Building energy performance assessment using volatility change based symbolic transformation and hierarchical clustering

Zhenjun Ma[a,*], Rui Yan[a,*], Kehua Li[a], Natasa Nord[b]

[a]Sustainable Buildings Research Centre, University of Wollongong, NSW, 2522, Australia

[b]Department of Energy and Process Engineering, Norwegian University of Science and Technology, Norway

*Email: zhenjun@uow.edu.au; ry721@uowmail.edu.au

Abstract: This paper presents the development of a symbolic transformation based strategy with interpretability and visualisation for building energy performance assessment. The strategy was developed using shape definition language based symbolic transformation and hierarchical clustering. Advanced visualisation techniques including dendrogram, heatmap and calendar view were used to assist in understanding building energy usage behaviours. A comparison of this proposed strategy with a Symbolic Aggregate approximation (SAX) based strategy was also performed. The performance of the proposed strategy was tested and evaluated using the three-year hourly heating energy and electricity usage data of a higher education building. The result demonstrated that the proposed strategy can identify distinct building energy usage behaviours. The visualisation techniques used also assisted the information discovery process. The discovered information helped to understand building energy usage patterns. The comparison of the proposed strategy with the SAX based strategy showed that that the proposed strategy outperformed the SAX based strategy for the case building tested in terms of the variations in building energy usage. This proposed strategy can also be potentially used to evaluate the operational performance of building heating, ventilation and air-conditioning (HVAC) systems.

Keywords: Buildings; Performance evaluation; Symbolic transformation; Visualisation; Hierarchical clustering

1. Introduction

The operation of buildings and building Heating, Ventilation and Air-conditioning (HVAC) systems may suffer from various issues such as equipment malfunctions, sensor reading faults, inappropriate operating procedures, incorrectly configured control systems and equipment performance degradation [1, 2]. Building energy performance assessment is therefore essential to understand building energy performance levels and timely assist in identifying the potential operational issues that may influence building energy efficiency and indoor thermal comfort.

Over the last several decades, many efforts have been made on the development of appropriate methods for effective building energy performance assessment [3]. Pang et al. [2], for instance, proposed a framework to facilitate the comparison between the building actual performance and the expected performance predicted by an EnergyPlus model. Based on a set of performance indicators, Kosai and Tan [4] developed a framework for quantitative analysis of energy performance of zero energy buildings. Yan et al. [5] developed a multi-level strategy for energy performance diagnosis of buildings with limited energy usage data available. Through a case study, Dascalaki et al. [6] concluded that building typologies can be considered as a useful tool in assessing the energy performance of residential buildings.

Data mining, as an interdisciplinary subfield of computer science, is attracting increasing attention and is now being considered as an alternative solution to address the challenges faced by conventional building energy performance assessment methods [7-12]. Gao and Malkawi [8], for instance, presented a methodology using *k*-means clustering for building energy performance benchmarking. The methodology consisted of four steps, including feature selection, cluster analysis, cluster validation and interpretation. Raatikainen et al. [9] described a method using self-organizing maps, U-matrix representation, Sammon's mapping, *k*-means clustering and Davis-Bouldin index to analyse the energy consumption of school buildings. do Carmo and Christensen [10] used *k*-means cluster analysis to identify the typical heating load profiles of Danish single-family detached homes in order to facilitate the development of cost effective demand side management solutions. The use of Partitioning Around Medoids clustering algorithm and Pearson Correlation Coefficient based dissimilarity measure to identify the typical heating load profiles of higher education buildings was presented in [11], in which the typical daily load profiles were identified on the basis of the variation similarity. A clustering method using k-shape algorithm was used by Yang et al. [12] to identify the shape patterns of time series building energy usage data in order to improve the accuracy of forecasting models. From the above studies, it can be seen that cluster analysis is the primary data mining algorithm used in building energy performance assessment and the results showed the effectiveness of using data mining algorithms in the identification of the hidden information from the massive amount of building operational data.

In data mining strategies, data transformation is often used to transform the time series data into suitable formats to support the data mining process. Symbolic transformation is one of the common families of a time series representation approach which converts numeric time series data into symbolic forms [13]. There are two types of symbolic transformation methods, which were developed based on the means of the time segments and the volatility change, that are commonly used [13]. Symbolic Aggregate approXimation (SAX) was used by Miller et al. [14] to transform building energy usage data into alphabets to identify discords and create performance motifs. SAX was also used by Fan et al. [15] to develop a methodology for temporal knowledge discovery of big data collected from building automation systems (BASs). Based on the operational cycle of a chiller identified using a *k*-means clustering algorithm, Habib et al. [16]

first transformed the operational cycles into symbols using SAX and the symbolic representation was further transformed into a bag of word representation for hierarchical clustering. The performance of an air handling unit was studied by Dedemen et al. [17], in which SAX was used to detect the frequently occurring patterns and unexpected patterns in the sensor data provided by the BAS. An extension of SAX was used by Kalluri et al. [18] to extract the features that are characteristic of individual appliance transient states in an office. From the above studies, it can be seen that SAX is the main methods used for symbolic transformation of the time series data to facilitate the data mining process.

With the wide deployment of building management systems and smart meters, a massive amount of high-resolution energy usage data from buildings can now be readily available. This provides a great opportunity to better understand building energy usage characteristics and operational performance through discovering the hidden information behind this massive amount of data. However, without advanced data analytic techniques, the valuable information underneath the massive data may not be properly extracted. This paper presents a strategy for building energy performance assessment using shape definition language based symbolic transformation and hierarchical clustering. Different from the majority of the previous studies used cluster analysis with a focus on the load magnitude for building energy performance assessment, this study used the volatility change based symbolic transformation to convert the time series data into symbolic forms and the typical building energy usage profiles were identified based on the energy usage variations. The advanced visualisation techniques including dendrogram, heatmap and calendar view were used to assist in building energy performance assessment. A comparison of the proposed strategy with a SAX based symbolic transformation strategy was also performed. The performance of this proposed strategy was tested and evaluated using the three-year hourly district heating energy and electricity usage data collected from a higher education building in Norway.

2. Development of the building energy performance assessment strategy

2.1 Outline of the proposed strategy

The outline of the proposed symbolic transformation based strategy to examine the building energy performance is presented in Fig. 1. It mainly consists of four steps, including data collection, data pre-processing, data mining, and an evaluation and interpretation of the results. The first step is the collection of building energy usage data from BASs. The collected data is then pre-processed in the second step, which consists of five main tasks including outlier removal, data segmentation, small variation segments removal, data normalisation and symbolic transformation of the time series data. In this study, the generalised Extreme Studentised Deviate (ESD) test method was used to identify and remove the outliers from the raw data as it can detect one or more outliers in a univariate data set that follows an approximately normal distribution [19]. The details of this test method can be found in [19, 20]. Data segmentation is to transform the data into 24-hour segments in order to form daily load profiles. In order to identify the typical daily energy usage profiles that have distinct patterns, the segments with a small difference between the daily maximum and minimum energy usage were discarded. In this study, 5.0% of the segments with the least difference among all the daily segments were considered as the small difference and were discarded. The daily load profiles were then normalised to a range of 0-1, where 1 is the daily maximum, and 0 is the daily minimum. The last step in the data pre-processing is to transform the segments of the normalised data through the symbolic representation which will be introduced in Section 2.2.

The data mining process starts to identify the pre-defined symbols and shapes and then summarises the distribution of the symbols and shapes to provide a preliminary understanding of the building energy usage behaviour. The Dice coefficient between each pair of the daily load profiles is then calculated to determine the dissimilarity measure for clustering the daily load profiles, which will be introduced in Section 2.3. A hierarchical clustering technique is used to determine the structure and the number of the clusters with the assistance of the heatmap and dendrogram based visualisation techniques. Typical daily load profiles are then formed by calculating the mean value of all the load profiles in each cluster. The distribution of the typical daily load profiles is further plotted as a calendar view to better understand the temporal distribution of the typical daily load profiles identified.

2.2 Symbolic transformation

In this study, a volatility change based method was used to capture the variations in the building energy usage data. The normalised daily load profiles were transformed into a symbolic representation form based on the Shape Definition Language (SDL) proposed by Agrawal et al. [21]. SDL is a small language which allows a variety of queries about the shapes found in histories and has the capability for blurry matching to give the primary focus on overall shape rather than the specific details [21]. Table 1 summarises the symbols used in this study for symbolic transformation, and the corresponding description and definitions. The values used in Table 1 were determined by referring to Agrawal et al. [21]. It is worthwhile to note that these values used might not be the optimal values. In this method, the symbols were defined according to the difference between the value at the i^{th} time step and the corresponding value at the $(i-1)^{th}$ time step. For instance, the value at the i^{th} time step is transformed to the symbol “stable” if the difference between the values at the i^{th} time step and the $(i-1)^{th}$ time step is between -0.05 and 0.05.

