2017

A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings

Zhenjun Ma
University of Wollongong, zhenjun@uow.edu.au

Rui Yan
University of Wollongong, ry721@uowmail.edu.au

Natasa Nord
Norwegian University of Science and Technology

A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings

Abstract

This paper presents a variation focused cluster analysis strategy to identify typical daily heating energy usage profiles of higher education buildings. Different from the existing cluster analysis studies which were primarily developed using Euclidean distance as the dissimilarity measure and tended to group the daily load profiles with similar magnitudes, Partitioning Around Medoids (PAM) clustering algorithm with Pearson Correlation Coefficient-based dissimilarity measure was used in this study to group the daily load profiles on the basis of the variation similarity. A comparison of the proposed strategy with a k-means cluster analysis with Euclidean distance as the dissimilarity measure was also performed. The performance of the proposed strategy was tested and evaluated using the three-year hourly heating energy usage data collected from 19 higher education buildings in Norway. The results demonstrated the effectiveness of the proposed strategy in identifying the typical daily energy usage profiles. The identified typical heating load profiles provided the information such as the peaks and troughs of the daily heating demand, daily high heating demand period and daily load variation. The identified profiles also helped to categorize multiple buildings into different groups in terms of the similar energy usage behaviors to support further energy efficiency initiatives.

Disciplines

Engineering | Science and Technology Studies

Publication Details

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

A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings

Zhenjun Ma[a,*], Rui Yan[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 a variation focused cluster analysis strategy to identify typical daily heating energy usage profiles of higher education buildings. Different from the existing cluster analysis studies which were primarily developed using Euclidean distance as the dissimilarity measure and tended to group the daily load profiles with similar magnitudes, Partitioning Around Medoids (PAM) clustering algorithm with Pearson Correlation Coefficient-based dissimilarity measure was used in this study to group the daily load profiles on the basis of the variation similarity. A comparison of the proposed strategy with a k-means cluster analysis with Euclidean distance as the dissimilarity measure was also performed. The performance of the proposed strategy was tested and evaluated using the three-year hourly heating energy usage data collected from 19 higher education buildings in Norway. The results demonstrated the effectiveness of the proposed strategy in identifying the typical daily energy usage profiles. The identified typical heating load profiles provided the information such as the peaks and troughs of the daily heating demand, daily high heating demand period and daily load variation. The identified profiles also helped to categorize multiple buildings into different groups in terms of the similar energy usage behaviors to support further energy efficiency initiatives.

Keywords: Cluster analysis; Load profile; Pearson Correlation Coefficient; Higher education buildings

Nomenclature

C	set of the identified clusters
cov	covariance
d	distance
D	Dunn index
k	number of clusters
n	number of observations
N_d	number of days belongs to a typical daily load profile
$N_{d,max}$	maximum number of days belongs to a typical daily load profile
o	data point identified as a medoid
p	tail area probability
PCC	Pearson Correlation Coefficient
q	data point
R	studentized deviate
RP	relative proportion
t	t-distribution
X, Y	vectors
x,y	values of individual dimension

Greek letters

α	significance level
λ	critical value
σ	standard deviation

ϕ	identified clusters
Subscripts	
ED	Euclidean distance
PCC	Pearson correlation coefficient

1. Introduction

Building energy efficiency is essential for reducing global energy usage and promoting environmental sustainability, as the building sector contributes to a large proportion of the total energy usage worldwide [1, 2]. With the development of automatic meter reading systems, massive high-resolution energy usage data from buildings can now be easily collected with a reasonably low cost [3]. This massive amount of data provides a great opportunity to assist in better understanding building energy usage characteristics and operational performance, and in extracting the useful and hidden information to support the areas including but not limited to building energy performance assessment and benchmarking, building load estimation and demand side management, occupant behavior prediction, and fault detection and diagnosis of heating, ventilation and air-conditioning systems.

Identification of typical building load profiles based on the collected massive energy usage data has been proved to be an effective way to understand building energy usage characteristics and help to develop cost effective load shifting and peak demand control strategies [4, 5]. Cluster analysis, as a data mining technique to discover the natural grouping(s) of a set of patterns, points, or objects [6], has been used in a number of studies to identify typical building load profiles [4, 5, 7, 8]. Jota et al. [4], for instance, used an agglomerative hierarchical clustering algorithm with Euclidean distance (ED) to identify the typical building load profiles, which were further used to predict the accumulated energy usage at the end of the day and the daily peak demand. Typical heating load profiles of Danish single-family detached homes were studied by do Carmo and Christensen [5] using the k-means algorithm. Three types of typical load profiles, i.e. high demand, medium demand and low demand, were identified for the buildings operated during weekdays and weekends, respectively. A binary regression analysis was also performed to identify the explanatory factors governing the different heating load profiles. The implementation and evaluation of a cluster analysis approach for smart meter data were reported by Flath et al. [7], in which the k-means algorithm was used to identify typical building daily and weekly load profiles of a business intelligence environment. Symbolic Aggregate approXimation (SAX) method was used by Miller et al. [8] to transform building energy usage data into alphabets while the k-means algorithm was used to identify the typical daily load profiles. Fuzzy c-means (FCM) was adopted by Fernandes et al. [9] to identify the typical gas consumption profiles of residential buildings. It was found that the gas consumption peaks were related to the upper-middle social class with a high income and the highest daytime off-peak gas usage was related to the ageing population. Panapakidis et al. [10] utilized several clustering algorithms, including k-means, k-means++, minimum variance criteria, FCM and self-organizing map (SOM), to identify typical building electricity usage profiles. It was concluded that SOM and k-means++ in the frequency domain outperformed the other clustering techniques in terms of the clustering error.

Cluster analysis distinguishes data vectors based on a certain type of dissimilarity measures [11]. Although different cluster analysis algorithms have been proposed for different scenarios, to the best knowledge of the authors, the existing studies on the identification of building typical load profiles using cluster analysis were primarily developed using ED as the dissimilarity measure. Cluster analysis using ED-based dissimilarity measure tends to identify the daily load profiles that are similar in terms of the intensity rather than the variation. In other words, the typical daily load profile identified using cluster analysis with ED as the dissimilarity measure is more related to the load magnitude. For example, do Carmo and Christensen [5] labeled the identified load profiles as high demand, medium demand, and low demand. ED-based dissimilarity measure is also difficult to identify building daily load profiles with similar variations but with different magnitudes, which will be elaborated in Section 2.2.

Higher education buildings have an important role in the minimization of greenhouse gas emissions from the built environment and in assisting the mitigation and adaptation of our society to climate change [12]. This paper presents a strategy using Partitioning Around Medoids (PAM) clustering algorithm to identify typical daily heating energy usage profiles of a group of higher education buildings. The novelty of this paper is to use Pearson Correlation Coefficient (PCC) as the dissimilarity measure to cluster daily heating energy usage profiles, in which the typical energy usage profiles are identified based on the load variation instead of the load magnitude, which is different from the majority of the previous studies used cluster analysis with ED as the dissimilarity measure. Based on the identified typical load profiles, a hierarchical clustering was used to group the buildings with similar heating energy usage characteristics. A comparison of the proposed strategy with an ED-based k-means cluster analysis strategy was also performed. The performance of the proposed strategy was evaluated using three-year hourly district heating energy usage data collected from 19 higher education buildings in Norway.

2. Development of the variation focused cluster analysis strategy

2.1 Outline of the variation focused cluster analysis strategy

The outline of the proposed variation focused cluster analysis strategy is illustrated in Fig. 1, which was developed following the standard Knowledge Discovery from Database (KDD) process [13]. It mainly consisted of four steps, including data collection, data pre-processing, data mining, and results evaluation and interpretation.

The collection of hourly energy usage data of individual buildings was the first step and the necessary data can be generally collected from building management systems. There were four tasks in the data pre-processing step, including outlier removal, data standardization, data segmentation and the removal of the weekend data and the data segments with small variations. In this study, the generalized Extreme Studentized Deviate (ESD) test method was used to identify and remove the outliers in the collected raw data. As the magnitude of the energy usage varied from building to building, to avoid the influence of identifying typical daily energy usage profiles, the processed data of each building was standardized to zero mean and one standard deviation. Data segmentation was then performed to transform the data into 24 hours segments in order to form daily load profiles. As the primary focus of this strategy was to identify the typical daily energy usage profiles during the building occupied periods with distinctive variation patterns, the segments during the weekends and the segments with small variations were discarded. The segments with small variations refer to the segments with a small difference between the daily maximum and minimum energy usages. In this study, a threshold of 5.0% was used, which means that 5.0% of the segments with the least difference among all daily segments were discarded.

In the data mining step (see Fig. 1), Pearson Correlation Coefficient (PCC) was first calculated to measure the dissimilarities among different daily load profiles. The Partitioning Around Medoids (PAM) clustering algorithm was then applied to cluster the daily load profiles with similar variations based on the PCC-based dissimilarity measure calculated. A boxplot was used to remove the daily load profiles with the large aggregated dissimilarities (i.e. the sum of the dissimilarities to all other daily load profiles in the same cluster) in each cluster, in order to reduce the influence of the extreme daily load profiles on the identification of typical daily load profiles. The daily load profiles with the aggregated dissimilarity measure beyond Q3+1.5IQR, where Q3 is the third quartile and IQR is an inter-quartile range between Q1 and Q3, were discarded. The typical daily load profiles were then determined by averaging the remaining daily load profiles in each cluster. Lastly, a hierarchical clustering was used to group the buildings with similar load characteristics.

