy efficiency, so that 1 is the lowest and 4 is the highest. The same with windows, where "very poor" is value 1, corresponding to non-isolated glasses, and 4 is double glazed. At the same time, we decided to change the order of the numbers of categories

Table 1 Investigation items and methods

Method	Number dwellings	Survey items	Energy sources	Measuring time
Questionnaire survey	166	Lifestyle, utilisation of equipment, annual income and relevant socio-economic information	None	Once only
Inquiring survey and visit	166	Other issues, such as basic building information (type of windows, number appliances)	None	Once only
Monthly energy bills (last 24 months)	59	Monthly and yearly final energy consumptions	Electricity natural gas	Every month or 2 months (depending on the house) for the last 2 years
Field measurement -Hole in façades -Infrared camera -Blowing door test	1 tower, 6 dwellings	Estimation of U values U values and surface thermal bridges Infiltration rate ACH (1/h)		Once only
Monitoring data -Indoor temperature and humidity -Méteo station -Electric consumption	1 building 6 dwellings	Comfort conditions and reference indoor temperature (two rooms) Outdoor temperature Hourly electric consumption	None Electricity	Measured every 15 min. Two periods of 3 months (summer, winter) Measured every 15 min

for some variables in order to avoid the same linear dependency (from lower to upper) with energy consumption (EUI) for all variables. This is the case of variable X_9, where "never use AC", which is the most energy efficient option, is the lowest (see Table 2).

A close scrutiny of the data, starting by eliminating variables linked to questions with unclear answers, or those not directly related to energy consumption was carried out. Different data gap-filling procedures were applied in case of missing data depending of the type of variables: (i) for monthly energy consumption data, the gap filling procedure was to take the monthly consumption of the same month (or bimonth) of previous year as the first estimates, taking into account that in Spain, it was usual at the time to measure every 2 months or even more; (ii) in case of categorical variables related to perception of energy consumption, comfort and use of appliances, missing data was filled with the option equal or equivalent to "do not know" or with the most usual answer, like "whole day" "no thermostat", "medium temperature", "do not care"

Table 2 List of variables selected in this study to analyse their influence in energy consumption

Code	Relevant	Name	Category	Range value	Value
X_1	Yes	Space heating	Heating use	Yes/no	(1-0)
X_2		Type of space heating	Heating use	Electric stove/gas stove/gas boiler/heat pump	(1 to 4)
X_3		Type of windows	Thermal comfort	Very poor/poor/ normal/good	(1 to 4)
X_4	Yes	Number of months heating	Thermal comfort and heating use	Number (month)	(0 to 6)
X_5	Yes	Heating schedules	Thermal comfort and heating use	Little/morning-afternoon/lunch-dinner/afternoon/night/all day	(1 to 6)
X_6	Yes	Degree of comfort	Thermal comfort	Very low, low, medium, high	(1 to 4)
X_7	Yes	Number of rooms unheated	Thermal comfort	All/all bedrooms/one bedroom/only dining room/only kitchen/none	(1 to 6)
X_8		Air conditioning (AC)	Cooling use	Yes/no	(1,0)
X_9		Use of AC	Cooling use	Never/occasionally/few/ noon-night/always	(1 to 4)
X_{10}		Number of adults and children	Socio-economic	Number (person)	(1 to 6)
X_{11}	Yes	Total monthly income	Socio-economic	Number (€/month)	1160–6000
X_{12}	Yes	Degree of good practices in heating	User behaviour	Little awareness/normal awareness/high awareness	(1 to 3)
X_{13}	Yes	Use of awnings	User behaviour	Much use/none	(1-0)
X_{14}		Type of cooking facilities	Kitchen	Gas/ceramic hob	(1-0)
X_{15}	Yes	Type of fridge	Appliances	Large/medium	(1-0)
X_{16}		Use of washer	User behaviour	Very inefficient/inefficient/normal/ efficient/very efficient	(1 to 5)
X_{17}	Yes	Number of appliances	Appliances	Number (units)	(0–27)
X_{18}	Yes	Switch off appliances by night	User behaviour	Yes/no	(1,0)
X_{19}		Number of energy-saving lamps	Lighting	Number (units)	(0 to 20)
X_{20}		Number of fluorescent tubes	Lighting	Number (units)	(0 to 5)
X_{21}		Type hot water system	Hot water	Electric storage, NG heater, butane gas heater	(1 to 3)
X_{22}	Yes	Shower time	Hot water	Number (minutes)	(5–40)
X_{23}		Decrease of consumption	Awareness	Yes/no	(1-0)
X_{24}	Yes	Number of appliances to switch off	User behaviour		
X_{25}		Time at home	Socio-economic	Morning, morning and night, evening, all day, no routine, lunch and dinner, night	(1 to 6)

and "no routine"; and (iii) for numerical values (i.e. economic data, number of people, data of the household), a second round of phone questionnaires was performed focused on this missing data. All household with gaps and with no second round of questionnaires were rejected. The scrutiny of the 166 surveys revealed that only 146 sets (from 166) of socio-economic data and 51 (from 59) sets of energy monthly data were complete. As a result of this process, a data set of 25 variables (from 43) for 146 dwellings was obtained (see Table 2).

Subsequently, a data transformation was applied to variables shown in Table 2 to deal with the differences in scale and in categories of the obtained data set. Specifically, *min–max* normalisation was performed to scale the values so that they fell within a predetermined range. This technique of linear normalisation has the advantage of preserving the relationships between the original data. In this study, the new range is defined as (0, 1).

For the second phase of data acquisition (only in six reference dwellings), we preceded to the estimation of the basic parameters for the energy simulation of the hourly thermal demand of those reference dwellings. Other parameters such as number of people, internal gains, use of night ventilation, comfort temperature and schedule time at home were estimated according to relevant variables of each cluster and then refined with the monitoring process. As outlined above, the adjustment of an overall infiltration rate for each dwelling was defined according to blowing door tests

Selection of relevant variables

After typification of the 25 initial variables, two approaches have been selected to obtain the relevance of the variables, first the relevance with respect to the EUI and second the relevance of correlations between variables. For the first item, the relevance corresponds to a weighting scheme that returns the squared value of the correlation between variables and the EUI as the attribute weight. In the correlation analysis of this project, correlation Pearson's coefficient is used to assess the strength of the linear relationships between variables and EUI. Correlation analysis in this case is performed using calculation of squares of correlations and so we can verbally describe the

strength of the correlation using the guidelines firstly proposed by Cohen (Cohen 1988) and modified by Evans (Evans 1996) that are widely use in behavioural science. Evans proposed "as a convention" for the absolute value of *r*, next guidelines (0–0.19) "very weak", 0.20–0.39 "weak", (0.40–0.59) "moderate", (0.60–0.79) "strong" and (0.80–1.0) "very strong". In our study, we are looking for a classification of dwellings based only on variables that have an impact in energy consumption, so that we are interested in variables showing at least a weak linear relationship with EUI; therefore, the selected threshold is less to 0.2 to reject variables.

The second approach is related to the sample size and the relationship between variables. It should be avoided to use an abundance of clustering variables with high correlation in the clustering analysis, since specific aspects covered by these variables will be overrepresented in the clustering solution (Mooi et al. 2014). In a subsequent step, the correlation matrix is calculated in order to quantify the rank correlations between variables. From previous projects, we can also affirm that absolute correlations with above 0.80 are always problematic, so that a threshold value of 0.8 was set as a minimum criterion to consider two variables as highly correlated. These two processes yield a selection of the most relevant (typical) variables (13 variables), as it can be seen in Table 2.

Data clustering with the Generative Topographic Mapping and *k*-means algorithms

Clustering is a process in which we aim to infer data grouping structure that is unknown beforehand. It is often used as an exploratory strategy that attempts to partition the data into groups that are internally homogeneous and different enough from other groups. Unlike in classification, no groups are predefined and there is no explicit modelling of the relationship between data and class labels. In this study, we combine two clustering methods, namely Generative Topographic Mapping (GTM) (Bishop et al. 1998) and the well-known *k*-means (Jain 2010), in the exploratory process of data clustering. GTM is chosen in this study because it simultaneously generates a data partition for clustering and an intuitive visualisation of the data in 2-D. It

is a probabilistic alternative to the well-known self-organising map (SOM) (Kohonen 2001), which has successfully been applied to energy use profiling. GTM is preferred to the more standard SOM in this study because its probabilistic definition ensures the convergence towards a minimum of a properly defined error function, as well as the adaptive estimation of the optimum values of some of its variables.

Formally, GTM is a non-linear latent variable model (Bishop 1998) of the manifold learning family, and as such, data are modelled through a low-dimensional manifold embedded in the data space. Such manifold is defined as a mesh whose knots are the centres of probability distributions (usually Gaussians) that become prototype representatives (cluster centres) of groups of data. In different variants, GTM has been used for missing data imputation (Vellido et al 2007, 2011), outlier detection (Tosi and Vellido 2013) or time series analysis (Tosi et al. 2014), as well as applied in areas such as medicine (Cruz and Vellido 2011) or e-learning (Etchells et al. 2006), amongst others.

