CLUSTERING THE OCCUPANT BEHAVIOR IN RESIDENTIAL BUILDINGS: A METHOD COMPARISON

N. Carbonare[1,2], T. Pflug[2] and A. Wagner[1]

[1]Fachgebiet Bauphysik und technischer Ausbau, KIT, Karlsruhe, Germany
[2]Energy efficient buildings, Fraunhofer ISE, Freiburg, Germany
Contact: nicolas.carbonare@ise.fraunhofer.de

ABSTRACT

The aim of this paper is to investigate possible patterns of the occupant behaviour in residential buildings. Measurements were taken in multifamily buildings where several occupant-related variables were recorded. We chose and compared two different clustering methods: whole time series and features clustering (k-means algorithm). The mentioned methods were performed selecting two variables (window opening and indoor temperature), and tested with supervised learning methods. Results suggest that features clustering can perform better than whole time series. The representation of the occupant behaviour through features is meant to be applied in future work regarding the optimization of control strategies in ventilation systems.

INTRODUCTION

The world's increasing energy demand has led in the last twenty years to a raised interest on energy efficiency. The efforts towards the consumption reduction in the residential sector have brought up the retrofit of buildings as a solution in European countries, in which the high air tightness is a characteristic, as it contributes to reduce the heating energy consumption. Within this frame, mechanical ventilation systems gain relevance to maintain a desirable indoor air quality (IAQ) in low-energy residential buildings.

On the other hand, the evaluation of these aforementioned technologies reveals that the performance is lower than expected in practical applications. It is already clear that the diversity of the occupant plays a key role on this underperformance, generating the so called 'Rebound effect' (Galvin 2014). Besides, one conclusion from the IEA EBC Annex 53 (Polinder et al. 2013) was that taking control of the systems out of the hands of the occupant (i.e. automatic window opening) leads to higher dissatisfaction with the indoor environment. Therefore, the forthcoming technology development should be optimized in order to be compatible to user-defined adjustments in residential buildings.

This paper aims therefore at providing the first steps for a user-orientated control system for mechanical ventilation. As stated in Carbonare et al. (2017), the correct characterization of the user-dependent variables in residential buildings plays a huge role to obtain improvements on the current state-of-the-art. We define a methodology for future research and will be tested on two variables (window opening and indoor temperature), which were already studied by many researchers on the field of building simulation (Calì et al. 2016; D'Oca and Hong 2014; Haldi and Robinson 2009). For reasons of practicality, only the window opening variable will be described, and in the case of the indoor temperature only the results will be presented.

BACKGROUND

Clustering is the process of classifying data into different groups, aiming at finding similarities among them. A cluster is then defined as a

subset of objects in the database that belong to the same group. Similarity is often calculated through distance measures. The main challenges of a clustering process are (Mitsa 2010):

- The attributes that differentiate one cluster from another are unknown
- The data is unlabelled. This means, there is no knowledge on how to distinguish if one object belongs to a certain cluster or another one (except from a priori knowledge provided by domain experts)
- The more data, the more complex the problem becomes
- Algorithms are usually strongly influenced by noisy data, missing values and outliers. Hence the importance of an appropriate pre-processing of the data is highlighted

In the literature review, three main conventional clustering methods were found to be applied for time series clustering, namely shape-based, feature-based and model-based (Aghabozorgi et al. 2015). A shape-based approach means the straightforward comparison of raw time series data (all the points on the same time step are directly compared). In addition, the shape of these objects is matched as well as possible. The shape-based process presents a higher simplicity than the others (since only pre-processing of data is required to perform the clustering), although being usually more computationally expensive due to the number of compared data points. On the other side, a feature-based approach refers to the selection of features that represent as close as possible the characteristics of the time series, reducing the number of data points. The best performing features are usually extracted from the a priori knowledge about the data, as well as from some typical statistical indicators (Guyon and Elisseeff 2003). The main advantage of this method lies on the rapid calculation process and its adaptability to machine learning processes (Guyon et al. 2002). A major drawback could be the potential loss of information, in case of not carefully selecting the features vector. Model-based methods will not be covered in this paper.