In the last step, the identified typical daily load profiles and building groups were visualized, evaluated and interpreted.

2.2 Outlier removal with the generalized Extreme Studentized Deviate (ESD) test method

Generalized ESD test method has been applied for identifying and removing outliers in building energy usage data in a number of studies [14-16]. This method detects outliers through comparing the studentized deviate R of n extreme observations to a critical value λ. The extreme observations are the observations with the first n largest differences compared to the mean value $\bar{x}$. The R_i of the i^{th} extreme observation $x_{e,i}$ is determined using Eq. (1) and the corresponding λ_i is defined in Eq. (2) [14]. The generalized ESD test method starts with the most extreme observation and compares its R_i to the corresponding λ_i. If R_i is greater than λ_i, the extreme observation is then identified as an outlier and removed from the dataset. The same process is applied to the next extreme observation until all the n extreme observations are examined. More details of the generalized ESD test method can be found in [14]. If an outlier is identified and removed, its position will be filled through the linear interpolation.

$$R_i = \frac{\left|x_{e,i} - \bar{x}\right|}{\sigma} \tag{1}$$

$$\lambda_i = \frac{(n-i)\, t_{n-i-1,p}}{\sqrt{(n-i+1)(n-i-1+t_{n-1-1,p}^2)}} \tag{2}$$

where σ is the standard deviation, $t_{n-i-1,p}$ is the t-distribution with $n-i-1$ degrees of freedom and p is the tail area probability and is defined in Eq. (3) [14].

$$p = \frac{\alpha}{2(n-i+1)} \tag{3}$$

where α is the significance level.

2.3 Pearson Correlation Coefficient (PCC)-based dissimilarity measure

Cluster analysis groups the data by minimizing the inter-cluster dissimilarity while maximizing the intra-clusters based on a certain type of the dissimilarity measures [17]. In the proposed strategy, the distance between the two daily load profiles (d_{PCC}) determined by Eq. (4) was used to measure the dissimilarity between the two daily load profiles, in which the PCC is defined in Eq. (5).

$$d_{PCC}(X,Y) = 1 - PCC \tag{4}$$

$$PCC(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{5}$$

where *d* means the distance, *cov* stands for the covariance, *X* and *Y* represent the vectors, and *x* and *y* stands for the values of the individual dimension.

A comparison between the use of the PCC-based and ED-based dissimilarity measures is illustrated in Fig. 2, where ED was calculated using Eq. (6). The data used in Fig. 2 was given only for illustration purpose.

$$d_{ED}(X,Y) = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \tag{6}$$

It can be seen that the ED of Profiles 1 and 2 (d_{12}) and ED of Profiles 1 and 3 (d_{13}) were 38.91 kWh and 8.178 kWh, respectively. Compared to Profile 2, Profile 3 was closer to Profile 1 in terms of the ED dissimilarity measure. However, the variation of Profile 2 was more similar to that of Profile 1, as shown in Fig. 2(a). The PCC of Profiles 1 and 2 (PCC_{12}) and PCC of Profiles

1 and 3 (PCC_{13}) were 0.978 and 0.173, respectively. A higher PCC indicated a higher similarity between the two profiles in terms of the daily load variation. Therefore, PCC-based dissimilarity measure can better identify the daily load profiles with similar variations.

2.4 Partitioning Around Medoids clustering algorithm

Partitioning Around Medoids (PAM) clustering algorithm [18] was used to cluster daily load profiles using the PCC-based dissimilarity measure. In PAM, a medoid is a data point in a particular cluster which has a minimized aggregated distance to all other data points in that cluster. The objective of PAM clustering algorithm is to find a subset $\{o_1,o_2,\ldots,o_k\}\in\{q_1,q_2,\ldots,q_n\}$ which minimizes the objective function as shown in Eq. (7) [18, 19].

$$\sum_{i=1}^{n} \min_{m=1,\ldots,k} d(q_i, o_m) \tag{7}$$

where n is the number of the data points, k is the number of the clusters, q is the data point, and o is the data point identified as a medoid.

PAM consists of two major steps, i.e. build and swap. The first step is to build initial medoids by selecting the first medoid as the data point with the minimum sum of the distance to all other points and selecting the subsequent medoids by finding the points which minimize Eq. (7). The second step repeatedly swap $i\in\{o_1,o_2,\ldots,o_k\}$ with $j\in\{q_1,q_2,\ldots,q_n\}$ if the swap decreases the objective significantly until reaching the convergence [18, 19].

PAM requires users to provide the number of clusters k as an input parameter. In the proposed strategy, Dunn Index was used to validate the clustering result and determine the optimal value of k. Dunn Index is expressed as the ratio of the smallest inter-cluster distance to the largest intra-cluster distance and is defined in Eq. (8) [20].

$$D(\phi) = \frac{\min\limits_{C_k, C_l \in \phi, C_k \neq C_l} \left(\min\limits_{i \in C_k, j \in C_l} d(i,j) \right)}{\max\limits_{C_m \in \phi} \left(\max\limits_{i,j \in C_m, i \neq j} d(i,j) \right)} \tag{8}$$

where C_k and C_l are the clusters belong to the set of the identified clusters ϕ. A higher Dunn Index means a better clustering result. The optimal number of clusters k was determined based on the highest Dunn Index within the defined range of the number of clusters.

2.5 Buildings classification with hierarchical clustering

A hierarchical clustering with the heat map visualization technique was used to group buildings that share the similar daily load characteristics. Hierarchical clustering is a bottom-up strategy, which starts with placing each object in its own cluster and then merges the atomic clusters into larger clusters until all objects are in a single cluster [21]. Complete-linkage, which is the maximum ED of the data objectives in two clusters, was used to measure the distance between the clusters.

An advantage of the hierarchical clustering is that the overall process can be represented by a tree structure graph called a dendrogram. The dendrogram can help to visualize the cluster structure and assist in determining the optimal number of clusters. Fig. 3 illustrated a dendrogram with three data points, where the ordinate axis indicated the distance between the data points/clusters. The split points indicated the distance between the two data points/clusters. The higher the split point, the less similarity between the data points/clusters [4]. Clusters can be determined by the dashed line shown in Fig. 3, which is a user-defined threshold. The data points under the same split point below the dashed line can be merged into a cluster while the split points above the dashed line are kept unchanged. For instance, the data points #1 and #3 were under the same split point and below the dashed line and they will be merged into the same cluster while the data point #2 formed another cluster. The threshold can be determined graphically or based on the cluster validation index such as Dunn Index. More details of the hierarchical clustering can be found in [21].

3. Performance evaluation of the proposed strategy

In this study, the proposed strategy was implemented in R [22] while PAM algorithm was implemented using the R package cluster [23]. The majority of the figures presented in this study were generated using R package ggplot2 [24].

3.1 Description of the case study buildings

The performance of the proposed strategy was evaluated based on the heating energy usage data collected from 19 higher education buildings, with a total floor area of approximately 200,000 m^2, at Norwegian University of Science and Technology in Trondheim, Norway. The hourly building operational data were collected through a web-based Energy Monitoring System.

Most of these 19 buildings were built before the year 2000, and the buildings built between 1960 and 1970 accounted for a large part. The energy certificate of the buildings indicated that the U-values of the exterior walls of the majority buildings were in the range of 0.4-0.60 W/m^2K, which failed to comply with the current energy efficiency regulations. Table 1 summarizes the major information of the 19 buildings studied. More information on these buildings can be found in [25].

The heating demand of these higher education buildings was supplied through a district heating network and each individual building was equipped with a dedicated heating energy usage meter. The three-year hourly heating energy data collected from September to April in 2011-2013 were used in this study for performance evaluation of the proposed strategy.

3.2 Data pre-processing

The generalized ESD test method was first used to detect and remove outliers. Fig. 4 illustrates the three-year hourly heating energy usage data collected from building 03 with the outliers identified (i.e. red circles). It can be seen that there is a large variation in the heating demand annually. The highest heating demand generally occurred in January and February. It should be noted that there was a small heating demand from May to August but this amount of heating demand was significantly lower than that during the main heating period and was therefore not considered in this study.

The data were then standardized to zero mean and one standard deviation and transformed to daily segments. After removing the daily load profiles with small variations and daily load profiles in the weekends, a total of 9,062 daily heating energy usage profiles were generated after the completion of the data pre-processing step.

3.3 Identification of typical daily heating energy usage profiles

The number of clusters selected will directly influence the identification of the typical daily load profiles. A too small cluster number might result in meaningless typical daily load profiles while a large cluster number requires a large computational cost and increases the difficulties in the results evaluation and interpretation. In this study, the optimal cluster number k (i.e. the number of the typical daily load profiles) was selected between 5 and 15. Fig. 5 presents Dunn Index calculated when using different numbers of the clusters. It is shown that the highest Dunn Index resulted when the cluster number was 11, which was therefore determined as the optimal cluster number in this study.

