Identification of Energy Use Time Patterns of Occupied Dwellings using Smart Meter Data

Eline Himpe[1,*], and *Arnold* Janssens[1]

[1]Department of Architecture and Urban Planning, Ghent University, Sint-Pietersnieuwstraat 41-B4, B-9000 Ghent, Belgium

Abstract. The increasing application of smart and digital energy meters leads to an increasing availability of frequent -e.g. hourly- and long-term measurements of the actual energy use in occupied buildings. In the resulting energy use time series, the diurnal fluctuations in energy use are recognised and similarities between diurnal profiles for various days are observed. These recurring profiles are called energy use time patterns and they are a result of various phenomena, such as patterns in the building use, occupational schedules, settings of the system control, short-term weather dynamics etc. These energy use time patterns can provide a better understanding of the energy use, which is useful in many fields including energy feedback, fault detection and energy auditing. In order to identify and characterise energy use time patterns for large data-sets, an automated approach is needed. This paper proposes a methodology for automated mathematical recognition of energy use time patterns based on cluster analysis. Secondly, a methodology to characterise the identified patterns in function of external variables is proposed, using classification analysis techniques. The methodologies allow an automated identification and characterization of energy use time patterns, allowing a better understanding of the variations and changes in building energy use and their relation to weather conditions and calendar aspects.

1 Introduction: energy use time patterns

Reduction of energy use in buildings is a major challenge faced by society today. The implementation of the Energy Performance of Buildings Directive leads to an increasing energy-efficiency in both new and refurbished buildings, at least in theory. While calculations of the building energy use in the design stage are useful to estimate the building energy performance in reference conditions, these theoretically estimated energy use figures often differ significantly from the actual measured energy use of the occupied building. In reality, construction deficiencies may occur, the actual operation and control of the building services may differ, and the building users interact with the building envelope and services according to their personal comfort needs and activities, which in turn may also be driven by energy-efficiency objectives. In order to estimate the actual and case-specific impact of energy-efficiency measures, there is an urgent need to better understand the actual energy use in occupied buildings.

While calculations of building energy use have become a general practice when designing or refurbishing buildings, the characterisation of the actual energy use in operational buildings – especially in residential buildings - is less common. Today, the application of energy monitoring and smart metering systems leads to an increasing availability of frequent, hourly or daily, and long-term measurements of the actual energy use in occupied buildings, thus allowing a long-term monitoring of the actual energy use for specific buildings and their users.