In order to cluster data, the distance between the different points must be defined. Regarding the different metrics available for the clustering of time series, several authors have expressed their opinions about reliability and performance (Iglesias and Kastner 2013; Mitsa 2010). Following the results obtained by Iglesias and Kastner (2013), and due to its widespread use on research activities, the Euclidean distance is selected as the similarity measure for this study. It is defined as the distance between the ith *x* and *y* points:

$$D_{eucl}(x,y) = \sqrt{\sum_{i=0}^{n-1}(x_i - y_i)^2} \qquad (1)$$

Analyzing the clustering algorithm selection, researchers established lately that the use of conventional algorithms in the clustering of static data generates results with acceptable quality and efficiency, in terms of time and accuracy (Aghabozorgi et al. 2014). Centroid-based K-Means was selected among different algorithms analyzed on the literature, and it is applied following the K-Means++ application in Raschka (2015). A disadvantage is that the k-means method required the number of clusters as an input, which is typically (and this is no exception) unknown. The elbow method described by Raschka (2015) is also quite popular due to its simplicity. This method consists on the calculation of the percentage of variance explained for every set number of clusters, and to observe in which number of clusters the relative increase of the explained variance by adding a new cluster becomes negligible. Since there is no quantifiable threshold, this method can be combined with other indexes.

The Dunn Index (DI) presents a widely-used measurement technique of cluster validity. The DI was selected on this study because it presents the best performance regarding the k-means clustering procedure (Kovács et al. 2006). The defining equation is then presented:

$$DI = \min_{i=1\dots n_c}\left\{\min_{j=i+1\dots n_c}\left\{\frac{d(c_i,c_j)}{\max_{k=1\dots n_c}(diam(c_k))}\right\}\right\} \qquad (2)$$

$$d(c_i,c_j) = \min_{x\in c_i, y\in c_j}\{d(x,y)\} \qquad (3)$$

$$diam(c_i) = \max_{x,y\in c_i}\{d(x,y)\} \qquad (4)$$

where n_c is the number of clusters, *d(x,y)* the distance between two elements and c_i the centre of each ith cluster. This index compares

directly the distance between clusters (inter-comparison) and the diameter of the clusters (intra-comparison). Therefore, a better clustering configuration means higher values of the Dunn index (larger space among clusters and smaller cluster diameters). The calculation of the DI is usually time consuming and sensitive to a noisy database (Kovács et al. 2006). The chosen implementation of DI compares the distance between the two closest points among clusters (minimum) with the maximum distance between cluster-centroids altogether, which does not collide with single-dwelling clusters whose cluster diameter is zero.

In order to evaluate the quality of the obtained clusters, the task becomes challenging due to the unlabelled data. In this paper a Support Vector Machines classifier (SVM) method is proposed (Hastie et al. 2009), in order to evaluate how a test data adjusts to the training data set. SVM methods are a class of supervised learning algorithms which train the classifier function using labelled data. Given the training data set where each point has a corresponding label, the objective of the problem is to define a hyperplane that separates two points of different classes with a maximal possible margin (the original SVM is defined for two-class classification problems, and in this paper is addressed as a multi-class classification). Since perfect separation between the two classes is often infeasible, errors are allowed through auxiliary variables in the classification of the data that may not be linearly separable. The objective function balances between maximizing the separation margin and minimizing the classification error given an error weight. More about the method can be found in the literature (Raschka 2015).

METHODOLOGY

The data selected corresponds to the measurements in 2012 and 2013 of a high rise building retrofitted to passive house standard in the city of Freiburg, South Germany (Carbonare et al. 2017), in which 27 dwellings were monitored in great detail. Two measured variables from this project were taken, namely window opening and indoor temperature. The year 2013 was used to train the model (6-minute interval), and the data from 2012 as test set (hourly data). This time step mismatch is not expected to generate a relevant impact on the clustering results. Full year measurements are considered for window opening, while only the winter period (October-April) is taken into account for the indoor temperature, due to its high dependence on the outdoor temperature during summer (direct correlation analysed (Nguyen et al. 2014) – Pearson's R coefficient = 0.4 for outdoor temperatures below 13°C, and R = 0.91 above 13°C).