The boxplot of the aggregated dissimilarity measure of the identified clusters is illustrated in Fig. 6 for visualization and removal of the daily load profiles beyond the threshold. It can be observed that the number of the daily load profiles in all clusters ranged from 474 to 1413, indicating that there was no cluster formed with few daily load profiles. It can also be seen that all clusters contained the extreme daily load profiles (i.e. black dots) with the aggregated dissimilarity beyond the threshold (i.e. Q3+1.5IQR) and these extreme daily load profiles were removed in subsequent analysis. A total of 8,521 daily load profiles remained after removing the identified outlier (i.e. extreme daily load profiles) from the dataset. The removal of this small fraction of the extreme daily load profiles could enhance the visualization of the identified typical daily load profiles without significant loss of the information.

Fig. 7 shows the identified typical daily load profiles by averaging all daily load profiles in each cluster after the removal of the extreme daily load profiles. The red curves in the figure showed the typical daily load profiles identified while the gray curves were all corresponding daily load profiles in this cluster. It can be found that there was a clear boundary in the heating demand between the working hours and non-working hours in some typical daily load profiles such as the load profiles 2 and 8 while that in some typical daily load profiles (e.g. the load profiles 1 and 5) were not very clear. There was no obvious boundary in the load profile 10. Moreover, the nighttime from 22:00 to 03:00 of next day was the lowest heating energy usage period for the majority of the typical daily load profiles identified except the typical load profiles 6, 7 and 10 with a noticeable high heating demand during the nighttime which is worthwhile for further investigation.

Fig. 8 shows the weekday distribution of the building daily load profiles in the identified clusters, in which y-axis represents the percentage of the number of days belongs to each weekday to the total number of days in each cluster. It was shown that the daily load profiles on Tuesday, Wednesday, Thursday and Friday in each cluster were almost evenly distributed. In some clusters such as the clusters 4 and 11, the number of days on Monday was obviously different from that on the other weekdays and the reason behind this is presented in Section 4. Therefore, this weekday load profile distribution can assist in determining whether a specific load profile existed only in some specific days of a week.

Table 2 summarizes the key characteristics and the estimated high heating demand period of the typical daily load profiles identified. To understand the knowledge and information discovered by the proposed strategy, the profiles with a relatively high demand in the early morning and late night as well as those with clear heating demand peaks and troughs will be further investigated in Section 4. These include the typical daily load profiles 4, 6, 7, 9 and 11. The rest of the typical load profiles were either similar to the typical daily load profiles mentioned above or did not contain interesting characteristics and were therefore not further investigated in this study.

3.4 Building classification based on the identified typical daily load profiles

In this section, 19 case study buildings were grouped according to the typical daily heating load profiles identified. In order to eliminate the influence from the insignificant profiles, the first two most dominant profiles of each building (see Table 3) were selected as the features for building classification. From Table 3, it can be seen that for some buildings such as buildings 02, 14 and 17, the most dominant profile accounted for a large proportion of the total number of days remained for the typical daily load profile identification. For instance, 436 days out of 490 of building 02 were in the most dominant profile, demonstrating that the daily load variation of this building was consistent. In contrast, the number of days in the most dominant profiles of some buildings such as building 10 and 15 were relatively small, which indicated that these buildings did not have a consistent daily load profile during the time period investigated (2011-2013).

The percentages of the first two most dominant profiles were then used to group the buildings that share the same daily energy usage characteristics based on the hierarchical clustering. Fig. 9 presents the dendrogram of building classification results, in which the buildings in the same cluster were marked with the same color. In this study, the threshold (i.e. dashed line in the figure) was visually selected due to the small number of the data points (i.e. buildings) used. It can be seen that some clusters were formed with a single building while some clusters were formed with several buildings. For instance, building 02 was identified as an individual cluster and buildings 01, 05, 10, 11 and 12 were grouped into one single cluster.

In order to better visualize and confirm the clustering results, the number of days belongs to a typical daily load profile of different buildings were plotted as a heat map and are shown in Fig. 10. In this figure, the relative proportion (*RP*) was determined using Eq. (9) and the same order of the building number as illustrated in Fig. 9 was used. It was visually shown that the majority of the buildings had one significant dominant profile.

$$RP = \frac{N_d}{N_{d,max}} \tag{9}$$

where N_d stands for the number of days belongs to a typical daily load profile of an individual building and $N_{d,max}$ stands for the maximum number of days belong to a typical daily load profile of the same building.

4. Interpretation of the identified typical daily load profiles

In order to understand the reasons behind the main characteristics of the typical daily load profiled identified, buildings 02, 14, 17, 08 and 03 were selected based on the clustering results and used to represent the typical daily load profiles of 4, 6, 7, 9 and 11 presented in Fig. 7, respectively.

4.1 Building 02 – Typical daily load profile 4

Building 02 is an office and laboratory building which was built in 1965. A recent survey indicated that this building was poorly insulated with a U-value of 0.91 W/m^{2}K for the exterior wall insulation and a U-value of 0.59 W/m^{2}K for the roof insulation. Different from many other buildings using hot water radiators for space heating, the heating of this building was supplied through ventilation without using heat recovery. However, the heat recovery has been mandatorily required for decades in Norway in ventilation.

Fig. 11(a) shows the heating energy usage of building 02 in the two consecutive days. It was clearly shown that the high heating demand started at 04:00 in the morning, which was consistent with the typical daily load profile 4. However, it was much earlier than the normal building occupied hours. The feedback from the building operator indicated that the occupants in this building continuously complained about the thermal comfort during the morning time. The heating period was therefore extended in order to satisfy the occupant thermal comfort and to provide freezing protection [26].

4.2 Building 14 – Typical daily load profile 6

Building 14 is a sports center, which was usually operated till midnight. The heating demand of this building in the two consecutive days is illustrated in Fig. 11(b). The major characteristics of the two-day heating demand matched well with that of the typical daily load profile 6. The highest heating demand generally occurred around 19:00. This high heating demand was probably related to the hot water usage for the shower requirement. The water usage data of this building in the same two days are presented in Fig. 12. It was clearly shown that there was a high peak of the water usage at around 19:00, which was in line with the heating energy usage profiles. It was also found that the water usage of this building dropped to zero at 01:00 which also matched with the heating demand variation.

4.3 Building 17 – Typical daily load profile 7

Building 17 is a multi-functional building with offices, educational rooms and laboratories, which was constructed around the year 1996. As shown in Fig. 11(c), the two-day heating load profile of this building was similar to that of the typical daily load profile 7 identified. The high heating demand period lasted till to 23:00. The feedback from the building operator indicated that the building occupants required the building to be heated till to 23:00 for special activity requirements.

4.4 Building 08 – Typical daily load profile 9

Building 08 is an old building constructed in 1924 and is also a multi-functional building with offices, educational rooms, and laboratories. A clear peak and a clear trough can be observed in Fig. 11(d) at 05:00 and 21:00 respectively, which were consistent with the information presented in the typical daily load profile 9. The heating demand peak and trough were found to be mainly caused by the sudden change of the supply water temperature. The recorded data showed that the hot water was supplied at about 70^{o}C during the daytime and 40^{o}C during the nighttime. The sudden rise of the supply water temperature in the early morning resulted in the heating demand peak of the building while the sudden drop of the supply water temperature in the nighttime led to the occurrence of the trough in the heating load profile. This relationship between the heating energy usage and the variation in the supply water temperature was also observed in a previous study [27].

The building operator was also approached for the reason why the high heating demand started at around 05:00. However, no information on this was recorded. This is probably also due to the poor insulation of the building (i.e. U-value of 1.0 W/m^{2}K for the exterior wall insulation and U-value of 0.7 W/m^{2}K for the roof insulation), which might result in a longer pre-heating period before the building was occupied.

4.5 Building 03 – Typical daily load profile 11

Building 03 is a mix of offices and laboratories, which was constructed in 1951. The typical daily load profile 11 was very similar to the typical daily load profile 9. However, in the typical daily load profile identified, there were very few days from Monday. Fig. 11(e) illustrates the heating demand of building 03 in two days of Monday and Tuesday. It was clearly shown that there was a heating demand peak at 07:00 and a trough at 17:00 in the daily heating load profile on Tuesday, which matched well with the typical daily load profile 11. However, on Monday, the heating demand peak occurred at 06:00. This is mainly due to the fact that, during the weekend, the heating system was either not running or running with a lower supply water temperature, resulting in a lower indoor temperature than during the weekdays. In order to achieve a desirable thermal comfort on Monday morning, the building was therefore pre-heated earlier than that during the weekdays.

5. Comparison between the use of ED-based and PCC-based clustering

In this section, the results of using the ED-based and PCC-based clustering were compared and presented. The same data pre-processing used for the PCC-based clustering was performed for the ED-based clustering while the commonly used k-means and ED-based dissimilarity measure were used to replace PAM and PCC-based dissimilarity measure. The optimal number of clusters for the ED-based clustering determined was 10, as shown in Figure 13, which was also determined based on Dunn index.

