Energy & Buildings 231 (2021) 110601

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data

Xue Liu [a,b], Yong Ding [a,b,*], Hao Tang [a,b], Feng Xiao [c]

[a] Joint International Research Laboratory of Green Buildings and Built Environments (Ministry of Education), Chongqing University, Chongqing 400045, China
[b] National Centre for International Research of Low-carbon and Green Buildings (Ministry of Science and Technology), Chongqing University, Chongqing 400045, China
[c] School of Business Administration, Southwestern University of Finance and Economics, Chengdu, China

ARTICLE INFO

Article history:
Received 14 July 2020
Revised 25 September 2020
Accepted 1 November 2020
Available online 5 November 2020

Keywords:
Building energy management
Time series clustering
Decision tree
Knowledge discovery
Electricity usage pattern
Data mining

ABSTRACT

With the development of advanced information techniques, smart energy meters have made a considerable amount of real-time electricity consumption data available. These data provide a promising way to understand energy usage patterns and improve building energy management. However, previous studies have paid more attention to methodologies for the identification of energy usage patterns and are limited in the interpretability and applications of the patterns. In this context, this paper proposes a general data mining-based framework that can extract typical electricity load patterns (TELPs) and discover insightful information hidden in the patterns. The framework integrates multiple data mining techniques and mainly consists of three phases: data preparation, identification of TELPs and knowledge discovery in the patterns. A new clustering method with a two-step clustering analysis is proposed to identify the TELPs at the individual building level. Before clustering, five statistical features that represent the shapes of electricity load profiles are first defined to reduce the dimensions of daily electricity load profiles. The first clustering step aims at detecting outliers of daily electricity load profiles (DELPs) by using the density-based spatial clustering application with noise (DBSCAN) algorithm clustering technique, which addresses the data quality issues for electricity consumption data derived from energy consumption monitoring platforms (ECMPs). The second clustering step aims at grouping similar DELPs by means of the k-means algorithm to extract TELPs. The effectiveness of the proposed clustering method is demonstrated by a comparison with two single-step clustering techniques. Furthermore, a classification and regression tree (CART) algorithm is employed to discover insightful knowledge on TELPs and improve the interpretability of clustering results, namely, to explain the relations between dynamic influencing factors related to electricity consumption and TELPs. The proposed framework is applied to analyze the time-series electricity consumption data of three practical office buildings in Chongqing, and its effectiveness has been confirmed. A potential application of discovered knowledge is presented: early fault detection of anomalous electricity load profiles. The proposed framework can provide building managers with an efficient way to understand the characteristics of building electricity usage patterns and detect anomalies therein.

1. Introduction

1.1. Background and motivation

The building sector consumes approximately 40% of global primary energy and produces more than 30% of CO_2 emissions [1]. Reducing the overall energy consumption associated with building operation is crucial for sustainability goals. Energy management plays a key role in improving energy efficiency and reducing total energy usage and operation costs. Advanced metering infrastructure (AMI), which can collect and store massive electricity consumption data in near-real time, has been developed during the last decade, provides more information on building operation than annual energy consumption data, and assists in understanding the characteristics of energy use behaviors and detecting potential energy waste. Data mining (DM), a promising approach to discovering patterns behind building operation energy consumption

* Corresponding author at: Joint International Research Laboratory of Green Buildings and Built Environments (Ministry of Education), Chongqing University, Chongqing 400045, China.
E-mail address: dingyongqq@163.com (Y. Ding).

https://doi.org/10.1016/j.enbuild.2020.110601
0378-7788/

Nomenclature

Abbreviations

DM	Data mining
ECMP	Energy consumption monitoring platform
DELP	Daily electricity load profile
TELP	Typical electricity load pattern
CART	Classification and regression tree
DBSCAN	Density-based spatial clustering application with noise
kNN	K-nearest neighbor
CVI	Clustering validation index
SSE	Sum of squared errors

Symbols

$p(t)$	Actual electricity load data
$p_{nor}(t)$	Normalized actual electricity load data
$p_{max}(t)$	Daily maximum actual electricity load data
eps	Radius
MinPts	Minimum number of observations in the *eps* region
N	A vector representing daily features
μ	A vector representing the cluster center
cp	Cost complexity
minsplit	Minimum number of samples for a node
k	Number of clusters
n	Number of daily profiles

Subscripts

i	i^{th} cluster of cluster center
j	j^{th} day of daily profile

data, has received increasing attention in recent years. Compared with traditional statistical or physical principle-based methods, DM is effective and efficient in handling massive datasets, capable of discovering potentially useful yet previously unknown knowledge, and less dependent on domain expertise than other methods [2]. Load profiling is some of the most important information discovered by DM techniques. It refers to the process of grouping temporal subsequences of measured electricity data to identify typical electricity consumption patterns of a building [3]. However, the raw forms of typical electricity consumption patterns can be difficult to interpret [2]; therefore, the interpretation of the clustering results which is called "knowledge discovery" is a more attractive and valuable step and make it applicable for practical applications.

Smart meters have been installed over 70 million in the United States by 2016 where only 13% of the installations were commercial customers [4]; in China, energy consumption monitoring platforms (ECMPs) for non-residential buildings have been gradually implemented in various cities, including Chongqing, since 2007 [5] and are able to collect and display real-time electricity consumption data at whole-building and subsystem levels. The aim of ECMPs is to improve the energy management of individual buildings during the operation phase and provide necessary information to policy makers to formulate scientific policies for energy savings [6]. However, the electricity consumption data collected from ECMPs have not been widely used in China. Current applications of ECMP data are largely based on simple statistical data analysis, such as calculating the statistical values of annual or monthly energy usage density of buildings across different functions for energy disclosure or formulating energy consumption quota standards. These applications take advantage of only the annual or monthly electricity consumption data instead of higher-resolution time series data. Therefore, this study aims to discover more insightful information behind the time series data from ECMPs and explore more potential practical applications of the obtained knowledge.

1.2. Literature review

Unsupervised learning approaches are powerful at identifying building electricity usage patterns from operational data [7]. The identification of typical electricity load patterns (TELPs) has been considered an important way to understand the characteristics of daily electricity load profiles (DELPs) of buildings [8]. Clustering techniques, commonly used unsupervised learning techniques that determine inherent patterns in datasets, have been widely used to extract building electricity load patterns [9,10]. The various clustering algorithms that have been used for investigating electricity load profiles can be divided into partition methods, hierarchical methods, density-based methods and model-based methods [11]. For example, Ma et al. applied the partitioning around medoids algorithm to identify typical daily heating load profiles of higher education buildings, where the Pearson correlation coefficient was used instead of Euclidean distance to measure the dissimilarity of cluster heating load profiles [12]. Agglomerative hierarchical clustering with combined dissimilarity measures was proposed to discover electricity load profiles of two university library buildings [13]. A density-based clustering technique was developed to obtain typical dynamics of electricity consumption behaviors, which worked for massive high-dimension electricity consumption data [14]. In addition, the Gaussian mixture model, which is a model-based clustering algorithm, was utilized to extract temperature-related and people behavior-related patterns for the heating load of a district heating system, where the obtained clusters were proven to improve the prediction accuracy of heating load prediction models [15].

The k-means algorithm, as a classic partitioning clustering method, is generally the most frequently used in the DM literature due to its easy implementation and high efficiency [9]. Park et al. [16] compared k-means, bisecting k-means and Gaussian mixture model algorithms and found that k-means was the most appropriate to investigate building electricity load patterns based on a dataset containing 1910 residential and 1919 non-residential buildings. Three typical clusters were obtained, and they showed the differences in electricity load profiles of residential and non-residential buildings. Meanwhile, the results were also shown to have potential for application in establishing benchmarking systems. Wen et al. [17] proposed an improved k-means clustering method with optimal initial cluster centers combined with principal component analysis to improve the convergence speed based on large-scale smart meter data. Carmo et al.[18] employed the k-means algorithm to identify daily heating electricity load profiles of 139 Danish dwellings and there were two main clusters, namely, one representing the characteristics of heating electricity load profiles for weekdays and another representing those for weekends.

However, for massive time series data with a high number of dimensions (24 or higher), some clustering algorithms, such as the k-means algorithm, become intractable and may not be appropriate to group similar electricity load profiles. This problem is also called the "Curse of Dimensionality" [19]. A distance measurement method, dynamic time warping [20], was designed for specifying similarity between time series to address this problem, but unfortunately, this algorithm can be computationally expensive [21]. Alternatively, researchers have proposed strategies to reduce the data dimension to overcome these problems. Feature definition is

a potential way to describe each electricity load profile by a limited number of features defined by experts without introducing additional parameters [9]. Xuan et al. [22] defined three load shape parameters extracted from raw time series data, including the peak-base load ratio, working/nonworking day load ratio and on-hour duration. Then, the k-means algorithm was applied to cluster similar DELPs based on these features. In reference [23], the authors defined seven statistical features to represent raw time series, the mean, standard deviation, skewness, kurtosis, chaos, energy, and periodicity, and then adopted k-mean clustering. Haben et al. [24] divided a day into four time periods: overnight, breakfast, daytime and evening periods. Then, the relative average electricity consumption data of each period were calculated. These studies have shown that compared with raw time series, feature-based clustering can improve the clustering performance and reduce time and computation costs.

Time series clustering for investigating typical patterns of building electricity load profiles has been sufficiently discussed in the current literature. In fact, the clustering algorithm itself does not denote the significance or meaning of TELPs, and merely clustering results are not capable to reflect the characteristics of TELPs [25]. Therefore, knowledge discovery from clustering analysis is a necessary and valuable step to fill the gap between users and clustering results, namely, exploring external dynamic factors such as weather, day type, and occupancy behaviors that can result in different identified TELPs to obtain better insights into managing building operational energy usage. However, there are only a few works making an effort toward this objective, while the majority do much less, typically simply identifying TELPs. The adopted methods mentioned in the studies of knowledge discovery of clustering analysis can be summarized as association rule mining, decision tree, and logistic regression. For example, an enhanced Apriori algorithm was proposed to investigate the relation between identified TELPs for households by the k-means algorithm with thirty-five household characteristics [26]. In reference [27], a novel methodology combined with an adaptive symbolic aggregate approximation method and the classification and regression tree (CART) algorithm was proposed to identify infrequent and unexpected building energy patterns. Two practical public buildings were used for case study analysis. Moreover, logistic regression was applied to link household characteristics and typical energy usage patterns where the clustering analysis were used for pattern recognition [28,18]. Most existing research mainly relies on DM techniques to discover the relation between inherent characteristics of households related to electricity usage behaviors and TELPs of households, such as physical characteristics of the dwellings, socioeconomic situations, etc., instead of the dynamic characteristics related to electricity use (e.g., daily weather parameters). These above studies encourage the potential and feasibility of time series DM in knowledge discovery of electricity data from ECMPs for operation management in non-residential buildings.

1.3. Contributions and paper structure

Based on the aforementioned literature, the majority of the current studies focus on the methods of extracting TELPs, while the knowledge discovery of the patterns, which is a more valuable step, has been seldom discussed, which can make the extracted patterns difficult for practical applications. In this view, this paper aims to develop a general DM-based framework to recognize typical patterns of the electricity load profile and discover the knowledge from the resulting patterns by determining the association between dynamic influencing factors and patterns. The main contributions of this study can be summarized as follows.

1) A general framework integrating both unsupervised clustering analysis and supervised decision tree was proposed, which provides an automatic and efficient way to extract TELPs and mine temporal information hidden in the patterns as well as improve the interpretability of the results of clustering analysis for data analytics in building electricity consumption data.
2) A new clustering method to identify the TELPs at the individual building level was proposed in this study. Different from the majority of the previous studies merely using a single-step clustering technique for grouping similar electricity load profiles, the proposed clustering strategy contains two clustering steps. Before clustering, the dimensions of the raw daily electricity consumption data were reduced by defining several statistical features representing the daily profiles. The first clustering step aims at detecting outliers of DELPs by using the DBSCAN clustering technique, which addresses the data quality issues for electricity consumption data derived from ECMP, such as data loss and instantaneous outliers [29]. The second step aims at grouping the similar DELPs by means of the k-means algorithm to extract TELPs. The effectiveness was demonstrated based on a comparison of the proposed clustering method with two single-step clustering techniques.
3) The framework was evaluated in three practical office buildings. A potential application of the framework was also demonstrated in terms of how to detect anomalous electricity load profiles at the early stage for building managers. Though the application of the proposed framework was validated in three practical buildings, the generalizability and robustness should be further explored in a larger dataset.