Data pre-processing results of utmost importance to improve the performance of the whole procedure, given that clustering is sensitive to noisy data. Several pre-processing methods are presented in the former literature (Aghabozorgi et al. 2015; Mitsa 2010; Raschka 2015). In this case, two variables are handled: one categorical (window opening – binary) and one continuous (indoor temperature). Therefore, they require different treatments. Firstly, using the same rules as developed by Carbonare (et al. 2017), the data corresponding to absence periods was neglected. Since this study considers the clustering of occupant behaviour, it is not the same to consider a window intentionally closed than a window left closed during absence, for example. Secondly, the window opening profiles are processed; the measurements were performed with contacts, which output values are 1 (closed window) and 0 (open window). However, for visualization simplicity, the values are switched (0 to closed and 1 to open). The removal of faulty data (sensor errors) in all variables is besides carried out. In the case of the indoor temperature profiles, no further data pre-processing was performed. Finally, the data is standardized by z-score normalization. The module applied is provided by the Python sklearn package (Pedregosa et al. 2011), for each data point X_i:

$$X_{i,norm} = \frac{X_i - \mu_X}{\sigma_X} \quad (5)$$

with μ_X as the mean of variable X in the training data set, and σ_X as their respective standard deviation. The key point of data normalization in data science is to analyse relative ranges of all measurements, instead of an absolute value. Thus, each data point is characterized by its distance to the sample mean of the training data set.

After pre-processing the data, the clustering takes place. For each variable, a shape-based

and a feature-based clustering are performed. As mentioned above, a K-means clustering algorithm is applied, using the Euclidean distance as similarity metric. When performing a whole time series clustering, the time steps considered are only the ones in which all the analysed dwellings have data (considering also the pre-processing). This reduces the total of 87,590 points to approximately 24,500 points. On the other hand, the feature clustering enables the utilization of every available data point for each variable.

A different set of features for each variable must be defined, depending on its characteristics. Following the literature (Haben et al. 2015), some features regarding *a priori* knowledge from the occupant behaviour in residential buildings were selected. Another set of features regarding statistical analysis of time series is also included, namely mean, standard deviation, skewness and kurtosis (Hastie et al. 2009). More specific, trend and seasonality indicators regarding time series decomposition (Wang et al. 2005) were also considered. Table 1 summarizes the feature selection ($\bar{X}$ means the mean value of the corresponding variable in each case).

Table 1: Proposed features and their suitability for each variable. WO = Window opening, IT = Indoor temperature

Feature	**Variable**	**Definition**
Weekend score (WkS)	WO	$\sum \frac{\bar{X}_{Weekday} - \bar{X}_{Weekend}}{\bar{X}}$
Seasonal score (SS)	WO	$\frac{\bar{X}_{Summer} - \bar{X}_{Winter}}{\bar{X}}$
Day-night score (DNS)	WO - IT	$\sum \frac{\bar{X}_{Day} - \bar{X}_{Night}}{\bar{X}}$
Hour change score (HCS)	WO - IT	$\sum_{h=0}^{8760} \frac{\bar{X}_{h+1} - \bar{X}_h}{\bar{X}_h}$
Average state changes score (ACS)	WO	$\frac{\sum_{h=0}^{8760} Stchanges_h}{8760\ hs}$

Not knowing if the proposed features will be representative for the whole data set, a comparative method is proposed, in order to obtain the optimal feature combination regarding dimensionality, representativeness and cluster structuring. The process follows these steps:

1. For a determined combination of features, calculate the minimal number of clusters that explain a selected threshold of 80% of the variance (K-Means algorithm), aiming at the minimization of the within cluster sum of squares (Raschka 2015)

2. Calculate the Dunn index for the different number of clusters between the obtained minimum and an imposed limit of 12 clusters, as it was considered sensible for 27 dwellings

3. Selection of the best combination of features and number of clusters which result in the highest DI – there is a preference for *a priori* defined features (Guyon and Elisseeff 2003)

4. Analysis of results and final selection of optimal combination considering DI, number of clusters and number of features involved

5. Labelling of the data with the resulting cluster structure and observe how the test data set fits to it through SVM. Comparison of results. Presentation of the obtained cluster structure.

ANALYSIS OF RESULTS

Features clustering

In this publication, the sleeping room of each dwelling is taken as example, due to the clarity of the measured profiles. The detailed procedure is described using the window opening variable of the measured dwellings. After calculating iteratively the DI for all the possible combinations of number of clusters and features, the analysis of the results was performed. The highest DI values are presented in Table 2 and later analyzed in detail to determine the optimal feature combination.