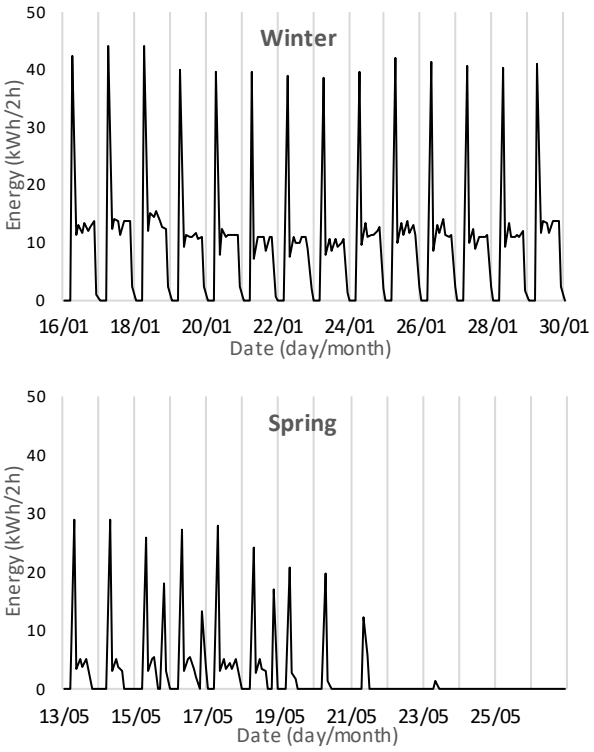

Fig. 1 Case 1: Energy use time series

* Corresponding author: eline.himpe@ugent.be

In measurements of the energy use at a sub-daily time interval, the diurnal fluctuations in energy use can be closely observed. For example in Fig. 1 energy meter data of a Belgian dwelling are presented with a time step of 2 hours, for two weeks in winter and spring season. The data originates from the gas meter of the house, in which natural gas is used only for the purpose of space heating. When observing the winter data, a repetitive daily pattern can be recognised, consisting of a period of zero energy use at night, a peak in the morning and moderate energy use during the rest of the day. This resembles the effect of the night set-back of the space heating system, followed by the morning start up and the continued space heating demand during the day. These daily recurring profiles are called energy use time patterns in this study. In the intermediate season data it can be observed that the heating also reduces during the day, as exterior temperatures are closer to comfort temperatures. Thus the energy use time patterns differ as a result of the weather. In Fig. 2 the energy use of a second case-study house is presented during two weeks in winter, spring and summer season. In winter it can be observed that a different energy use time pattern occurs for weekdays and weekend days, most probably a result of the occupants schedules. The more irregular spikes in the pattern are also seen in summer and spring time, and are related to the gas use for domestic hot water.

Both examples illustrate that the energy use time patterns, like the energy use, are a result of the interaction between the building, building services and occupants ('the occupied building'), and the interaction of this system with external conditions such as the weather, which influences especially the energy use for space heating, but also the calendar (e.g. the day of the week), which may influence user behaviour.

By mere observation of the sub-daily energy use time series plots, insights in the energy-related behaviour of the observed system can be acquired, for example settings and properties of the heating system, occupational characteristics etc. This kind of insights can be informative for the occupant of the house, e.g. for awareness raising and information via energy feedback on when most gas is consumed or assessing the effect of a change in control or behavior. It can also be useful for energy auditors, commissioners and researchers, who try to gain a better understanding of the energy-related properties of the occupied house, to detect anomalies in the heating system, to size new systems for the specific house, or to detect changes in energy use of the house as a result of a renovation or behavioural change.

The visual inspection of often long energy use time series and the visual search for recurring patterns and anomalies is a difficult and time-consuming task. If instead a computer could automatically recognise these patterns and group them according to their similarity, then the characteristic energy use time patterns of the house can easily be summarized, without the need for visualising the entire time series. Therefore a first objective of this study is to develop a methodology that allows to mathematically identify the different groups of energy use time patterns that occur in energy use time series.

In a next step, the relationship between these energy use time patterns and external aspects such as the weather conditions or the day of the week can be investigated. This relationship can be expressed in a statistical model, which can be used to detect changes in energy use time patterns over time or to predict the patterns for other periods. Therefore the second aim of this study is the development of a methodology that allows to mathematically characterize groups with similar energy use time patterns in function of specific weather conditions and/or calendar information. This model allows to analyse energy use time patterns in different conditions and may be useful for energy use analysis, normalization and prediction.

In the next section of this paper, the developed methodologies are summarised. Then, the methodologies are applied to the two case study buildings and the results and perspectives are discussed.