The paper is organized as follows. Section 2 briefly describes the methodology adopted for the proposed framework. Section 3 details the results of applying the proposed framework to three practical office buildings. In Section 4, a comparison of the proposed clustering method with two single-step clustering techniques is presented. Section 5 provides a discussion, and Section 6 highlights the conclusions.

2. Methodology

2.1. Outline of the proposed framework

The proposed framework consists of three phases as shown in Fig. 1. In the data preparation phase, there mainly three tasks in terms of data preprocessing, data segmentation and data normalization. The next phase is identification of TELPs, which mainly contains three steps. Step 1 is feature definition. Specifically, five statistical features representing the shape characteristics and replacing the raw time series of DELPs (i.e., 24 h with 24 dimensions) are defined and used in further clustering steps. Then, the outliers of DELPs are identified by means of DBSCAN algorithm and then removed in the second step. Therefore, the effect of outliers can be eliminated when k-means algorithm is further conducted. The aim of step 3 is recognizing TELPs through k-means clustering algorithm. Phase 3 aims to discover useful decision rules from the relation between the obtained TELPs and influencing factors by means of decision tree (i.e., CART algorithm). Finally, a potential application in anomaly detection is demonstrated. The following sections will provide details of each phase.

2.2. Data preparation

Data preparation covers three main steps: data preprocessing, data segmentation and data normalization. There are two main

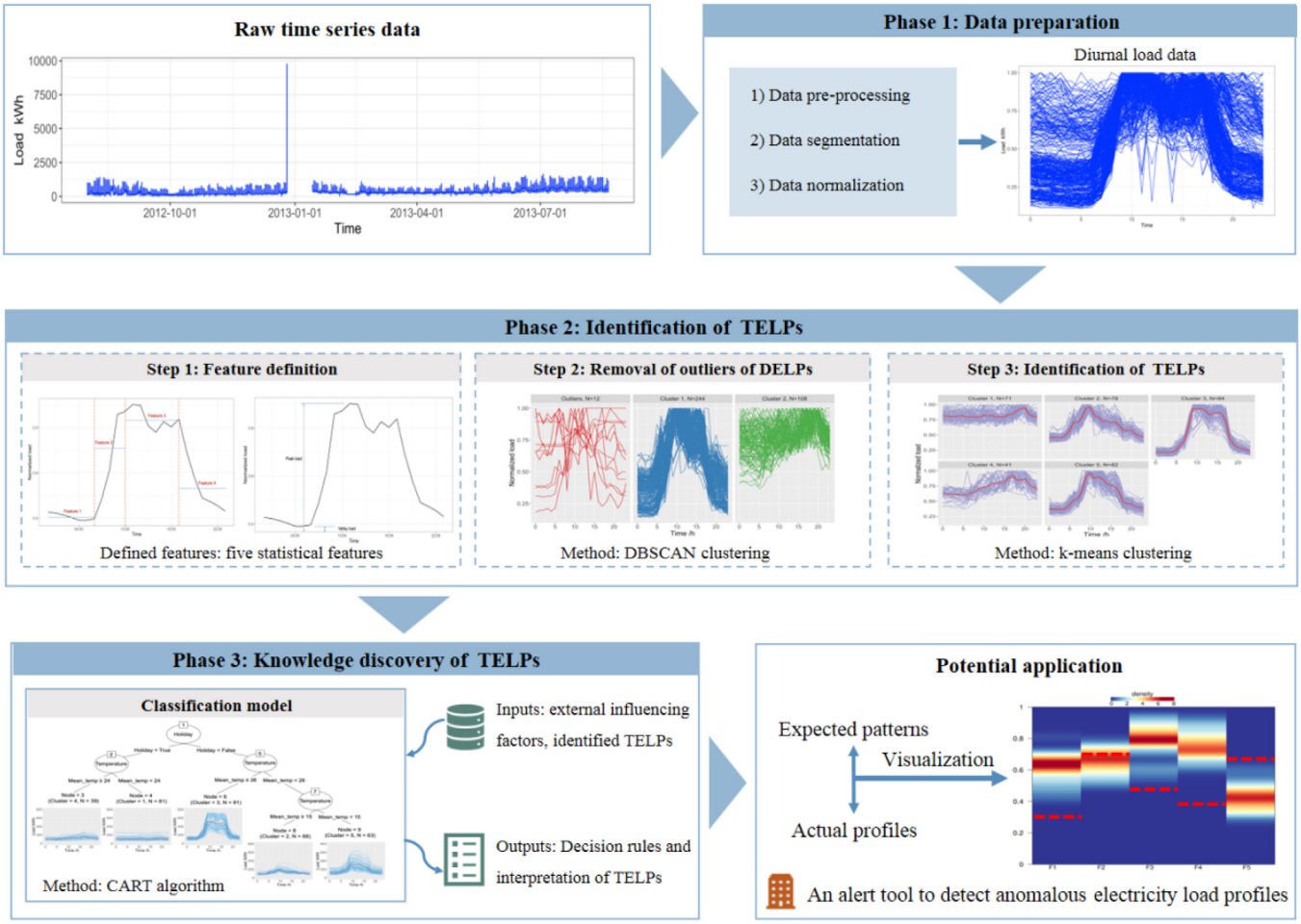


Fig. 1. The framework proposed in this study.

tasks in the data preprocessing process: detecting and removing missing values and outliers in the raw time series dataset. The outliers whose values are significantly higher or lower than the normal range were identified based on the threshold $Q3 + 1.5IQR$ (where $Q1$ and $Q3$ are the first and third quartiles, respectively, and IQR is an interquartile range between $Q3$ and $Q1$) and then labeled as missing values. Then, the missing values (including the inherent and labeled missing values) were filled by using a simple moving average method with a window size of 3.

Data segmentation means reshaping the raw time series data (i.e., annual hourly electricity consumption data) into DELPs. Data normalization aims to capture the temporal variation rather than the magnitude difference and preserve the information in the original DELPs, thereby the hourly electricity consumption data of DELPs were normalized to the daily maximum electricity load for each individual building for further analysis as given by equation (1):

$$p_{nor}(t) = \frac{p(t)}{p_{max}(t)} \tag{1}$$

where $p_{nor}(t)$ and $p(t)$ represent the normalized and actual electricity load data at time t $(t = 1, 2, .., 24)$, respectively; $p_{max}(t)$ means the daily maximum electricity load.

2.3. Feature definition of daily electricity load profiles

Time-series electricity load data often have high dimensions that can bring challenges to clustering algorithms based on the distance function (i.e., Euclidean distance), causing issues such as producing poor clustering results and increasing computational costs [30,31]. Feature definition was therefore used for dimensionality reduction in this study.

We first divided daily profiles into four segmentations according to the working schedule of case study buildings. Specifically, we defined four key time periods, namely, 00:00–06:00, 07:00–10:00, 11:00–17:00, 18:00–23:00, representing off time, rise time, daytime and evening, respectively. Then, the mean value of electricity consumption for each segmentation was calculated and used for the first four features. To better capture the shape characteristics of the daily electricity load curve, the daily peak-to-valley difference rate was introduced as a fifth feature and defined as the ratio of the difference between the daily maximum and the minimum load to the daily maximum load. A higher daily peak-to-valley difference rate reflects a clearer shape of peak electricity load. The five features are displayed in Fig. 2. Therefore, a 24-dimension dataset was reduced to a 5-dimension dataset. The dataset of the five features replaced the original DELPs and were prepared for further clustering analysis.

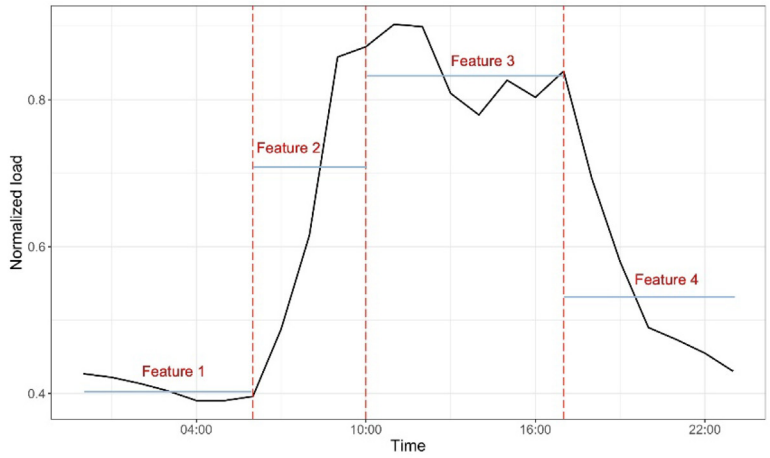
2.4. Removal of outliers from daily electricity load profiles

Before TELPs are identified, small parts of electricity load profiles with an irregular shape called "outliers" can exist in the DELPs for each building, such as dead values that refer to the load values remaining the same for a certain long period [25] and some unknown errors generated in a long-term process [29]. These out-

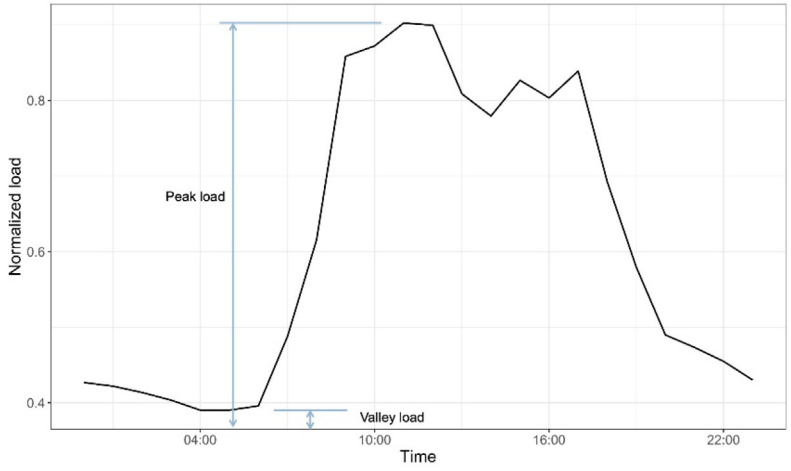
liers may influence the accuracy of extracting TELPs when conducting k-means clustering analysis and are therefore required to be recognized and deleted. To efficiently recognize these irregular electricity load profiles, DBSCAN algorithm, which clusters the observations that are closely packed together and marks the observations in low-density regions as outliers, was selected in this study. Compared with partitioning clustering algorithms, the advantages of DBSCAN algorithm are as follows: 1) it can partition the data clusters with arbitrarily shaped clusters, while partitioning clustering algorithms are suitable for dealing with spherical-shaped clusters or convex clusters; 2) the number of clusters is not required to be pre-assigned; and 3) it is capable of recognizing outliers in the low density areas.

The central concept of DBSCAN algorithm is that a neighborhood of a given radius has to contain at least a minimum number of observations [32]. Therefore, there are two important parameters that must be determined for DBSCAN algorithm: *eps* and *MinPts*. The parameter *eps* denotes the radius, and *MinPts* represents the minimum number of observations in the *eps* region. Given a specific *eps* and *MinPts*, DBSCAN selects a random point and checks its *eps*. If there are more than *MinPts* points in the neighborhood, the points are marked as core points, and the algorithm will start a new cluster that is expanded by assigning all points within this neighborhood to the cluster. The points that are reachable from core points but are not core points are marked as border points and are also included in the cluster. If these new points are still core points, they will be included in the neighborhood. The next step is to randomly select another point that has not been visited in previous steps and apply the same procedure. After all the points are processed, the points that are not assigned to any cluster are labeled noise.

The difficulty of using the DBSCAN algorithm lies in choosing appropriate values for *eps* and *MinPts*. There is no general rule to determine these two parameters. An overly small *eps* can lead to a large part of the data being considered noise, while an excessively large *eps* can partition the majority of the data into the same cluster. To effectively select proper *eps* and *MinPts* values, fast *K*-nearest neighbor (kNN) search and fixed-radius nearest neighbor search were applied in this study [33]. *MinPts* was set to the dimensionality of the data for clustering plus one, as suggested



a) The first four defined features of a DELP (time intervals=4, 00:00-06:00, 07:00-10:00, 11:00-17:00, 18:00-23:00)



b) Explanation of the fifth feature of a DELP

Fig. 2. The five defined features of daily electricity load files.

by the literature [33]. The *eps* value was determined by means of *K*-nearest neighbor distances, that is, the mean distances of each point to its *K-th* nearest neighbors were calculated, and the value of *K* corresponded to the *MinPts* that we previously specified. Then, *K*-distances were plotted in ascending order to find the knee, which corresponds to *eps*. Fig. 3 shows an example of a kNN plot where we set the *MinPts* to 6, and the knee in the figure (red dashed line) is 0.13; thus, eps was determined to be 0.13.