Based on the optimal number of clusters determined, the typical daily heating load profiles can then be identified after removal of the extreme daily load profiles based on the box plot analysis. Fig. 14 presents the clustering results and the identified typical daily heating load profiles using the ED-based clustering. It can be seen that the profiles identified using the k-means clustering with ED-based dissimilarity measure can still provide some useful information in the identified typical daily heating load profiles. For instance, a morning peak was observed and the building was heated till to midnight in the typical daily load profile 7, which was very similar to the typical daily load profile 6 identified using the proposed strategy.

However, some profiles identified such as the typical daily load profiles 3 and 9 were too flat and cannot provide useful information for further analysis. Some important information, e.g.

05:00 heating demand peak (corresponding to the load profile 9 in Fig. 7), 04:00 high heating demand start time (corresponding to the load profile 4 in Fig. 7), 17:00 low trough (corresponding to load profile 11 in Fig. 7), identified by the proposed strategy cannot be identified using the k-means clustering with ED-based dissimilarity measure. In addition, some profiles such as the load profiles 2 & 7, and the load profiles 6 & 8 presented in Fig. 14 showed very similar trends but with different magnitudes. This further demonstrated that the ED-based dissimilarity measure tends to identify daily load profiles that were similar in terms of the intensity.

The heat map in Fig. 15 illustrated the number of days belongs to a typical daily load profile of different buildings when using the ED-based clustering. The order of the buildings in the map was also determined based on the hierarchical clustering. Compared to the results presented in Fig. 10, it was clearly shown that the building heating energy usage cannot be characterized by the most dominant load profiles as there was no clear difference between the number of days belong to the most dominant typical daily load profile and that belongs to the other typical daily load profiles. It was also demonstrated that it is difficult to use the ED-based clustering identified typical daily load profiles for building classification.

6. Conclusions

Understanding multiple buildings energy performance requires advanced data analytics. This paper presented a variation focused Partitioning Around Medoids (PAM) cluster analysis strategy to identify the typical daily load profiles of higher education buildings, in which Pearson Correlation Coefficient was used as the dissimilarity measure to group the daily load profiles on the basis of the variation similarity instead of the magnitude similarity.

The performance of the proposed strategy was evaluated using the heating energy usage data of 19 higher education buildings in Norway collected from 2011 to 2013. The results showed that the proposed strategy can identify and discover the information related to building daily heating energy usage characteristics, including daily high heating demand start time and end time, the peaks, troughs and variations of daily heating energy usage. The results obtained also confirmed the effectiveness of the proposed strategy in identifying the typical daily heating energy usage profiles in terms of the variation similarity.

The identified daily heating energy usage characteristics can be used to assist in the development of advanced building control and fault detection & diagnosis strategies, and cost-effective demand side management techniques. The information discovered is also useful to support the energy planning and retrofitting of higher education buildings. This method could be adapted to identify the daily energy usage characteristics of other types of buildings.

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, P. Cooper, D. Daly, L. Ledo. Existing building retrofits: Methodology and state-of-the-art. Energy and buildings. 55 (2012) 889-902.

[2] L. Pérez-Lombard, J. Ortiz, C. Pout. A review on buildings energy consumption information. Energy and Buildings. 40 (2008) 394-8.

[3] H. Gadd, S. Werner. Heat load patterns in district heating substations. Applied Energy. 108 (2013) 176-83.

[4] P.R.S. Jota, V.R.B. Silva, F.G. Jota. Building load management using cluster and statistical analyses. International Journal of Electrical Power & Energy Systems. 33 (2011) 1498-505.

[5] 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.

[6] A.K. Jain. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 31 (2010) 651-66.

[7] C. Flath, D. Nicolay, T. Conte, C. van Dinther, L. Filipova-Neumann. Cluster analysis of smart metering data: An implementation in practice. Business & Information Systems Engineering. 4 (2012) 31-9.

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

[9] M.P. Fernandes, J.L. Viegas, S.M. Vieira, J.M. Sousa. Analysis of residential natural gas consumers using fuzzy c-means clustering. 2016 IEEE International Conference on Fuzzy Systems (FUZZ). pp. 1484-91.

[10] I.P. Panapakidis, T.A. Papadopoulos, G.C. Christoforidis, G.K. Papagiannis. Pattern recognition algorithms for electricity load curve analysis of buildings. Energy and Buildings. 73 (2014) 137-45.

[11] L. Kaufman, P.J. Rousseeuw. Finding groups in data: An introduction to cluster analysis. John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 2008.

[12] L. Ledo. Energy efficiency and thermal comfort upgrades for higher education buildings. University of Wollongong, PhD thesis, 2015.

[13] P.N. Tan, M. Steinbach, V. Kumar. Introduction to data mining. Pearson Addison Wesley, USA, 2006.

[14] J.E. Seem. Using intelligent data analysis to detect abnormal energy consumption in buildings. Energy and Buildings. 39 (2007) 52-8.

[15] C. Fan, F. Xiao, S. Wang. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Applied Energy. 127 (2014) 1-10.

[16] I. Khan, A. Capozzoli, S.P. Corgnati, T. Cerquitelli. Fault detection analysis of building energy consumption using data mining techniques. Energy Procedia. 42 (2013) 557-66.

[17] H. Blockeel, L. De Raedt, J. Ramon. Top-down induction of clustering trees. Proceedings of the Fifteenth International Conference on Machine Learning, 1998. pp. 55-63.

[18] L. Kaufman, P. Rousseeuw. Clustering by means of medoids. The First International Conference on Statistical Data Analysis based on the L_1-Norm and Related Methods, Neuchatel, Switzerland, 1987.

[19] A. Struyf, M. Hubert, P. Rousseeuw. Clustering in an object-oriented environment. Journal of Statistical Software. 1 (1997) 1-30.

[20] G. Brock, V. Pihur, S. Datta, S. Datta. clValid: An R package for cluster validation. Journal of Statistical Software. 25 (2008) 1-22.

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

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

[23] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik. cluster: Cluster analysis basics and extensions. 2015.

[24] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer, New York, 2009.

[25] J. Guan, N. Nord, S. Chen. Energy planning of university campus building complex: Energy usage and coincidental analysis of individual buildings with a case study. Energy and Buildings. 124 (2016) 99-111.

[26] N. Nord, M.H. Martin. Freeze protection method in ventilation system using two hydronic circuits. The 12th International Conference on Air Distribution in Rooms, Trondheim, Norway, 2011.

[27] N. Djuric, V. Novakovic, F. Frydenlund. Heating system performance estimation using optimization tool and BEMS data. Energy and Buildings. 40 (2008) 1367-76.

Table 1 Major information of the case study buildings

Building NO.	Construction year	Main functions#	Floor area (m^2)	Building NO.	Construction year	Main function	Floor area (m^2)
01	1962	O/E/L	15,026	11	1968	O/L	12,861
02	1965	O/L	3,030	12	1910	O	3,375
03	1951	O/L	2,215	13	1981	O/E/L	3,955
04	1960	O/E/L	7,598	14	1966	S	4,046
05	1966	O/E/L	11,400	15	1975	O/E/L	18,175
06	1958	O/E/L	12,600	16	1951	O/E/L	5,053
07	1965	O/E/L	9,168	17	1996	O/E/L	2,476
08	1924	O/E/L	4,116	18	2002	E/L	4,312
09	1960	O/L	5,028	19	2000	O/E/L	52,773
10	1961	O/L	17,936				

O: office; E: educational room; L: laboratory; S: sports complex.

Table 2 Key characteristics of the identified typical daily heating energy usage profiles

Typical load profile No.	Est. high heating demand period	Weekday load profile almost evenly distributed	Main characteristics
1	07:00-15:00	Yes	There was a high heating demand from 07:00 to 10:00. The heating demand was then gradually decreased till to 16.00 and then kept relatively stable.
2	07:00-17:00	No	The high heating demand occurred during the office hours. A clear heating demand peak can be observed at 07:00.
3	07:00-18:00	No	There was a clear heating demand peak at 07:00 and the heating demand was then gradually decreased till to 18:00.
4	04:00-17:00	No	A high heating demand started at around 04:00 and then kept relatively stable till to 17:00.
5	07:00-18:00	Yes	The daily heating demand variations were similar to that of the typical load profile 1.
6	09:00-24:00	Yes	There was a small peak at 06:00. A high heating demand started at 09:00 and lasted till to the midnight.
7	09:00-23:00	Yes	Similar to the load profile 6 but the heating demand during the high heating demand period was more stable.
8	06:00-18:00	No	Similar to the load profile 2. However, there was a clear trough at 19:00.
9	05:00-20:00	No	Similar to the load profiles 2 and 8 but there was a clear peak at 05:00 and a clear trough at 21:00.
10	Not clear	Yes	The heating demand during 24 hours was relatively stable. However, the demand in the early morning was slightly higher than the rest of the day.
11	07:00-16:00	No	Similar to the load profiles 2, 8 and 9. There was a clear heating demand peak at 07:00 and a clear trough at 17:00.