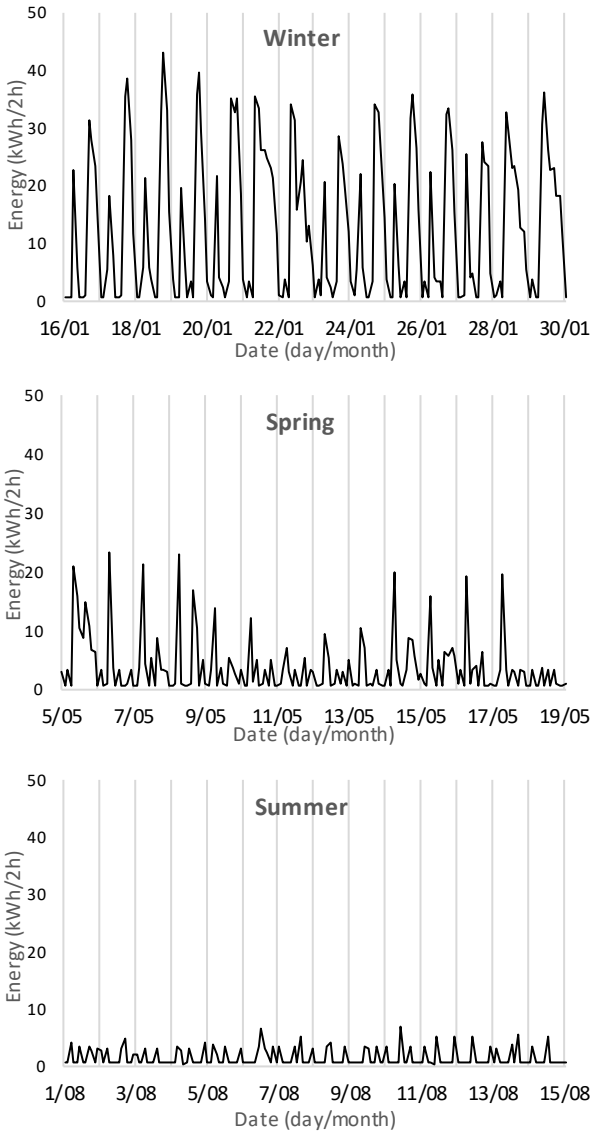

Fig. 2 Case 2: Energy use time series

2 Methodology

The identification of energy use time patterns from energy use time series relies on cluster analysis techniques. The identification methodology is summarised in section 2.1. For the second objective, the characterisation of the obtained energy use time patterns in function of external conditions (weather, calendar), classification methods are applied. The classification methodology is explained in section 2.2. The data manipulation and calculations are executed in R, a language and software environment for statistical computing and graphics [1]. A complete explanation of the methodologies and how they were developed can be found in Himpe [2].

2.1 Identification of energy use time patterns

Cluster analysis techniques are used to identify energy use time patterns in the energy use time series. As pointed by Kaufman & Rousseeuw, '*Cluster analysis is the art of finding groups in data*' [3]. It is a pattern recognition technique that comprises the grouping of a collection of objects in clusters in such a way that the objects within a cluster are more similar to each other than those found in different clusters [4]. The clustering procedure applied in this study, is summarised in Fig. 3.

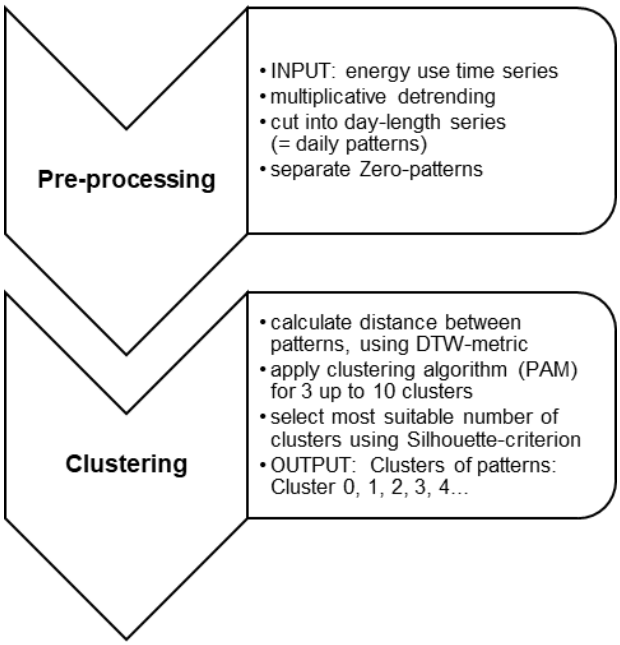

Fig. 3: Clustering of energy use time patterns: methodology

2.1.1 Time-series pre-processing

In the current study, each object is an energy use time series of length one day (and time step of 2 hours) and the various day-length objects are grouped, based on their similarity in shape. Therefore the original long-term energy use time series needs preprocessing. As a first step, the original time series is decomposed using multiplicative time series decomposition techniques, and the series is detrended [5]. This manipulation allows to make sure that the day-length objects will be clustered based on their shape rather than on the absolute magnitude of the energy use. As a second preprocessing step, the detrended energy use time series is cut into shorter time series with length of one day. These day-length objects are the clustering objects. At this point, the objects for which the energy use is zero at each time step, can already be separated into a separate cluster Zero, that thus contains the days with zero energy use. All remaining objects are used in the clustering.

2.1.2 Clustering

From the variety of existing clustering algorithms, in this study the 'Partitioning Around Medoids' (PAM or k-medoids) algorithm is selected, which is more robust to outliers than the well-known k-means method. The algorithm tries to find k representative objects (medoids) that are central in their clusters, by minimising the average distance to the medoid [3]. For calculating the dissimilarity or distance between the different objects, a distance measure or metric is needed. In this study, the distance between two objects is calculated using Dynamic Time Warping (DTW) distances: the series are stretched or compressed locally until they resemble each other, and the path to do this that minimizes the total cumulative distance between the two series, results in the actual 'distance' between the series [5, 6]. In contrast to the often applied Euclidean metric, this technique allows to recognize the similarity between two patterns with a similar shape that are shifted one timestep.
During the clustering process, also the amount of clusters k needs to be decided. In this study, the clustering is repeated for k: 3 -> 10 and the Silhouette index is used to select the best amount of clusters for the given data-set. For energy use data sets this typically leads to a partitioning in 3 to 5 clusters of distinct energy use time patterns, apart from the Zero-cluster that was already separated in the pre-processing stage.

