8th International Conference on Sustainability in Energy and Buildings, SEB-16, 11-13 September 2016, Turin, ITALY

Clustering of household occupancy profiles for archetype building models

Giuseppina Buttitta[a], William Turner[a], Donal Finn[a,]

[a]*University College Dublin, Energy Institute, Dublin, Ireland*

Abstract

The continued penetration of renewable energy sources in electricity generation and the de-carbonization of the domestic space heating and hot water sectors is increasing the importance of demand side management (DSM). The development of end-use energy consumption models that can be easily integrated with electricity dispatch models is crucial for the assessment of the integration of supply and demand. The energy consumption of the domestic building stock is highly correlated with occupant behaviour, however the inclusion of occupant behaviour in energy models is challenging due to its highly variable nature. Nevertheless, in order to obtain reliable models of domestic energy consumption at high time resolution, the analysis of occupant behaviour patterns is fundamental. This paper aims to develop a new methodology to generate realistic occupancy patterns that can be representative of large numbers of households. This method is based on the clustering of household occupancy profiles using the UK 2000 Time Use Survey data as a case study. The occupancy profiles that result from this method can be used as input to residential building energy end-use models, thereby giving improved overall model performance.

Peer-review under responsibility of KES International.

Keywords: building model; occupancy profile; k-mode clustering

1. Introduction

Due to the increasing penetration of renewable energy sources in electricity generation portfolios and the de-carbonization of the domestic space heating and domestic hot water sectors, demand side management (DSM) is assuming a fundamental role in the development of new energy policies. The assessment of the integration between demand and supply of electric energy is fundamental and it can be achieved by integrating unit commitment and dispatch models with large scale simulation models of the building energy sector [1]. The simulation of building energy demand at large scale is challenging and it is approached in different ways by different researchers [2]. In general, two main methodologies can be identified: top-down and bottom-up. The top-down approach is based on the analysis of the whole residential sector energy consumption, and it does not distinguish individual end-users. In contrast, the bottom-up approach is based on knowledge of the energy consumption of each single building. For this reason, it is useful when the aim of the simulation model at large scale is the assessment of the contribution of each building towards the aggregate energy consumption of the domestic stock. If the model must be used for the assessment of DSM, it is important to monitor customer engagement at an individual level, especially when the electricity demand is modified. In this case, the choice of the bottom-up approach is essential.

The bottom-up approach can be sub-divided into a statistical approach and an engineering approach [2]. A key feature of engineering approaches is that they do not need any historical data, so they are preferred to the statistical approach in case of a lack of past information of building energy consumption. The engineering approaches include: distributions, archetypes and samples [2] techniques. Between these three techniques, the methodology that appears to assure a high degree of accuracy with lower computational effort is the archetype technique. It is based on the classification of buildings in groups with similar characteristics called archetypes, and on the simulation of these buildings instead of the overall buildings belonging to the building stock. Each archetype is defined by specific features into four main areas: form, envelope, system and operation [3] . The total energy end-use demand can be scaled up by multiplying the results of each archetype by the number of houses represented by each archetype.

In the case of residential buildings, it seems that while a considerable number of researchers have focused on the division of the building stock according to form, envelope and system characteristics of the building [4–8], just a few have focused on the relevance of occupancy patterns, which can define the archetype in the operation area. Richardson et al. [9] proposed a method based on the use of Monte Carlo simulations to produce random occupancy patterns based on the behaviour of large and heterogeneous populations. The profiles that can be obtained from this methodology are the results of combined occupancy profiles, and they cannot be used to characterise an archetype. Yao and Steemers [10] argued that the main characteristics influencing residential building energy consumption are the number of occupants and the length of the periods in which the houses are occupied. In their research, five fixed common occupancy profiles for UK households were proposed, based on authors experience. These profiles were used to characterise the archetypes in the operational area. This approach has a lower computational cost because it eliminates the necessity to implement a Monte Carlo simulation for each household, but it introduces errors due to the arbitrary choice of the occupancy scenario, which are not based on quantitative data.

The aim of the methodology presented in this paper is the development of characteristic occupancy profiles that can be applied to characterise archetypes considering the operational area. The profiles are obtained using a new approach based on a statistical clustering technique, which groups together the households with similar daily occupancy profiles. Once the characteristic occupancy profiles are identified, they can be coupled with the archetypes identified considering just the form, envelop and system characteristics, to create a complete archetype, with specific physical and also operational characteristics, as described in [3].

In this paper, the methodology, which is presented in Section 3, is applied on the data available from the UK 2000 Time Use Survey, which is described in the Section 2. The occupancy profiles that are obtained by applying the methodology to the UK building stock are shown in the Section 4.