Table 3 Summary of the first two most dominant profiles of individual buildings

Building No.	Total number of days	The most dominant profile			The 2nd most dominant profile		
		Typical daily load profile No.	Total days	Percentage (%)	Typical daily load profile No.	Total days	Percentage (%)
1	457	5	240	53	3	79	17
2	490	4	436	89	3	17	3
3	486	11	252	52	8	57	12
4	458	11	207	45	4	58	13
5	436	5	271	62	7	36	8
6	437	3	246	56	1	93	21
7	471	8	300	64	11	65	14
8	471	9	285	61	7	87	18
9	448	10	124	28	3	118	26
10	439	7	107	24	5	103	23
11	471	5	175	37	7	81	17
12	371	5	172	46	10	83	22
13	449	3	148	33	2	127	28
14	495	6	440	89	7	35	7
15	382	1	94	25	3	73	19
16	486	2	316	65	4	67	14
17	480	7	386	80	5	40	8
18	367	7	152	41	6	115	31
19	427	1	116	27	5	112	26

Figure Captions

Fig. 1 Outline of the variation focused cluster analysis strategy.

Fig. 2 Comparison between the PCC and ED-based dissimilarity measures (a) ED; (b)&(c) PCC.

Fig. 3 Illustration of the dendrogram with three data points.

Fig. 4 Illustration of the building heating energy usage and outliers identified - building 03.

Fig. 5 Dunn Index calculated for different numbers of the clusters - PCC-based clustering.

Fig. 6 Boxplot of the aggregated dissimilarities of the identified clusters.

Fig. 7 Typical daily heating load profiles (red) identified using the proposed strategy with all corresponding daily load profiles (gray).

Fig. 8 Weekday load profile distribution in different clusters identified.

Fig. 9 Dendrogram of building classification results.

Fig. 10 Heat map of the typical daily load profiles in different buildings - PCC-based clustering.

Fig. 11 Illustrations of the heating energy usage of the buildings in two consecutive days.

Fig. 12 Water usage of building 14.

Fig. 13 Dunn Index calculated for different numbers of the clusters - ED-based clustering.

Fig. 14 Typical daily heating load profiles (red) identified using the ED-based clustering with all corresponding daily load profiles (gray).

Fig. 15 Heat map of the typical daily load profiles in different buildings - ED-based clustering.

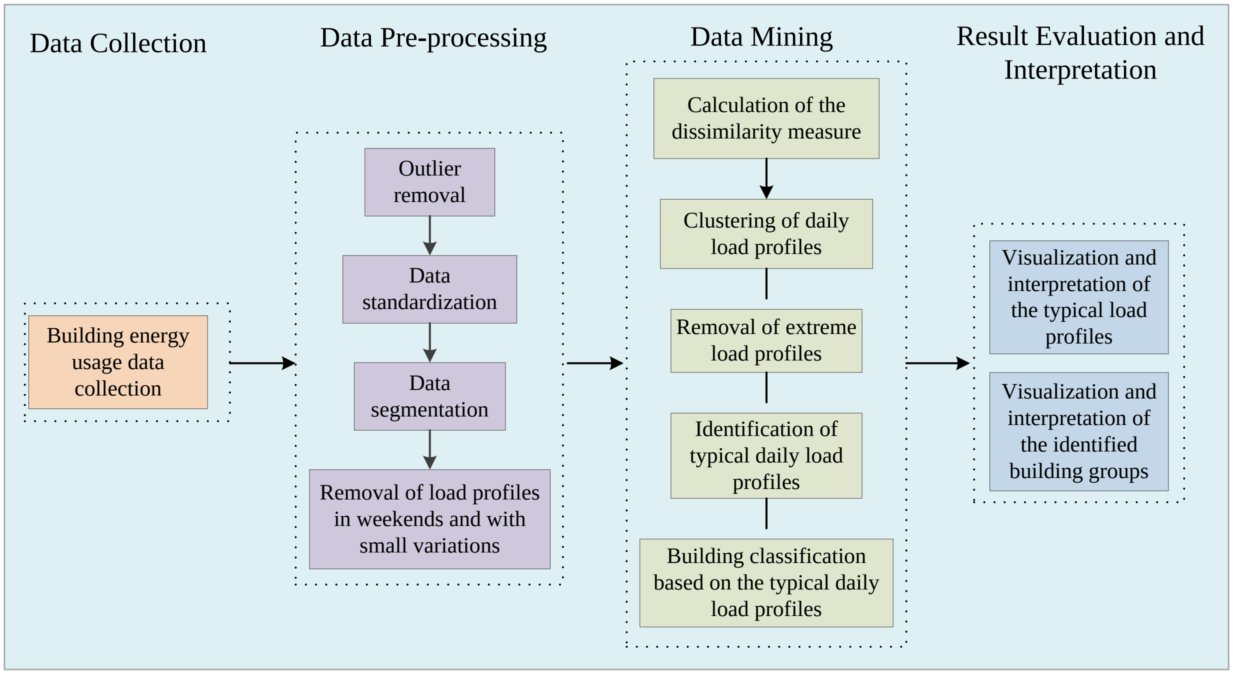

Fig. 1 Outline of the variation focused cluster analysis strategy.

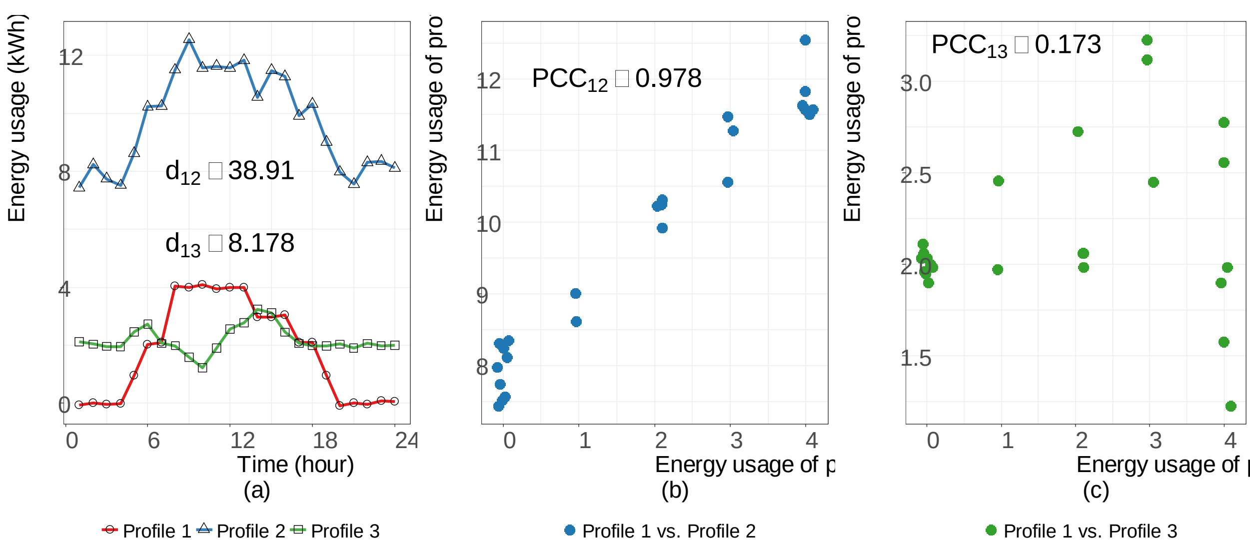

Fig. 2 Comparison between the PCC and ED-based dissimilarity measures (a) ED; (b)&(c) PCC.

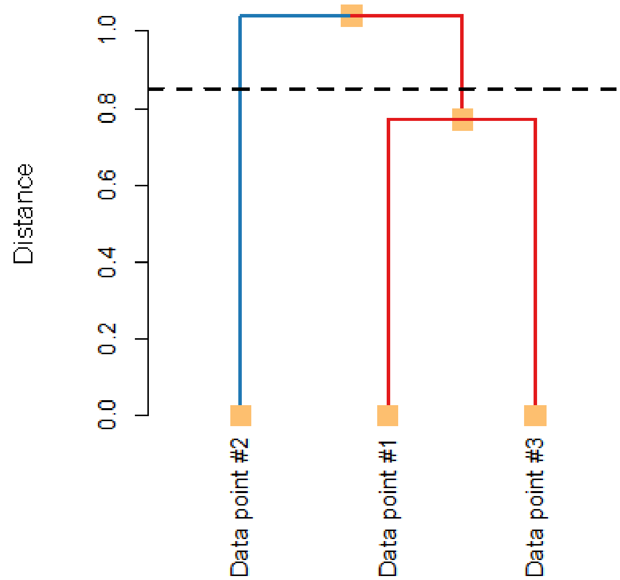

Fig. 3 Illustration of the dendrogram with three data points.