2.2 Classification of energy use time patterns

The clusters obtained in the previous step contain energy use time patterns that are similar in shape. When observing these clusters, it is seen that the objects within one cluster often appear at similar external conditions, e.g. certain weather conditions or certain days of the week. This second part of the methodology aims at mathematically describing the relationship between the clusters of energy use time patterns on one hand, and external variables, that are weather and calendar variables, on the other hand. The relation between the clusters and the external variables is described using a classification method. In this study multinomial logistic regression analysis is used. This is a type of regression analysis that is applicable when the dependent variable – in this case the cluster variable - is a categorical variable [7, 8].

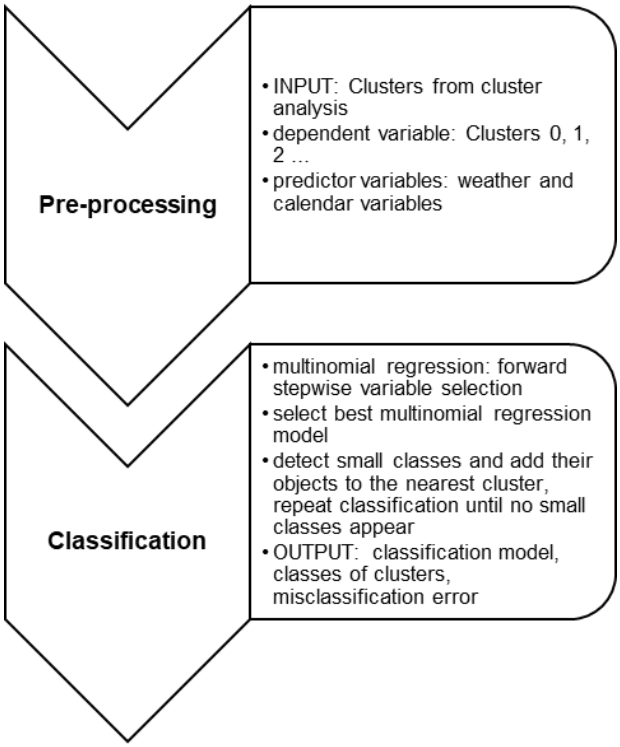

Fig. 4: Classification of energy use time patterns: procedure summary

2.2.1 Cluster pre-processing

In this study the dependent variable is the clusters of energy use time patterns (nominal categorical variable) and the predictor variables are weather variables (continuous) and calendar variables (categorical). The regression model is fitted by maximising the likelihood function.

2.2.2 Classification

As with linear regression models, if multiple predictor variables are available, a variable selection process is performed. In this study a stepwise forward selection process is constructed, that uses preliminary insights in the data. During the stepwise regression procedure, following variables are tested consecutively: the equivalent exterior temperature, the day of the week, solar radiation and wind speed. The model selection is a case-specific automated procedure. Only the variables for which the coefficients are significantly different from zero, remain in the model. The model selection relies on the likelihood ratio test for nested models, with 99% confidence level. At each step of the selection procedure, the suitable models are stored, and the final model is selected by selecting the model with the lowest deviance. The misclassification error documents the amount of cases that are misclassified when comparing the original energy use time patterns with the in-sample predicted energy use time patterns for each day of the energy use time series.

During the classification process it happens that clusters containing a relatively small amount of objects and occurring for a range of external variables that overlaps with the range occurring for larger clusters, tend to be absorbed by the larger clusters and remain (almost) depleted in the classification model – no matter whether the larger cluster is a neighbouring cluster or not. Therefore the classification approach is modified: after a first classification round, the small and depleted clusters are detected and the objects from the smallest cluster are allocated to their original neighbouring cluster; then the classification is repeated and this procedure is repeated for each of the empty or small clusters. As a result, the clusters become more balanced, the overall classification quality improves, and objects of small clusters are no longer absorbed by dissimilar large clusters. The classification procedure is summarised in Fig. 4.