2.5. Extraction of typical electricity load patterns

TELPs can be used to understand the characteristics of building operating electricity consumption. We used k-means algorithm to cluster similar electricity load profiles based on the defined five features after removing the outliers of DELPs. K-means clustering algorithm was selected in this study since it is capable of handling big data because its time complexity is close to linear [34]. Meanwhile, k-means and its variants [16,35] are commonly used for load profile clustering. The main idea of k-means algorithm is partitioning a dataset of daily features $N = \{N_1, N_2..., N_j\}$ into k sets to minimize the within-cluster sum of squared errors (SSE). Eq. (2) illustrates the objective function of the k-mean clustering algorithm. The Euclidean distance was used for dissimilarity measure, which is the most commonly used distance for k-means algorithm due to its high competitiveness [36].

$$SSE = \sum_{i=1}^{k} \sum_{j=1, j\in i}^{n} \|N_j - \mu_i\|^2 \tag{2}$$

where N_j is a vector that represents the *j-th* daily feature, $j = 1, 2, 3, \cdots, n$; and μ_i is the vector represents the *i-th* cluster center, $i = 1, 2, 3, \cdots, k$.

K-means algorithm proceeds as follows: 1) a prior k is determined; 2) randomly create k number of initial centroids; 3) calculate the distance between each daily feature and the cluster center, then assign the object to the cluster with the minimum distance; 4) update the cluster center by calculating the mean value of all the daily features in the cluster; 5) steps 3 and 4 are repeated until the centers do not change.

There are two core inputs that must be determined before k-means algorithm is implemented: the initial cluster centroids and the number of clusters k. For choosing optimal initial centroids, k-means is conducted 50 times with different initial cluster centroids. The record is chosen based on the best *SSE* among all the iterations. In terms of determining the optimal number of clusters,

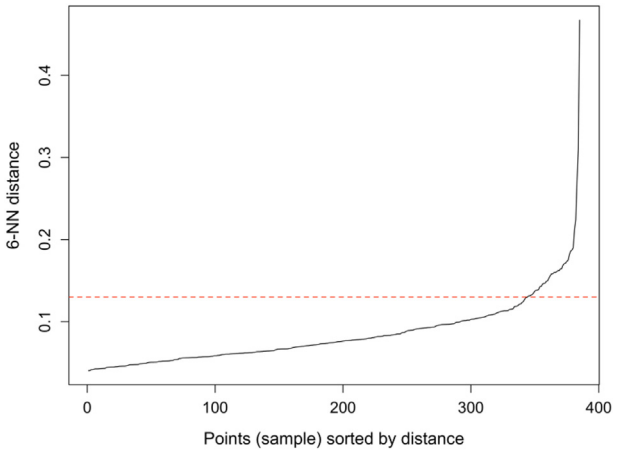


Fig. 3. *K*-nearest neighbor distance plot for determining the suitable *eps* for a sample building.

most existing clustering validation indexes (CVIs) have their limitations and none can always perform robustly, thus the Dunn index was selected to evaluate the clustering results based on a comparison with the other four CVIs, as presented in Appendix A. The Dunn index defines the ratio between the minimal inter-cluster distances to the maximum intra-cluster distance [37]. A larger Dunn index indicates that clusters are compact and well separated.

2.6. Knowledge discovery by CART

After identification of TELPs for each building, knowledge discovery aims to find the relations between TELPs and potential influencing variables to better interpret the identified TELPs and explore potential applications of the discovered knowledge. There are two potential DM techniques that are suitable for achieving this objective: association rules mining (ARM) and decision tree. ARM is widely used in discovering associations between different transactions in a large Boolean transactional database [38]. However, the commonly used influencing variables related to electricity load, such as temperature, humidity, and occupancy rate, are numeric variables, and it is difficult to transform the numeric data into categorical data since the intervals for the categories of "high", "medium" or "low" are difficult to determine. Therefore, ARM was not considered in this study. The CART algorithm is a supervised machine learning technique that can be used to conduct a predictive modeling task of multiclass classification. Moreover, CART enables to handle both numeric and categorical data and automatically selects the most important variables to generate classification trees, which is suitable to deal with the dataset in this study. In addition, compared with other popular supervised learning algorithms such as random forest or extreme gradient boosting, CART has a higher interpretability of resulting classification models. Thus, the CART algorithm was selected to generate classification trees in this study. The target variable (i.e., TELPs) is categorical and classification trees can be used to determine the expected pattern based on given input predictors.

The procedure of developing a classification tree by CART algorithm is shown as follows. A classification tree is generated via binary recursive partitioning. At the beginning, all the observations are partitioned at the root node for binary partitions. At each node of the tree, the algorithm chooses a predictor variable from the factors influencing building electricity consumption to partition electricity load profiles into two nodes. The variable and the location of the split are selected to minimize the impurity of the node. The *Gini* index is used to assess the minimization of node impurity. A smaller *Gini* index indicates a higher purity and a more accurate classification model. Each of the two groups obtained from initial splits continues to be split into smaller subsets, and the process continues until it is no longer possible to generate additional splits or the user-set stop conditions are met. Hence, the final binary trees reflect the most important predictor variables of typical patterns and the partition locations of predictor variables.

To restrict a classification tree to an appropriate size, we first set an early stop condition by providing the minimum number of samples for a node split (*minsplit*), which is used to control the tree size and avoid over partitioning. Even when the early stop condition is satisfied, the tree can be large and complex. Thus, a post-pruning process should also be conducted by setting a cost complexity (*cp*) parameter. The *cp* value is determined by the one-standard error rule during a 10-fold cross-validation [39]. Specifically, the cross-validation error representing the average classification error from 10-fold cross-validation is summed with its corresponding standard error. Then, the cross-validation error with the minimum number of partitions below this summed value is selected, and the corresponding *cp* is the optimal value for developing the final clas-

sification tree. This selection rule for pruning the tree provides a trade-off between the simplicity of the tree and minimizes the error rates of classifications.

3. Case study applications

In this study, the proposed framework was implemented in the R programming language. DBSCAN clustering, k-means clustering, and CART algorithm were implemented by using the R packages *dbscan* [33], *cluster* [40] and *rpart* [41], respectively.

3.1. Description of the case study buildings

The dataset that we used for this study comes from ECMP in Chongqing. ECMP in Chongqing was established in 2012 by Chongqing Municipal Commission of Urban-Rural Development to collect hourly electricity consumption data for non-residential buildings in Chongqing [42]. ECMP mainly includes five parts: measurement meters for collecting electricity consumption data, a field bus, a communication network for transmitting energy data, the server's data center and an electricity consumption platform [43]. ECMP can not only collect hourly electricity consumption data but also demonstrate the results on the web in the form of tables or figures, which can be visited by the authorized users.

Though we cannot obtain access to visit ECMP directly, with the help of Chongqing Municipal Commission of Urban-Rural Development, three office buildings with relatively complete and normal electricity consumption data were selected for the case study. The data of each building were downloaded from ECMP in Chongqing and provided in the format of comma-separated values (CSV) files and contain approximately 300 daily profiles, ranging from August 1, 2012 to August 20, 2013. Some basic information of each building was also collected from the ECMP, including the building floor area, number of floors, construction year, and heating and cooling system type, which are given in Table 1. In addition to the aforementioned information, weather data from China Meteorological Administration (http://data.cma.cn/), including daily average outdoor air temperature and relative humidity, were also collected for further analysis and are presented in Fig. 4. It is noted that the three buildings are located in Chongqing, where the climate zone is hot summer and cold winter zone; therefore, we did not consider the climate differences across the buildings, i.e., the case study buildings share the same outdoor average air temperature and relative humidity data.

3.2. Outlier detection for daily electricity load profiles

In this process, DBSCAN algorithm was adopted to identify the outliers of DELPs based on the defined features of DELPs. The number of raw and cleaned DELPs of each building is summarized in Table 2. No more than 5% of the total daily profiles were identified as outliers for each sample building.

Figs. 5 and 6 present two types of visualizations outliers and general clusters of DELPs for three buildings. Fig. 5 presents a convex hull plot with two dimensions to visualize the relations between identified clusters and outliers, where the two dimensions are generated by principal component analysis. For instance,

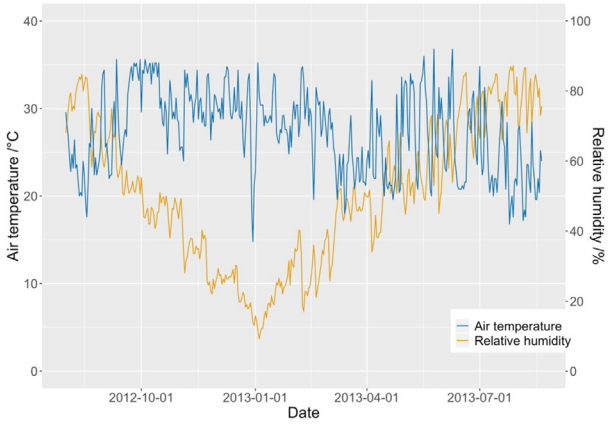


Fig. 4. The daily mean outdoor air temperature and relative humidity during the observed period.

in Fig. 5(a), Dim 1 and Dim 2 are two principal components and represent 58.8% and 22.5% of the variation in the clustering dataset, respectively. Each convex hull represents a cluster identified by the DBSCAN algorithm, and the red crosses refer to the outliers of DELPs and are expected to be relatively far away from the convex hulls, i.e., outliers. Note that although some outlier points are contained within the convex hull of a different cluster for building A033, the noise points are basically well recognized according to Fig. 6(a). In addition, the clusters have different densities (especially for building A169), which makes k-means clustering algorithm underperform and outputs inadequate clustering results if the outliers are not removed since it is difficult to identify clusters with widely different sizes or densities.

In Fig. 6, the outliers and general clusters of the daily profiles are displayed. The irregular curves are identified in the cluster called "outliers". Note that there are only two electricity load patterns identified by DBSCAN algorithm in addition to the cluster of outliers for the three buildings, indicating that the result of DBSCAN algorithm is likely not sufficient to represent all the potential TELPs. Consequently, it is necessary to conduct a second-step clustering analysis to further identify the TELPs after removing the outliers.

3.3. Identification of typical electricity load patterns

After the outliers of DELPs were removed, TELPs for three buildings were identified through k-means clustering method. The optimal number of clusters should be first decided before the k-means algorithm is implemented. Using too few clusters could not be useful to discover the TELPs, while using too many clusters could result in insignificant differences across some of the patterns. Therefore, the optimal number of clusters was selected to be from 2 to 8, and the results were evaluated by Dunn index. A higher Dunn index represents a better clustering result.

Fig. 7 shows the calculated Dunn indexes for different numbers of clusters. The optimal numbers for buildings A033, A155 and A169 were 5, 4 and 4, respectively, as the corresponding Dunn index is the highest in each case.

Table 1
Basic information of case study buildings.

Building name	Total floor area (m^2)	Construction year	Number of occupants	Number of floors	Heating and cooling system type
A033	11,579	2004	260	8	VRV (variable refrigerant volume) system
A155	20,652	2008	250	7	Chiller + gas boiler
A169	47,579	2003	300	12	Centralized heat pump system

Table 2
A comparison between the number of raw and cleaned daily electricity load profiles.

Building name	Number of raw daily profiles	Number of cleaned daily profiles after outlier removal
A033	364	352
A155	385	374
A169	299	284

The TELPs identified for three buildings by k-means clustering are given in Fig. 8. The purple lines in the figure represent DELPs in the cluster, and the red curves are the TELPs calculated by averaging the electricity load profiles in the same cluster. Fig. 9 displays the electricity usage intensity (EUI) reflecting electricity consumption levels of each cluster for three buildings, where the EUI was calculated by the ratio of hourly electricity consumption to total floor area.

For building A033, there are five TELPs representing five electricity usage characteristics. Clusters 2, 3 and 5 have clear peaks during the daytime but different time durations. Specifically, for cluster 3, the peak electricity loads occur from 9:00 am to 16:00 pm and then start to fall, corresponding to high-level electricity consumption according to Fig. 9. For clusters 2 and 5, the peak electricity loads appear at 10:00 am and then drop gradually. On the other hand, clusters 1 and 4 have relatively flat and high curves, revealing a small variation in electricity consumption during the day, where the peak time often occurs in almost the evening, and the curves of cluster 1 are relatively gentle compared with those of cluster 4.