Table 2: Best feature combinations for window opening on the sleeping room

Features	**Clusters**	**Dunn index**
Mean, Seasonality, Skewness	3	1.8373
Mean, HCS.	8	1.8025
Mean, HCS, ACS.	8	1.7924
Mean, Skewness	3	1.7773
Mean, Seasonality, Trend, Skewness	3	1.7706

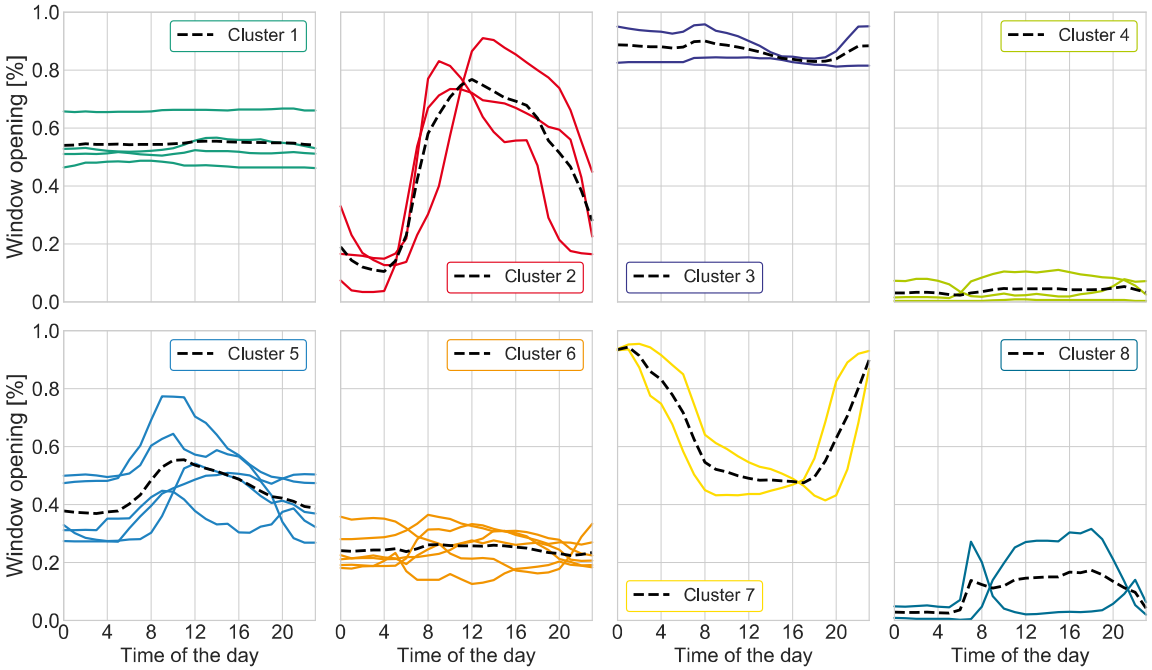

Figure 1: Mean profile sleeping room window opening - Cluster structure after optimal features clustering with training dataset. Dotted line: mean profile of the cluster.