2. Time-Use Survey data

The UK 2000 Time Use Survey [11] (UK TUS 2000), is a national survey which was conducted in UK in 2000, to record the everyday routine of 11700 UK citizens belonging to 6500 households. One household is defined as a person or group of people who have specified the accommodation as their only or main residence and either share at least one meal a day or share the living accommodation. The survey respondents occupied private houses which were evenly spread into the 5 regions of Great Britain: South East (excluding London) and South West; London; North West, North East and York/Humberside; Wales and Scotland; Northern Ireland. The respondents were asked to complete a household questionnaire, an individual questionnaire and two diaries, respectively one for a weekday and one other during the weekend, where the activities conducted were recorded every ten minutes.

The household questionnaire allows the determination of the number of members of each household, which defines the household size, and the type of accommodation in which they live. From these data it is possible to understand the breakdown of different size households by type of accommodation, which can be associated to a defined archetype. Figure 1 shows the percentage of the different accommodation types, identifies in the UK TUS 2000, occupied by households with different number of members. This figure, for example, shows that the percentage of accommodation occupied by one occupant is more than 60% (62.6%) for the flat/apartment, but it does not exceed 35% for the other accommodation types. Otherwise, the percentage of accommodations occupied by 4 people is very low when the flats are considered (3%), but it increases gradually from the terrace house (13.3%) to the detached house (18.2%).

The following methodology is based on the data contained in the diaries, where the daily activities of all the occupants aged above 8 are registered during a whole day. In particular, each respondent was asked to complete two diaries, one during a weekday and another during a weekend. The data were collected in the period between June 2000 and September 2001; in particular, all the members of the same household described the activities conducted on

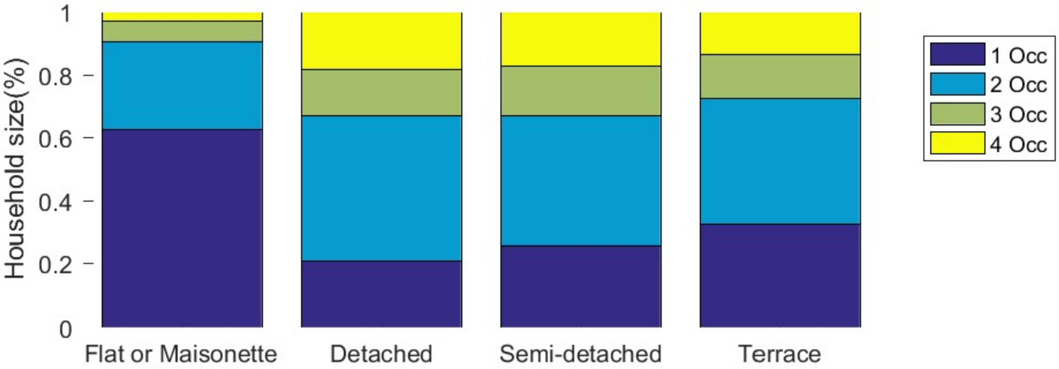

Fig. 1. Distribution of different size households in each accommodation type

the same day, and for each month a representative number of households was considered in each region. Each day was divided into 144 ten-minute time slots starting at 4:00 am. In each time slot the occupants indicated their primary and secondary activities, the location of the activity and whom the respondent was with. In this way, a complete database of the activities [12] from 4:00 am until 3:50 am the next day was created.

Unfortunately, in many cases not all the occupants of the same household completed the diary. For this reason, the data of some households were discarded, and just the data from 4588 households were used for weekdays and 4595 for weekends.

3. Methodology

The electric energy consumption of residential buildings is strongly influenced by the amount of time that the occupants spend at home [10], and by their state (awake or asleep) during this time. This is due to the fact that only when the occupants are home is it necessary to assure their thermal comfort in the house. Moreover, most of the electric appliances are used only when the occupants are awake. For this reason, it is important to identify three possible states of the occupants: home and awake, home and asleep, and away [13]. Other researchers [14], suggested that building electric energy consumption increases with the age of people, the number of occupants and their socio-economic and cultural state. In order to analyse how the interactions between occupants of the same household affects energy consumption, the household diaries of the UK TUS 2000 were classified according to the number of occupants. For the same reason, the state of each user must be correlated with the state of the users which belong to the same household. The combination of the state of each occupant of the same household determines the household state. The daily sequence of the household states defines the daily occupancy profile of the household. As a common behaviour, the occupancy profiles during the weekdays differ from the occupancy profiles during the weekend. For this reason, the household diaries were grouped also according to the day type.

After the grouping of the household diaries, the classification of the households can be performed by clustering the household occupancy profiles of each group according to their state, using the data available from a TUS (Time of Use Survey).