According to Fig. 8(b), building A155 has four TELPs. The patterns of clusters 2 and 3 share a similar peak time, with peak electricity loads from 9:00 am to 17:00 pm, representing a high-level electricity usage during building operation, and these two clusters account for the largest proportion of all the electricity load profiles. In addition, the pattern for cluster 1 is relatively unchanging, with two peaks appearing at 12:00 am and 18:00 pm, respectively. However, the pattern for cluster 4 is high and smooth without a clear rising or falling trend, and its low electricity usage level indicates that the building is rarely occupied during the day.

For building A169, peak loads are from 10:00 am to 15:00 pm for clusters 1 and 4, respectively, while the pattern of cluster 4 additionally has a small peak at 21:00 pm, indicating that overtime working is likely to occur in the evening in this cluster. For cluster 2, the peak electricity loads appear at 10:00 am and 21:00 pm. Different from the patterns of other three clusters, cluster 3 pattern remains stable during the daytime and evening except for two peaks at 10:00 am and 21:00 pm.

Overall, the identified TELPs from the proposed clustering method for three buildings are reasonable since each cluster has a unique electricity load pattern with clear variations compared with the patterns of other clusters.

3.4. Knowledge discovery by CART

To further understand the reasons leading to different identified TELPs based on the proposed clustering analysis, i.e., improve the interpretability of the clustering results, a classification model was developed using CART algorithm to explore the underlying relations between the identified patterns and potential influencing factors. The potential influencing factors considered in this study are summarized in Table 3.

The pre- and post-pruning processes were conducted to simultaneously partition data into smaller subsets and reduce the complexity of the final classification tree to avoid overfitting. The two parameters *cp* and *minsplit* were determined by users. The *minsplit* value was set to 10, meaning that a node can be further split to achieve purity when the number of observations in each split is more than 10. The *cp* value was determined by the one-standard error rule. The final optimal *cp* and the corresponding cross-validation error and total accuracy of three buildings after pruning

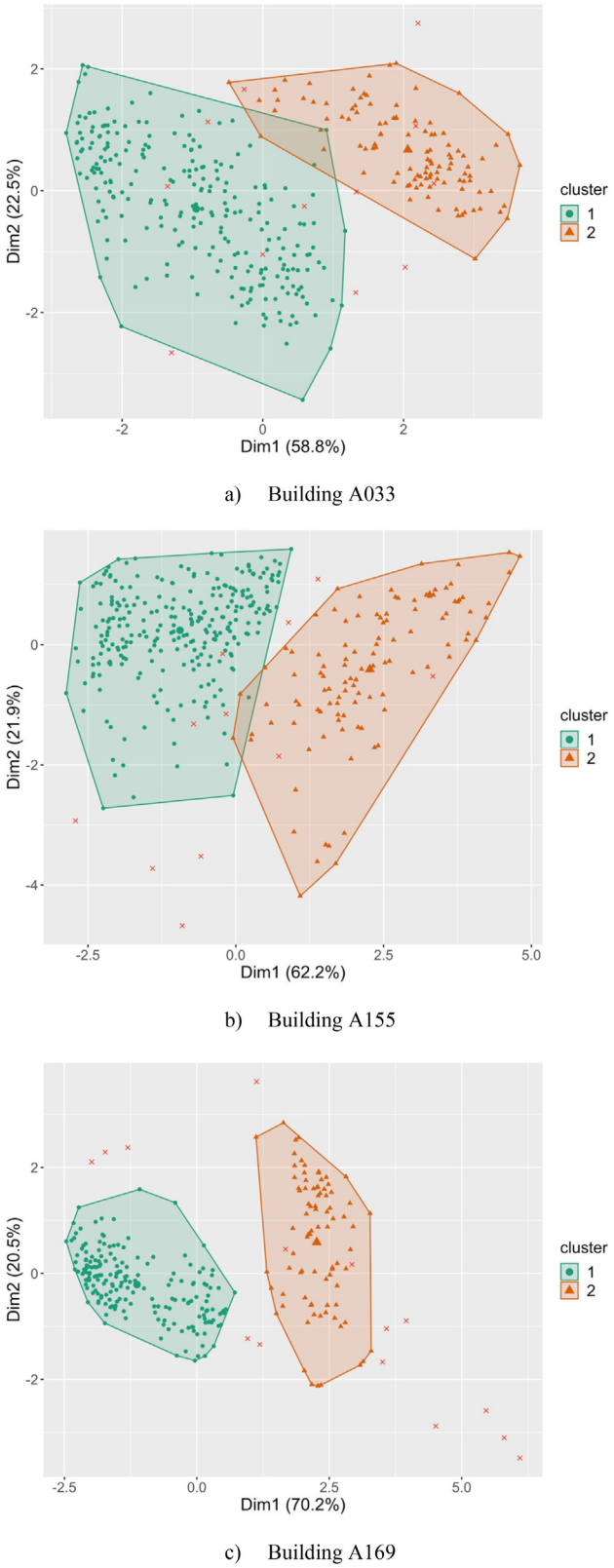
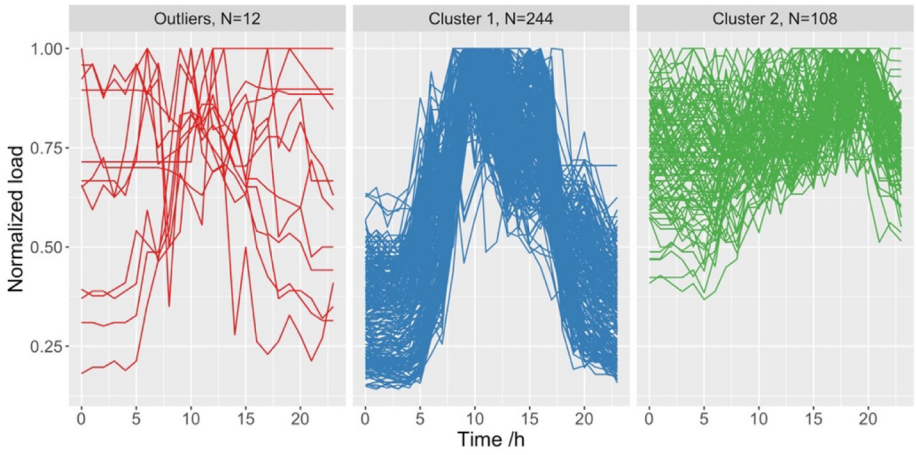
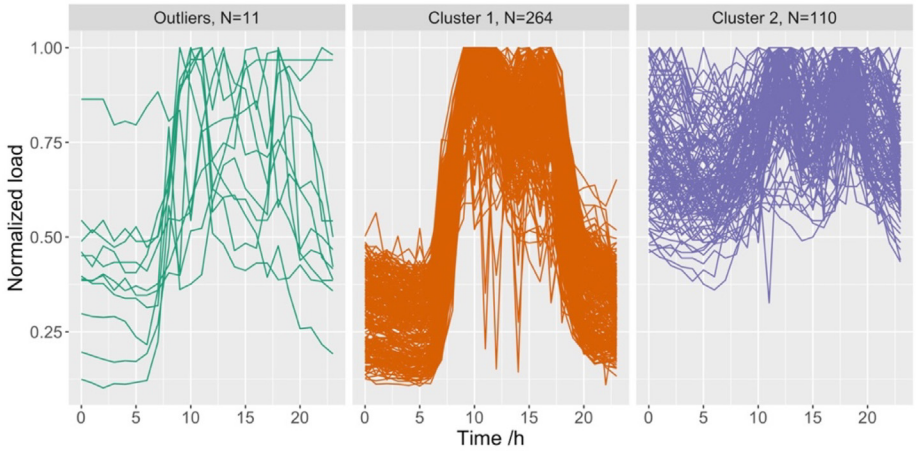


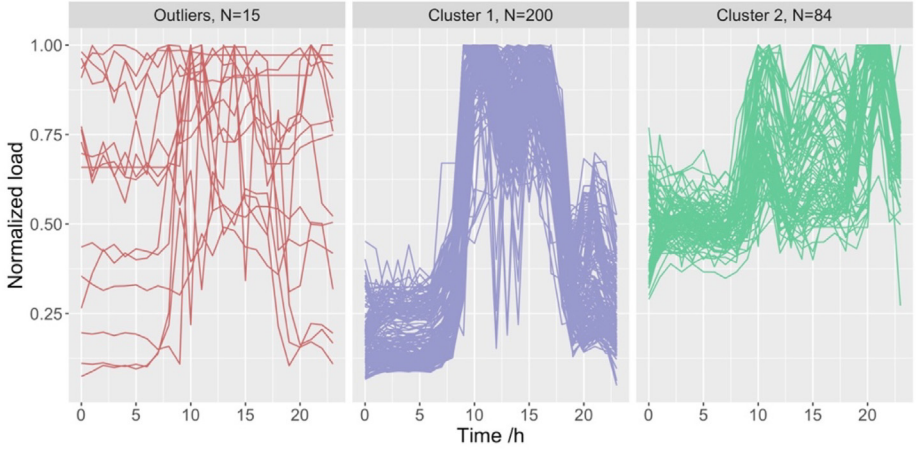
Fig. 5. Visualization of the results of DBSCAN clustering by convex hull plots for three buildings.



a) Building A033



b) Building A155



c) Building A169

Fig. 6. Visualization of the results of detecting outliers of daily electricity load profiles for three buildings.

are shown in Table 4. Total accuracy means the accuracy rate calculated by the ratio between the predicted classification and the identified TELPs. The result shows that the cross-validation error ranges between 0.1 and 0.3, indicating that the classification models basically produce reliable results and are able to explain the clustering results for the three buildings.

The structure of the classification tree for three buildings are visualized in Fig. 10. Note that the original electricity load profiles

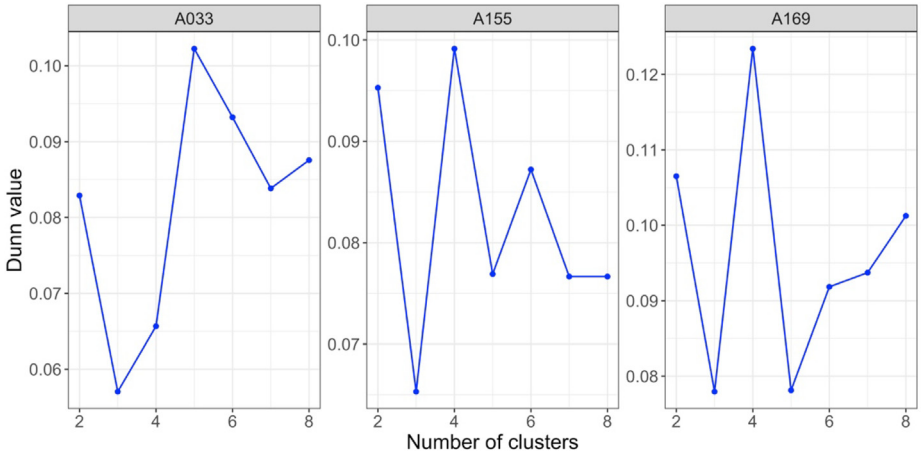


Fig. 7. Clustering performance for different numbers of clusters for three buildings.

without data normalization are provided in this figure in order to directly display the electricity consumption levels of each class and help elucidate the characteristics of TELPs. Holiday and mean_temp were the two main variables that contributed to the different TELPs identified for each building. The detailed information of the classification tree is presented as follows.

For building A033, there are five terminal nodes (i.e., nodes 3, 4, 6, 8 and 9) representing five different electricity usage patterns, and the five decision rules generated by the classification tree are summarized in Table 5. CART algorithm selects "holiday" as the splitting variable at the root node (node 1) and then automatically separates the "Temperature" into two groups: nonworking and working days. The highest electricity consumption is associated with working days (Holiday = False) and high outdoor air temperature (mean_temp $\geq$ 26). This result is reasonable because the building is highly occupied during weekdays for normal business use; additionally, the cooling demands are needed when the outdoor air temperature is high. It can also be observed that relatively higher electricity consumption also appears when the outdoor air temperature is low (mean_temp $<$ 15) during working days but that is lower than the electricity consumption in hot season. This suggests that heating demand of this building is likely to be lower than the cooling demand. On the other hand, lower electricity consumption is related to nonworking days (Holiday = True) and a mild season (15 $\leq$ mean_temp $<$ 26). It is interesting to find that there are still two different patterns during nonworking days, which are distinguished by daily mean outdoor air temperature. The possible reason could be that the building is still partly occupied on nonworking days in the hot season (Figs. 8(a) and 10(a)).