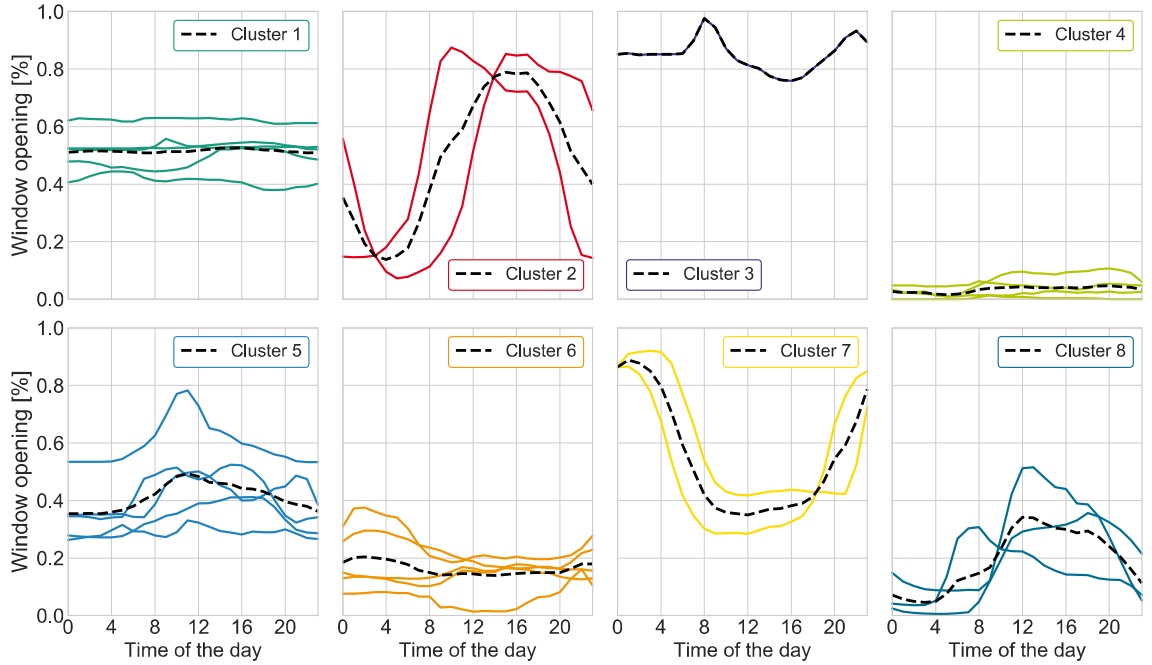

Figure 2: Mean profile sleeping room window opening - Cluster structure with optimal features after SVM with test dataset. Dotted line: mean profile of the cluster.

The five best DI are close to each other; hence it is reasonable to analyze them in further detail. Since simplicity is a desired condition, the models with fewer features are preferred. The models resulting in three clusters were discarded, since a threshold of 4 clusters has been defined, as it is the point in which the explained variance stops improving significantly. Both lasting models from Table 2 generate the same cluster structure in this case, represented on Figure 1. Therefore, the model with the mean and hour change score as key features is selected.

As it can be observed, the resulting clusters have distinctive characteristics:

- Cluster 1: almost no changes during the day, with around 50% window opening (probably open in warm days and closed during cold ones).
- Cluster 2: open during day and closed while sleeping.
- Cluster 3: almost always constantly open.
- Cluster 4: almost always constantly closed.

- Cluster 5: same concept of cluster 2 but with smaller changes between day and night profile, and higher night mean.
- Cluster 6: small changes and low mean value without a typical profile.
- Cluster 7: closed during day and open while sleeping.
- Cluster 8: similar to cluster 2 but with lower mean values during day.

Moreover, we applied data labelling and SVM classification process to the test set, to observe to what extent the conformed structure in the previous step is valid. Figure 2 shows how the test set is classified.

19 out of 27 dwellings were classified at the same category (70 %). Nevertheless, it must be said that those who changed category presented as well a different profile, which are more compatible to the newly assigned clusters. The description of the obtained clusters with the training data set suits the new ones obtained with the test data set. The DI for the test set is 1.5086, which is lower than the original one as expected. Nevertheless it is still higher in comparison with the obtained ones in the iterative process of feature selection.

In the case of the indoor temperature, the resulting cluster structure consists of four clusters, with a DI of 1.5930. The optimal combination of features selected was also mean and hour change, as it was ranked in first place.

The classification of the test set data with SVM showed a correct classification rate of 77.78 % and a DI of 1.3478.

Whole time series clustering

On the contrary, the whole time series clustering method presents significantly lower DI values. After computing the values iteratively, the highest DI (1.0525) was reported with a six-cluster structure, which corresponds to 50% of the variance explained. Figure 3 shows the resulting cluster composition of the whole time series with six clusters and K-means algorithm.

Comparing the results, two cluster structures are identical (Clusters 2 and 7 from Figure 1 against Clusters 3 and 5 in Figure 3), while Cluster 7 from features in Figure 2 was split third was here split into two single-dwelling categories in Figure 3 (Clusters 3 and 4). The two remaining clusters present significant differences among each other, although the Cluster 1 has a tendency to present lower mean values, and Cluster 6 results difficult to understand.

In addition, indoor temperature clustered with the whole-time-series method obtains also a four-cluster structure, with a DI of 1.1996 and 68 % of variance explained. A minimum threshold of 80% of the variance explained in this procedure would have shown a higher number of clusters (16 and 9 for window opening and indoor temperature) and lower DI (0.8452 and 0.7695) respectively.