The GTM is mathematically defined as a non-linear mapping from points u in a low-dimensional latent space (used for visualisation) onto the prototypes y residing in data space, with a functional form described as $y = \Phi(u)W$, where Φ is a set of $\mathbf{M}$ basis functions $\Phi(u) = \{\phi_1(u), \ldots, \phi_M(u)\}$ and $\mathbf{W}$ is a matrix of adaptive weights that defines a specific mapping. The probability distribution for data point $\mathbf{x}$ in a data space $X = \{x_1, \ldots, x_N\}$ with $x \in \mathfrak{R}^D$, being generated by a latent point u, is defined as an isotropic Gaussian noise distribution with common inverse variance β, from which the likelihood of the model can be defined. The adaptive parameters of the model (W, β) can then be estimated through maximum likelihood using, for instance, the expectation-maximisation (EM) algorithm (Dempster et al. 1977). In order to use GTM for visualisation, the relation between each data point $\mathbf{x}$ and each latent space point u is quantified as a conditional probability $p(u_k|x_n)$ and its calculation is a by-product of the maximisation step of EM. This probability is known as the *responsibility* r_{kn} of each latent point u_k for the generation of each data point x_n. Each data point x_n can therefore be visualised by its assignment to the location in the latent space (to the cluster), where the mode of the corresponding conditional probability is highest; that is, $\mathbf{u}_{\text{mod}}^n = \text{argmax}_{\mathbf{u}_k} r_{kn}$

This type of visualisation, known as *mode projection*, was used in the experiments reported in the "Experimental results" section. The fact that clustering is somehow subordinated to visualisation in GTM means that the resulting clustering solution is often too detailed for practical purposes. To overcome this limitation, the well-trodden k-means algorithm, which, as SOM, has been used for energy use profiling, was used to cluster the prototypes resulting from GTM. K-means aims to partition our GTM prototypes $\{\mathbf{u}_k\}_{k=1}^K$ into $L/L \leq K$ clusters $\mathbf{S}_i$, in such a way that the within-cluster sum of squares $\text{argmin}_{\mathbf{S}_i} \sum_{i=1}^L \sum_{\mathbf{u} \in \mathbf{S}_i} \|\mathbf{u} - \mathbf{\mu}_i\|^2$, where $\mathbf{\mu}_i$ are the second-stage cluster centres, is minimised.

This becomes, de facto, a two-stage clustering procedure; first, the original household data are clustered by GTM, which defines prototypes as cluster centres, and second, k-means clusters these prototypes to end up with a more parsimonious final cluster partition that can be interpreted in terms of the original data variables with the assistance of the GTM visualisation maps.

Definition of reference dwellings

The definition of reference dwellings is then carried out by identifying those which meet the following two main criteria: first, having the values of relevant variables closest to the values of the centroid of each cluster in terms of Euclidean distance and second, having the monthly EUI (kWh/m^2 · month) of gas and electricity closest to the median monthly value of each cluster, also in terms of Euclidean distance.

Thermal simulation and refinement of the reference dwellings

The methodology for this step of simulation and refinement was mainly based on the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) recommendations (AHSRAE 1999) for energy modelling and calibration of existing buildings. It was considered that following the ASHRAE's recommendations [8] was the best choice for the project in order to follow widely recognised procedures when working with building's energy-related issues. Nevertheless, the method did not follow exactly the mentioned suggestions for energy modelling and calibration, as it needed to be adapted to the specific characteristics of

this project (ASHRAE's recommendations are more focused in non-residential buildings).There was mainly two differences: (i) occupation schedule, in the present project, an occupancy schedule was first derived from results of data analysis of surveys (see "Definition of reference dwellings" section). Afterwards, when the monitored indoor temperature and hourly electrical energy consumption monitoring data was available, an accurate occupancy schedule was defined (see "Energy simulation and refinement of the reference buildings of each cluster" section), (ii) the project's parameter for model acceptance, in contrast to proposed in previous research (Pedrini et al. 2002) and in ASHRAE methodology, where calibration was made against monthly energy consumption; in the present project, the model calibration and the simulation is made against the indoor hourly temperature. This difference is mainly due to the fact that the calibration of the dwellings' existing systems was not our goal. In our case, it was intended to characterise the buildings' fabric, the occupancy and its related parameters of energy use, in order to obtain a more realistic description of the energy loads of the building. In other words, for this project, the dwellings' inside temperature and relative humidity were monitored and a meteorological station (measuring temperature, relative humidity and solar radiation) was installed on the building's roof. Our aim was to be able to use the hourly inside temperature as the parameter for model evaluation and calibration. Given the specific characteristics of the project, it was considered that this could be a more accurate analysis for model calibration. Nevertheless, ASHRAE's premises for model acceptance were taken as reference; simulation results with accuracy within 10 % of real measured data were considered adequate.

In order to check the real indoor conditions in the selected reference dwellings and to calculate their related heating and cooling demands, a thermal simulation was performed with Energy Plus software (Energy Plus 2009).The measured indoor temperature, electricity hourly consumption and outdoor temperature in the representative dwellings (see Table 1) were chosen as reference for refinement of the energy demand simulation. In the current project, two different dwellings, each with its specific characteristics, were firstly modelled, one of them having no heating or cooling systems (non-AC or non-air-conditioned residence or dwelling from now on) and the other one having both (AC residence or dwelling from now on). From these results of

calibration, four more dwellings corresponding to the reference dwellings were then modelled to obtain their heating and cooling loads. The procedure is based on four consecutive steps: (i) first, generation of the basic models for the two apartments with information from building design, ad hoc tests (see "Data acquisition and treatment" section) and results from surveys; (ii) monitoring and calibration of the non-AC residence through the collection of more detailed data from energy audits and indoor T monitoring temperature (two periods of monitoring); (iii) monitoring and calibration of the AC residence through the collection of more detailed data from energy audits and hourly indoor T and hourly electricity consumption; and (iv) then, to obtain the more realistic values of comfort conditions (T set point for heating and/or cooling) from these calibration and proceed by simulate for the whole year, the heating and cooling need to reach these comfort conditions.

Results of the different steps of the procedure are shown in "Energy simulation and refinement of the reference buildings of each cluster" section, where all parameters for simulations of the six reference dwellings are summarised. Some of these parameters, like U values, infiltration rate (ACH) and type of windows, were obtained from ad hoc tests and measures as explained above (see "Data acquisition and treatment" section). The adjustment of the other parameters such as natural ventilation rate, use of shading devices, indoor set point temperature, hours/day being at home, internal energy demand (number of electrical appliances, nominal power and artificial lighting) and use of heating and AC (in air-conditioned dwellings) were preliminary estimated according to relevant variables of each reference dwelling of each cluster and subsequently changed (iterative process) to obtain the minimum relative error between T_{indoor} and $T_{simulated}$ for the monitoring period (summer and winter).

Extrapolation of results for the entire neighbourhood

Extrapolation of results and calculation of disaggregated energy consumption for the whole district were carried out considering a tower of 60 dwellings with the same distribution of types of households as those obtained from the clustering procedure as a pilot. This pilot tower was considered to be an appropriate representative of the 12 towers of the district. A proportional aggregation according to the surface area of dwellings in each cluster was implemented to estimate the total energy demand of

the neighbourhood. We used the so-called weighting coefficient, which is the number of buildings of the stock which are represented by each archetype building, as presented in Mata et al. (2014).

Experimental results

Data collection and preprocessing

As explained in "Data acquisition and treatment" section, a close scrutiny of the data from the 166 surveys, for which 59 have a yearly period of monthly energy data, revealed that only 146 sets of socio-economic data and 51 sets of energy monthly data were complete. As previously explained, aggregation and data transformation of related questions from surveys were carried out, resulting on a data set of 25 variables for the 146 dwellings. Table 2 shows the complete list of variables extracted from the questionnaires.

Selection of influencing variables

Correlation between variables and EUI was calculated for the 51 samples with monthly bills, under the assumption that this result will apply to the rest of samples, as explained in "Data acquisition and treatment" section. The covariance matrix was then calculated over the remaining variables of the 146 dwellings. This resulted in a selection of 13 variables for clustering (marked with yes in Table 2).

From those variables (see Table 1 for coding), five are related to heating use and comfort (X_1, X_4, X_5, X_6 and X_7), two are related to electricity consumption (X_{15}, X_{17}), four to energy behaviour and awareness (X_{12}, X_{13}, X_{18} and X_{24}), one to the economic situation (X_{11}) and one to hot water consumption (X_{22}). Note that none of the variables are related to summer comfort, air conditioning or the kitchen. The reason for this is that only a few dwellings have air-conditioning systems (around 25 % of the total) and their use is low (as will be confirmed in the next section).

Two-stage clustering results

The selected data (13 relevant variables from 146 dwellings) underwent a two-stage, fully unsupervised, clustering process, in which GTM was first used to obtain a loose data partition into natural groups with a focus on exploratory visualisation.

For the experiments reported in this study, the visualisation grid of GTM latent nodes was fixed to square layouts of 10×10 nodes (i.e. 100 constrained mixture components) Figure 3 (right) shows the 2-D representation of the 146 13-D points on the GTM 2-D visualisation space, according to their *mode projection* as described in "Data clustering with the GTM and *k*-means algorithms" section. Each square corresponds to one of the latent points in the 10×10 grid, and its relative size corresponds to the ratio of cases (dwellings) assigned to that point. The different square sizes and the empty spaces in some areas are a clear indication that the analysed data have some intrinsic cluster structure. The *k*-means algorithm was then applied to the obtained prototypes (the functional images of the latent points) in order to further group the different visualisation regions into a specific number of clusters.