3 Cases

The data used in this study consists of smart meter readings of the gas meters in single-family houses in Flanders (Belgium) from June 2011 to June 2012. The data belongs to the Flemish Distribution System Operator (Eandis). The methodology was developed and assessed on data-sets of 27 case-study buildings [2], out of which 2 are selected for illustrating the method for the purpose of this paper. The first case-study dwelling corresponds to the time series in Fig. 1. In this house gas is used only for space heating purposes. According to a survey, the terraced dwelling is occupied by one elderly person who is usually at home during day-time and night-time, and the house has a programmable thermostat. In the second case-study single-family house, corresponding to the time series in Fig. 2, gas is used both for space heating and domestic hot water generation. No further information about the type of dwelling and its occupants is available.

4 Analysis and Results

Fig. 5 and Fig. 6 summarise the results of the identification and characterisation of the energy use time patterns for case-study 1 and 2 respectively. For both dwellings four different energy use time patterns were identified by using the cluster analysis methodology, apart from the Zero-cluster, that was separated in the pre-processing stage. The four graphs in Fig. 5 (a) and Fig. 6 (a) represent the four non-Zero clusters for both cases. The black lines are the various detrended day-length time series that belong to a certain cluster, and the coloured lines are the so-called representative objects or medoids of these clusters.