Four shapes, including rise, fall, spike, and sink, were also defined based on certain combinations of the symbols to assist in understanding the variations in building energy usage. Table 2 provides a description and definition of the shapes used in this study.

2.3 Dice coefficient based dissimilarity

Similarity/dissimilarity is fundamental to the definition of a cluster [22]. Various similarity and dissimilarity measures such as Euclidean distance, Pearson correlation coefficient, Dice coefficient, Hausdorff distance, probability-based distance, edit distance and dynamic time warping distance have been used in cluster analysis [11, 14, 23-24]. In this study, Dice coefficient based dissimilarity measure as shown in Eq. (1) was used to measure the dissimilarity between the two symbolically represented daily load profiles. Dice coefficient is defined as the ratio of the number of *n-gram*s that are shared by two strings to the total number of *n-grams* in both strings and is shown in Eq. (2) [25].

$$d_{Dice}(X,Y) = 1 - Dice \tag{1}$$

$$Dice = \frac{2 \times \left| n-grams(X) \cap n-grams(Y) \right|}{\left| n-grams(X) \right| + \left| n-grams(Y) \right|} \tag{2}$$

where *n-grams* is a function which divides the original string into substrings with a length of *n*. In this proposed strategy *n* was selected as one. *X* and *Y* are the strings, which are the symbols representing the daily load profiles with a size of 24 in this study.

2.4 Hierarchical clustering

Hierarchical clustering is a relatively simple and unbiased method that is often used to determine whether a given set of data from one group closely resemble another group [26]. There are two hierarchical clustering methods, i.e. agglomerative and divisive, depending on whether the hierarchical decomposition is formed in a bottom-up or top-down approach [27]. One advantage of hierarchical clustering is that the overall process can be represented by a tree structure graph called a dendrogram. The dendrogram can help to visualise the cluster structure and assist in determining the optimal number of clusters.

In this study, clustering the symbols to represent the daily load profiles was achieved based on an agglomerative hierarchical clustering with the complete linkage as shown in Eq. (3) [27]. In the complete linkage, the measurement of the distance between two clusters is the maximum distance between any daily load profile in cluster A and any daily load profile in cluster B. At a specific point, the two clusters that have the smallest complete linkage will be merged into a larger cluster [28].

$$max\{dist(a,b): a \in A, b \in B\} \tag{3}$$

where *a*, *b* are the two daily load profiles belong to the clusters *A* and *B*, respectively, and *dist* is the distance represented by the Dice coefficient-based dissimilarity.

3. Performance test and evaluation of the proposed strategy

In this study, the proposed strategy was implemented in R [29]. The majority of the figures presented in this study were generated using R package ggplot2 [30].

The energy usage data of a higher education building at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway, were used to test and evaluate the performance of this proposed strategy. The building concerned was built in 1965 and is used for laboratory and office purposes with a total floor area of 3,030 m^2. The building heating was provided from district heating. The hourly building heating energy and electricity usage data were collected through a web based Energy Monitoring System.

Fig. 2 presents the collected heating energy and electricity usage data from 2011 to 2013. It can be seen that the heating energy usage varied significantly with the variation in the weather conditions. More heat was generally required from September to April, and there was almost no heating demand during the summer periods (i.e. May to August) in 2012 and 2013, but a small amount of heating energy was still required during the summer periods in 2011 mainly for the domestic hot tap water purpose. The observed building implemented three retrofit measures during 2012, including a) the building was connected to the local university district heating ring with a lower supply water temperature; b) an electric boiler for the domestic hot tap water purpose was installed and; c) the air recirculation was introduced in one of the ventilation systems. These three measures might lead to different heating energy usage patterns in 2012 and 2013 from 2011. In order to identify the typical daily heating energy usage profiles, the heating energy data collected from May to August were discarded and the analysis was focussed on the high heating energy demand periods. From Fig. 2, it can also be observed that the electricity usage data in the first few months of 2011 differed from those during the rest of the time, which should be further investigated. After discussion with the building operator, it seems that the data logging was not working appropriately during that time period. Unlike the heating energy usage data, there was no clear relationship between the variation in the building electricity usage and the seasonal changes. In order to make the analysis be consistent, the electricity data from May to August were also discarded in the following analysis.

3.1 Performance test results based on the heating energy usage data

The generalised ESD test method was first used to detect and remove outliers, the data were then segmented into the daily load profiles and the segments with small variations were removed. The daily load profiles were then normalised to the range of 0-1 and transformed into pre-defined symbolic representations. After the data pre-processing, a total of 686 daily load profiles remained and were used in the following analysis.

The temporal distribution of the symbols and shapes was presented in Fig. 3 to provide a preliminary understanding of the building heating energy usage characteristics. Each bar represents how many times the symbols and shapes appeared at a specific hour. It can be seen that the symbols “up”, “down” and “stable” appeared at the most hours during a day. The symbol “jump” mainly appeared at 03:00, 04:00, 05:00, 07:00, 09:00 and 11:00, while the symbol “plunge” mainly appeared at 08:00, 10:00, 17:00 and 18:00. Moreover, there was a considerable amount of spikes at 05:00, 07:00 and 09:00 and a large number of sink at 10:00. These symbols and shapes indicated that there was a significant change in the heating energy usage at these hours. The rise and fall shapes showed that the increase in the heating energy demand was mainly occurred at around 03:00-5:00 and 09:00 and the decrease in the heating energy demand was occurred at around 16:00-19:00.

Fig. 4 demonstrates the dendrogram of the hierarchical clustering and how the symbols were distributed in the daily load profiles as ordered by the dendrogram. It can be seen that four major clusters were formed when a threshold of 0.97 was used to ensure a relatively uniform distribution of the symbols in each cluster and the clusters 2, 3, 4 were formed with distinct features. For instance, the most daily load profiles in the cluster 2 had the symbol “jump” appeared at 09:00 and 11:00 and the symbol “plunge” appeared at 10:00. The most daily load profiles in the cluster 3 had the symbol “jump” appeared at 04:00 and 07:00 and the symbol “plunge” appeared at 08:00 and 17:00.

The typical daily load profiles formed by the identified clusters are shown in Fig. 5. The number on the top right-hand corner indicated the total number (*T*) of the daily load profiles in the cluster. The boxplot at each hour showed the variance of all daily load profiles at this specific hour, and the width of the box was the significance of the variance. The typical daily heating load profile 1 only represented 38 daily load profiles and the variance of the represented profiles during the main heating demand period was also significant. The major heating demand period shown in the typical daily heating load profile 2 was shorter than the other three typical daily load profiles, and there was a significant spike followed by a sink in the morning. The typical daily load profiles 3 and 4 showed a significant spike at 07:00 and 05:00, respectively, and the main heating demand period occurred from the early morning to about 16:00. Table 3 summarises the main characteristics of the identified typical daily heating load profiles.

Fig. 6 shows the distribution of the typical daily heating load profiles in a calendar view, where the profile 0 represented the days with small variations that were excluded from the analysis. It was shown that the typical daily heating load profiles 2, 3, and 4 mainly represented the daily load profiles of the weekends, and from Tuesday to Friday and Monday, respectively. The typical daily heating load profile 1 mainly appeared in October 2013, and the potential causes are therefore worthwhile to further investigate. The patterns in October 2013 also showed the disordered energy usage with a mixture of different types of typical daily heating load profiles.

3.2 Performance test results based on the electricity usage data

The electricity usage data of the case study building was also analysed. A total of 673 daily electricity load profiles were transferred to a symbolic representation form after the data pre-processing. Fig. 7 shows how symbols and shapes were distributed over the 24 hours. It can be seen that the electricity demand obviously increased at around 09:00-12:00 and decreased at around 17:00-23:00.

Fig. 8 shows a dendrogram of the electricity usage clustering result and the distribution of the symbols in the daily load profiles as ordered by the dendrogram. The threshold of 0.94 was also determined by visualising the distribution of the symbols in order to have a relatively uniform distribution of the symbols in each cluster. A total of seven clusters were formed. It can be seen that the symbol distribution in some clusters showed very distinct characteristics. For instance, in the cluster 3, the symbols "jump" and "plunge" appeared alternately, which indicated that the electricity usage fluctuated significantly and is therefore worthy of further investigation.

The typical daily electricity load profiles identified are presented in Fig. 9. The typical daily electricity load profiles 1, 2, 4, 5 and 6 had a similar trend where the electricity usage started to increase at about 09:00 and decrease at about 17:00. The typical daily electricity load profiles 3 and 7 showed significant fluctuations during the most hours, indicating that the building might be operated under the abnormal conditions. Table 4 summarises the key features of the typical daily electricity load profiles identified.