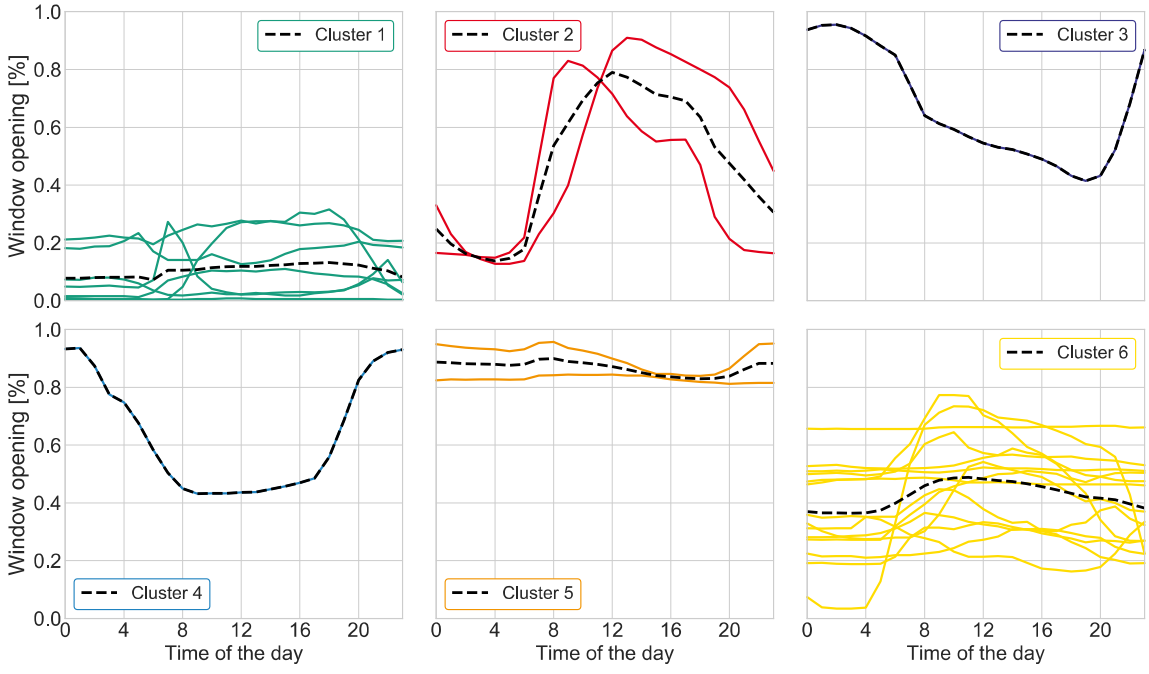

Figure 3: Mean profile sleeping room window opening - Cluster structure after whole time series clustering with training dataset. Dotted line: mean profile of the cluster.

Method comparison

Figure 4 shows the percentage of variance explained for both methods in increasing number of clusters. As it can be seen, the complexity of the data structure while analyzing a complete time series prevents the formation of an “elbow” that defines the potential optimal number of clusters and the variance steadily increases. The optimal features method shows a negligible explained variance increase from five to nine clusters. This justifies that the whole time series method presents weaknesses when obtaining a reliable structure of clusters.

Table 3: Dunn index (DI) for every clustering method

Variable	**Window opening**	**Indoor temperature**
Features training set	1.8025	1.5930
Features test set	1.5026	1.3478
Whole series training set	1.0525	1.1996
Whole series test set	0.9423	0.8727

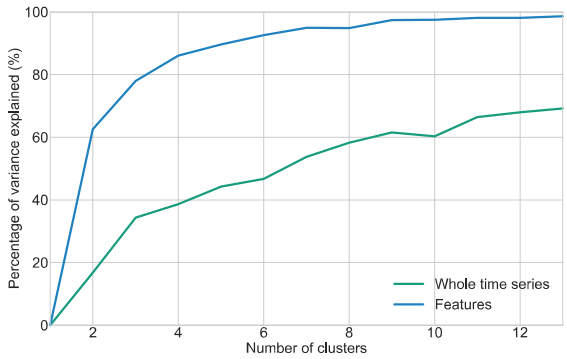

Figure 4: Percentage of variance explained with increasing number of clusters.

In addition, Table 3 summarizes the obtained Dunn indexes, which are significantly higher while dealing with feature methods than with whole time series clustering. This indicator proves that the cluster structure obtained with the optimal features is better posed than the results obtained while comparing each step of the time series, taking into account the above cluster structure described for each method. The same procedure will be applied in future work to other variables regarding the occupant behavior to simplify its representation.