The overall methodology adopted can be divided in three steps: classification of the survey, clustering and determination of the number of clusters. The aim of the first step is to select a code for the identification of the state of each household for each time step, in order to create the objects about which the clustering can be performed. Then the k-modes clustering method is applied to cluster together all the households with similar states during the day. At the end the results are evaluated using the root-mean-squared standard deviation (RMSSTD) and R-squared (RS) coefficients. An overview of the methodology is given in Figure 2.

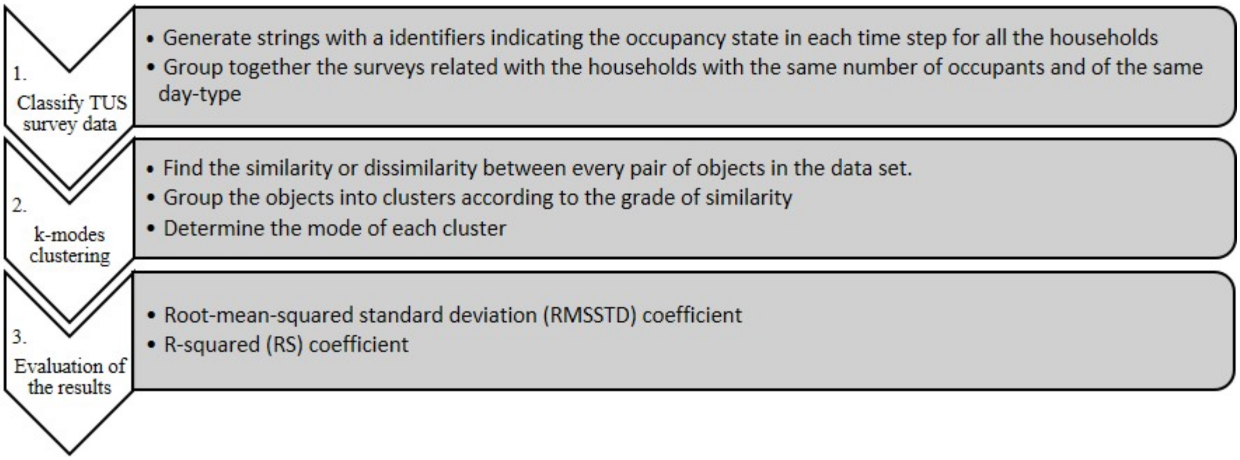

Fig. 2. Methodology

3.1. Classification of Time Use Survey - UK data

From the UK TUS 2000 the state of the occupants (home and awake, home and asleep and away) are available in each ten-minute time slot. The household state is given by the combination of the states of all the occupants of the same households. Consequentially, the number of states which can be assumed by the households is a function of the number of occupants. This can be calculated from Equation (1) [15], where r is the number of occupants and n is equal to the number of states which can be related to a single occupant, and it is equal to three.

$$n_{householdstates} = \binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!} \tag{1}$$

The number of states are 3, 6, 10, 15 for households with 1, 2, 3 and 4 occupants, respectively. Each household state is associated with an identifier. The sequence of the household state identifiers during one day defines a string which represents the daily occupancy profile of the household. In the case in which there are two occupants the identifiers associated with each state are reported in Table 1. Considering the daily household profiles of all the households with two occupants, it is possible to obtain the spectrum of the household occupancy profiles during one day, which is shown in Figure 3(a). The x-axis indicates the households, while the y-axis corresponds to the time of the day, which is based from 4:00 am to 4:00 am of the successive day. Each colour of the image identifies the household state in each ten-minute time slot. From the spectrum it is possible to observe that in the majority of the households the occupants are away from around 9:00 am to around 5:00 pm and this trend is confirmed by the mode of all the occupants with two occupants during the weekdays (Figure(3(b)).

Table 1. Identifiers used to categorise the state of the household in each time slot (Two occupants)

State of the household (Two occupants)	Identifier
Two occupants sleeping	2Sl
One occupant sleeping/One occupant home	1Sl,1Ho
One occupant sleeping/One occupant away	1Sl, 1Aw
Two occupants home	2Ho
One occupant away/One occupant home	1Aw, 1Ho
Two occupants away	2Aw

3.2. k-modes clustering

The aim of clustering is to discover groups of households with similar characteristics, i.e., households with similar occupancy profiles. There are several clustering approaches proposed in the literature [16]. The most appropriate