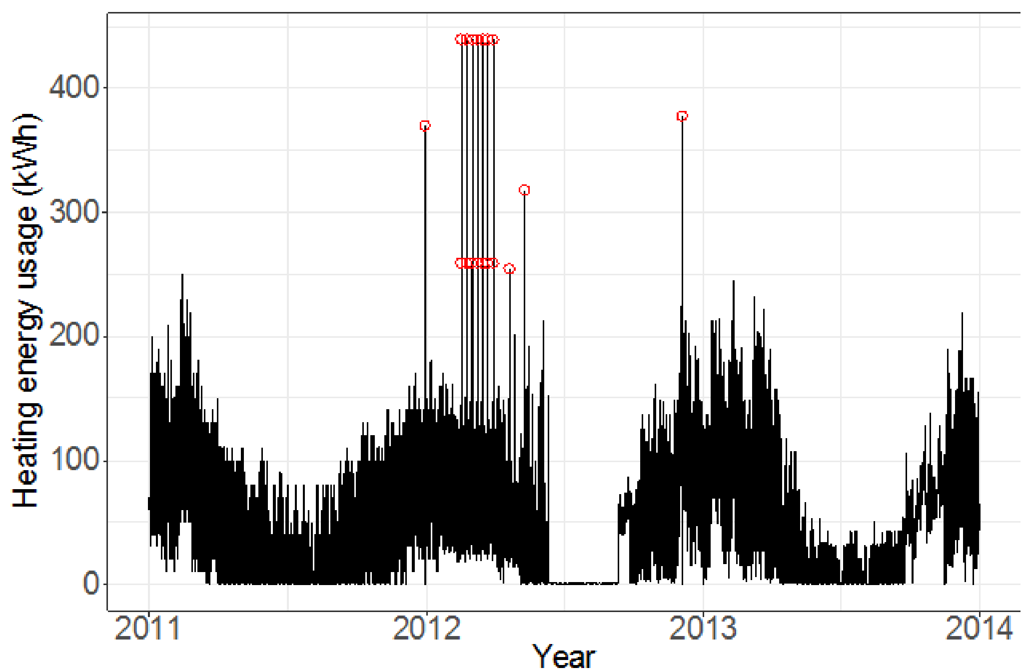

Fig. 4 Illustration of the building heating energy usage and outliers identified - building 03.

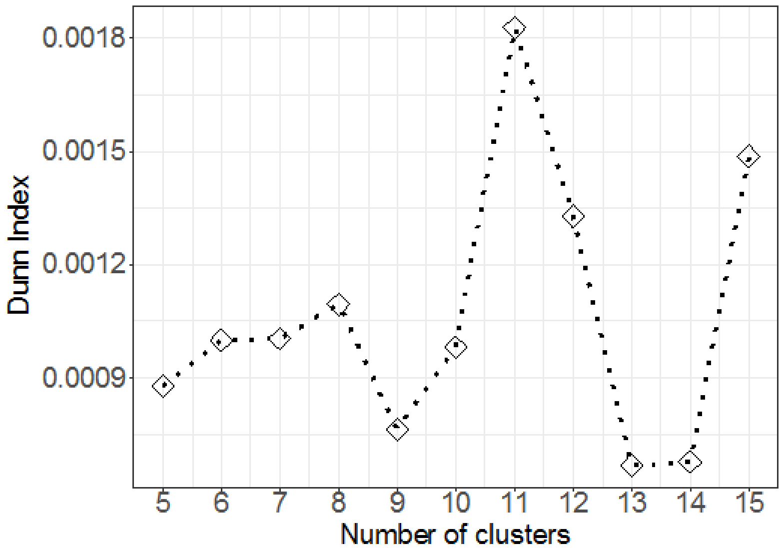

Fig. 5 Dunn Index calculated for different numbers of the clusters - PCC-based clustering.

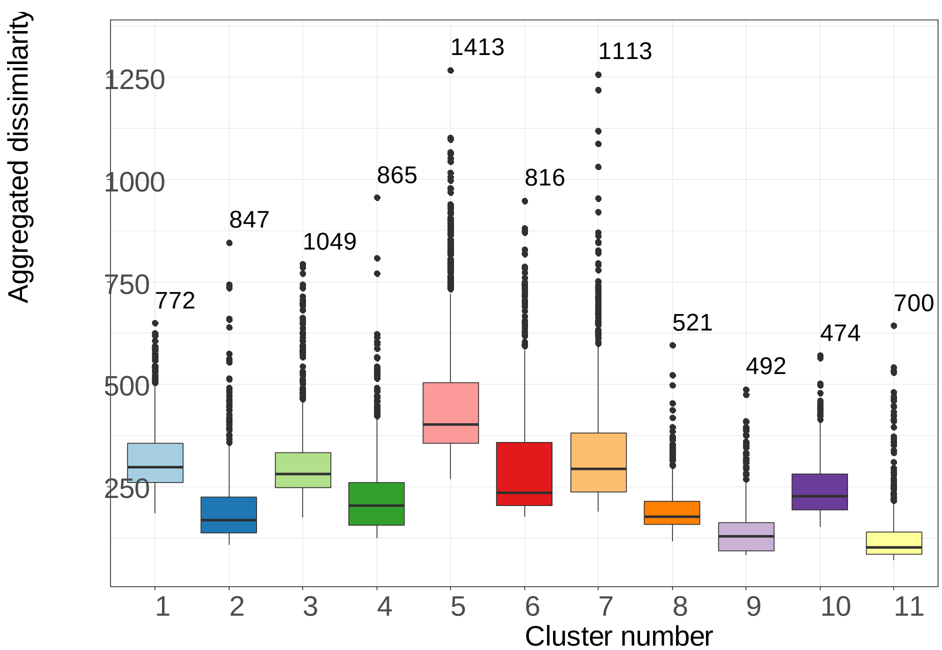

Fig. 6 Boxplot of the aggregated dissimilarities of the identified clusters.

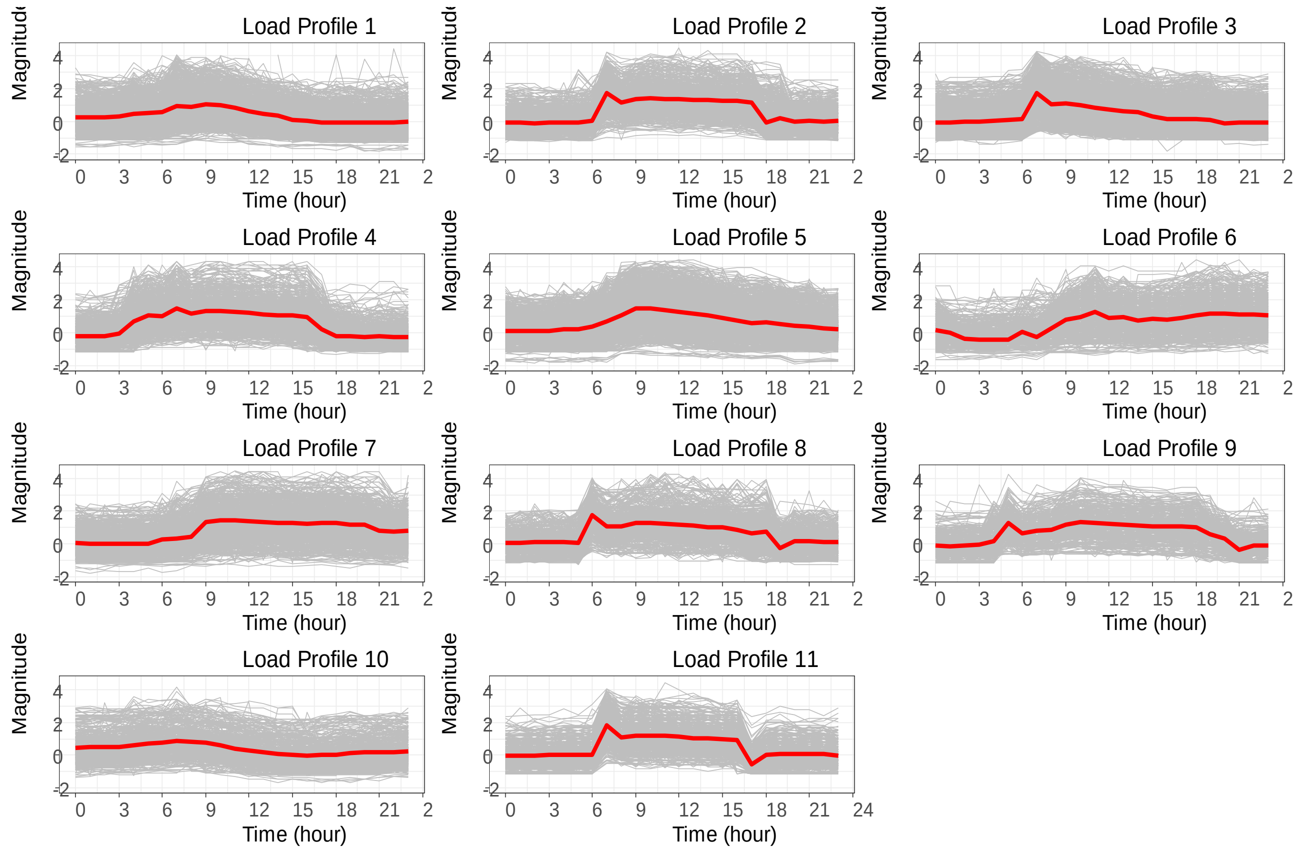

Fig. 7 Typical daily heating load profiles (red) identified using the proposed strategy with all corresponding daily load profiles (gray).