Summarizing, the features clustering method presents advantages over the whole time series clustering at the following points:

- Clustering accuracy and prediction, given by the higher DI in train and test data sets
- Variables representation, as the variance explained is significantly higher with lower number of clusters
- Computational resources, due to dimensionality reduction by feature extraction

CONCLUSION AND SUMMARY

The representation of different variables of the occupant behavior in residential buildings was successfully carried out through features definition. We found different optimal combinations of these features, in connection with the type of room and the variable analyzed. Regarding window opening and indoor temperature on a sleeping room, the combination of the mean and the hour change score previously defined showed the best performance.

Features clustering process presents better results than whole time series clustering when representing the some aspects of the occupant behavior in residential buildings. Better clusters shapes could be found when carrying out an optimization of previously selected features, identified through higher Dunn indexes, higher percentage of variance explained and at the same time well-defined profiles with a manageable number of features and clusters. The application of resulting cluster structures to test sets resulted in smaller DI, but within acceptable values. The representation of different variables of the occupant behavior through selected features is therefore acceptable for future applications. The next research step involves the application of the methodology to different building data, to discuss its transferability by repeating the analysis with other data sets.

Different features following the presented methodology will be selected in future research for other variables of the occupant behavior, in order to perform a multi-variable characterization process, with the objective of developing new control strategies for ventilation systems under consideration of the occupant behavior in low-energy residential buildings.

ACKNOWLEDGEMENT

The study presented in this paper is funded by the German Ministry of Economics and Labour BMWi under the reference FKZ03ET1401A.

REFERENCES

Aghabozorgi, S., A. Seyed Shirkhorshidi, and T. Ying Wah. 2015. "Time-series clustering – A decade review." *Information Systems* 53:16–38.

Aghabozorgi, S., T. Ying Wah, T. Herawan, H. A. Jalab, M. A. Shaygan, and A. Jalali. 2014. "A hybrid algorithm for clustering of time series data based on affinity search technique." *The Scientific World Journal*:562194.

Calì, D., R. K. Andersen, D. Müller, and B. W. Olesen. 2016. "Analysis of occupants' behavior related to the use of windows in German households." *Building and Environment* 103:54–69.

Carbonare, N., F. Coydon, A. Dinkel, and C. Bongs. 2017. "The influence of occupancy behaviour on the performance of mechanical ventilation systems regarding energy consumption and IAQ." *Proceedings of 38th AIVC Conference*.

D'Oca, S., and T. Hong. 2014. "A data-mining approach to discover patterns of window opening and closing behavior in offices." *Building and Environment* 82:726–739.

Galvin, R. 2014. "Making the 'rebound effect' more useful for performance evaluation of thermal retrofits of existing homes: Defining the 'energy savings deficit' and the 'energy performance gap'." *Energy and Buildings* 69:515–524.

Guyon, I., and A. Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3:1157–1182.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. 2002. "Gene Selection for Cancer Classification using Support Vector Machines." *Machine Learning* 46:389–422.

Haben, S., C. Singleton, and P. Grindrod. 2015. "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data.":1–19.

Haldi, F., and D. Robinson. 2009. "Interactions with window openings by office occupants." *Building and Environment* 44:2378–2395.

Hastie, T., R. Tibshirani, and J. Friedman. 2009, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer New York: USA.

Iglesias, F., and W. Kastner. 2013. "Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns." *Energies* 6:579–597.

Kovács, F., C. Legány, and A. Babos. 2006. "Cluster Validity Measurement Techniques.".

Mitsa, T. 2010, *Temporal Data Mining*, Chapman & Hall/CRC: USA.

Pedregosa, F., G. Varoquax, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

Polinder, H., M. Schweiker, A. van der Aa, K. Schakib-Ekbatan, V. Fabi, R. K. Andersen, N. Morishita, C. Wang, S. P. Corgnati, P. Heiselberg, D. Yan, B. W. Olesen, T. Bednar, and A. Wagner. 2013, *IEA EBC Annex 53 - Occupant behavior and modeling - Separate Document Volume II. Total energy use in buildings analysis and evaluation methods*.

Raschka, S. 2015, *Python Machine Learning. Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics*, Packt Publishing: Birmingham B3 2PB, UK.

Wang, X., K. A. Smith, and R. J. Hyndman. 2005. "Dimension Reduction for Clustering Time Series Using Global Characteristics." *ICCS 2005* 3516:792–795.