Fig. 10 presents a calendar view of the temporal distribution of the typical daily electricity load profiles. It is noted that the daily electricity load profile 0 represented those days with small variations that were excluded from the analysis. It can be seen that there was a very uniform distribution of the typical daily electricity load profiles in the first four months of 2011, where Monday to Wednesday were under the typical daily electricity load profile 4, Thursday and weekends were under the typical daily electricity load profile 3, and Friday was under the typical daily electricity load profile 6. Over the remaining time, the operation was mainly under the typical daily electricity load profile 1 during the majority of the days, especially the weekdays. The typical daily electricity load profile 2 mainly occurred in September and the first week of October 2012, which is also an interesting pattern for further investigation. In summary, the calendar view of the typical daily electricity load profiles did not show a uniform electricity usage distribution during a week.

4. Interpretation of the information discovered

As shown in the calendar view (Fig. 6), during the majority of the weeks, the heating energy usage data had a uniform distribution pattern. In order to confirm this, the heating energy usage data from one week starting from the third Monday of November 2013, which was considered as a typical heating week, were presented in Fig. 11(a). Another week starting from the second Monday of October 2013, which was considered as a non-typical heating week, was presented in Fig. 11(b). The shaded areas were the time periods with the heating load higher than 40% of the daily maximum heating load, while the orange and light blue colours represented the weekdays and weekends, respectively.

From Fig. 11(a), it can be observed that the general trend of the daily heating load profile was in line with that of the typical daily load profiles identified. The heating demand profiles on Monday to Friday were very similar but the heating demand on Monday was one hour earlier than those from Tuesday to Friday, which was consistent with that presented in the typical daily heating load profiles identified.

The data from the non-typical week indicated that the main heating demand period during the weekdays was the same as the weekdays of the typical week, but with a different trend. For instance, the heating load profiles of the non-typical week from Tuesday to Friday did not show a spike at 07:00. The heating load profile at the weekends was significantly different from those in the typical week. These heating energy usage patterns were not presented by the identified typical daily load profiles, which indicated that the patterns of the disorder in the calendar view can help to identify the abnormal heating energy usage.

Four weeks of the electricity usage data with interesting patterns were presented in Fig. 12. The first two consecutive weeks extracted starting from the first Monday of January 2011 as the similar patterns lasted for four months in 2011. The third interesting week started from the first Monday of November 2013, corresponding to the typical daily electricity load profile 1 during the weekdays and the typical daily electricity load profile 6 during the weekends. The fourth interesting week started on the third Monday of September 2012 and the electricity usage was in line with the typical daily electricity load profile 2.

The electricity usage data of the first two interesting weeks (Fig. 12a) showed that there was a large fluctuation in the electricity usage, especially on weekends. This means that the data from this period cannot reflect the typical electricity usage of the building. However, the reason behind this is worthwhile to investigate.

The third interesting week (Fig. 12b) represented the typical weekdays where the daily load profiles were in line with the typical daily electricity load profile 1. The electricity energy usage began to increase significantly at around 10:00 and dropped significantly in the late afternoon at around 18:00. However, the actual daily electricity load profiles of each weekday were different. The daily electricity load profiles during the weekends were consistent with that of the typical daily electricity load profile 6 with a small amount of electricity usage although they shared the similar trends.

The daily load profiles of the fourth interesting week (Fig. 12c) from Monday to Saturday were in line with that of the typical daily electricity load profile 2, while the daily electricity load profile on Sunday was consistent with that of the typical daily electricity load profile 1. The trend of the electricity usage for the first six days had different patterns, but they all shared a common feature that a significant increase in the electricity usage started at 09:00 except on Saturday and the main electricity usage lasted until the late night. This is an interesting point for the further investigation to understand why the electricity usage behaviour for this week was different from the others.

The electricity usage was related to the activities of the occupants. The building has a wind tunnel that was used about 200 days per year for the research purpose. The wind tunnel was used randomly based on the research requirement. The analysis showed that the high electricity demand started at around 09:00. On some days, the large electricity demand remained until the late night even though the heating supply significantly decreased at 18:00. The extended electricity demand might be due to light and computer use, because researchers might work longer than the typical working time. The heating and ventilation systems were usually scheduled to provide a higher temperature and a larger amount of air during the typical working time.

Unlike the heating energy usage, the electricity usage data showed more variations. This is reasonable as the electricity was used by various facilities such as lighting, computers, and laboratory equipment. This means that all behaviours in the three-year electricity usage data cannot be fully represented by the typical daily electricity load profiles identified. However, the identified typical daily electricity load profiles can be used to assist in understanding the electricity usage behaviours and to provide the guidance on the electricity usage data analysis as well as to detect the abnormal electricity usage behaviours.

5. Comparison between the proposed strategy with a Symbolic Aggregate approXimation based strategy

Symbolic Aggregate approXimation (SAX) [31] is another commonly used symbolic transformation method for time series data analysis and has been used in a number of building energy studies [14, 15]. In this section, a comparison between the proposed strategy with a SAX based symbolic transformation strategy was presented to confirm the effectiveness of the proposed strategy.

In this comparison, all the other steps in the SAX based strategy were the same as that of the proposed strategy, but SAX was used to replace the SDL to transform the time series data. In the SAX based strategy, the original time series data were transformed into the segments with a length n and the mean value of a segment was represented by A letters such as a, b and c. Since this data was normalised to a range of 0-1, the equal size breakpoints were used, which means that each letter represents *1/A* in the range of 0-1. In the comparison, each hour data was considered as a segment (i.e. n=1) and the five letters were used to represent the time series data (i.e. A = *5 with* the letters of *a, b, c, d* and *e*). Fig. 13 shows the dendrogram of the SAX based clustering and the corresponding heatmap by using the heating energy data. It can be seen that there were four clusters identified by using the same method used in the proposed strategy. The heatmap showed that the two clusters with the most daily heating load profiles (i.e. orange and red) had different and distinctive patterns.

Fig. 14 shows the typical daily heating load profiles formed by the identified clusters. The typical daily heating load profiles 1 and 2 only accounted for a small number of the daily load profiles, while the typical daily heating load profiles 3 and 4 were similar to the typical daily heating load profiles 3 and 2 presented in Fig. 5 identified by the proposed strategy. The temporal distribution of the typical daily heating load profiles is shown as a calendar view in Fig. 15. It can be seen that the SAX based strategy successfully isolated the weekend and weekday load profiles, but the unique heating energy usage pattern on Monday identified by the proposed strategy was not identified by the SAX based strategy.

This comparison demonstrated that both strategies can identify the key patterns related to the building heating energy usage, but the proposed strategy can identify more features and better reflect the unique energy usage behaviours from the perspective of the energy usage variation, in comparison to the SAX based strategy.

6. Conclusion

This paper presented a combination of symbolic transformation and cluster analysis based strategy to evaluate the building energy performance. In this strategy, the building daily load profiles were first transformed into the volatility change based symbols. The symbols were then grouped to represent the daily load profiles through hierarchical clustering and Dice coefficient based dissimilarity measure to identify the typical daily load profiles. A key advantage of this strategy is that it can utilise the advanced visualisation techniques to help understand the information extracted from the raw data.

The performance of this strategy was evaluated using the three-year heating energy and electricity usage data from a higher education building in Norway. The results showed that the proposed strategy can discover the information related to the building energy usage behaviour. The visualisation techniques also helped to discover the hidden information and better understand the typical patterns of energy usage as well as identify the unique energy usage behaviours. The results from this study can be further used to assist in the fault detection and diagnosis. It was shown that the proposed strategy worked better with the heating energy usage data than the electricity usage data mainly due to the fact that the electricity was consumed by different equipment and varied considerably in daily operations. During some weekends, the electricity usage was much lower than that of the weekdays but with a similar trend and it was therefore classified into the same cluster. This means that the magnitude of the energy usage should be considered as a factor in the further improvement of the proposed strategy. The results also showed that proposed strategy showed a better performance to identify the characteristics of energy usage behaviours of the case study building in comparison with a SAX based strategy in terms of the variations in building energy usage. In order to capture more information from building energy usage data, it might be worthwhile to develop advanced strategies which can take both magnitude similarity and variation similarity into consideration simultaneously. The proposed strategy has a potential to be used to evaluate the operational performance of building HVAC systems.

Acknowledgement

This research work was made possible through an Endeavour Research Fellowship. The first author would like to thank the support of Australian Government - Department of Education and Training.

References

[1] Z. Ma, S. Wang, Online fault detection and robust control of condenser cooling water systems in building central chiller plants, Energy and Buildings 43 (2011) 153-165.