The adequate number of clusters must be estimated according to some criterion. In our experiments, we used the *silhouette* index, which provides a succinct graphical representation of how well each data item (a dwelling in our experiments) lies within its cluster. It was first described by Rousseeuw (1987), and a value near 1 indicates the simultaneous near-maximum levels of intra-cluster homogeneity and inter-cluster heterogeneity, whereas a value near −1 indicates the corresponding near-minimum ones. This index is a way to quantitatively compare different clustering solutions, that is, partitions of the same data that differ in the number of clusters, with the best solution yielding the highest silhouette value. In the case of our experiments, a value of 0.3 was the highest and it corresponds to a six-cluster solution, which becomes our partition of choice. In real data sets of relatively small size (in the sense that only a small number of cases are available; in our case, only 146), clusters are bound to be, on average, smaller and less internally homogeneous, with individual atypical data or outliers having a higher impact on the index than when large data samples are available. For all these reasons, we reckon that a value of 0.3 is acceptable. The distribution of colours in the map of Fig. 3 (right) reflects the results of this second stage of the clustering process. Importantly in terms of usability and data consistency assessment, the six cluster solution is shown to partition the data in mostly self-contained independent map areas with minimal cluster overlapping and very little cluster discontinuity.

Reference maps and feature-based interpretation

Visualisation is of little use unless we make it interpretable (Vellido et al. 2011). In order to make practical sense of these six clusters, an interpretation on the basis of the original 13 variables is needed. GTM provides such interpretation through the so-called *reference maps*, shown in Fig. 3 (left). Each reference map displays a variable's relative contribution over the representation map (and thus how it contributes to the clustering solution as a whole). We can do this because each of the 100 elements of these grids corresponds to one of the GTMs generated in the first-stage clusters and, as a result, to the reference vector (centroid) of each of these clusters. We can therefore visualise the distribution of each of the elements of these vectors (which corresponds to each of the data variables) over the GTM reference map. Sometimes, the distribution of values of the variables over the reference maps at least partially explains the distribution of second-stage clusters, so that the latter can be summarily described according to the former.

The reference maps are coded in grey scale, from black (lowest values) to white (highest values), allowing a straightforward interpretation. It can be seen, for instance, that the reference map relative to variable X_1 (space heating), which reflects whether the household has central space heating system (yes) or stoves (no), is neatly partitioned vertically according to low/high values. Its correspondence with the GTM map in Fig. 3

(right) reveals that almost all households with only stoves (electric or gas) are located on the left-hand side of the map, which corresponds strictly to cluster 3 (in yellow). This cluster also seems to be neatly characterised by low values (black colour on the left) of X_4 and X_5 (*months of heating* and *heating schedule*), and therefore, variable X_1 dominates the first level of the partition. Similar exploratory interpretations can be carried for other clusters using the reference maps.

Through visual interpretation, different variables seem to contribute in different degrees to explaining the final cluster partition in Fig. 3.

Characterisation of groups of households

Beyond visual exploration, we would like to quantify the specificity of the clusters according to actual characteristics. For this, we could display, in the categorical variables, the percentage of households that exhibit a specific categorical value, whilst, for real-valued variables, we could display the quartile values of their distribution in each of the identified clusters. We select from the 13 relevant variables used for clustering those variables that are highly distributed over the clusters. The selected nine "representative variables" are X_2 = type of space heating, X_4 = number of month heating, X_5 = heating schedule, X_7 = number of rooms unheated, X_{11} = total month income, X_{13} = use of awnings in summer, X_{17} = number of appliances and X_{18} = switch off appliances by night (see
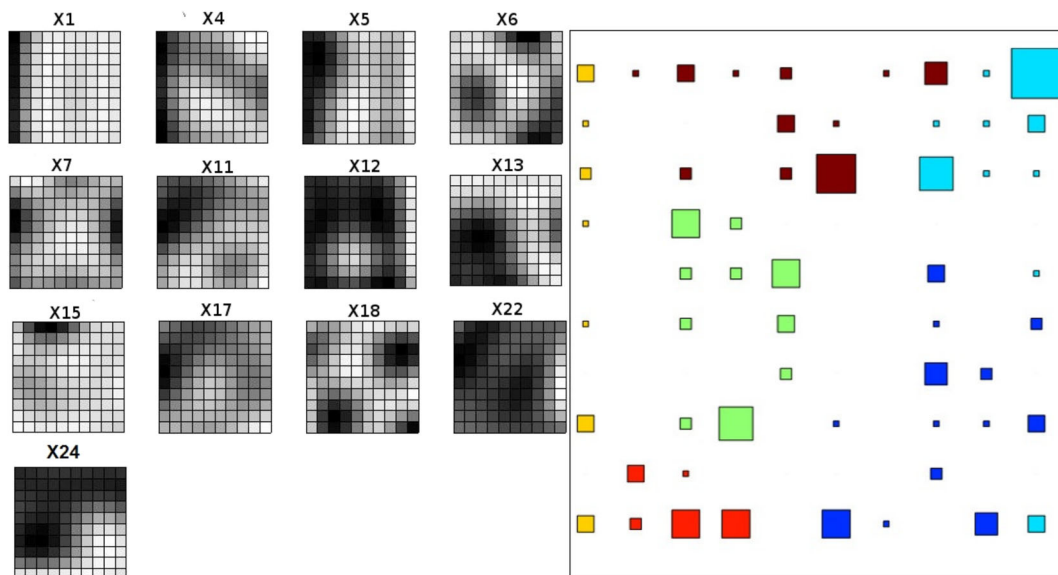

Fig. 3 **a**, *right* GTM 10 × 10 cluster map, *blue* cluster 0, *light blue* cluster 1, *green* cluster 2, *yellow* cluster 3, *red* cluster 4 and *brown* cluster 5. **b**, *left* Reference maps of single 13 variables

Fig. 4). It should be noted that, rather than selecting X_1, which can only discriminate those households without centralised space heating, we have selected X_2 as representative variable (it is in fact highly correlated to X_1) because households without centralised space heating will be determinant in making decisions about energy improvements at building level due to their low comfort, consumption and low incomes (they represent 10 % of the whole analysed sample).

Beyond the subset of relevant variables, there are others that are unevenly distributed over the different clusters. Due to this different distribution over groups, they could also be considered when trying to understand the characteristics of groups. This second set of variables could be denoted as complementary features of the different groups of households and might be considered as "complementary variables". It includes X_3 = quality of windows, X_8 = type of AC, X_9 = use of AC, X_{10} = number of people and X_{26} = time at home (see Fig. 5, top). Additionally, the distribution of the yearly aggregated monthly bills (gas and electricity collected in some of the households) can be obtained for each cluster (Fig. 5, bottom). A description of the main categories of the representative and complementary variables, according to the analysis of graphics in Figs. 4 and 5, together with the mean values of energy consumption, income and energy costs for each cluster are shown in Table 3. This table can be seen as a characterisation of the groups

of households. In this table, we can see that 12 variables have finally been selected (both representative and complementary) as the main representative to characterise the different clusters. In the practical description of groups, we have also included the number of people living at home but only to increase the description of each group.

In relation to energy consumption, Table 3 shows the mean consumption of each cluster as well as the standard deviation. We can affirm that households of people spending long periods at home (mainly elderly people) but in poor comfort conditions (cluster 5) and small families spending very limited time at home (cluster 3) are representative of the groups with lowest-energy consumption rates (both gas and electricity). The type of space heating in cluster 3 (only gas stoves) and the little use of heating and small number of rooms usually heated in cluster 5 (only bedrooms in the majority of dwellings), together with the small number of appliances (6 and 8, respectively, to cluster 3 and cluster 5), are the main influencing factors that explain these low-energy consumptions. In contrast, cluster 4 represent the highest-energy consumption. This group is mainly formed by families with more people (usually adults with children), many rooms heated (70 % of dwellings have only dining room and kitchen unheated), intensive use of heating (although average is 2 months of heating, the majority of dwellings use heating from half to all
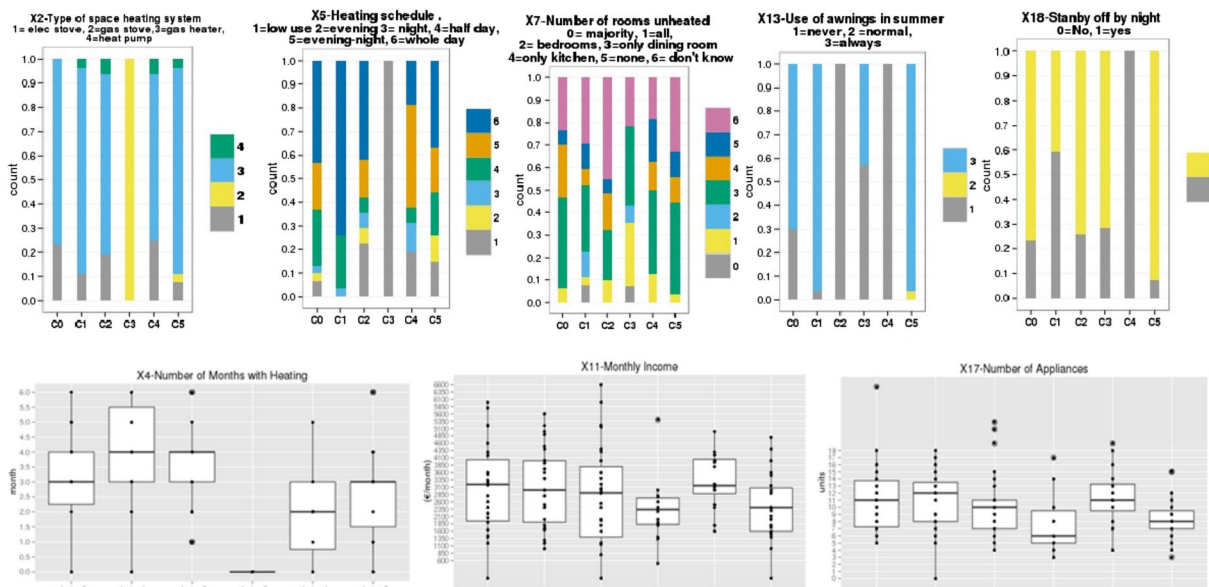