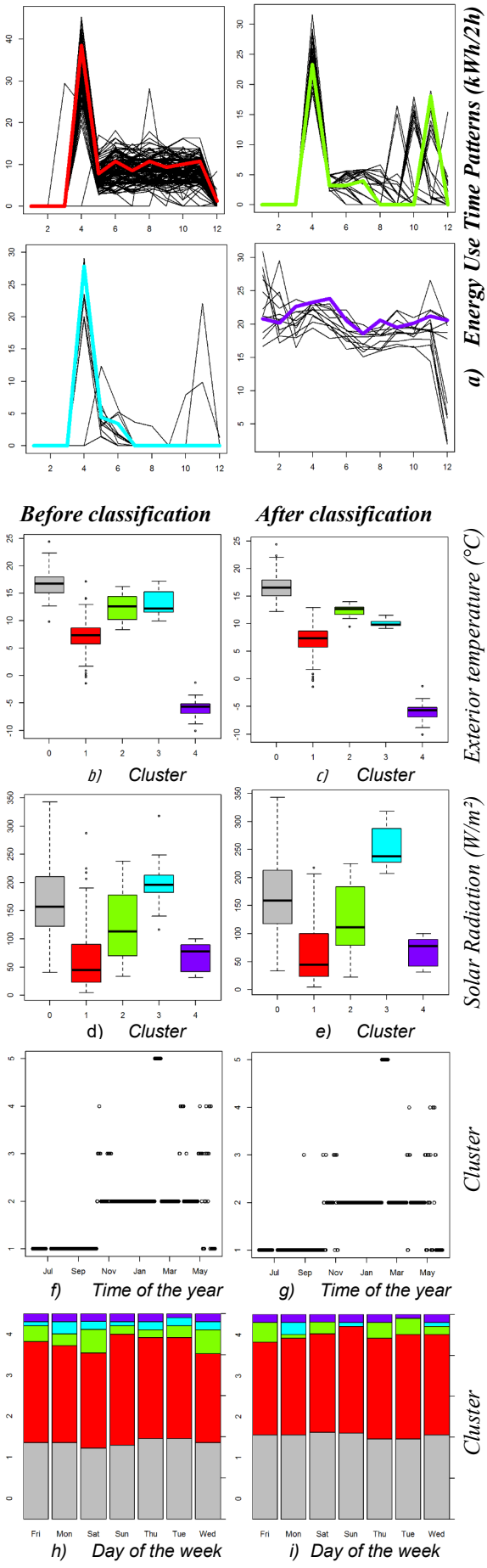

Fig. 5: Case 1

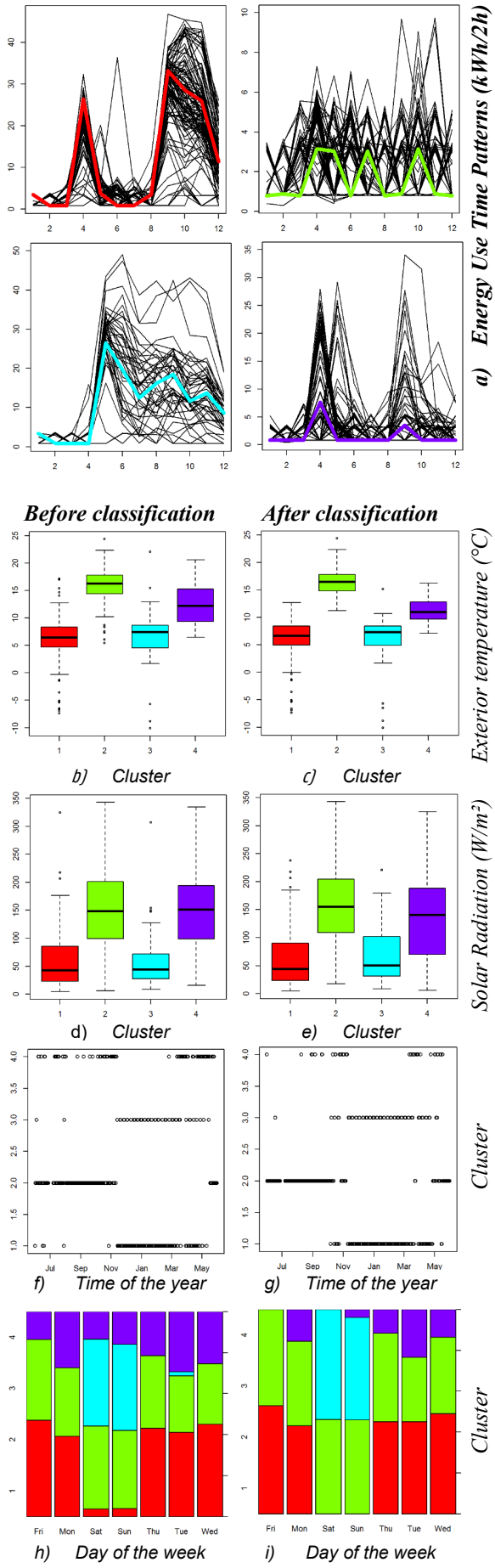

Fig. 6: Case 2

For case 1, in Fig. 5 (a), we identify a big (red) cluster (1) containing the profiles in which the house is heated during day-time and a night set-back is applied, resulting in a heating peak when the system restarts in the morning. The green (2) and blue (3) clusters represent days with a heating peak in the morning, and (in the green cluster only) a peak in the evening. The purple (4) cluster contains all profiles of days when the space heating is activated day and night. We can see that the clustering algorithm succeeded in grouping the patterns according to their similarity in shape. Looking further at the 8 plots in the lower part of Fig. 5, the clusters are represented as function of the key external variables to which they are related: exterior temperature (b-c), solar radiation (d-e), time of the year (f-g) and day of the week (h-i). The left four graphs are simply the identified clusters in function of the variables (after clustering). In Fig. 5 (b) we can see that for case 1, the purple cluster only appears when exterior temperatures are below -2°C, corresponding to the case that the night set-back is not activated because of the heavy cold. The main red cluster mostly appears in the heating season (0°C – 15°C) and the green and blue clusters in the mid-season. Remark that the blue cluster occurs on days with higher solar radiation, which explains the absence of space heating need in the evening. The grey cluster appears mostly at higher temperatures, in the summer time.

The right four graphs low in Fig. 5 contain the relationship between the clusters and external variables after the classification process has been accomplished. The clusters are now modelled mathematically in function of (a selection of) the external variables. For case 1, only exterior temperature and solar radiation are taken up as variables in the model. As can be seen in Fig. 5 (h), there is no clear relationship between the clusters and the day of the week as a variable. After classification, some objects have moved from one cluster to the nearby cluster, because of the relationship with the external variables. In case 1, about 10% of the objects is misclassified, which is a rather low number.