[2] X. Pang, M. Wetter, P. Bhattacharya, P. Haves, A framework for simulation-based real-time whole building performance assessment, Building and Environment 54 (2012) 100-108.

[3] S. Wang, C. Yan, F. Xiao, Quantitative energy performance assessment methods for existing buildings, Energy and Buildings 55 (2012) 873-888.

[4] S. Kosai, C. Tan, Quantitative analysis on a zero energy building performance from energy trilemma perspective, Sustainable Cities and Society 32 (2017) 130-141.

[5] C. Yan, S. Wang, F. Xiao, D. Gao, A multi-level energy performance diagnosis method for energy information poor buildings, Energy 83 (2015) 189-203.

[6] E.G. Dascalaki, K.G. Droutsa, C.A. Balaras, S. Kontoyiannidis, Building typologies as a tool for assessing the energy performance of residential buildings – A case study for the Hellenic building stock, Energy and Buildings 43 (2011) 3400-3409.

[7] R. Yan, Z. Ma, Y. Zhao, G. Kokogiannakis, A decision tree based data-driven diagnostic strategy for air handling units, Energy and Buildings 133 (2016) 37-45.

[8] X. Gao, A. Malkawi, A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm, Energy and Buildings 84 (2014) 607-616.

[9] M. Raatikainen, J.P. Skön, K. Leiviskä, M. Kolehmainen, Intelligent analysis of energy consumption in school buildings, Applied Energy 165 (2016) 416-429.

[10] C.M.R. do Carmo, T.H. Christensen, Cluster analysis of residential heat load profiles and the role of technical and household characteristics, Energy and Buildings 125 (2016) 171-80.

[11] Z. Ma, R. Yan, N. Nord, A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings, Energy 134 (2017) 90-102.

[12] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S.E. Lee, C. Sekhar, K.W. Tham, k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement, Energy and Buildings 146 (2017) 27-37.

[13] T.C. Fu, A review on time series data mining, Engineering Applications of Artificial Intelligence 24 (2011) 164-181.

[14] C. Miller, Z. Nagy, A. Schlueter. Automated daily pattern filtering of measured building performance data, Automation in Construction. 49 (2015) 1-17.

[15] C. Fan, F. Xiao, H. Madsen, D. Wang, Temporal knowledge discovery in big BAS data for building energy management, Energy and Buildings 109 (2015) 75-89.

[16] U. Habib, K. Hayat, G. Zucker, Complex building's energy system operation patterns analysis using bag of words representation with hierarchical clustering, Complex Adaptive Systems Modeling 4 (8) (2016) 1-20.

[17] G. Dedemen, M. Vakilinezhad, S. Ergan, Using data driven methodologies to identify patterns in BAS data to support facility operations, Computing in Civil Engineering (2017) 282-289.

[18] B. Kalluri, A. Kamilaris, S. Kondepudi, H.W. Kua, K.W. Tham, Applicability of using time series subsequences to study office plug load appliances, Energy and Buildings 127 (2016) 399-410.

[19] NIST/SEMATECH, e-handbook of statistical methods, http://www.itl.nist.gov/div898/handbook/, accessed 20 01 2017.

[20] J.E. Seem, Using intelligent data analysis to detect abnormal energy consumption in buildings, Energy and Buildings 39 (2007) 52-58.

[21] R. Agrawal, G. Psaila, E.L. Wimmers, M. Zaït, Querying shapes of histories, Proceedings of the 21st International Conference on Very Large Data Bases, Zürich, Switzerland, (1995) 502-514.

[22] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: A review, ACM Computing Surveys 31 (3) (1991) 264-323.

[23] E.S. Ristad, P.N. Yianilos, Learning string-edit distance, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (5) (1998) 522-532.

[24] S. Aghabozorgi, A.S. Shirkhorshidi, T.Y. Wah, Time-series clustering – A decade review, Information Systems 53 (2015) 16-38.

[25] G. Kondrak, N-gram similarity and distance, Proceedings of the 12th International Conference on String Processing and Information Retrieval, Buenos Aires, Argentina, (2005) 115-126.

[26] Y. Vodovotz, G. An, Translational systems biology: Concepts and practice for the future of biomedical research, Academic Press, Elsevier, USA, 2014.

[27] J. Han, M. Kamber, Data mining: Concepts and techniques (Second edition), Morgan Kaufmann Publishers, USA, 2006.

[28] D. Krznaric, C. Levcopoulos, Optimal algorithms for complete linkage clustering in d dimensions, Theoretical Computer Science, 286 (1) (2002) 139-149.

[29] R Development Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2008.

[30] H. Wickham, ggplot2: Elegant graphics for data analysis. Springer, New York, 2009.

[31] J. Lin, E. Keogh, L. Wei, and S. Lonardi, Experiencing SAX: A novel symbolic representation of time series, Data Mining and Knowledge Discovery, 15 (2007) 107-144.

Table 1 Description and definition of the symbols

Symbol	Description	Lower bound	Upper bound
stable	virtually no variation	-0.05	0.05
jump	significant increase	0.20	1.00
up	slight increase	0.05	0.2
down	slight decrease	-0.20	-0.05
plunge	significant decrease	-1.00	-0.20

Table 2 Description and definition of the shapes

Shape	Description	Definition
rise	continues the increasing trend	continuous up or jump symbols
fall	continues the descending trend	continuous down or plunge symbols
spike	a significant peak	a jump followed by a plunge
sink	a significant trough	a plunge followed by a jump

Table 3 Key characteristics of the identified typical daily heating energy usage profiles

Typical heating load profile No.	Estimated high heating demand period	Total number of days	Main characteristics
1	04:00-17:00	38	The heating demand greatly increased from 03:00 with two small peaks at 7:00 and 11:00.
2	09:00-17:00	190	The major heating period started with a significant spike at 09:00 and ended at 17:00.
3	04:00-17:00	344	The main heating period started at 04:00 and lasted until 17:00 with a significant spike at 07:00.
4	03:00-17:00	114	The main heating period started at 03:00 and lasted until 17:00 with a significant spike at 05:00.

Table 4 Key characteristics of the identified typical daily electricity load profiles

Typical electricity load profile No.	Estimated high electricity demand period	Total number of days	Main characteristics
1	10:00-19:00	373	The electricity demand increased significantly from 09:00 and reached the peak at around 13:00 and then started to decrease at 17:00.
2	09:00-20:00	63	The trend of the demand variation was similar to the profile 1 except that the decrease was not as sharp as the profile 1 at around 18:00.
3	Not clear	53	The profile had a large fluctuation, indicating very unstable electricity usage behaviour.
4	09:00-21:00	100	The trend was similar to the profiles 1 and 2 but with more fluctuations. A small peak occurred at 21:00.
5	10:00-21:00	28	The profile was similar to the profiles 1 & 2 except a small peak at 07:00.
6	09:00-21:00	46	The overall trend was similar to the profile 4 but with more fluctuations.
7	Not clear	10	The profile only represented a few days with the fluctuating electricity usage behaviour.

Figure Captions

Fig. 1 Outline of the proposed symbolic transformation based strategy.

Fig. 2 Illustration of building heating energy and electricity usage data.

Fig. 3 Temporal distribution of the symbols and shapes - heating energy usage data.

Fig. 4 Dendrogram of the hierarchical clustering result and distribution of the symbols in the daily load profiles ordered by the dendrogram - heating energy usage data.

Fig. 5 Typical daily heating load profiles formed by the identified clusters.

Fig. 6 Calendar view of the distribution of the typical daily heating load profiles.

Fig. 7 Temporal distribution of the symbols and shapes - electricity usage data.

Fig. 8 Dendrogram of the hierarchical clustering result and distribution of the symbols in the daily load profiles as ordered by the dendrogram - electricity usage data.

Fig. 9 Typical daily electricity load profiles formed by the identified clusters.

Fig. 10 Calendar view of the distribution of the typical daily electricity load profiles.

Fig. 11 The heating energy usage data of a typical week and a non-typical week.

Fig. 12 The electricity usage data of the four selected interesting weeks.

Fig. 13 Distributions of the symbols in the daily heating load profiles ordered by the dendrogram using SAX based method.

Fig. 14 Typical heating load profiles formed by the identified clusters using the SAX based strategy.

Fig. 15 Calendar view of the distribution of typical heating load profiles using the SAX based strategy.