Fig. 4 *top* Distribution of relevant categorical variables for each cluster and *bottom* maximum, minimum, median and 25 and 75 % quartiles of continuous variables within each cluster

day), high number of appliances (12 units) and medium time spent at home

Dwellings in cluster 2, cluster 1 and cluster 0 show very similar gas and electricity consumption rates, due to the similarities type and use of heating system, time spent at home, income and comfort (number of rooms unheated). The small increase in energy consumption observed for cluster 2 over clusters 0 and 1 is due to small variations in combinations of these variables related to energy use (mainly use of heating, income and comfort).

The average income per family ranges from 2300 to 3.150€ per month (in year 2008), as it is also shown in Table 3. The influence of differences in monthly income on energy consumption is only important at the extremes (cluster 4 with highest consumption and income and cluster 3 with lowest), where this difference leads to different type and use of their heating energy systems, as well as different number of appliances. That is to say, dwellings in cluster 3 do not have central heating, and only have 6 appliances (in average) and dwellings in cluster 4 use intensively their central heating system (at least half day) and have 11 appliances. An important aspect that is also highlighted in Table 3 is that these high differences in energy consumption are not reflected in same differences in energy cost, due to the structure of energy tariffs in Spain, where the fixed terms are very high (especially in electricity). All groups spent around 3 % of their monthly

income in energy consumption. However, this percentage may increase up to 11 % for cluster 3 and cluster 5 in months of high-energy consumption (winter).

Differences of energy consumption within clusters

In order to examine the variability in annual gas and electricity consumption within each cluster, the yearly EUI of gas and electricity was normalised and plotted (see Fig. 6). The normalisation is based on dividing all the EUI of dwellings by the median value in each cluster, thus highlighting the variability and allowing the EUI to be plotted together on the same scale. As shown in Fig. 6, a significant variability that ranges from close to zero to about one time upon the mean value is induced by some aspects of the user behaviour that have not been identified throw the questionnaires. Since the energy consumption (separately electricity and natural gas) in each building is normalised by the mean value of all the buildings in that cluster, the values range from zero to zero point five as many as the mean value was considered to be an insignificant variation. Accordingly, the threshold value for significant variation is defined as 0.5. Such important variability in electricity consumption implies that there still remains great potential for energy saving by improving occupant behaviour. Contrarily, considering the relatively narrow range of cluster 4 (gas consumption) and cluster 5
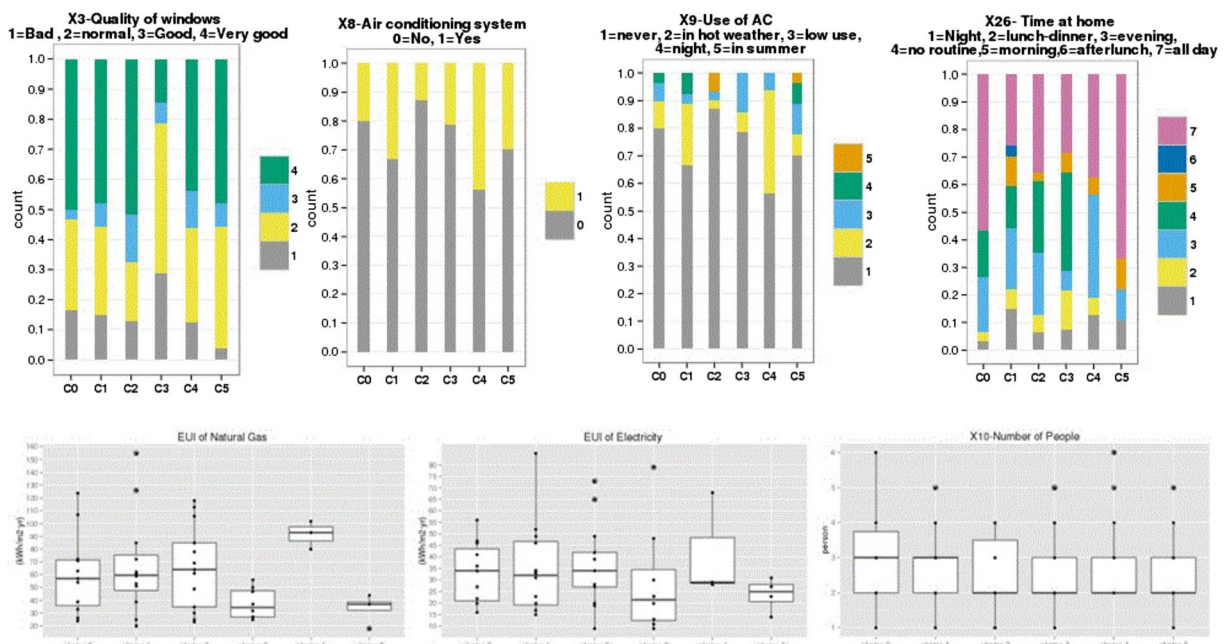

Fig. 5 *top* Distribution of complementary categorical variables within each cluster and *bottom* maximum, minimum, median and 25 and 75 % quartiles of EUI of gas and electricity

Table 3 Relevant variables, mean value and standard deviation of variables that define the different features of the groups of households (clusters)

| Cluster	Description of the majority value/category within the group	Income (€/month)	Natural gas (kWh/yr · m²)	Electricity (kWh/yr · m²)	Average energy costs (€/month)	Standby/off by night	Awning use
Variables	$X_2, X_{25}, X_7, X_9, X_4, X_5, X_{17}, X_8, X_3$	X_{11}	Mean ± SD	Mean ± SD	Mean ± SD	X_{18}	X_{13}
C_0	Majority central gas space heating (80 %, the rest with electric stoves). Long time at home (80 % all day or evening). Few rooms unheated (80 % only dining room, kitchen or none). Three to four members (middle-aged adults). Three months of heating and half-day to all-day use (70 %). Ten appliances (average). Low AC use (70 % never), mixed quality of windows (60 % good). Ten minutes of shower time (average)	3190 ± 1580	57 ± 32	34 ± 13	92.52	Mostly say yes (70 %)	Most of them say always (70 %)
C_1	Majority of central gas space heating (90 %, the rest with electric stoves). Medium time at home (60 % from evening to all day). Few rooms unheated (70 % only dining room, kitchen or none), 2–3 members. Longest heating period (4 months) with high use (75 % all day). Twelve appliances. Low AC use (90 % never or only hot weather), mixed quality of windows (60 % good). Ten minutes of shower time (average)	3000 ± 1327	60 ± 39	32 ± 20	91.55	Mixed (60 % say no)	Always (90 %)
C_2	Majority of central gas space heating (85 %, the rest with electric stoves). Medium time at home (70 % from half to all day). Three to four members (middle-aged adults). Few rooms unheated (80 % only dining room, kitchen or none). Long heating period (4 months) with medium use (60 % evening and all day). Nine appliances. No uses of AC. Good quality of windows (70 % good or very good). Ten minutes of shower time	2900 ± 1814	64 ± 33	34 ± 18	94.96	Mostly yes (70 %)	Never (100 %)
C_3	No heating, only gas stove (100 %). Little time at home (60 % few, lunch-dinner). Young couple or little family (2–3 members). Six appliances. No AC use. Bad or normal quality of windows (80 %). Six minutes of shower time (average)	2300 ± 1100	34 ± 12	21 ± 20	71.41	Mostly yes (65 %)	Mixed (50 % never; 50 % always)
C_4	Central heating (80 %). Medium time at home (60 % few time). Few rooms unheated (70 %). Three members. Short period of heating (2 month) but intensive (70 % half to all day). Eleven appliances. Almost no use of AC	3150 ± 950	93 ± 11	29 ± 22	100.03	Never	Never (100 %)
C_5	Central heating (95 %). Long time at home (90 %). Two to three members (family with elderly). Half of rooms unheated (70 % only bedrooms usually heated). Three-month heating, medium use (65 % from half to all day). Seven appliances. Some with AC (30 %) but low use	2400 ± 1100	37 ± 11	25 ± 17	76.48	Mostly yes (90 %)	Always (90 %)
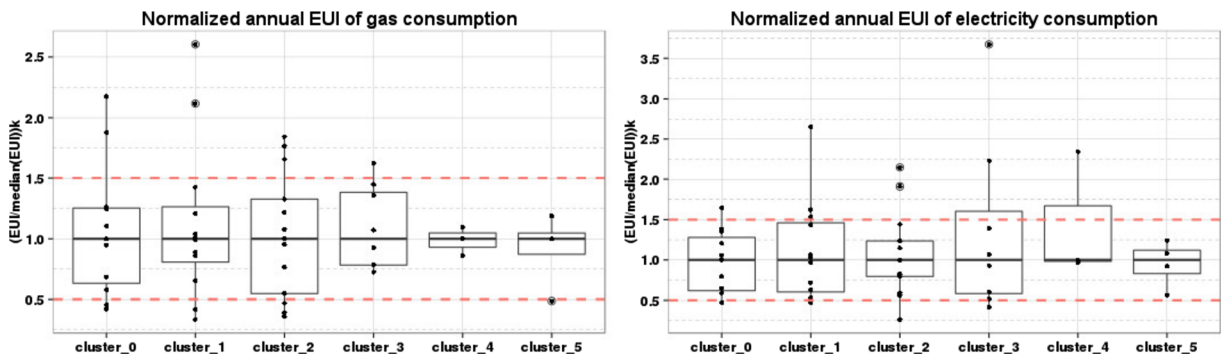
Fig. 6 Box plots of the normalised EUI of gas (*left*) and electricity (*right*) in each cluster

(both electricity and gas), there could be little expectation of reducing energy consumption in these clusters via improving occupant behaviour.