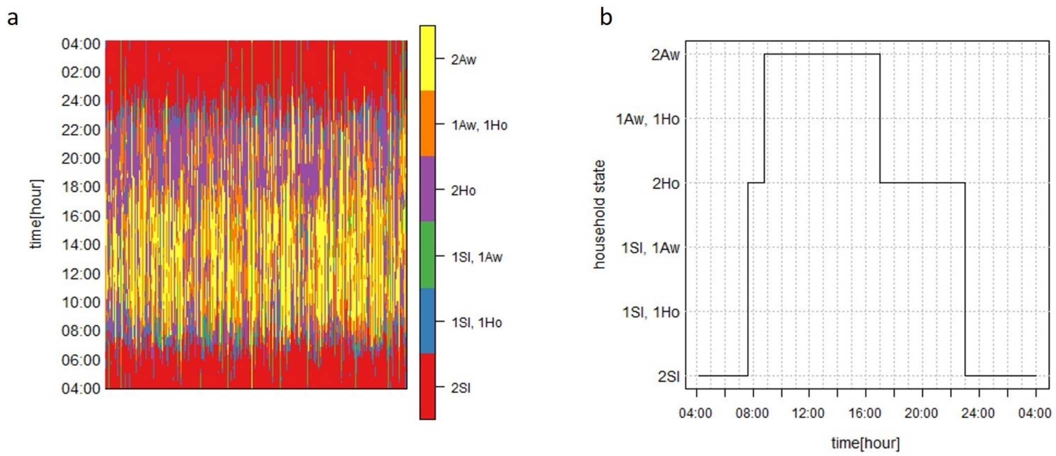

Fig. 3. (a) Spectrum of the occupancy profiles for the households with two occupants (weekdays); (b) Mode of the profile

method to use depends on the type of input data and objective of the analysis. In this paper, the inputs are the strings of all the households containing the identifiers of the household-state during one day. The string inputs are categorical variables rather than numerical. The number of algorithms that can be applied to these kind of variables is limited to two: the k-mode algorithm, which belongs to the partitional clustering category, and the ROCK algorithm, which belongs to the hierarchical clustering category [17]. The aim of the clustering in this work is to obtain a representative occupancy profile for each cluster, this means that the partitional method is the most appropriate and for this reason, the k-mode algorithm is used [18]. The aim of the algorithm is to find a set of modes $\{Q_1, Q_2, \ldots Q_k\}$ that can minimise the function (2).

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{i,l} d(X_i, Q_l) \tag{2}$$

where k is the number of clusters, n is the number of categorical objects in the data set $\mathbf{X}$, $X_i \in \mathbf{X}$,Q_l is the mode of the cluster l, and $y_{i,l}$ is an element of the partition matrix, which is a binary variable equal to one when the object X_i is allocated in the cluster C_l. The expression $d(X_i, Q_l)$ is the distance measure, which can be defined as the total mismatches between the mode of the cluster and the objects of the cluster [18].

The clustering is implemented using the statistical analysis package *R* [19], in particular the R function *kmodes* [20], which belongs to the package *klaR*, which performs k-modes clustering on categorical data. The algorithm of the function is initialised selecting k initial modes, one for each cluster. Then, each object is allocated to the cluster with the most similar mode, according to the value of the distance d. After each allocation the mode of the cluster is updated to minimise the distance between the categorical objects inside the cluster and the mode of the cluster.

When all the objects are allocated to clusters, the distance of all the objects against the current mode is calculated, and if an object is more similar to the mode of another cluster, it is reallocated and the mode of the clusters are updated again. This procedure is repeated until all the objects are assigned to the correct cluster. At the end of the implementation of the algorithm, the modes of each cluster and also the number of objects which belong to the same cluster are available.

3.3. Evaluation of the clustering results

The evaluation of non-hierarchical clustering results is usually performed through the use of two coefficients: the root-mean-squared standard deviation (RMSSTD) (3) and R-squared (4) (RS) [17]. The first index measures the non-homogeneity of the clusters, if its value decreases the obtained clusters are more homogeneous. RS is considered as the measure of the degree of difference between clusters. It is included in the range [0,1], and it is equal to zero when there is no difference between the groups, while when it is closer to one it means that there is a significant difference between the groups. Considering the partition C ={C_1,... C_l,... C_k} of the initial data set $\mathbf{X}$={X_1,... X_i,...,X_n}, with X_i={x_1,... x_j,... x_m}, where n are the elements of the initial data set and k is the number of cluster, and m the dimension of the space, the RMSSTD and the RS indexes are defined as:

$$RMSSTD = \left[\frac{\sum_{l=1}^{k} \sum_{X \in C_i} \sum_{j=1}^{m} \delta(\mathbf{x_j} - q_j)^2}{d(n-nc)} \right]^{1/2} \tag{3}$$