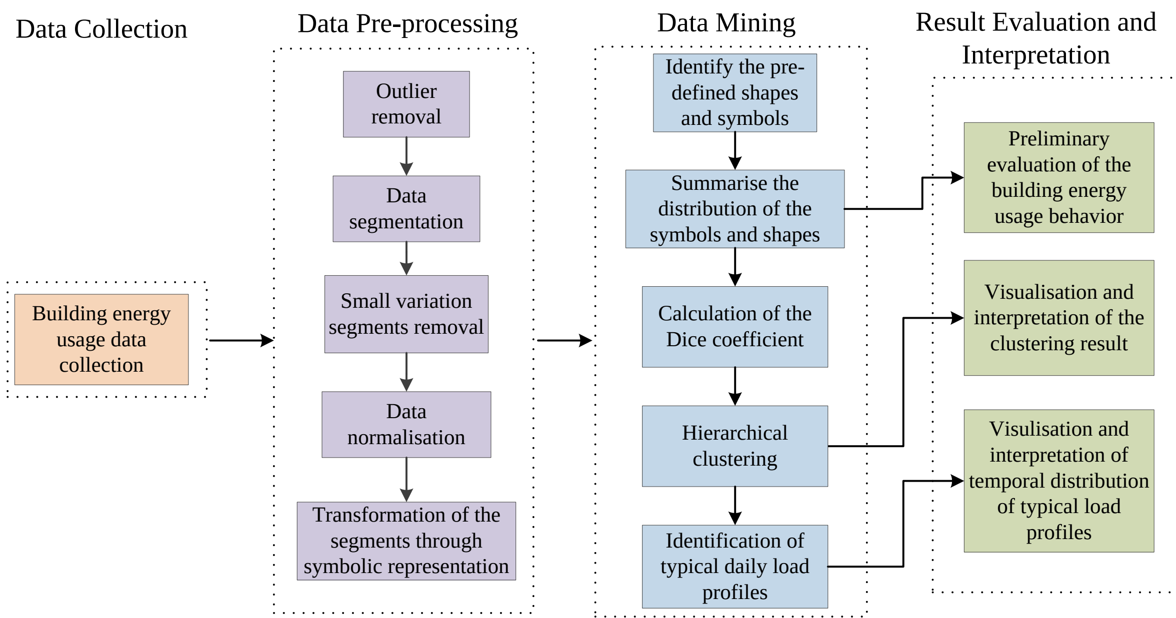

Fig. 1 Outline of the proposed symbolic transformation based strategy.

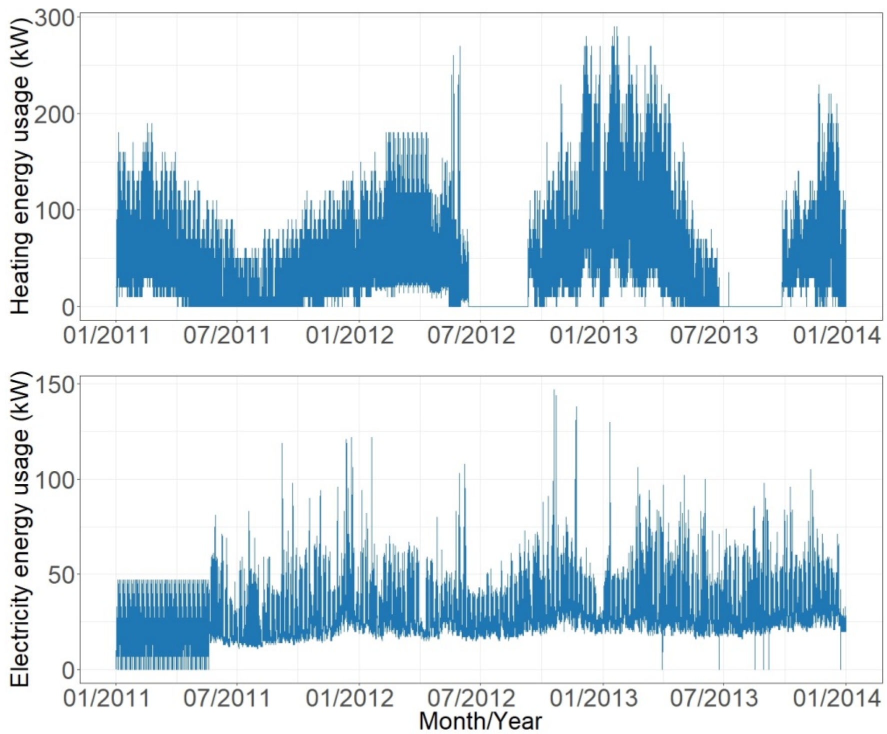

Fig. 2 Illustration of building heating energy and electricity usage data.

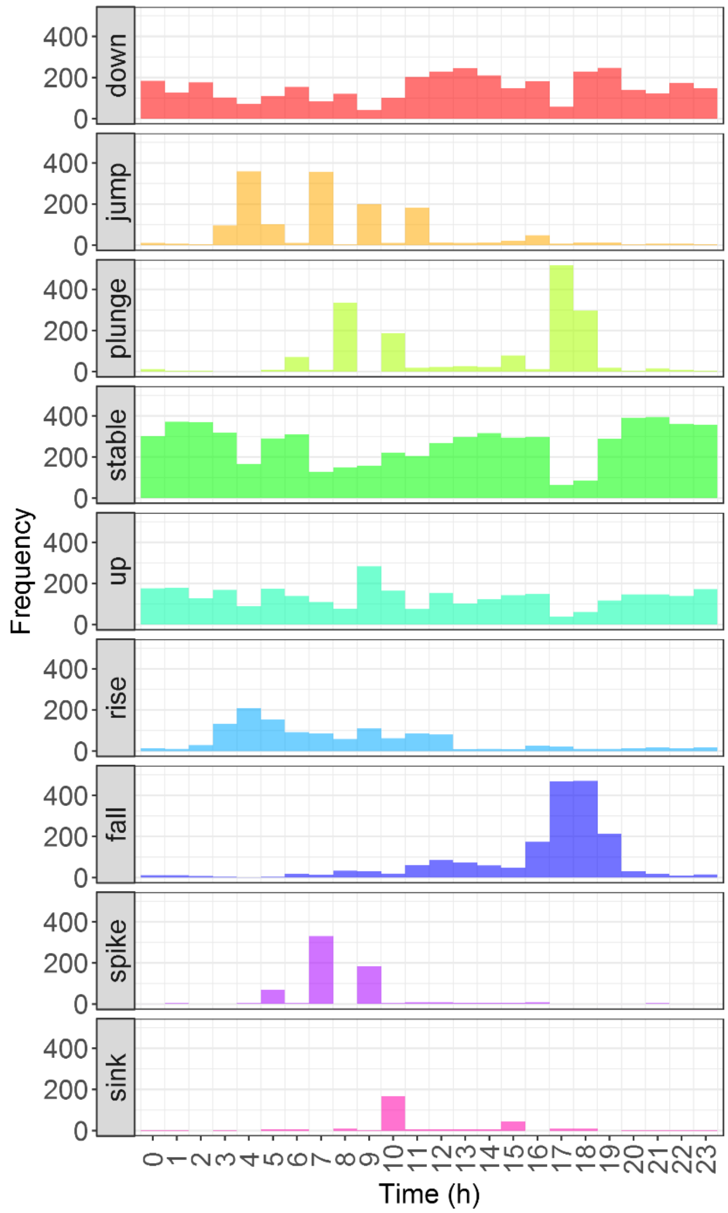

Fig. 3 Temporal distribution of the symbols and shapes - heating energy usage data.