Selection of reference dwellings

The next step of the analysis consists on the characterisation of the reference dwellings that can represent each cluster. This task was carried out by selecting those dwellings that complied with the two criteria previously defined in "Definition of reference dwellings" section (minimum Euclidean distance of all relevant variables and monthly energy consumption to the centroid of each cluster). An example for cluster 0 of visualisation of the monthly EUI values of natural gas and electricity (separately), with their corresponding median value, is shown in Fig. 7. The dashed grey line corresponds to the median value, whilst the red line corresponds to the reference building. It can be seen that the monthly distance between the median and the selected reference dwelling is acceptable, as significant differences are identified only in 2 months. According to this procedure, the ID of the six selected reference dwellings, together with their related

energy consumption, were obtained (see Table 4). The results of corresponding variables of each reference dwelling are summarised in Table 5.

Energy simulation and refinement of the reference buildings of each cluster

As outlined above, the first step of simulation consisted in generating a first model of the two apartments (basic models of an air-conditioned and a non-air-conditioned dwellings, corresponding to cluster 3 and cluster 4 reference dwellings). The apartments' distribution was always the same. In the model, every zone represented one room in the residence in order to obtain accurate inside temperature values (see Fig. 8). The rest of the dwellings in the building were used only for shadowing simulation purposes, as well as the rest of the buildings in the district. In the base case, all walls between dwellings as well as ceilings and floors are considered to be adiabatic. The simulations are done firstly with free temperature mode, and then fixing standard criteria of T set point (21 °C in winter and 24 °C in summer), the rest of parameters like occupancy rates or internal gains are estimated according
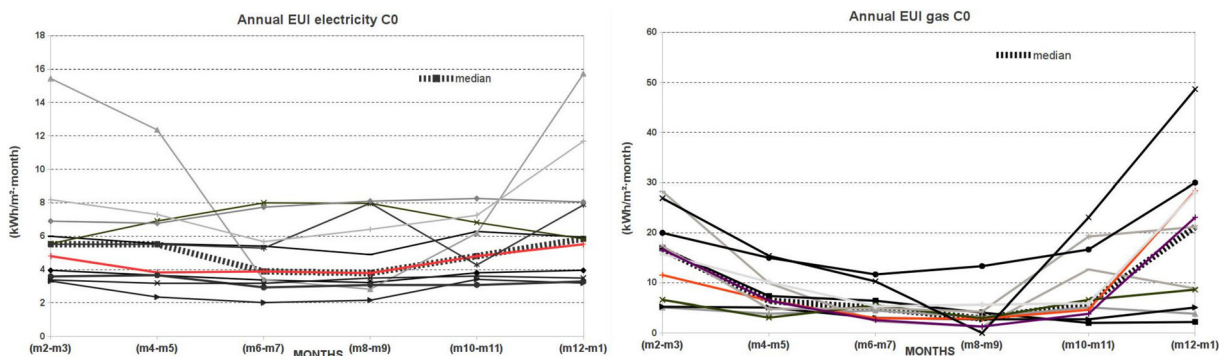

Fig. 7 Monthly gas and electricity consumption of households in cluster 0

Table 4 Yearly EUI of gas and electricity and energy thermal demand for cooling and heating of reference buildings

Group	Percent of dwellings	Surface (m^2)	ID reference dwelling	Gas consumption (kWh/yr · m^2)	Electricity consumption (kWh/yr · m^2)	Heating demand (kWh/m^2 · yr)[a]	Cooling demand (kWh/m^2 · yr)[b]
Cluster 0	22 %	89	38	57	27	38	0
Cluster 1	18 %	89	44	53	46	36	−16
Cluster 2	21 %	65	139	64	42	44	−10
Cluster 3	10 %	65	123	32	23	21	0
Cluster 4	11 %	89	32	93	68	63	-17
Cluster 5	19 %	65	127	37	27	24	0

[a] Heating demand is obtained by simulating with an ideal heating system that covers the defined equivalent winter T set point

[b] Cooling demand is obtained by simulating with an ideal AC system that covers the set point T when the household has air-conditioned system

to results of clustering or to usual values in Catalonia region. In the AC dwelling case, the relative error between the simulated and measured indoor temperatures (see "Thermal simulation and refinement of the reference dwellings" section) was higher than 10 % for the 46.85 % of the monitoring time. The non-AC residence's basic model also gave us an error bigger than 10 % for the

48.93 % of the total monitoring time. The big differences found between real and simulated inside temperatures showed that the models were not realistically representing the two studied dwellings. These results led us to step 2, which would provide us with more detailed information about the dwellings' inside conditions and its occupants' habits

Table 5 Input data values for the air and non-air-conditioned calibrated dwellings

| Description	Non-AC dwelling	AC dwelling
External walls (U value)	1.54 W/m^{2}K	1.54 W/m^{2}K
East Façade openings (double glazing; U value)	3.23 W/m^{2}K	3.23 W/m^{2}K
Internal façade openings (single glazing; U value)	6.14 W/m^{2}K	6.14 W/m^{2}K
Thermal bridges	4.51 W/m^{2}K	4.51 W/m^{2}K
Infiltration rates (empirical from blow door test)	0.40 ACH	0.40 ACH
Occupancy schedule (living room) adjusted with simulations. Values that lead to minimum error between T_{real} and $T_{simulated}$	11 months a year, 0.015 people/m^2 from 7:00 to 10:00 and 16:00 to 24:00	January to December, 0.02 people/m^2 from 8:30 to 9:30 and from 15:00 to 24:00
Ventilation schedule (whole dwelling) adjusted with simulations. Values that lead to minimum error between T_{real} and $T_{simulated}$	Winter, 0.6 ACH from 0 to 24 h. Spring and autumn, 1.5 ACH from 21 to 10 h. July, 3 ACH from 21 to 10 h. August, 1.5 ACH from 0 to 24 h	Winter, 0.5 ACH from 0 to 24 h. Spring and autumn, 0.9 ACH from 18:00 to 24:00. Summer, 2 ACH from 01:00 to 6:00
Lighting internal gains (living room) adjusted with simulations. Values that lead to minimum error between T_{real} and $T_{simulated}$	Winter, 2.4 W/m^2 from 7:00 to 9:00 and 4 W/m^2 from 17:30 to 24:00. Spring, summer and autumn (except August), 2.4 W/m^2 from 7:00 to 8:00 and 3.2 W/m^2 from 20:00 to 24:00	Winter, 2.4 from 7:00 to 9:00 and 4 from 17:30 to 24:00. Spring, summer and autumn (except August), 2.4 from 7:00 to 8:00 and 3.2 from 20:00 to 24:00
Equipment internal gains adjusted with simulations	45 W/m^2	45 W/m^2
Environmental control adjusted with simulations	No heated nor cooled, 24° by night, 26 °C by day (summer) and 15 °C by night, 18 °C by day (winter)	Summer, 25 °C (l3:00–21:00 h in living room and 22:00–24:00 h in bedroom). Winter, 21 °C (l3:00–21:00 h whole dwelling)
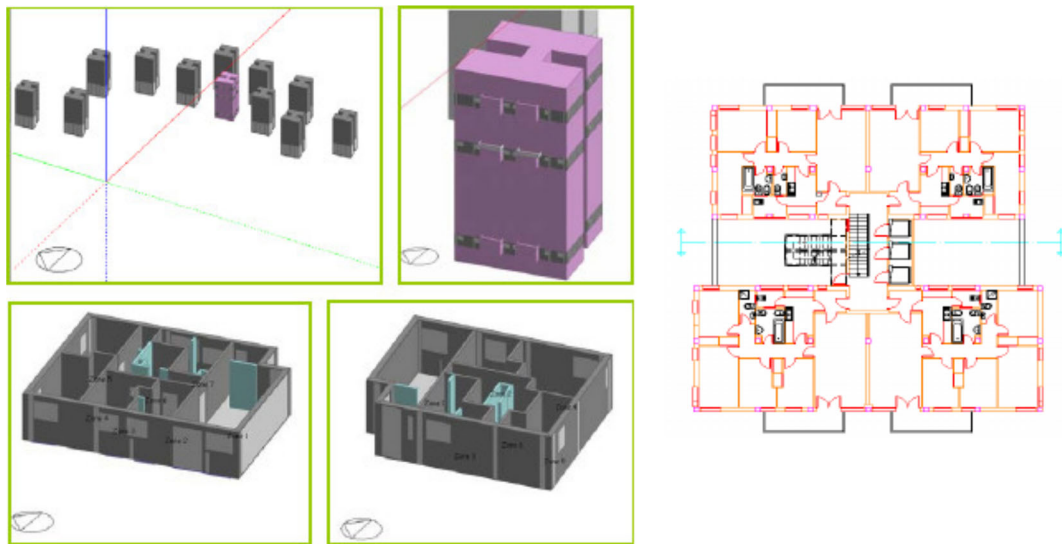
Fig. 8 Computer model. *top left* Complete district, *top right* complete building (15 stories, 60 dwellings) reference dwellings of each cluster in *grey shades*, *bottom left* WO dwelling and *bottom right* east oriented dwelling