In the second case (Fig. 2 and Fig. 6), the daily energy use time profiles are more variable than in the first case, especially in the intermediate and summer season. This is a result of the function of the heat use in these periods, that is domestic hot water generation, which is a much more user-dependent and thus variable. Looking at the clusters in Fig. 6 (a), the days in the heating season are mainly found in the red (1) and blue (3) clusters, apart from a few that have been ordered mainly in the purple (4) cluster. Clusters (2) and (4) mainly contain summer and intermediate season days, that are more variable as well. Looking now to the representation of the clusters in function of the calendar variables in Fig 6 (h), it can be seen that the blue and red clusters appear on very different moments of the week. The red cluster is dominant on weekdays, when the heating starts between 6h and 8h and typically turns out during the day. The blue cluster corresponds to weekend days, when the occupants activate the heating between 8h and 10h and it is more likely to be turned on for the rest of the day. This effect is also recognised in the classification process, and 'day of the week' is a significant variable in the classification model. After classification about 24% of the objects were misclassified. 15% of the objects were misclassified in cluster 2 instead of cluster 4.

5 Discussion and Perspectives

The two cases by which the clustering and classification methodology has been illustrated represent a rather orderly or repetitive, and a rather disorderly / more random energy use time series for residential dwellings. Even for the more random case, the clustering and classification allows to roughly cluster the energy use time patterns, and show their underlying relationships to the external variables (especially before clustering). The misclassification error is then a measure for how well the results of the classification can be trusted. In the larger study behind this paper, a larger data-set of 25 houses with gas use for space heating was analysed, and it was found that misclassification errors were between 5% and 35%, with the majority of cases between 15% and 25% [2]. This confirms that the two examples in this paper are a very 'good' and a more 'chaotic' case.

In this study, the external variables for the regression were limited to the main weather variables and the day of the week. Of course, the approach could be extended towards other variables, such as other calendar aspects (e.g. holidays etc.). The methodology could also be explored on other data-sets, including energy use time series for other building types (e.g. office buildings) or measurements of indoor temperatures.

Fig. 5 and Fig. 6 are examples of how a small selection of graphs can summarise the major properties of the energy use time patterns that can be identified in long energy use time series. The type of information that is represented in those, can be useful for both building users and building professions aiming to understand the actual energy use of their specific building. Therefore, applications such as personalised energy use feedback and case-specific commissioning are relevant. This study therefore also illustrates how analysis of 'big data' allows new ways of following up on the energy use in dwellings.

In further parts of this study, the identified energy use time patterns are successfully used to improve energy normalisation techniques, by integrating the clustering results into classical energy signature methods [2].

6 Conclusions

This paper presents and demonstrates a methodology for the mathematical identification and characterization of energy use time patterns from energy use time series. For the identification methodology, the long-term energy use time series is split into numbers of day-length time series that are partitioned into groups based on their similarity in shape. Cluster analysis is here used as a methodology for pattern recognition. Similarities in shape are calculated using dynamic time warping principles. For the characterization methodology, multinomial logistic regression methods are proposed to classify the obtained clusters of energy use time patterns in function of weather variables and calendar variables (day of the week). The application of the methodology on two example cases illustrates how the one year energy use time series can be summarised and characterized in a few informative visualisations and a classification model, that allow to identify/retract some energy-use related properties of the dwellings studied, related to their occupancy, building and building services, without the need for visiting the building or enquiring occupants. The identification and characterization procedures are automated, thus allowing to apply them to large and long data-sets. As such the methodology opens opportunities in applications such as automated and personalized energy feedback, and can be of assistance in energy auditing and commissioning.

Acknowledgement

The authors are grateful to Eandis, the Flemish Distribution System Operator, for providing access to the monitoring data of their smart metering proof-of-concept project.

References

[1] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (2015)
[2] E. Himpe, *Characterisation of residential energy use for heating using smart meter data*. (Ghent University, Ghent, Belgium, 2017)
[3] L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. (Wiley, USA, 1990)
[4] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, (Springer, New York, 2009)
[5] E. Himpe, A. Janssens, "Data-driven modelling of the energy use in dwellings using smart meter data", in *CLIMA 2016 - proceedings of the 12th REHVA World Congress*, **6,** (Aalborg, Denmark, 2016)
[6] E. Keogh, C. A. Ratanamahatana, "Exact indexing of dynamic time warping", *Knowl. Inf. Syst.* (2004)
[7] C. A. Ratanamahatana, E. Keogh, "Making Time-Series Classification More Accurate using Learned Constraints", in *Proceedings of the SIAM International Conference on Data Mining* (Florida, USA, 2004)
[8] A. Agresti, *An Introduction to Categorical Data Analysis* (Wiley, Hoboken, New Jersey, 2007)
[9] W. N. Venables, B. D. Ripley, *Modern Applied Statistics with S* (Springer, New York, 2002)