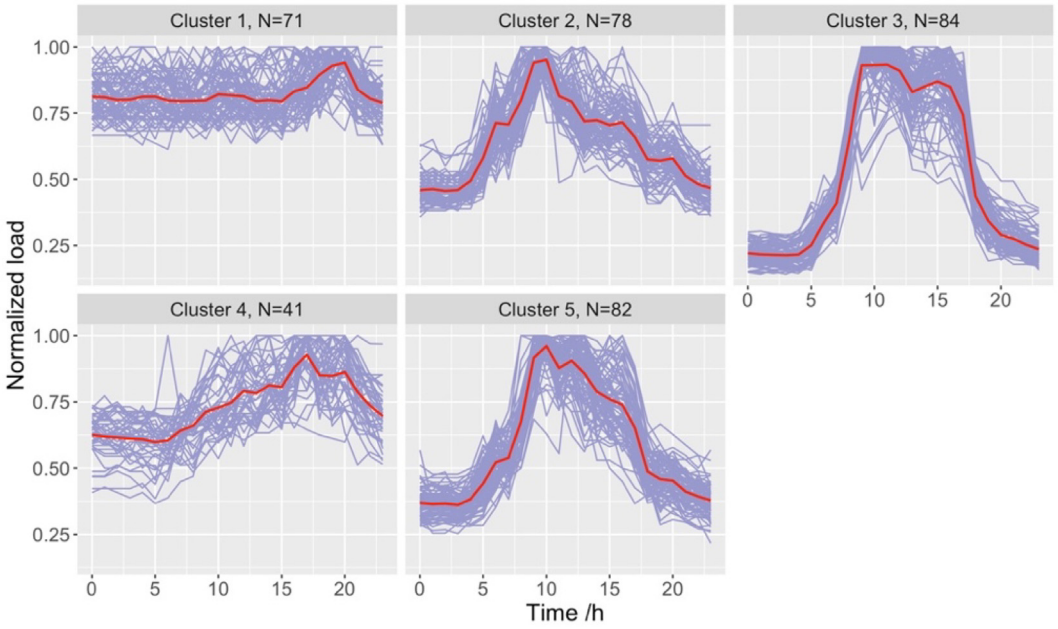
For buildings A155 and A169, some similar decision rules are summarized in Table 5. The right tree in Fig. 10(b) divides the group of node 7 into two subgroups by the decision rule associated with temperature (mean_temp $\geq$ 23 and mean_temp $<$ 23). This means that the electricity load profiles of hot season are partitioned into one group, and those of the cold and mild seasons are partitioned into another group during working days for building A155. This can be explained by the type of heating equipment installed in this building, that is, a gas boiler. Gas consumption data were not collected by the platform, leading to the heating energy consumption data excluded in the total electricity consumption; as a result, the electricity load patterns of the cold and mild seasons are almost the same.

The decision rules and predicted occurrence probability of each pattern extracted from the classification trees are given in Table 5. The occurrence probability of the expected patterns is marked in bold. It is possible to predict an expected pattern under a given boundary condition with a high probability based on a classification tree. For example, for building A169, there are four predicted occurrence probabilities for each TELP, corresponding to the patterns 1, 2, 3, and 4 in Fig. 8(c), respectively. The occurrence probability of pattern 1 reaches 90% when the mean_temp is lower than 13 °C during working days. Based on this rule, the occurrence probability of patterns 2, 3 and 4 can be very low and even close to zero under this boundary condition. In general, a higher occurrence probability also indicates a better reliability of the predicted pattern. Furthermore, it is found that the predicted occurrence probability for the patterns on working days tends to be higher than the probability on nonworking days, possibly because the overtime occurrence on nonworking days can be stochastic and uncertain [46] and thereby increases the uncertainties of predicted patterns. Thus, other input influencing variables regarding occupancy behaviors on nonworking days are expected to be further considered to enhance the accuracy of the classification trees.

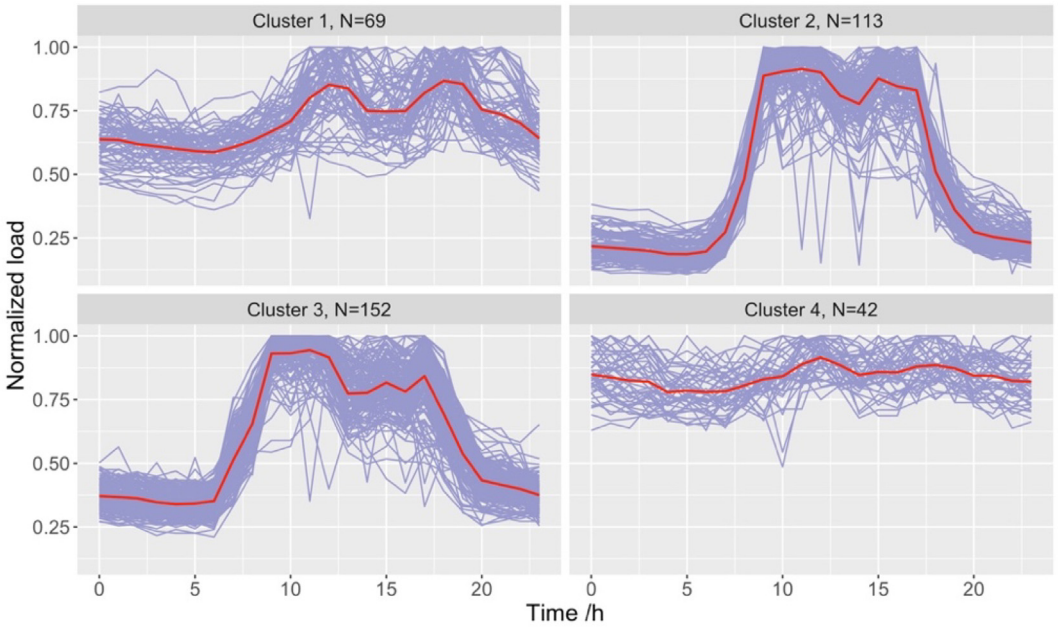
4. Comparison of the proposed clustering method with other two clustering techniques

As shown in the results, the TELPs were extracted by the proposed clustering method. To further demonstrate the effectiveness of this method, we compared the performance of the proposed clustering method with that of two other single-step clustering techniques, in which TELPs were directly extracted based on raw DELPs without removing outliers of DELPs or reducing dimensions. One is k-means clustering, and the other is the Gaussian mixture model (GMM) clustering algorithm. These two clustering techniques were selected because they are the most popular clustering methods and have been widely adopted in time series clustering analysis [15,22,8]. Moreover, different from k-means clustering, GMM clustering is a model-based clustering algorithm that enables accommodation of clusters with different sizes and correlation structures within them [47]. The optimal number of clusters for k-means and GMM clustering algorithm was also determined by Dunn index, selecting from 2 to 8. The results of the calculated Dunn index for different numbers of clusters are shown in Appendix B.

The TELPs identified by the k-means clustering technique are shown in Fig. 11. The red line represents the cluster centroid. The optimal cluster number of k-means clustering is two for all three buildings. The two identified typical patterns for each building can reflect only the electricity usage differences for working days and nonworking days. In comparison with the results



a) Building A033



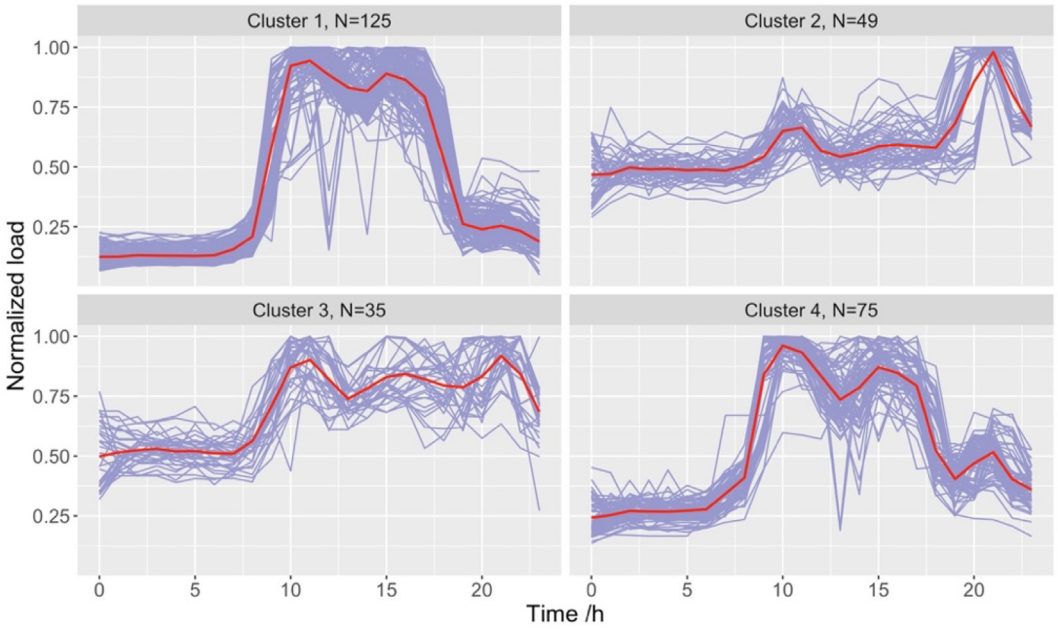
b) Building A155

Fig. 8. Typical electricity load patterns identified for three buildings by the k-means clustering method.

obtained from the proposed clustering method, other interesting information, such as the outdoor temperature effect on electricity usage patterns for each building, cannot be reflected.

The TELPs identified by using the GMM clustering technique are presented in Fig. 12. For building A033, only two types of TELPs are identified and reflect the characteristics of electricity usage for working and nonworking days, which is similar to the results from k-means clustering. In terms of the other two buildings, the number of TELPs identified by GMM is 6 for both buildings, which is more than the number from the proposed clustering method. However, some TELPs identified by the GMM clustering algorithm have relatively small differences, indicating that GMM may not be a suitable clustering method to extract distinct TELPs.

Notably, k-means and GMM clustering techniques seem to have poor performance in detecting outliers of DELPs and partitioning the outliers into a group. The above comparison demonstrated the effectiveness of the proposed clustering method in outperforming the other two single-step clustering methods in terms of detecting outliers and discovering typical electricity usage characteristics.



c) Building A169

Fig. 8 (*continued*)

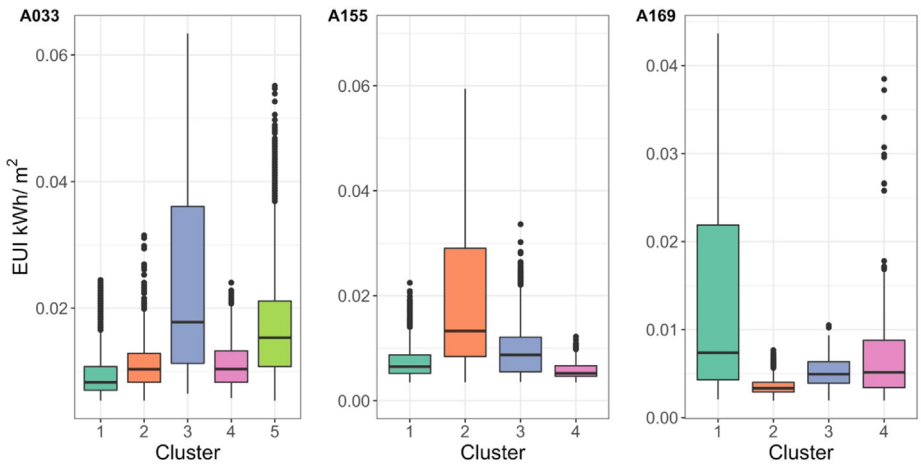


Fig. 9. EUI distribution across obtained clusters for sample buildings.

Table 3
Summary of selected potential influencing factors.

Variable	Type	Description
Weekday	Categorical	Day of the week
Holiday	Categorical	Chinese legal holidays including weekends
Month	Categorical	Month of the year
Season	Categorical	Season of the year
mean_temp	Numerical	Daily average outdoor air temperature (°C)
mean_hum	Numerical	Daily average outdoor air humidity (%)

Table 4
The optimal *cp* and the corresponding accuracy for the three buildings.