In this second step, the adjustment of an overall infiltration rate for each dwelling was defined according to blowing door tests. For the adjustment of *night ventilation*, *time at home*, *internal gains* and cooling *heating period* definition, the indoor temperature was simulated (without HVAC system) against the real *indoor T* during the monitoring period. The assumption of adiabatic walls, ceilings and floors was also changed to only consider as adiabatic ceilings and floors.

Table 5 shows the values of the different input variables affecting the energy balance of this residence after model calibration.

Results for both simulated and real indoor temperature (dining room) for this dwelling (reference dwelling of cluster 3) in August is shown in Fig. 9. These results show that the relative error was above 10 % only the 0.94 % of the total monitoring time, which reaches the acceptable accuracy threshold in adjustment of an energy simulation model defined by AHSRAE (1999). This adjustment procedure was also applied in cluster 4 reference dwelling (air conditioned). In this case, dwelling's envelope characteristics were assumed to be reliable, as it was one of the outputs of the non-AC dwelling model calibration. For the same reasons, internal gains rates could be quickly assessed adapting results to the specific characteristics of the AC dwelling (artificial lighting, equipment, occupancy rates and natural

ventilation rates). The main changes that could not be deduced from the non-AC case, as the energy use profile is absolutely different, is the quantity (number of equipment and nominal power) and type of existing air condition and electrical appliances. Also, the occupants' goal in terms of comfort was going to be absolutely different. Therefore, it was necessary to analyse the available electricity monitoring data and take some conclusions. The analysis involved the comparison of hourly electricity consumption against the difference of temperature between $T_{external}$ and T_{indoor} (ΔT). There is a close relationship between the temperature variation and changes in consumption. When (ΔT) rises, electricity consumption rises by the same proportion and the same when it drops. This analysis provided detailed information regarding occupancy schedules and energy use (specially the use of the air-conditioning system) in the residence.

The input of the new data permitted a refinement of the basic AC dwelling model, adjusting the parameters mentioned above. Table 5 shows the values of the different modified input variables after model calibration. Results also showed an error above 10 % in only 1.94 % of total hours (see Fig. 9, bottom). Small variations of these conditions were selected for the rest of reference dwellings according to their related influencing variables (cluster 0, cluster 1, cluster 2 and cluster 5). An extended simulation based on results obtained from
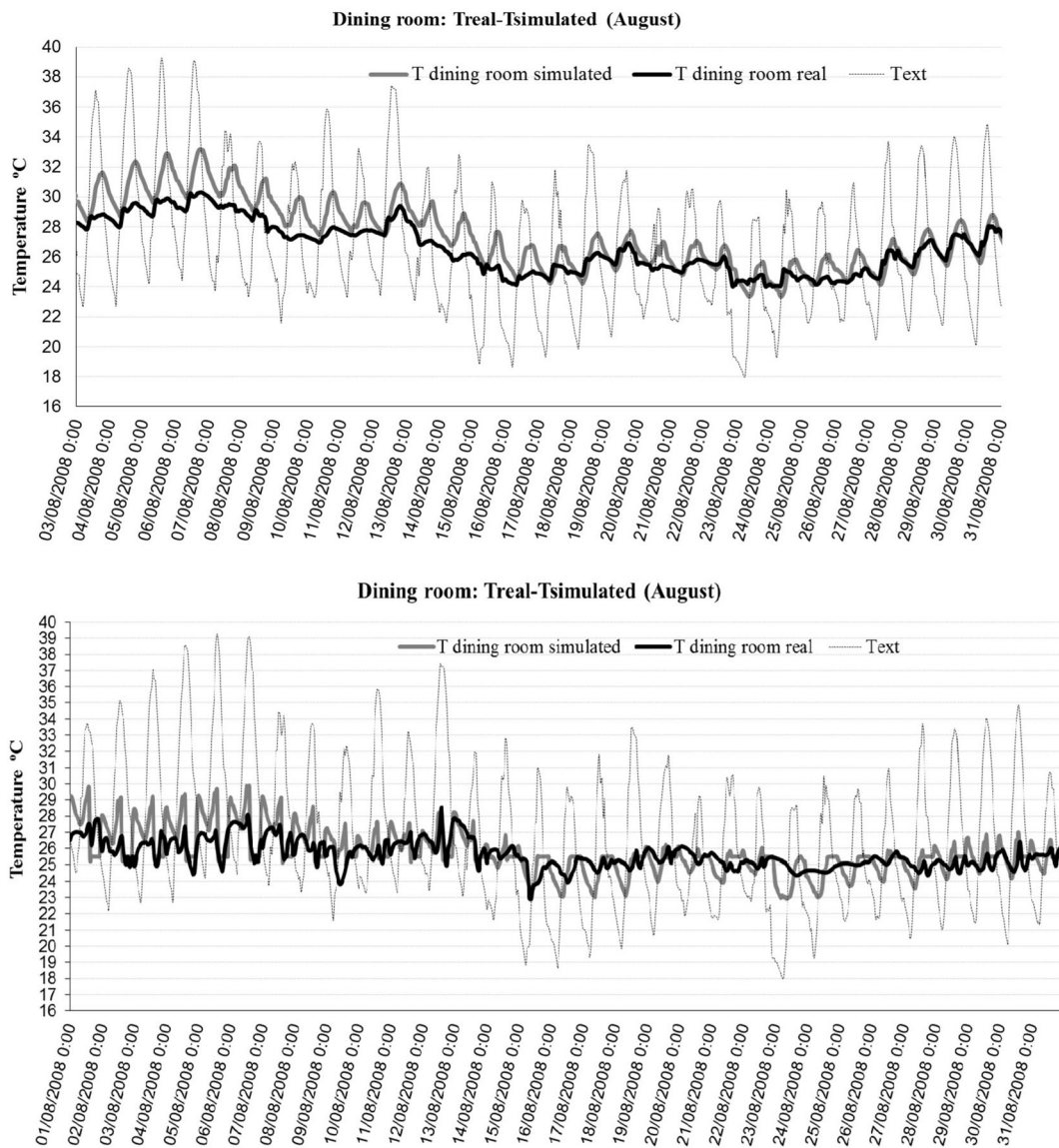
Fig. 9 Simulation results for indoor *T* for the cluster 3 reference non-air-conditioned dwelling in August (*top*) and the cluster 4 reference air-conditioned dwelling, also in August (*bottom*). Text = external temperature

the monitoring period was performed in order to obtain the energy demands for the whole year (also in winter).

Figure 9 shows the results of typical cluster 3 non-air-conditioned dwelling, where the hourly indoor temperature in August is around 26 °C during the day and around 24 °C at night (average in summer is 25.5 °C during the day and 23 °C during the night). In winter, the average of measured temperature is 18 °C for the complete day and 15 °C over night (these types of dwellings have butane gas or electric stoves as heating systems). Thermal energy demands for

dwellings considering these comfort conditions are 21.23 kWh/m² · yr for heating (see Table 4). Small variations in comparison to cluster 3 when simulating the cluster 5 reference dwelling were obtained. In the case of cluster 0, the higher heating demand in comparison to cluster 3 is due to the use of a centralised heating system that allows the inhabitants to get better thermal set points and comfort. Thermal energy demands for cluster 4 representative dwelling are 63.20 kWh/m² · yr for heating and −17.58 kWh/m² · yr for cooling, as shown in Table 4. Conditions of less

comfort, in comparison to cluster 4, specially in winter, were assumed for clusters 2 and 1.

Disaggregation of consumption and stock aggregation for the whole district

To obtain energy consumption and demand for the whole district, we assume that results obtained in the pilot tower may be extrapolated, after refinement, to the whole district. Taking the percentage of dwellings in each cluster together with their respective energy demands, the total energy demand of the tower was calculated. Then, some assumptions were made in order to disaggregate the gas and electricity consumptions; for the gas stoves and gas heater (centralised), performance were assumed to be, in turn, 75 and 79 %, and for split units, they were assumed to be 111.6 %, according to the official annex document of the national energy certification (Salmerón et al. 2009; CTE 1999). Results are shown in Table 6.