$$RS = \frac{\sum_{X \in \mathbf{X}} \sum_{j=1}^{m} \delta(\mathbf{x_j} - \overline{\mathbf{y_j}})^2 - \sum_{l=1}^{k} \sum_{X \in C_i} \sum_{j=1}^{m} \delta(\mathbf{x_j} - q_j)^2}{\sum_{X \in \mathbf{X}} \sum_{j=1}^{m} \delta(\mathbf{x_j} - \overline{\mathbf{y_j}})^2} \tag{4}$$

where $\overline{\mathbf{y_j}}$ is the mode of all the initial categorical objects and $\delta(\mathbf{x_j}, q_j)$and $\delta(\mathbf{x_j}, \overline{\mathbf{y_j}})$ are the distance measures between the objects of the cluster and the mode and the objects of the cluster and the mode of all the initial categorical objects.

Formulae (3) and (4) are an adapted version of the coefficients reported in [21], where all the objects are considered to have the same dimension. In this case if the value of the mode is not equal to the value of the data the spread between the two is assumed to be two to emphasise the difference between the data and the mode. This adaption was necessary because in this case categorical variables are considered, and the arithmetical difference between the numbers which identify the states in the mode and the data is meaningless. The most appropriate number of clusters corresponds to the one which determines a significant step in the value of the index. This value is indicated as significant knee [17], and it can be easily recognised plotting the validity indexes as a function of the number of the clusters.

4. Results and discussion

The aim of the paper is the generation of representative occupancy profiles that can be used to identify archetypes considering the amount of time spent at home, sleeping or away by all the occupants of the same household. The methodology presented above was applied to the data available in the UK TUS 2000. The number of clusters used for all the different households was determined using the RMSSTD and RS coefficients shown in the Table 2.

Table 2. Number of clusters

Number of occupants	Weekdays			Weekend		
	Number of clusters	RMSSTD	RS	Number of clusters	RMSSTD	RS
1	3	0.91	0.33	3	0.93	0.32
2	4	1.10	0.28	4	1.11	0.25
3	7	1.25	0.23	8	1.30	0.25
4	8	1.31	0.19	9	1.36	0.23

The variation of the RMSSTD index and the RS index as a function of the number of cluster, obtained by aggregating the households containing two occupants during a weekday are reported in Figure 4. In this case, the knee is observed at 4 clusters. The relative incremental improvement of the RS index is 10% when the number of clusters is increased from 3 to 4, while the relative improvement is just 4% when the number of clusters increases from 4 to 5. The RMSSTD index has the same behaviour. It means that in this case the consideration of more than 4 clusters is not beneficial.

The proposed methodology leads to the creation of 46 different household occupancy profiles: 22 for the weekdays and 24 for the weekend.

From Table 2 it is possible to observe the increase of the value of the RMSSTD index is combined with the decrease of the value of the RS index, when the number of occupants increases. This means that because of the larger number of states that can be associated with the household when the number of occupant increases, the clusters that are created are less homogeneous, and moreover the difference between the clusters is smaller.

For each cluster, the occupancy profiles are clearly visible by plotting the state of the households belonging to the same cluster as a function of time. In Figure 5, the spectrum of the occupancy profiles that can be associated with a household with two occupants during the weekdays is shown for each cluster. For all the clusters, it is possible to

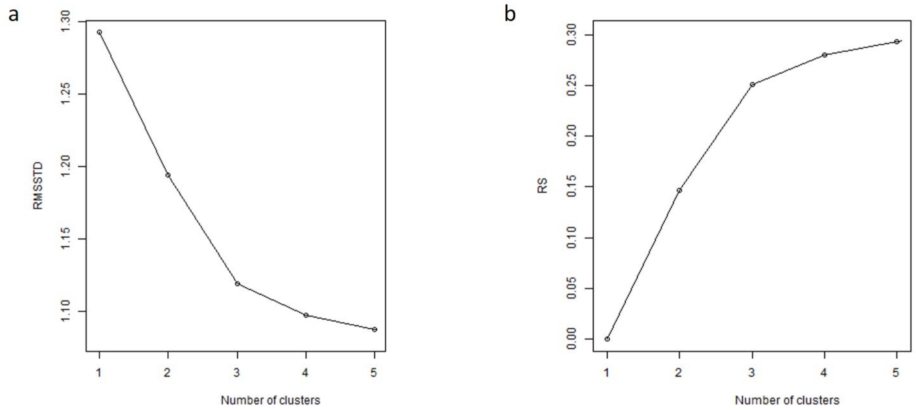

Fig. 4. (a) RMSSTD value considering different number of clusters; (b) RS value considering different number of clusters

identify a common period during which all the occupants are sleeping from around 11:00 pm to around 7:00 am. Then, during the day, the households belonging to different clusters are characterised by different occupancy behaviour. For example, it is clearly visible that in the case of Cluster 1 (Figure 5(a)) both occupants are mostly home during the day, while in Cluster 2 (figure 5(b)) just one of the occupants is mostly away during the day. In Clusters 3 (figure 5(c)) and 4 (figure 5(d)) there is also a different behaviour during the day, which changes at 6:00 pm for both. In Cluster 3 the occupants are mostly away before 6:00 pm but then they are home, while in Cluster 4 just one of the occupants is away until 6:00 pm and then both are mostly home. Using the k-modes method, the modes that represent the clusters are associated with each group of households (Figure 6).