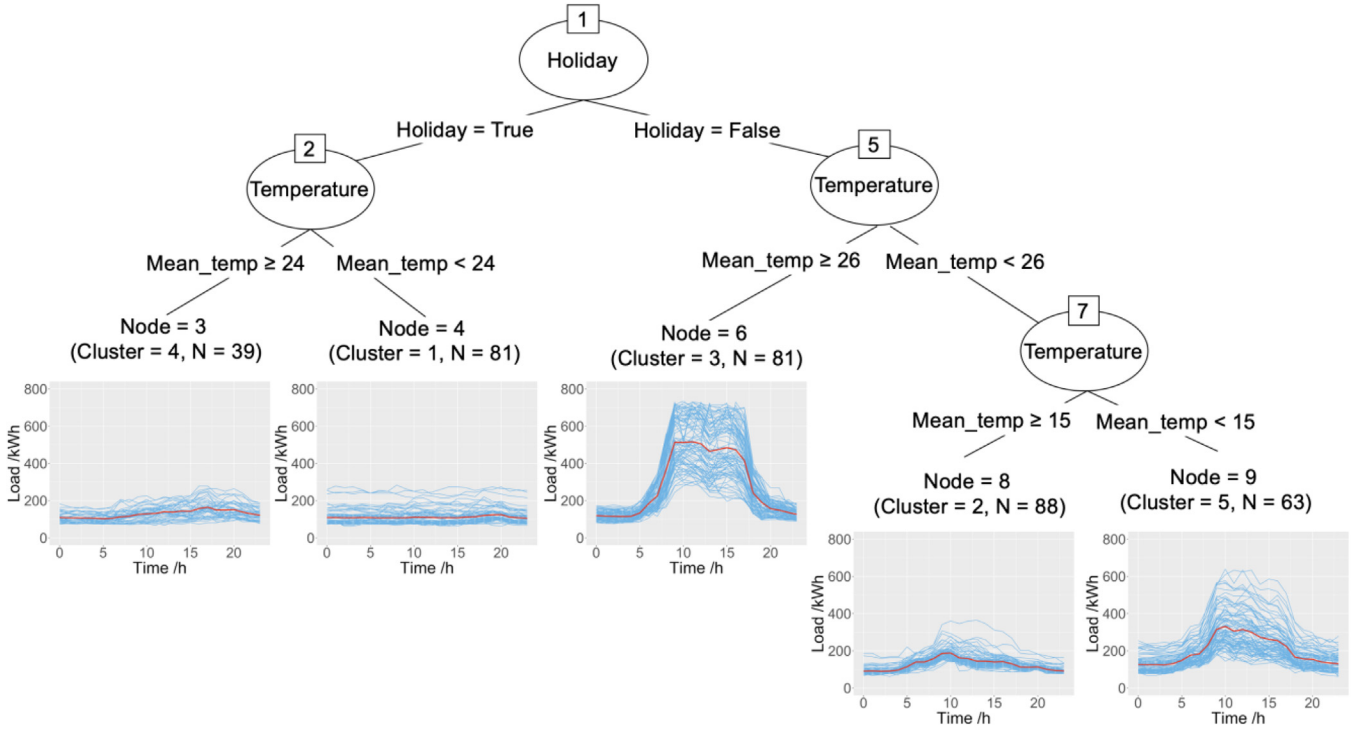
Building name	Optimal cp	Cross-validation error	Total accuracy
A033	0.037	0.265	0.817
A155	0.029	0.296	0.838
A169	0.062	0.182	0.915

5. Discussion

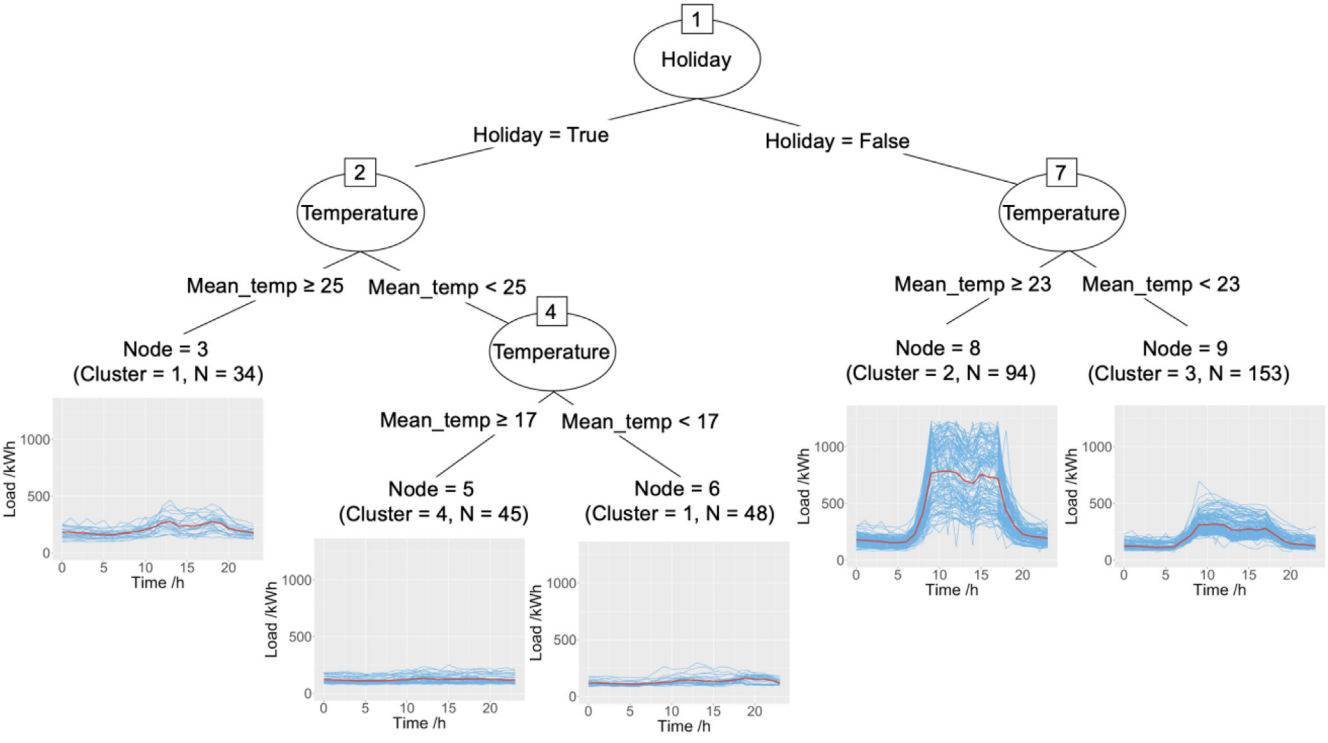
5.1. Potential application of the proposed framework

One potential application of the proposed framework is to detect anomalous changes in building electricity load profiles at the early stage for building managers. This application procedure contains two steps: first, using the obtained decision rules from the framework to predict the expected pattern with a high occurrence probability under a given certain boundary condition; and second, comparing and quantifying the differences between the expected and actual patterns.

Taking building A155 as an example, there are two given anomalous profiles (profiles 1 and 2), and the corresponding expected patterns are also shown in Fig. 13. For profile 1, the boundary condition is low temperature (lower than 17 °C) and a nonworking day, and the corresponding electricity usage pattern



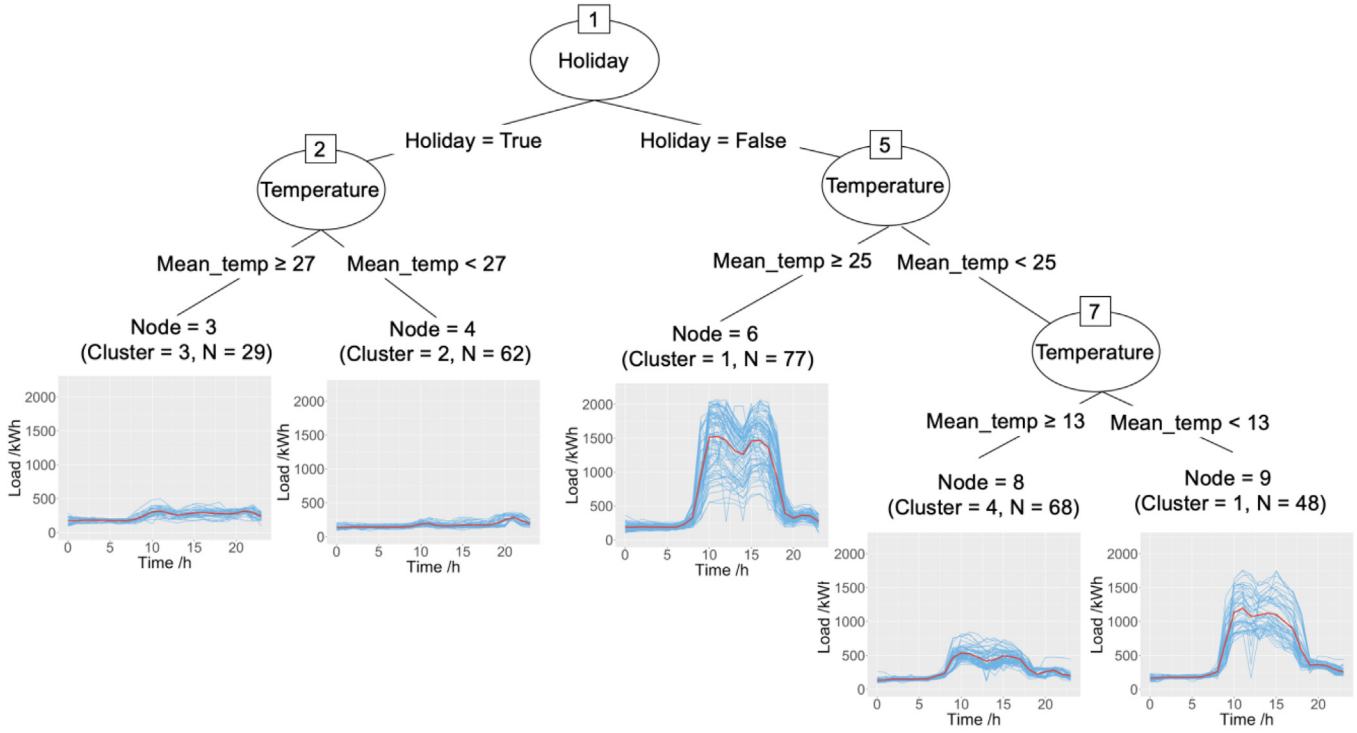
a) Building A033



b) Building A155

Fig. 10. Classification trees developed for typical electricity load patterns for three buildings.

is expected to be pattern 1 according to Table 5. Similarly, the boundary condition for profile 2 is high temperature and a working day, therefore, pattern 2 is expected. If the characteristics of a candidate electricity load profile are not in accordance with the expected pattern, the profile can be considered a potential abnormal event.



c) Building A169

Fig. 10 (*continued*)

Table 5
Decision rules generated by the classification trees for the three buildings.

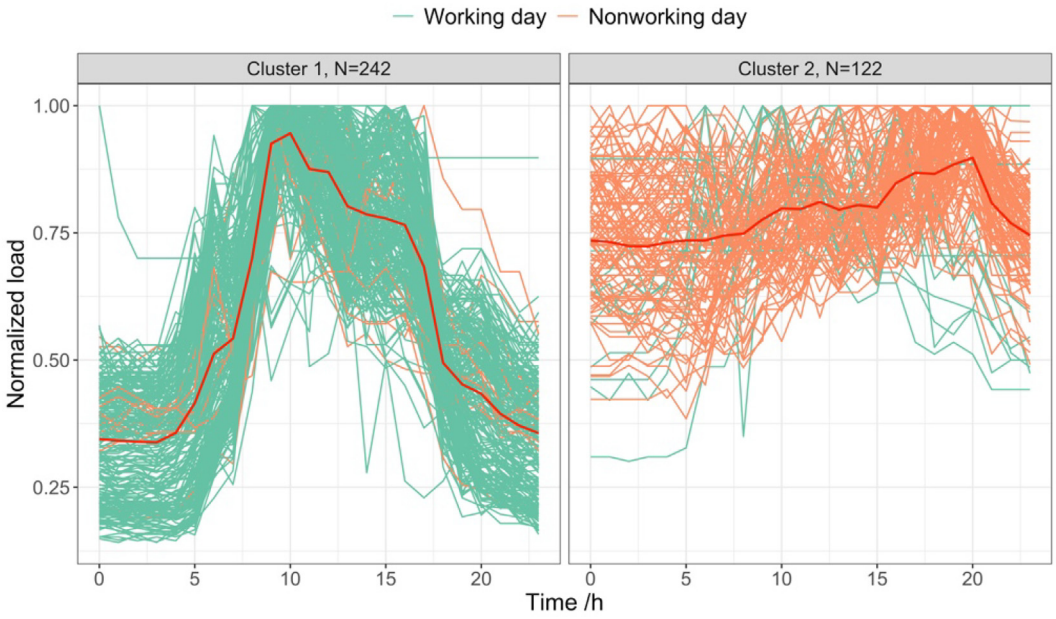
Building name	Decision rule	Boundary condition	Predicted pattern	Occurrence probability of each pattern
A033	Rule 1	If holiday = true, and mean_temp ≥ 24	4	(13%, 3%, 3%, **77%**, 5%)
	Rule 2	If holiday = true, and mean_temp < 24	1	(**80%**, 2%, 0%, 12%, 6%)
	Rule 3	If holiday = false, and mean_temp ≥ 26	3	(0%, 0%, **93%**, 1%, 6%)
	Rule 4	If holiday = false, and 15 ≤ mean_temp < 26	2	(0%, **77%**, 2%, 0%, 21%)
	Rule 5	If holiday = false, and mean_temp < 15	5	(0%, 10%, 10%, 0%, **81%**)
A155	Rule 1	If holiday = true, and mean_temp ≥ 25	1	(**76%**, 3%, 12%, 9%)
	Rule 2	If holiday = true, and mean_temp < 17	1	(**63%**, 4%, 8%, 25%)
	Rule 3	If holiday = true, and 17 ≤ mean_temp < 25	4	(29%, 7%, 4%, **60%**)
	Rule 4	If holiday = false, and mean_temp ≥ 23	2	(0%, **99%**, 1%, 0%)
	Rule 5	If holiday = false, and mean_temp < 23	3	(0%, 10%, **90%**, 0%)
A169	Rule 1	If holiday = true, and mean_temp ≥ 27	3	(3%, 7%, **90%**, 0%)
	Rule 2	If holiday = true, and mean_temp < 27	2	(0%, **77%**, 15%, 8%)
	Rule 3	If holiday = false, and mean_temp ≥ 25	1	(**100%**, 0%, 0%, 0%)
	Rule 4	If holiday = false, and mean_temp < 13	1	(**90%**, 0%, 0%, 10%)
	Rule 5	If holiday = false, and 13 ≤ mean_temp < 25	4	(3%, 0%, 0%, **97%**)