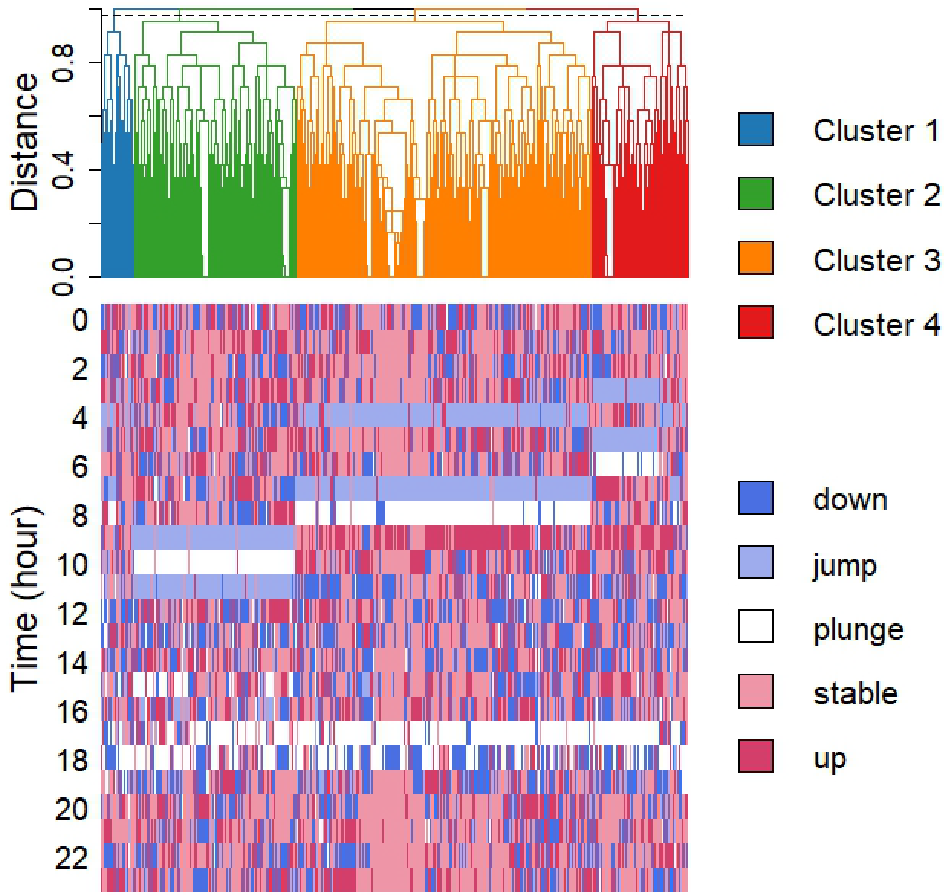

Fig. 4 Dendrogram of the hierarchical clustering result and distribution of the symbols in the daily load profiles ordered by the dendrogram - heating energy usage data.

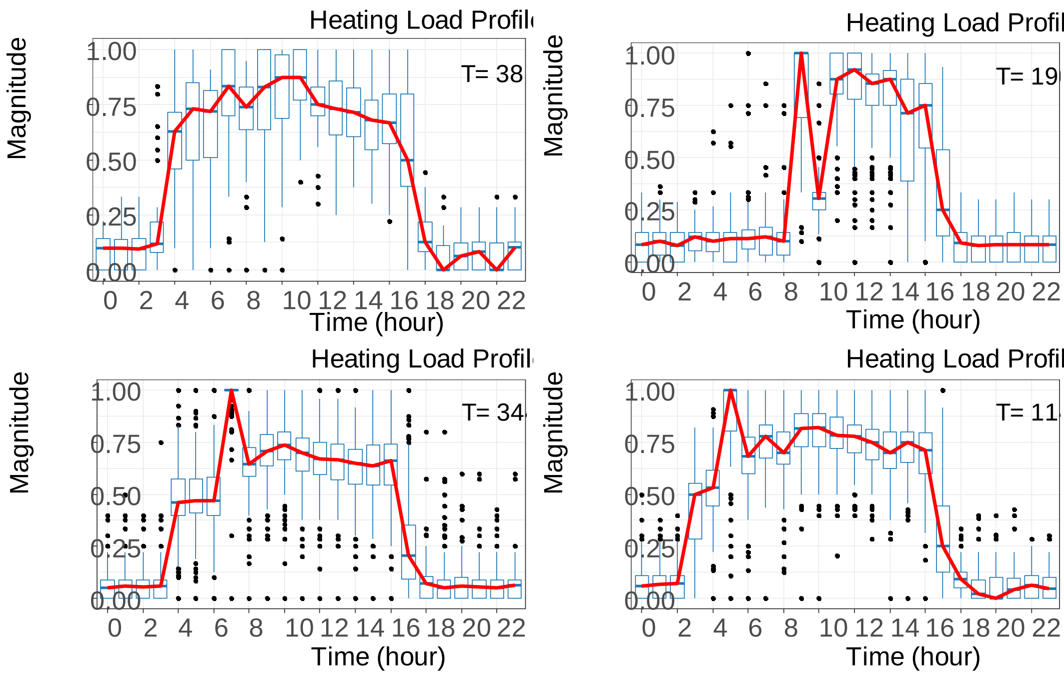

Fig. 5 Typical daily heating load profiles formed by the identified clusters.

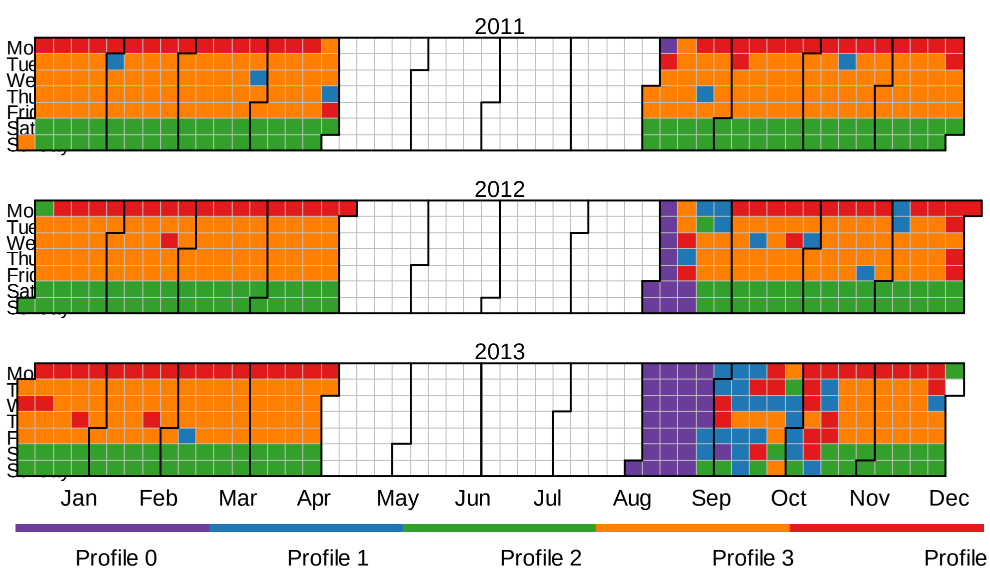

Fig. 6 Calendar view of the distribution of the typical daily heating load profiles.

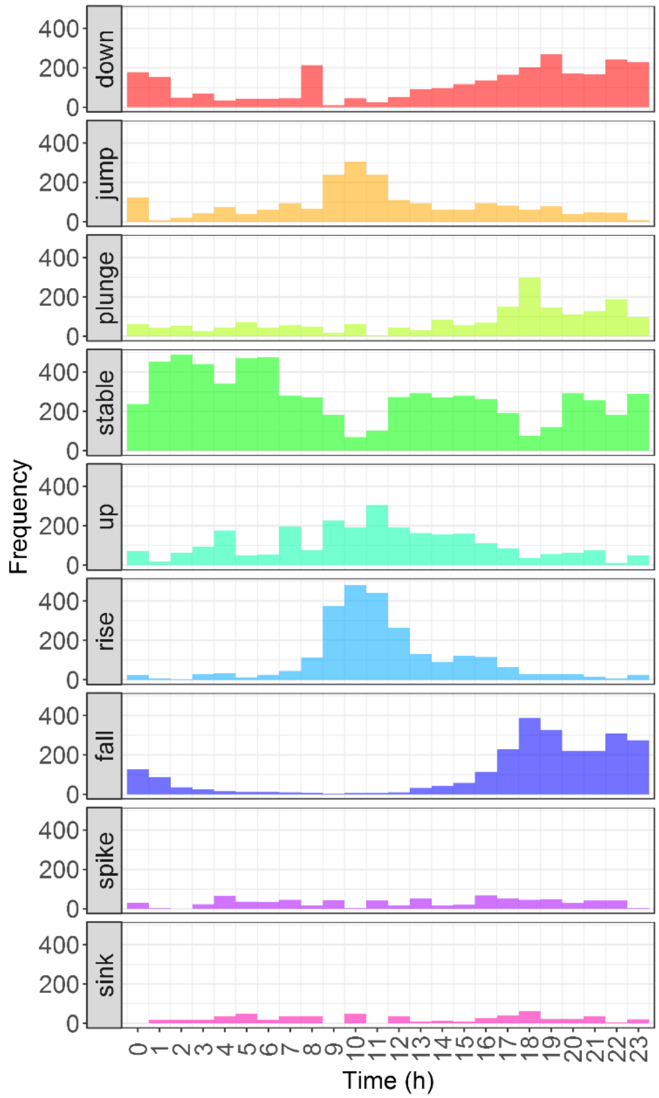

Fig. 7 Temporal distribution of the symbols and shapes - electricity usage data.