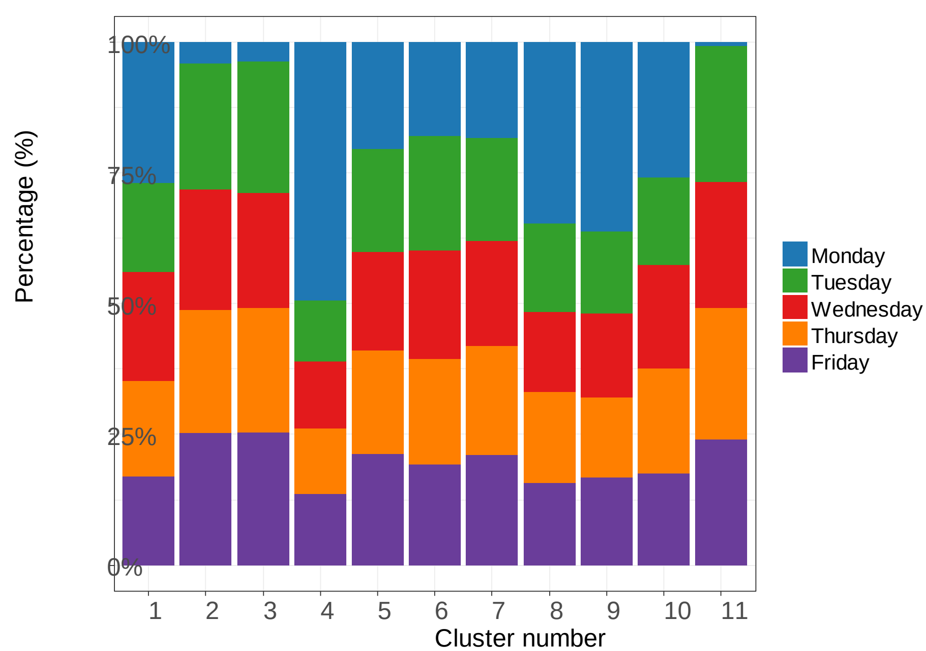

Fig. 8 Weekday load profile distribution in different clusters identified.

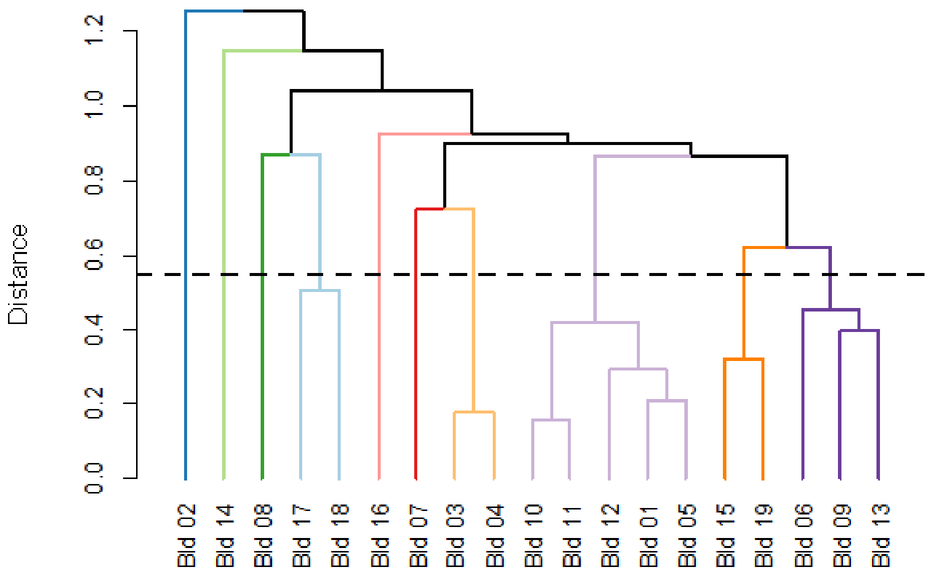

Fig. 9 Dendrogram of building classification results.

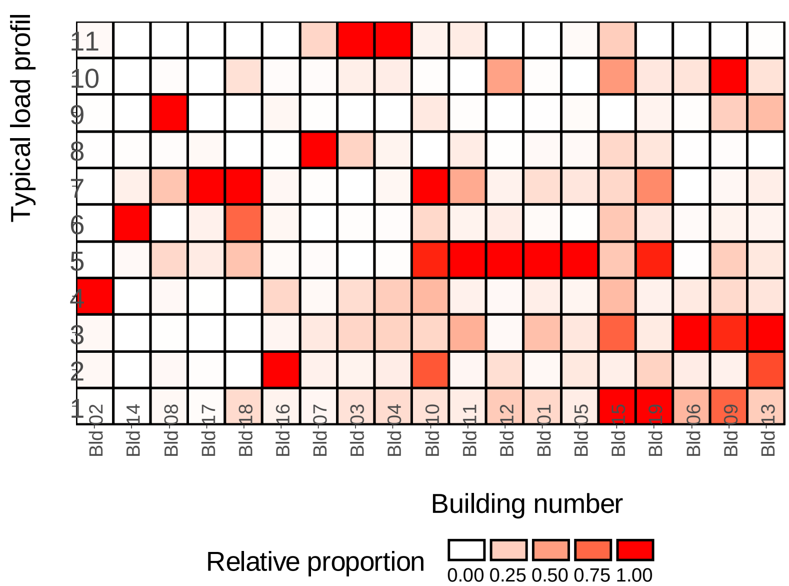

Fig. 10 Heat map of the typical daily load profiles in different buildings - PCC-based clustering.

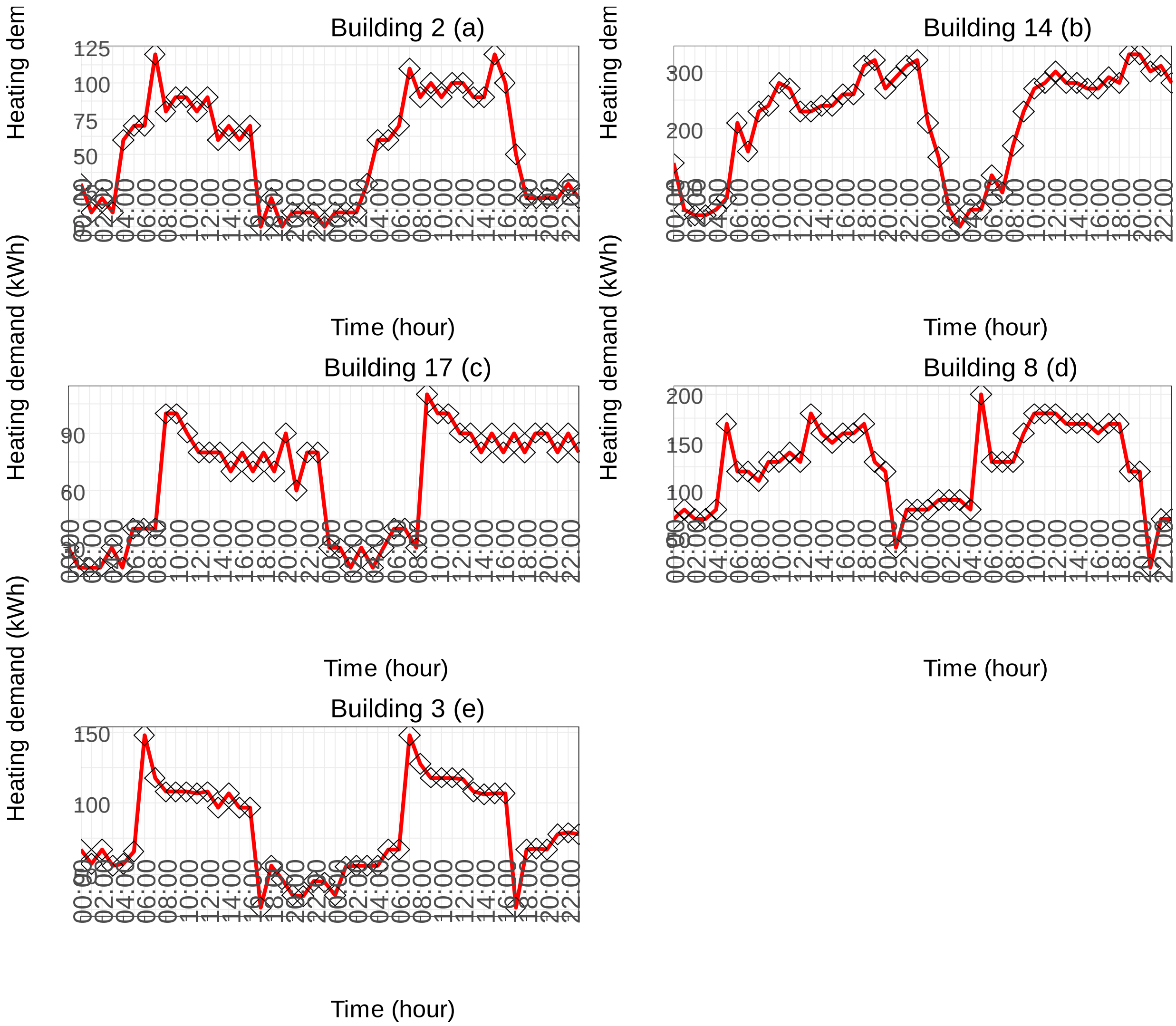

Fig. 11 Illustrations of the heating energy usage of the buildings in two consecutive days.