To quickly quantify the difference between the expected pattern and a potential anomalous profile, density heat maps of the five defined features of electricity load profiles are presented in Fig. 14, namely, mean values of normalized electricity load during four time periods and the peak-to-valley ratios of electricity load profiles. A hotter color represents a higher concentration of a feature, i.e., the higher occurrence frequency. Conversely, a cooler color denotes a lower occurrence frequency of a feature. The features of two potential anomalous profiles are also shown in the figure (pattern 1 and pattern 2). Specifically, features 1 and 2 of profile 2 are significantly higher than the frequent occurrence zone, which indicates that some electrical systems in the building were still operating from 12:00 am to 10:00 am, when the building should be unoccupied or partly occupied. Building managers are advised to determine whether the abnormalities happened due to building operational changes (e.g., individuals working overtime), equipment malfunction or data errors to take appropriate measures for better management.

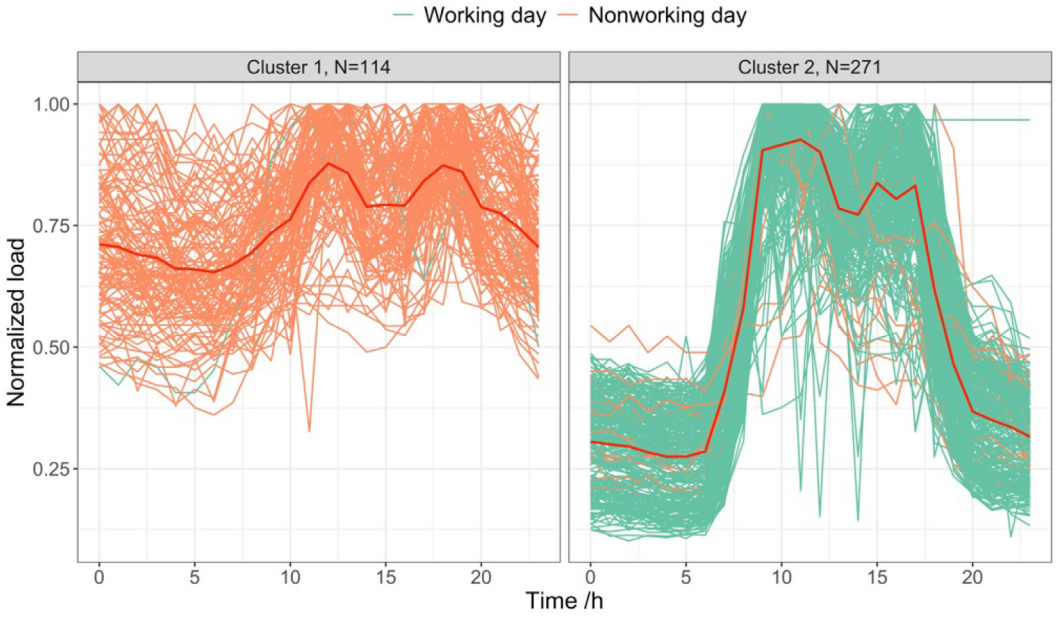
It is noted that the proposed framework is more appropriate to be an alert tool to detect anomalous trends in electricity load profiles based on the whole-building level at an early stage, while the specific locations where anomalies occur need to be examined with a further on-site diagnosis.

5.2. Conception of the framework

The proposed framework integrated unsupervised and supervised learning techniques. The combination of these two



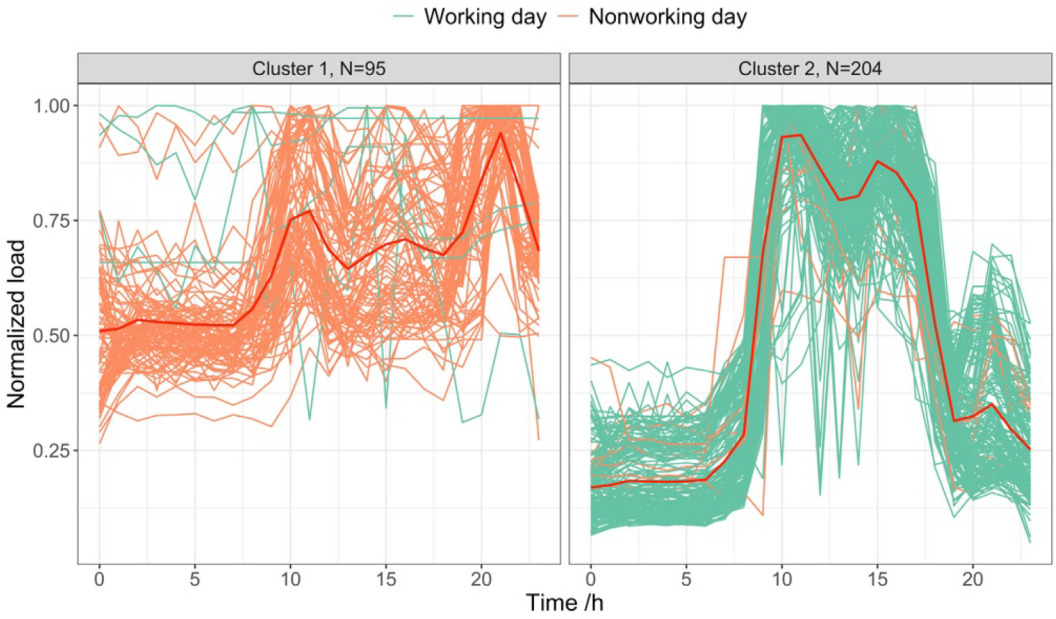
a) Building A033



b) Building A155

Fig. 11. Clustering results of the k-means clustering.

approaches provides a meaningful way to develop an interpretable framework for knowledge discovery from massive amounts of building electricity consumption data. It takes advantage of the strengths of unsupervised and supervised learning approaches and fills the gap between users and unsupervised learning results. Clustering is an unsupervised learning technique that focuses on discovering the typical electricity usage patterns and creates class labels for classification models; however, it does not provide explanations of the insightful knowledge behind the typical patterns. Thus, the clustering result requires experts in the field of building performance analysis to interpret the data partition. On the other hand, CART is a supervised learning technique and is used to build a classification model to quantitatively describe the relation between the known class labels and given predictors, which increases the interpretability of the resulting typical patterns. The proposed framework reduces the dependency on expertise domain for building energy analysis; thus, the time and labor costs can also be reduced.



c) Building A169

Fig. 11 (*continued*)

5.3. Limitations and future work

From a research perspective, this study has contributed to the development of a general framework to identify TELPs for individual buildings, producing deeper insights into the links between influencing factors related to electricity consumption and electricity usage patterns. In addition, a potential application was also provided to demonstrate an implication for early anomaly detection. However, this study is limited by the methods and the size of dataset.

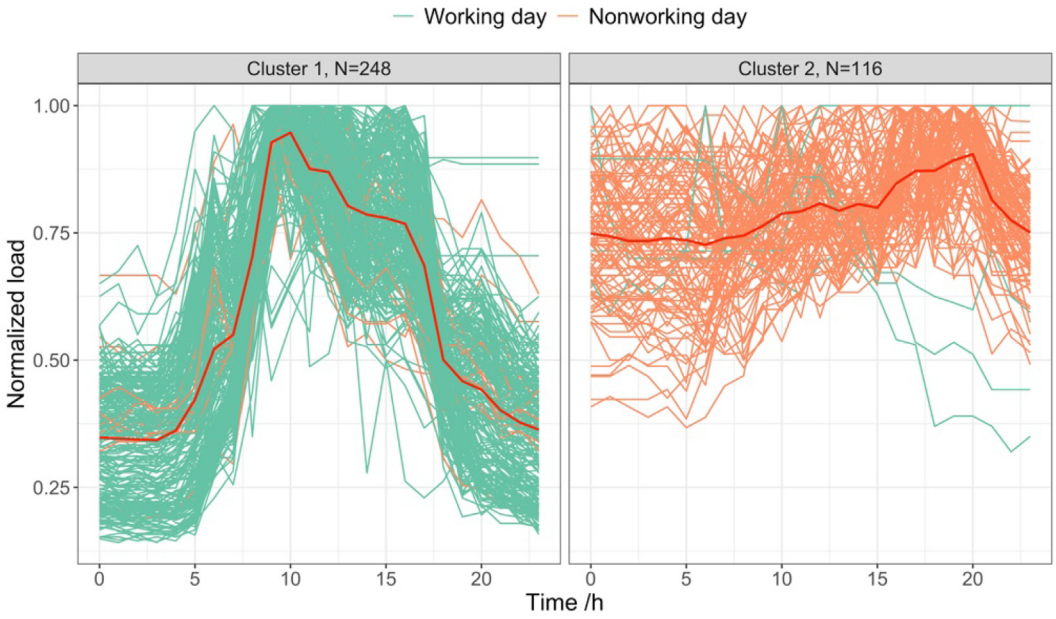
First, for the defined features, we partitioned time windows according to the working schedule of office buildings since the office buildings that we used for case studies have relatively regular daily routines. For a highly diverse dataset, the feature definition method in this study may not be suitable. Future work intends to propose a data-adaptive method to define the features of DELPs for diverse buildings to enhance flexibility. Another methodological limitation is that the framework did not consider the dynamic influencing factors related to occupants' behaviors, such as the daily occupants' presence or overtime working, to test whether the accuracy and explanatory power of classification model can be improved. Accordingly, more insightful knowledge could be gained from the obtained decision rules.

Regarding the limitations of the dataset, we tested and validated our framework based on only a small nonpublic dataset from the ECMP in Chongqing instead of a large dataset. We used this data source for the case study application mainly due to the two reasons. First, unlike developed countries, where there are available open datasets of time-series electricity consumption data for residential or non-residential buildings [16,44], while in China, reliable open and large datasets of building energy consumption data are still absent. Additionally, the privacy and security concerns are always the major barriers to collect smart meter data from individual buildings because building users are unwilling or unable to share their energy use data [45]. Therefore, the proposed framework may not be generalizable or sufficiently robust when it is applied to a larger dataset. Nonetheless, the proposed framework is systematic, and the procedure could be still applied to more individual buildings with available smart meters to extract TELPs as well as analyze the insights behind the patterns. However, future work is required to explore the applicability and generalizability of the proposed framework by using a larger dataset.