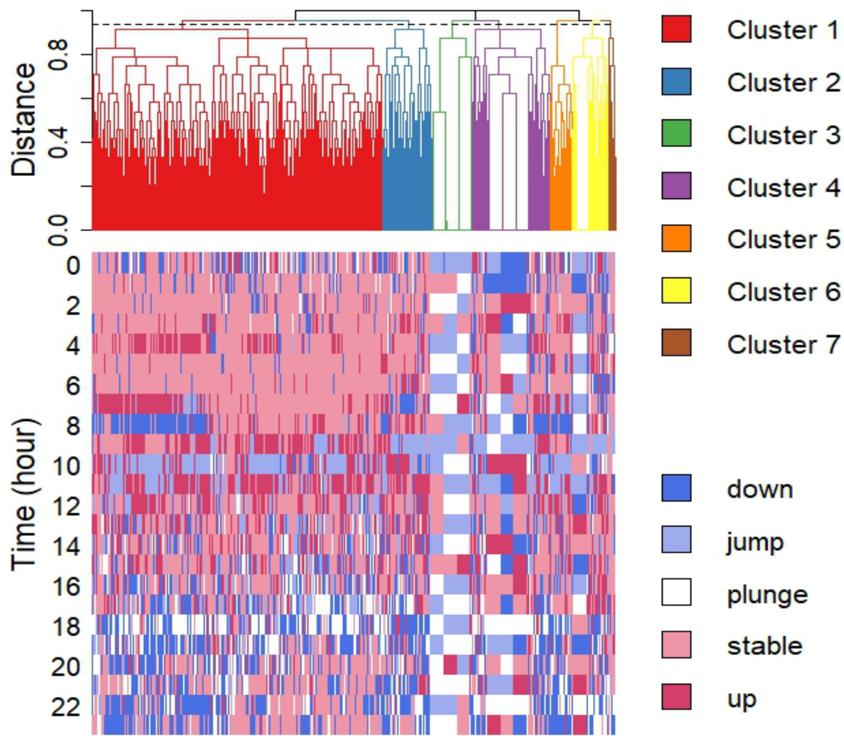

Fig. 8 Dendrogram of the hierarchical clustering result and distribution of the symbols in the daily load profiles as ordered by the dendrogram - electricity usage data.

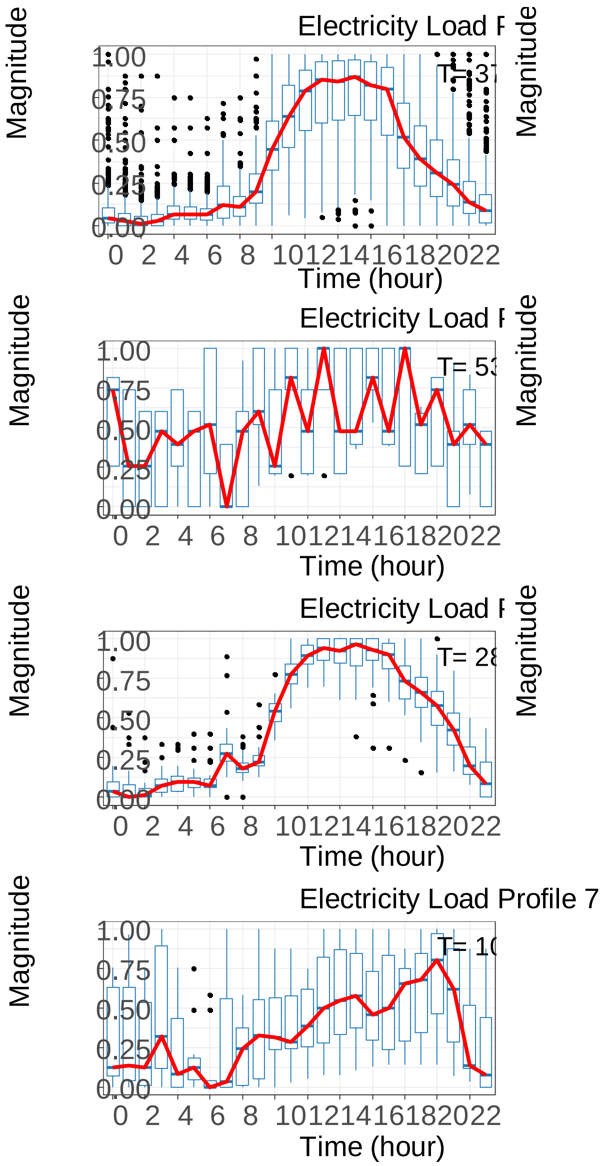

Fig. 9 Typical daily electricity load profiles formed by the identified clusters.

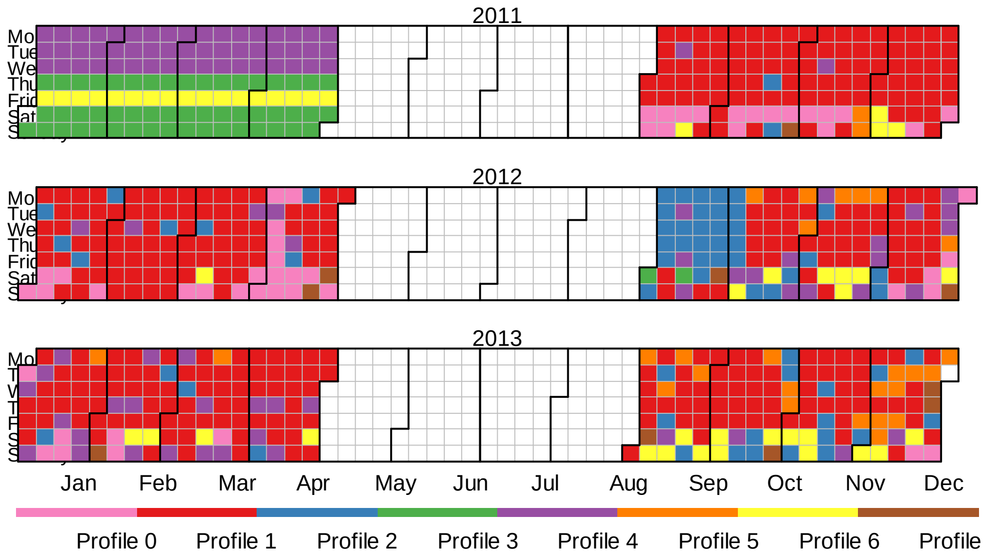

Fig. 10 Calendar view of the distribution of the typical daily electricity load profiles.

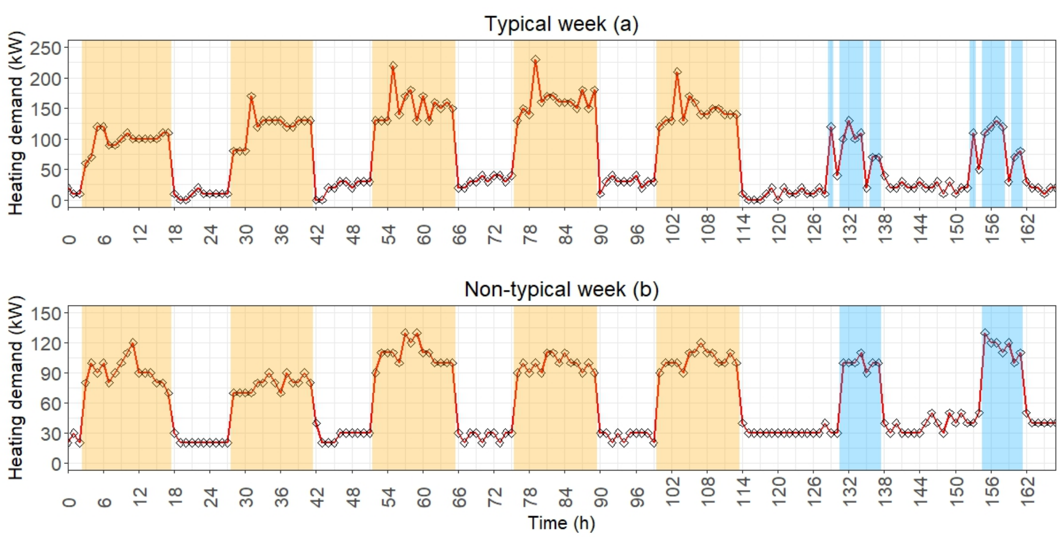

Fig. 11 The heating energy usage data of a typical week and a non-typical week.