Conclusions

It can be concluded that the methodology used in this study is relatively simple and reliable, as intended clustering of data obtained from surveys in combination with refined simulation models allowed the evaluation of the current situation and of the impact of socioeconomic aspects and tenants' behaviour in a realistic way. The increase in time effort, as compared to simulations based on standard characterisation of dwellings,

is compensated by the higher quality of results in to the objective of understanding the real situation. The proposed methodology is useful in evaluating the current situation of energy consumption of the different types of families and dwellings, and in foreseeing the possibilities of implementing a real project of refurbishment in existing districts, where the different impacts in the different groups of tenants are a key factor in decision making.

Results of energy disaggregation are quite different than expected; 29 % of the residences consumed in similar patterns to the two non-AC dwellings (clusters 3 and 5; see Tables 4 and 6), that is, very low, both in terms of total consumption and in terms of the main component of consumption, gas for heating. These two groups were originally supposed to be one of the minor energy consumers in the district, in proportion, due to the fact that previous hypothesis assumed that buildings were badly isolated with old thermal systems, therefore with high heating consumption. These dwellings are often occupied by one or two elderly people (high time at home), little families or couples spending little time at home, with low incomes and sacrificing their thermal comfort in winter to not to spend too much money in heating their homes.

On the other hand, dwellings with AC and central heating system, with higher incomes and with a high number of appliances, represent only 11 % of the district (cluster 4). Dwellings with centralised space heating but normal conditions of comfort (around 55–60 % of hours per year with comfort as it can be seen in the yearly simulation results of indoor T in the reference dwellings)

Table 6 Aggregated results for the whole pilot tower and district

	Number dwellings	Percentage	Total consumption (MWh/yr)	Heating consumption (MWh/yr)	HW + kitchen (MWh/yr)	Cooling consumption (MWh/yr)	Appliances + light (MWh/yr)
C_0	13	22 %	94	54	10.2	0.0	30.3
C_1	11	18 %	100	47	7.1	14.7	32.0
C_2	12	21 %	83	44	6.4	7.5	25.5
C_3	6	10 %	22	11	1.5	0.0	9.0
C_4	7	11 %	95	47	7.9	9.3	30.6
C_5	11	19 %	48	23	5.1	0.0	20.2
Tower	60		442	226	38.2	31.5	147.6
Whole district			5304	2712	458	378	1771
Percent				51 %	9 %	7 %	33 %

Table 7 Simulation parameters that define the different reference buildings of each group of households

	T comfort cooling[6]*	T comfort heating[6]*	Windows quality equivalent U (W/m²·K)	Use of AC	Natural ventilation	Infiltration rate ACH*** (1/h)	Time at home	Number of people	Internal gains (W/m²)	Lighting internal gains (W/m²)	Heating period
C_0	23° by night, 26 °C by day	18 °C by night, 20 °C by day	50 % double glass (U = 3.24), 50 % single glass (U = 6.14)	No	Winter, 0.4 ACH from 0 to 24 h. Spring and autumn, 1.5ACH from 21 to 10 h. July, 3 ACH from 21 to 10 h.August, 1.5 ACH from 0 to 24 h	0.4	11 months a year, 0.015 people/m² from 7:00 to 10:00 and 16:00 to 24:00	3	45	Winter, 1.5 from 8:30 to 9:30 and from 17:30 to 24:00. Summer, 1.2 from 20:00 to 24:00	Evening, mid-lunch and breakfast
C_1	23° by night, 24 °C by day	17 °C by night, 19 °C by day	55 % double glass (U = 3.24, 45 % single glass (U = 6.14)	Medium	Winter, 0.4 ACH from 0 to 24 h. Spring and autumn, 0.9 ACH from 18:00 to 24:00. Summer, 2 ACH from 01:00 to 6:00	0.4	January to December, 0.02 people/m² from 8:30 to 9:30 and from 15:00 to 24:00	4	55	Winter, 1.5 from 8:30 to 9:30 and from 17:30 to 24:00. Summer, 1.2 from 20:00 to 24:00	Mid-lunch and evening
C_2	23° by night, 24.5 °C by day	18 °C by night, 21 °C by day	50 % double glass (U = 3.24), 50 % single glass (U = 6.14)	Low	Winter, 0.4 ACH from 0 to 24 h. Spring and autumn, 0.9 ACH from 18:00 to 24:00. Summer, 2 ACH from 01:00 to 6:00	0.4	11 months a year, 0.015 people/m² from 7:00 to 10:00 and 20:00 to 24:00	3	60	Winter, 1.5 from 8:30 to 9:30 and from 17:30 to 24:00. Summer, 1.2 from 20:00 to 24:00	Only in the evening (from 20 to 24 h in weekdays and from 18 to 24 h in weekend)
C_3	24° by night, 26 °C by day	15 °C by night, 18 °C by day	15 % double glass (U = 3.24, 85 % single glass (U = 6.14)	No	Winter, 0.6 ACH from 0 to 24 h. Spring and autumn, 1.5 ACH from 21 to 10 h. July, 3 ACH from 21 to 10 h. August, 1.5 ACH from 0 to 24 h	0.6	11 months a year, 0.015 people/m² from 7:00 to 10:00 and 20:00 to 24:00	2	45	Winter, 2.4 from 7:00 to 9:00 and 4 from 17:30 to 24:00. Spring, summer and autumn (except August), 2.4 from 7:00 to 8:00 and 3.2 from 20:00 to 24:00	Only in the evening (weekdays and from 18 to 24 h in weekend)
C_4	25 °C (13:00–21:00 h)	21 °C (13:00–21:00 h), 18 °C by night	45 % double glass (U = 3.24,55 % single glass (U = 6.14)	medium	Winter, 0.5 ACH from 0 to 24 h. Spring and autumn, 0.9 ACH from 18:00 to 24:00. Summer, 2 ACH from 01:00 to 6:00	0.5	January to December, 0.02 people/m² from 8:30 to 9:30 and from 15:00 to 24:00	3.5	65	Winter, 2.4 from 7:00 to 9:00 and 4 from 17:30 to 24:00. Spring, summer and autumn (except August), 2.4 from 7:00 to 8:00 and 3.2 from 20:00 to 24:00	Mid-lunch and evening
C_5	24° by night, 26 °C by day	15 °C by night, 19 °C by day	50 % double glass, 50 % single glass (U = 6.14)	no	Winter, 0.4 ACH from 0 to 24 h. Spring and autumn, 1.5 ACH from 21 to 10 h. July, 3 ACH from 21 to 10 h. August, 1.5 ACH from 0 to 24 h	0.4	September to July and September to December, 0.01 people/m² from 7:00 to 23:00	2.5	45	Winter, 2.4 from 7:00 to 9:00 and 4 from 17:30 to 24:00. Rest of the year (except August), 2.4 from 7:00 to 8:00 and 3.2 from 20:00 to 24:00	All day (from 8 to 24 h)

*6 T comfort cooling and heating is obtained from calibration of the thermal simulation

*** Obtained from the blower test held in 3 representative dwellings according to their quality of windows and door

are the main group of the district (cluster 1, cluster 2 and cluster 0 represent 61 % of the total). Many of them (45 % of dwellings) have AC system but with low use in summer. They spend middle-range time at home and middle-range incomes.

The reference consumption of gas for heating in Barcelona Metropolitan region is 62.3 kWh/m$^2 \cdot$ yr (according to the national energy code; CTE 1999). The consumption in 89 % of the analysed dwellings is lower than this reference value. However, in this minority of households that consume more, the difference is considerable up to 35 % of the CTE. In these dwellings, the greatest consumption is due to heating, but followed by electricity consumption due to light and appliances, which represent 29 to 41 % of total energy consumption (within cluster 4). These results also strongly correlate with the level of income of the families, since 36 % of difference in monthly income between clusters 3 and 4 exist, which is enough to be the cause of big changes in the type of space heating systems and in number and use of appliances.

Families in cluster 3 and cluster 5 are at the limit of energy poverty; they can neither reduce their consumption more nor invest in improvements. A majority of dwellings (cluster 1, cluster 2 and cluster 0) have low or medium consumption and assume a certain thermal discomfort, and a small group of families (cluster 4) reach high levels of consumption. Therefore, retrofitting measures offering improvements in heating efficiency or in improving isolation of envelope, although implies benefits in indoor comfort, mostly benefit families with higher consumption and have little impact on the poorest families, in terms of economic savings. The vast majority of families are somewhere between these two extremes. Therefore, any improvement that implies a certain level of investment must be designed considering finance mechanisms to help make costs affordable and offer subsidies for the poorest families. On the other hand, improvements in heating systems and in thermal comfort will lead to an increase of the energy consumption in households with less comfort, as explained by the Jevons' paradox phenomena (Giampietro et al. 2011), so any technology that will be implemented should offset the increase with a further increase in energy efficiency or renewable energy contribution.

In relation to cooling loads, we can affirm that in non-air-conditioned dwellings, thermal conditions in summer are not as bad as expected prior to analysis. In visits, even during some especially hot summer days, indoor conditions were found to be comfortable. Even though the building's envelope characteristics were in general rather bad (low insulation level) and the buildings are under high insulation rates in summer, it was found that the building's cooling loads in the dwelling are extremely low, as a consequence of climate's characteristics in the area, high natural ventilation potential, good orientation of buildings, and good use of canopies and other shading devices and low internal gains. Instead, high conditions of discomfort were found in winter. This is a very interesting finding, since there are a lot of similar social-housing districts in the country, all with similar shapes, surroundings (in the outskirts of cities, free of the obstructions and, as a result, exposed to the wind) and envelope's characteristics.