They clearly define the general trend of the households which was observed from the figure 5. For example, while from the spectrum of occupancy profiles of the cluster 3 (figure 5(c)) it is just possible to deduce that the 2 occupants are mostly away from 8:00 am to 6:00 pm, from the mode of the cluster (figure 6(c)) it is possible to know that, as average, the two occupants wake up at 7:20 am, and they stay home until 8:00 am during the morning, when one of them leaves the house slightly before the other occupant who then leaves at 8:10 am. Both occupants return home at 6:00 pm, and remain awake until 10:00 pm.

From the k-mode clustering it is possible to determine also the percentage of households which belong to the same cluster. This information is useful to assign the correct percentage of households with similar occupancy profiles for each dwelling type characterised by specific characteristics. In this case, 39% of households with 2 occupants (759 households) belong to the cluster 3, which appears to be predominant also from the spectrum of the occupancy profiles of all the households (Figure 2(a)) and the correspondent mode (Figure 2(b)). Then 27% of the households with 2 occupants (535 households) are associated to the cluster 1, 20% (417 households) to the cluster 4 and the remaining 14% (246 households) to cluster 2.

In the case of UK, the distribution of the households with different number of occupants is known for each accommodation type (Figure 2). For each group of households with a defined number of occupants and a specific day-type, the percentage of households represented by a specific occupancy profile can be obtained from the methodology introduced in this paper. Combining these data, it is possible to associate the correct amount of households with a specific occupancy profile for each accommodation type. In this way it is possible to obtain complete archetypes considering their form and also their operational features.

5. Conclusion

In this paper a new methodology for the identification of representative daily residential occupancy profiles for households with different numbers of occupants is developed.

The methodology is applied to UK households using data available from the UK TUS 2000 and it identifies representative occupancy pattern for the households, correlating the occupancy behaviour of each occupant. It was found