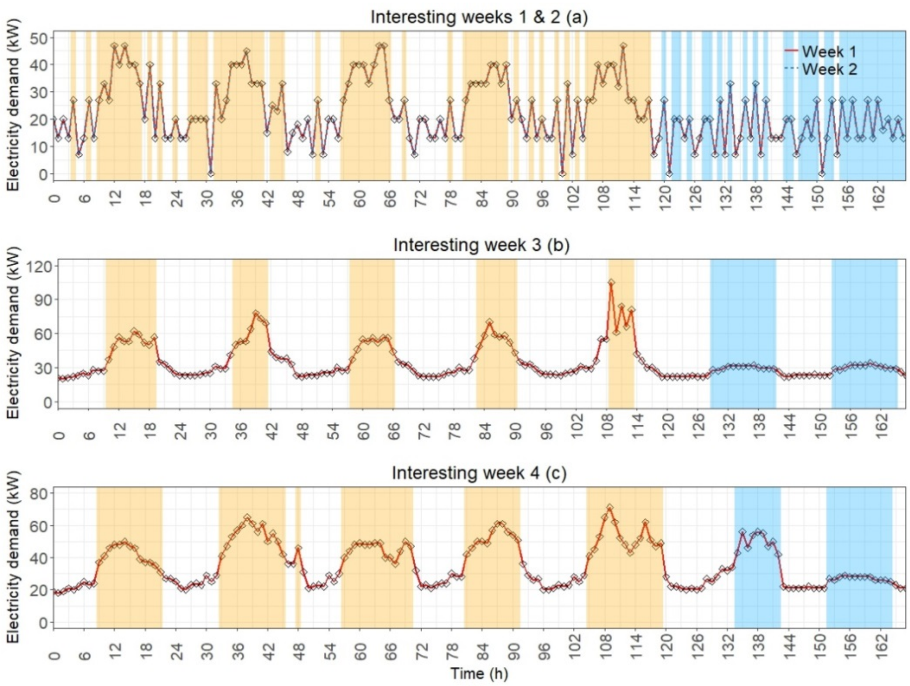

Fig. 12 The electricity usage data of the four selected interesting weeks.

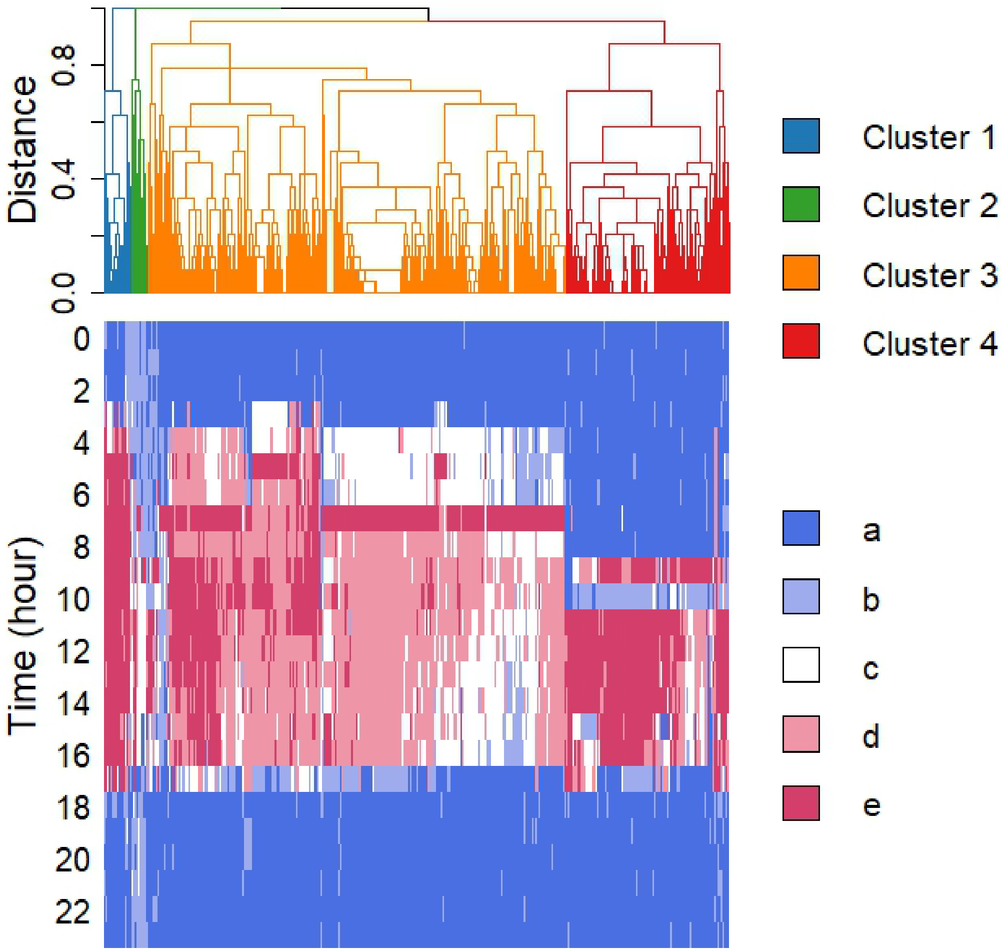

Fig. 13 Distributions of the symbols in the daily heating load profiles ordered by the dendrogram using SAX based method.

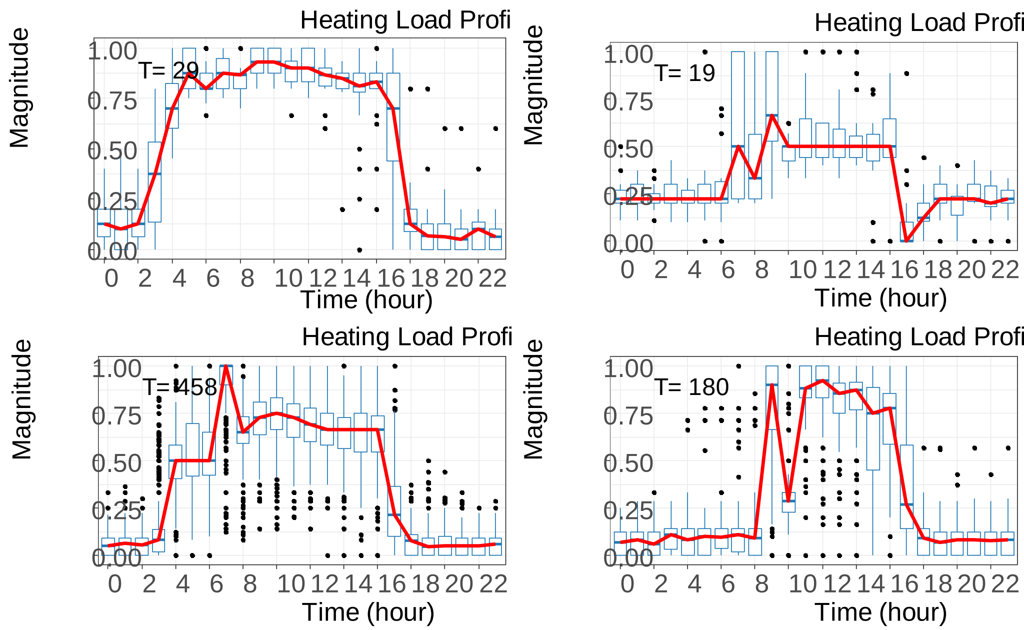

Fig. 14 Typical heating load profiles formed by the identified clusters using the SAX based strategy.

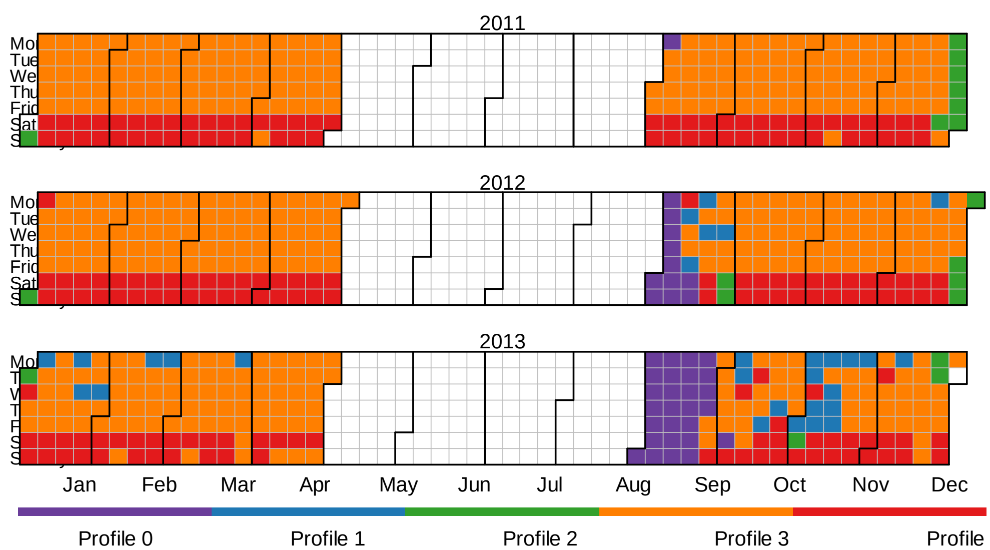

Fig. 15 Calendar view of the distribution of typical heating load profiles using the SAX based strategy.