Complementary to this, we can see that, although the greatest consumption is due to heating (about 50 %), lighting and appliance consumption represents from 33 % of the total. This is a rather interesting finding, as many of energy-saving actions in these consumptions imply low investments, with good payback periods, which could led to important impacts in terms of economic savings in the majority of medium- and low-income dwellings.

In future research, other actions beyond traditional isolation systems, such as energy management control systems, boiler replacement, micro-generation systems, solar heat water systems or freezer replacement, will be considered to obtain improvements with low or even none economical cost (Table 7).

References

American Society of Heating, Refrigerating and Air-Conditioning Engineers. (ASHRAE). (1999). *Handbook of HVAC applications. Chapter 38* (pp. 8–9).

Bishop, C. M., Svensén, M., & Williams, C. K. I. (1998). GTM: the Generative Topographic Mapping. *Neural Computation, 10*(1), 215–234.

Bishop, C. M. (1998). Latent variable models. In *Learning in graphical models* (pp. 371–403). Netherlands: Springer.

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

Chicco, G., Napoli, R., Postolache, P., Scutariu, M., & Toader, C. (2003). Customer characterization options for improving the tariff offer. *IEEE Transactions on Power Systems, 18*, 381–387. doi:10.1109/TPWRS.2002.807085.

Código Técnico de la Edificación (CTE). (1999). *Ley 38/1999 de 5 de Noviembre, de Ordenación de la Edificación (LOE)*. Spain: Ministerio de Industria.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cruz, R., & Vellido, A. (2011). Semi-supervised analysis of human brain tumours from partially labelled MRS information, using manifold learning models. *International Journal of Neural Systems, 21*, 17–29.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1–38.

Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings. European Commission (2010).

EnergyPlus Energy Simulation Software (v6), US Department of Energy (2009). http://apps1.eere.energy.gov/buildings/energyplus/energyplus_about.cfm

Etchells, T. A., Nebot, A., Vellido, A., Lisboa, P. J. G., & Mugica, F. (2006). *Learning what is important: feature selection and rule extraction in a virtual course* (Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN 2006), Bruges, Belgium, pp. 401–406).

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove: Brooks/Cole Publishing.

Giampietro, M., Mayumi, K., & Sorman, A. H. (2011). *The metabolic pattern of societies: where economists fall short*. London: Routledge.

Goldstein, D. B., & Eley. (2014). A classification of building energy performance indices. *Energy Efficiency, 7*, 353–375. doi:10.1007/s12053-013-9248-0.

International Energy Agency Energy Conservation in Buildings and Community Systems Programme, (IEA-CBCS) (2010) Annex 33: energy related environmental impact in buildings.

International Energy Agency Energy Conservation in Buildings and Community Systems Programme, (IEA-CBCS) Annex 53 (2013). Total energy use in buildings: analysis & evaluation methods. Final report of ST-C—statistical analysis and prediction methods, Chapter 10—statistical analysis of total energy use.

Jain, A. K. (2010). Data clustering: 50 years beyond *k*-means. *Pattern Recognition Letters, 31*(8), 651–666.

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z., & Djurovic-Petrovic, M. (2010). A review of bottom-up building stock models for energy consumption in the residential sector. *Building and Environment, 45*, 1683–1697. doi:10.1016/j.buildenv.2010.01.021.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Berlin: Springer-Verlag.

Li, C., Hong, T., & Yan, D. (2014). An insight into actual energy use and its drivers in high-performance buildings. *Applied Energy, 131*, 394–410. http://doi.org/10.1016/j.apenergy.2014.06.032.

Li, X., Bowers, C. P., & Schnier, T. (2010). Classification of energy consumption in buildings with outlier detection. *IEEE Transactions on Industrial Electronics, 57*, 3639–3644. doi:10.1109/TIE.2009.2027926.

Lopes, L., Hokoi, S., Miura, H., & Shuhei, K. (2005). Energy efficiency and energy savings in Japanese residential buildings—research methodology and surveyed results. *Energy and Buildings, 37*, 698–706. doi:10.1016/j.enbuild.2004.09.019.

Mata, E., Sasic Kalagasidis, A., & Johnsson, F. (2014). Building-stock aggregation through archetype buildings: France, Germany, Spain and the UK. *Building and Environment, 81*, 270–282. doi:10.1016/j.buildenv.2014.06.013.

McLoughlin, F., Duffy, A., Conlon, M., (2012). Analysing domestic electricity smart metering data using self organising maps. Lisbon: CIRED Workshop.

Mooi, E., & Sarstedt, M. (2014). A concise guide to market research: the process, data, and methods using IBM SPSS. Springer Books. ISBN: 978-3-642-53964-0 (Print) 978-3-642-53965-7 (Online)

Murray, S. N., Walsh, B. P., Kelliher, D., & O'Sullivan, D. T. J. (2014). Multi-variable optimization of thermal energy efficiency retrofitting of buildings using static modelling and genetic algorithms—a case study. *Building and Environment, 75*, 98–107. doi:10.1016/j.buildenv.2014.01.011.

Nakagami, H. (1996). Lifestyle change and energy use in Japan: household equipment and energy consumption. *Energy, 21*, 1157–1167.

Oca, S. D., & Hong, T. (2014). A data-mining approach to discover patterns of window opening and closing behavior in offices. *Building and Environment, 82*, 726–739. http://doi.org/10.1016/j.buildenv.2014.10.021.

Oca, S. D., & Hong, T. (2015). Occupancy schedules learning process through a data mining framework. *Energy & Buildings, 88*, 395–408. http://doi.org/10.1016/j.enbuild.2014.11.065.

Ourghi, R., Al-Anzi, A., & Krarti, M. (2007). A simplified analysis method to predict the impact of shape on annual energy use for office buildings. *Energy Conversion and Management, 48*, 300–305. doi:10.1016/j.enconman.2006.04.011.

Pedrini, A., Westphal, F. S., & Lamberts, R. (2002). A methodology for building energy modelling and calibration in warm climates. *Building and environment, Elsevier, 37*(8-9), 903–912.

Räsänen, T., Ruskanen, J., & Kolehmainen, M. (2008). Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. *Applied Energy, 85*, 830–840. doi:10.1016/j.apenergy.2007.10.012.

Ren, X., Yan, D., & Hong, T. (2015). Data mining of space heating system performance in affordable housing. *Building and Environment, 89*, 1–13. http://doi.org/10.1016/j.buildenv.2015.02.009.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics, 20*, 53–65. doi:10.1016/0377-0427(87)90125-7.

Salmerón, J. M., Cerezuela, A., & Salmerón, R. (2009). *Escala de calificación energética para edificios existentes*. Cuadernos de Eficiencia Energética: Publicaciones IDAE, Madrid, Spain.

Swan, L., Ugursal, V. I., & Beausoleil-Morrison, I. (2009). *Implementation of a Canadian residential energy end-use model for assessing new technology impacts* (pp. 1429–1436). Glasgow: Proceedings of Building Simulation.

Tosi, A., Olier, I., & Vellido, A. (2014). Probability ridges and distortion flows: visualizing multivariate time series using a variational Bayesian manifold learning method. 10th Workshop on Self-Organizing Maps (WSOM 2014). *Advances in Intelligent Systems and Computing, 295*, 55–64.

Tosi, A., & Vellido, A. (2013). *Robust cartogram visualization of outliers in manifold learning* (Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013), Bruges, Belgium, pp. 555–560).

Tsekouras, G. J., Kotoulas, P. B., Tsirekis, C. D., Dialynas, E. N., & Hatziargyriou, N. D. (2008). A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. *Electric Power Systems Research, 78*, 1494–1510. doi:10.1016/j.epsr.2008.01.010.

Ueno, T., Sano, F., Saeki, O., & Tsuji, K. (2006). Effectiveness of an energy-consumption information system on energy savings in residential houses based on monitored data.
Applied Energy, 83, 166–183. doi:10.1016/j.apenergy.2005.02.002.

Vellido, A., Martí, E., Comas, J., Rodríguez-Roda, I., & Sabater, F. (2007). Exploring the ecological status of human altered streams through Generative Topographic Mapping. *Environmental Modelling & Software, 22*(7), 1053–1065.

Vellido, A., Martín, J. D., Rossi, F., & Lisboa, P. J. G. (2011). *Seeing is believing: the importance of visualization in real-world machine learning applications* (Procs. of the 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2011), pp. 219–226).

Yang, Z., & Becerik-Gerber, B. (2014). Modeling personalized occupancy profiles for representing long term patterns by using ambient context. *Building and Environment, 78*, 23–35. doi:10.1016/j.buildenv.2014.04.003.

Yu, Z., Fung, B. C. M., Haghighat, F., Yoshino, H., & Morofsky, E. (2011). A systematic procedure to study the influence of occupant behaviour on building energy consumption. *Energy and Buildings, 43*, 1409–1417. doi:10.1016/j.enbuild.2011.02.002.

Yu, Z., Haghighat, F., Fung, B. C. M., & Yoshino, H. (2010). A decision tree method for building energy demand modelling. *Energy and Buildings, 42*, 1637–1646. doi:10.1016/j.enbuild.2010.04.006.