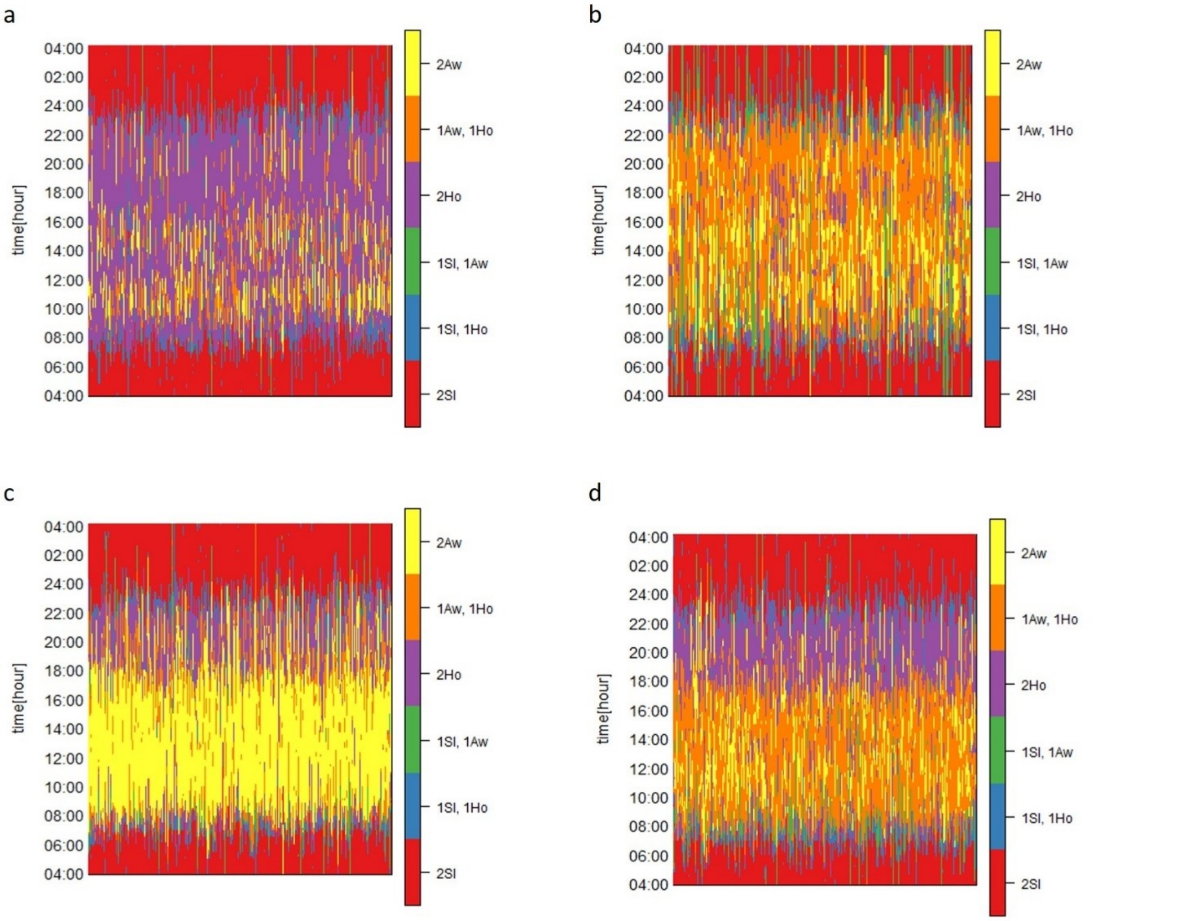

Fig. 5. Spectrum of occupancy profiles for the households with two occupants (weekdays). (a) cluster 1: 2 occupants mostly home; (b) cluster 2: 1 occupant mostly home/1occupant mostly away; (c) cluster 3: 2 occupants mostly home; (d) cluster 4: 1 occupant mostly home/1occupant mostly away during the morning and 2 occupants mostly home during the afternoon.

that 22 different household occupancy profiles can be attributed to UK households during the weekdays, and 24 during the weekend. This new methodology can be applied also to other countries where TUS data is available. For other applications the criteria used to cluster the households occupancy profiles can be changed according to the different goals of the clustering.

In future work, the presented methodology can be used to obtain realistic occupancy profiles which can be attributed to the archetypes for modelling building energy consumption. Then the results of the simulations of each archetype could be scaled and used to model the energy end-use demand at neighbourhood, city or national level.

The occupancy profiles generated by this model can contribute to planning and designing of energy production and energy dispatch systems for residential buildings, developing building retrofit packages or population-scale HVAC strategies, and other applications which require electric energy demand of large residential building stocks.

6. Acknowledgements

This work is part of the RealValue project. This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 646116.