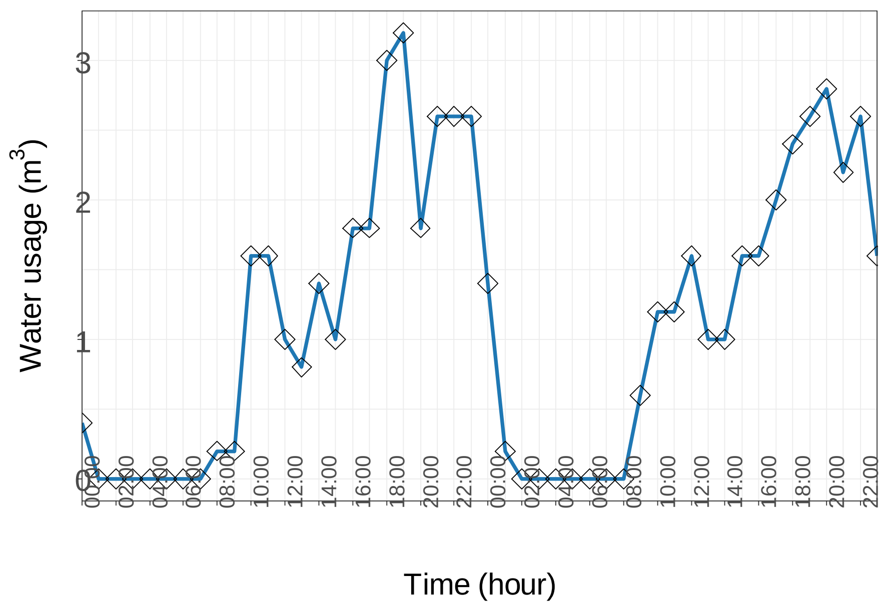

Fig. 12 Water usage of building 14.

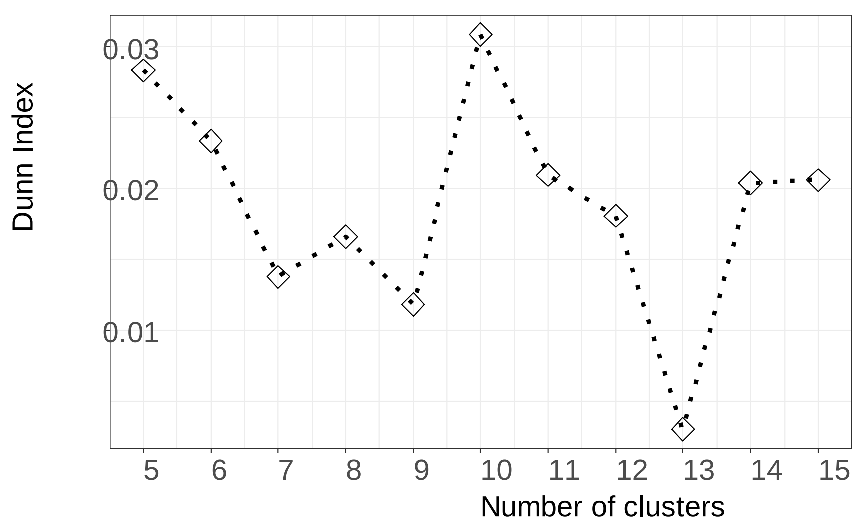

Fig. 13 Dunn Index calculated for different numbers of the clusters - ED-based clustering.

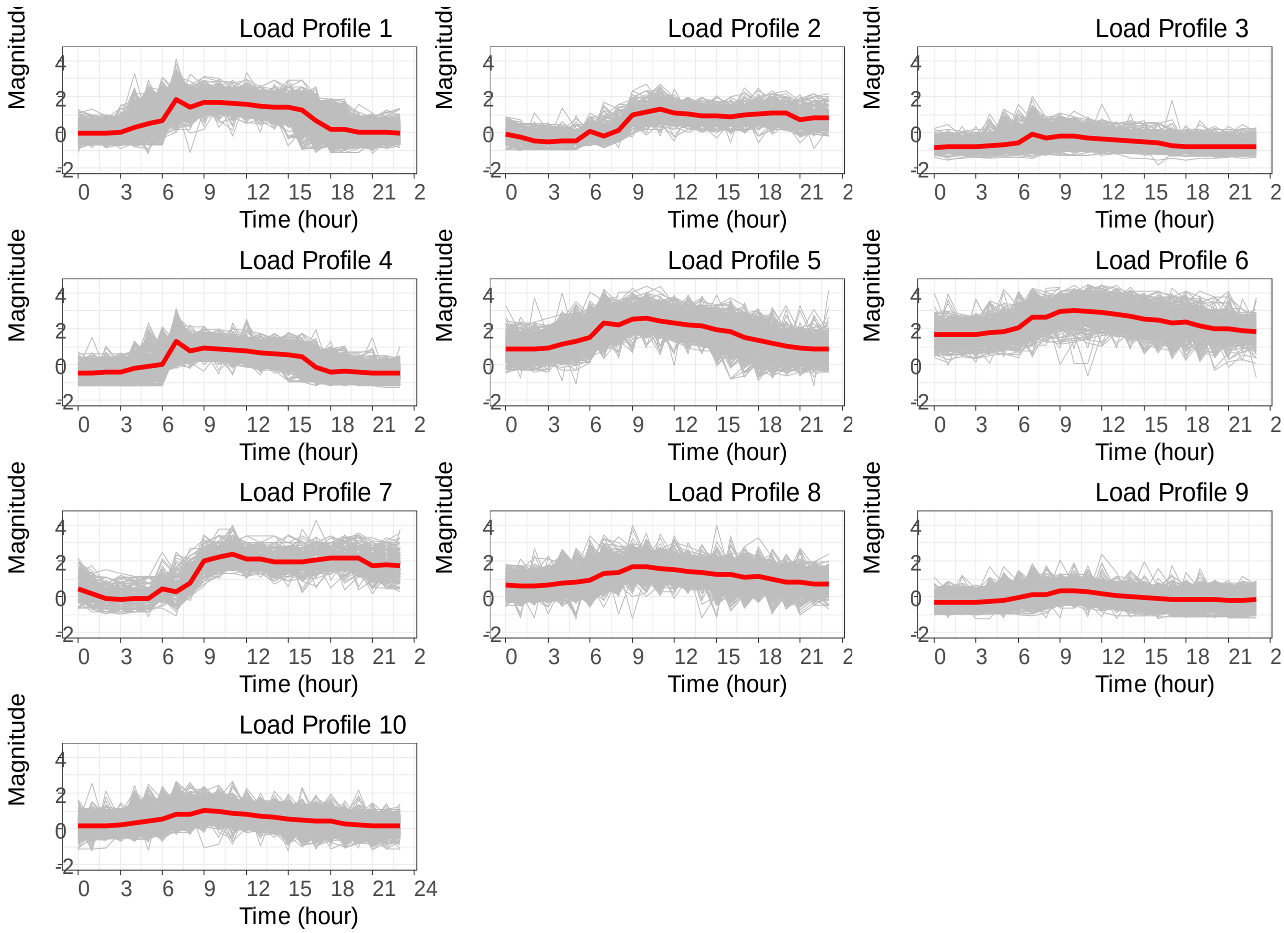

Fig. 14 Typical daily heating load profiles (red) identified using the ED-based clustering with all corresponding daily load profiles (gray).

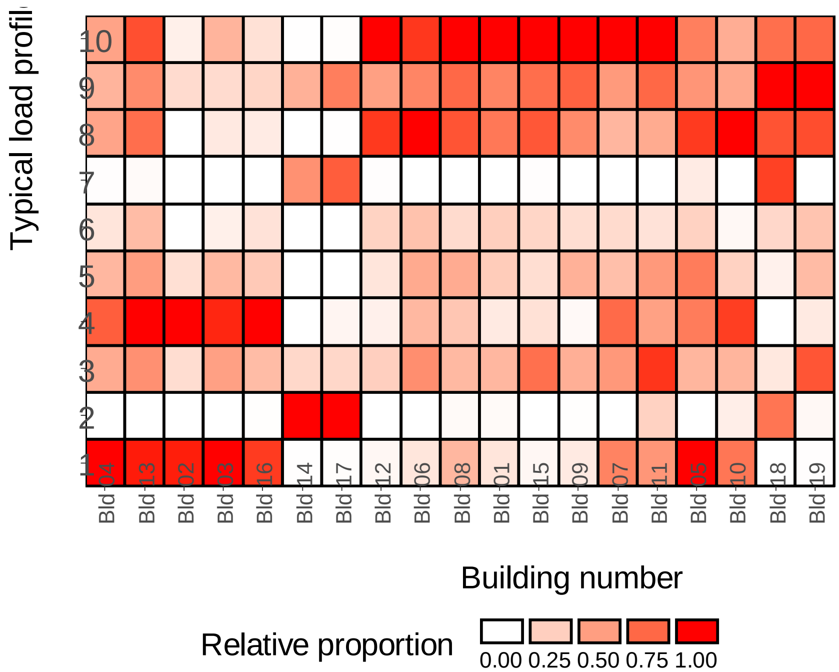

Fig. 15 Heat map of the typical daily load profiles in different buildings - ED-based clustering.