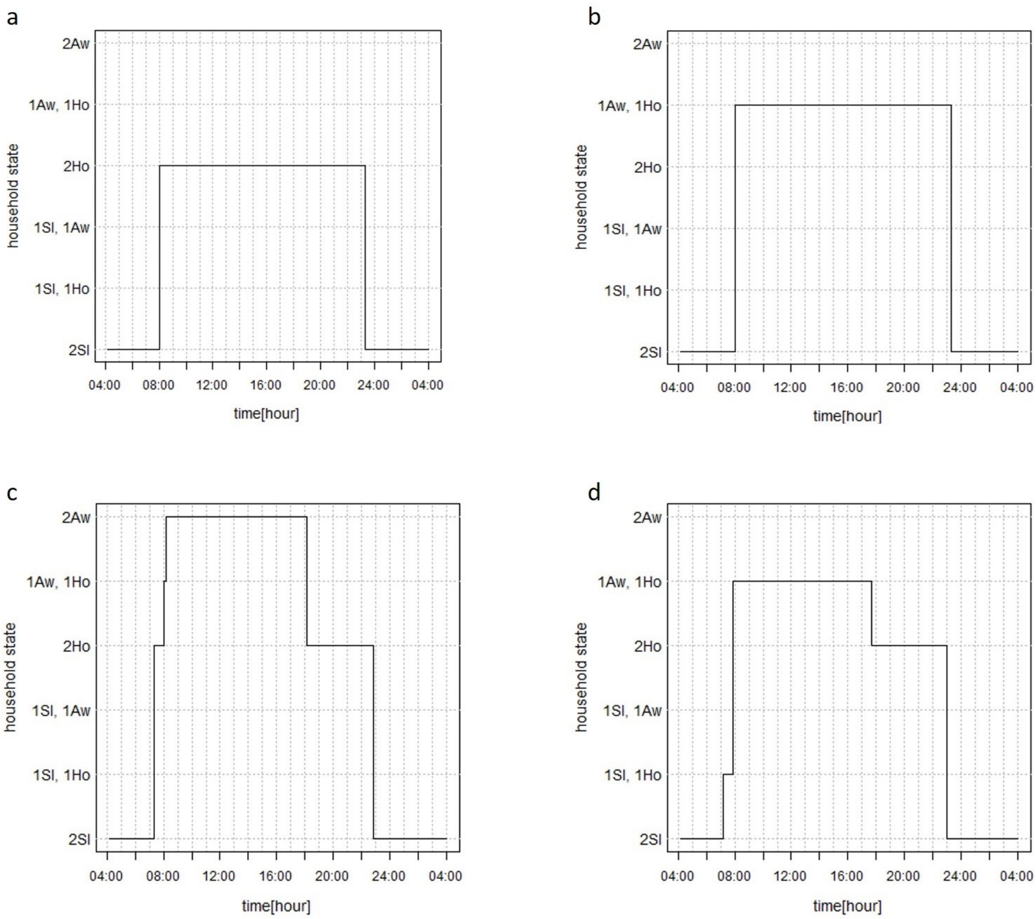

Fig. 6. (a) mode of cluster 1; (b) mode of cluster 2; (c) mode of cluster 3; (d) mode of cluster 4.

References

[1] Patteeuw, D., Bruninx, K., Arteconi, A., Delarue, E., D'haeseleer, W., Helsen, L.. Integrated modeling of active demand response with electric heating systems coupled to thermal energy storage systems. Applied Energy 2015;151:306–319.

[2] Swan, L.G., Ugursal, V.I.. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. Renewable and Sustainable Energy Reviews 2009;13:1819–1835.

[3] Corgnati, S.P., Fabrizio, E., Filippi, M., Monetti, V.. Reference buildings for cost optimal analysis: Method of definition and application. Applied Energy 2013;102:983–993.

[4] Neu, O., Sherlock, B., Oxizidis, S., Flynn, D., Finn, D.. Developing building archetypes for electrical load shifting assessment: Analysis of irish residential stock. The CIBSE ASHRAE Technical Symposium; Dublin; 2014,.

[5] Caputo, P., Costa, G., Ferrari, S.. A supporting method for defining energy strategies in the building sector at urban scale. Energy Policy 2013;55:261–270.

[6] Theodoridou, I., Papadopoulos, A.M., Hegger, M.. A typological classification of the Greek residential building stock. Energy and Buildings 2011;43:2779–2787.

[7] Capozzoli, A., Grassi, D., Piscitelli, M.S., Serale, G.. Discovering Knowledge from a Residential Building Stock through Data Mining Analysis for Engineering Sustainability. Energy Procedia 2015;83:370–379.

[8] Turner, W.J.N., Kinnane, O., Basu, B.. Demand-side characterization of the Smart City for energy modelling. Energy Procedia 2014;62:160–169.

[9] Richardson, I., Thomson, M., Infield, D.. A high-resolution domestic building occupancy model for energy demand simulations. Energy and Buildings 2008;40:1560–1566.

[10] Yao, R., Steemers, K.. A method of formulating energy load profile for domestic buildings in the UK. Energy and Buildings 2005;37:663–671.

[11] Ipsos-RSL and Office for National Statistics, . United kindgdom time use survey. Tech. Rep.; Uk Data Archive; 2000.

[12] Short, S.. Review of the UK 2000 time use survey. Tech. Rep.; Office for National Statistics; 2006.

[13] Aerts, D., Minnenb, J., Glorieux, I., Wouters, I., Descamps, F.. A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison. Energy and Buildings 2014;75:67–78.

[14] Santin, O.G., Itard, L., Visscher, H.. The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock. Energy and Buildings 2009;41:1223–1232.

[15] Goulden, I.P., Jackson, D.M.. Combinatorial Enumeration. Dover Publications; 2004.

[16] Jain, A., Murty, M., Flynn, P.. Data Clustering: A Review. ACM Computing Surveys 1999;31:264–323.

[17] Halkidi, M., Batistakis, Y., Vazirgiannis, M.. Clustering algorithms and validity measures. Thirteenth International Conference on Scientific and Statistical Database Management; Fairfax, Virginia; 2001,.

[18] Huang, Z.. A fast clustering algorithm to cluster very large categorical data sets in data mining. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery; The University of British Columbia, Canada; 1997, p. 1–8.

[19] R Core Team, . R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria; 2013. URL: `http://www.R-project.org/`.

[20] Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G., Zentgraf, M.. Package 'klaR': Classification and visualization; 2014. URL: `https://cran.r-project.org/web/packages/klaR/klaR.pdf`; r package version 0.6-12.

[21] Gan, G., Ma, C., Wu, J.. Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM Series on Statistics and Applied Probability; SIAM; 2007.