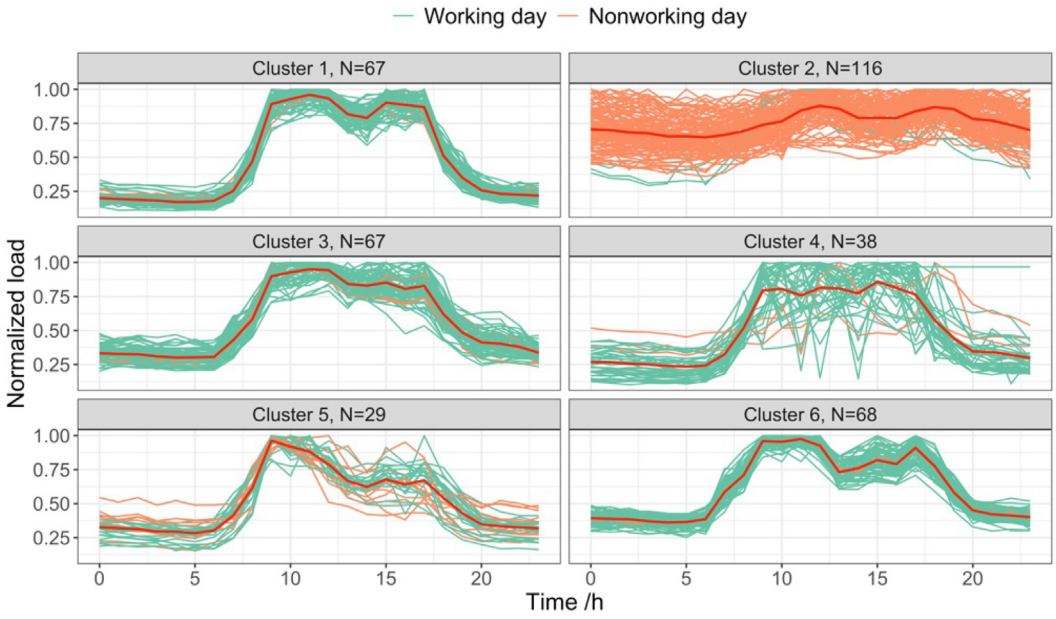
Furthermore, the analysis in this study considered only the whole-building hourly electricity consumption of individual buildings, while the electricity consumption data of electrical subsystems such as lighting, plug-in or heating, ventilation and air conditioning (HVAC) systems were not analyzed, which makes the fault detection difficult to discover the specific positions. Future work will explore the application of this framework in early anomaly detection for electricity consumption data of subsystems.

6. Conclusions

In this study, a general framework based on various DM techniques was proposed to extract TELPs and discover insightful information hidden in the patterns for individual buildings. There are three main phases for the framework: data preparation, identification of TELPs and knowledge discovery of TELPs. A new clustering method based on two-step clustering analysis was proposed to identify the TELPs at the individual building level. In this method, the dimensions of the raw daily electricity consumption data were first reduced by a feature definition method. The first clustering step aims at detecting outliers of DELPs by using the DBSCAN clustering technique. The second clustering step aims at grouping similar daily profiles to identify TELPs by means of k-means clustering. The effectiveness was demonstrated based on a comparison of the proposed clustering method with two single-step clustering techniques. Then, classification trees are developed by CART algorithm for knowledge discovery in terms of the associations between TELPs and dynamic influencing factors related to electricity consumption. Accordingly, the output of classification trees is a set of interpretable decision rules in the format of "if-then" rules. The framework provides a general and systematic procedure based



a) Building A033



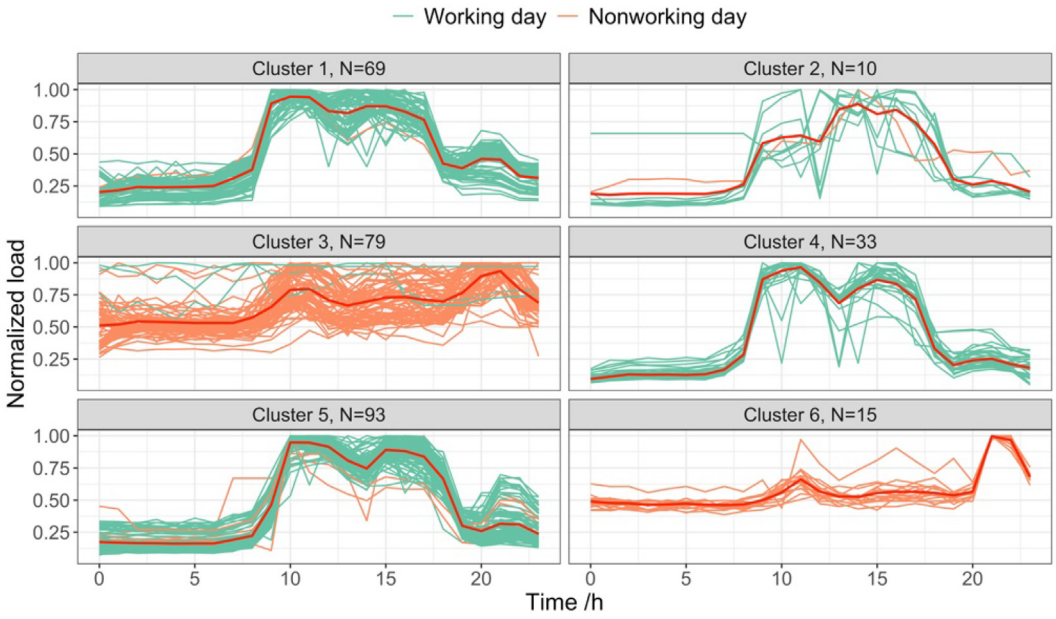
b) Building A155

Fig. 12. Clustering results of the GMM clustering.

on DM approaches for analyzing electricity usage patterns and knowledge discovery in time-series electricity consumption data.

The framework was implemented in analyzing time-series electricity consumption data from ECMP for three office buildings in Chongqing, China. It is found that the daily average outdoor air temperature and day type (working or nonworking day) were the main factors that distinguished TELPs according to the developed classification trees. Thanks to the classification trees, a set of decision rules can be generated to estimate the electricity usage pattern that are most likely to occur under a given boundary condition. Furthermore, a density heat map was presented to visualize the expected/frequent and unexpected/infrequent electricity usage patterns, which allows building managers to quickly recognize anomalous electricity load profiles.

The proposed framework in this paper provides a general solution to identify representative electricity usage patterns and discover deeper insightful knowledge behind the patterns, which increases the interpretability of clustering results and application value of discovered knowledge. This framework can also be an alert tool to detect anomalous electricity load profiles based on the whole-building level at an early stage. However, there are some limitations regarding to the dataset and methodology. The future works are listed as follows: 1) a more data-adaptive method to automatically extract features of DELPs is expected to be proposed



c) Building A169

Fig. 12 (*continued*)

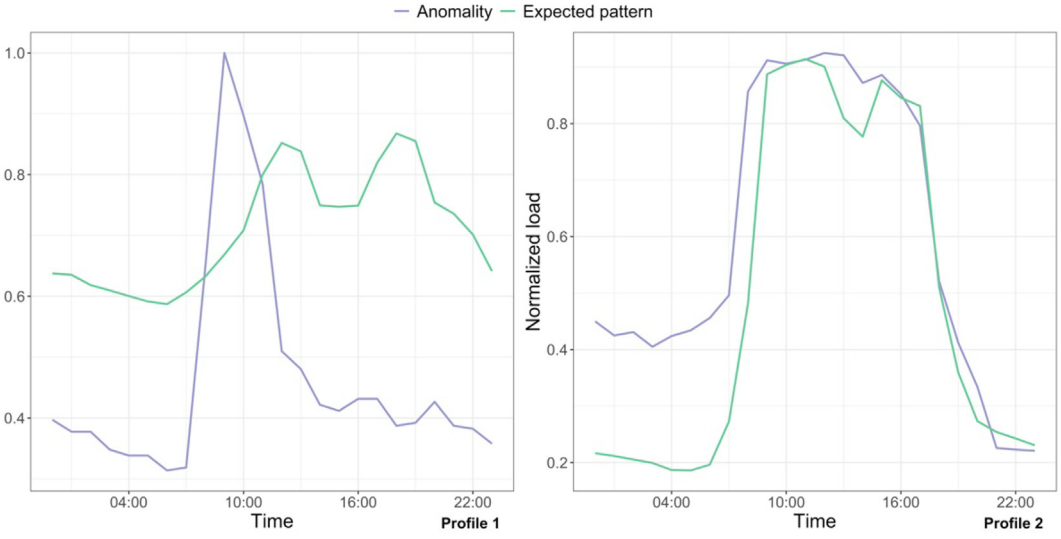


Fig. 13. Two possible anomalous profiles and the expected electricity load patterns.

for a highly diverse dataset in order to make the framework more flexible; 2) additional potential influencing factors, such as daily occupancy numbers and overtime working phenomenon, should be considered for classification trees; 3) the generalizability and applicability of the proposed framework should be further evaluated by a larger dataset; and 4) the framework should be evaluated by the electricity consumption data of building subsystems to explore more potential applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Dr. Samir Touzani at the Lawrence Berkeley National Laboratory for his many useful suggestions and inspiration for this study. The authors are thankful to the Chongqing Housing and Urban-Rural Committee for providing data supports of the energy consumption monitoring platform. This work is also supported by the project "Standard System for Post-assessment of Green Building Performance" (the National Key Projects in 13th Five-Year, project no. 2016TFC0700105), the National Natural Science Foundation of China (Grant number 51978095), and 111 Project (Grant number: B13041). Feng Xiao contributed to the manuscript preparation and revision, but not the study design; his involvement was supported by the National Natural Science Foundation of China (Grant number 71861167001).

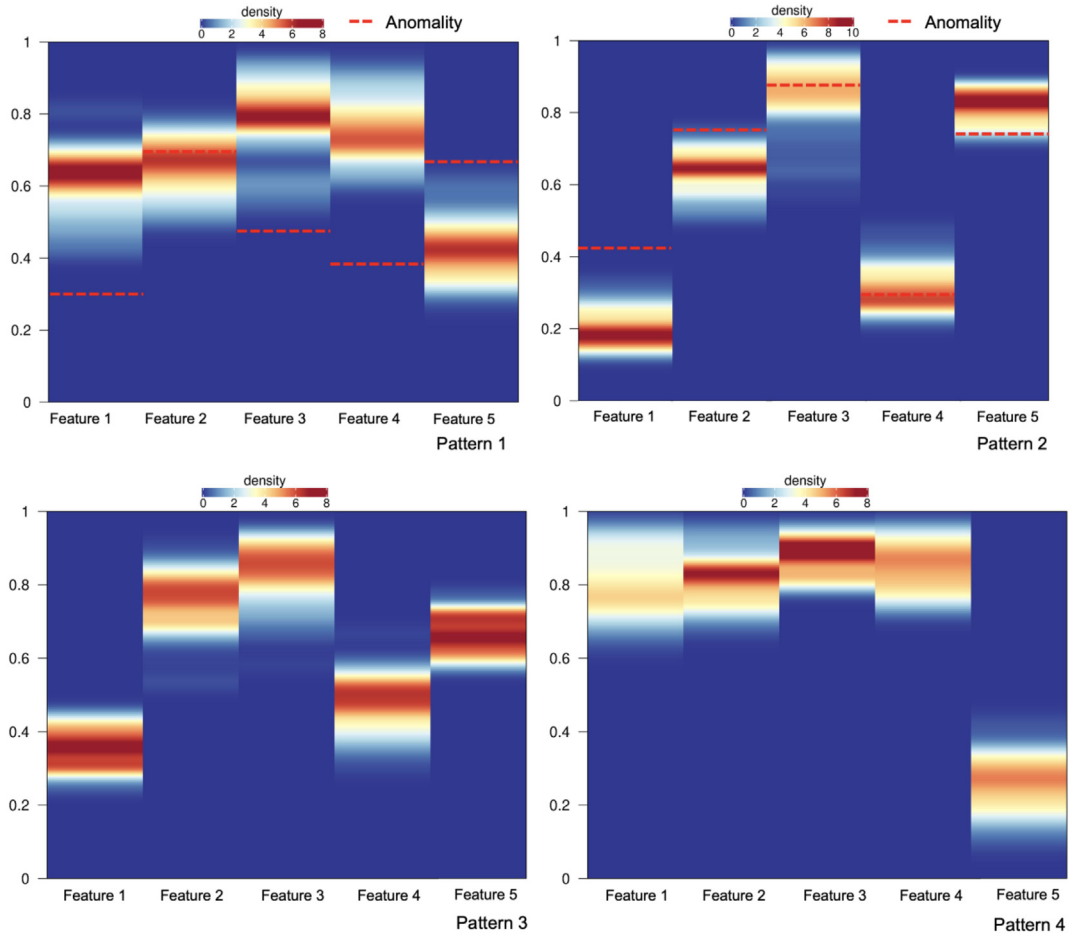


Fig. 14. Density heat maps of five features extracted from electricity load profiles for building A155.

Appendix A. Comparison of clustering validation indexes (CVIs)

In this section, five widely used CVIs are selected to determine the optimal CVI when the proposed clustering strategy is applied based on the three buildings: C-index [48], Calinski-Harabasz (CH) index [49], Dunn index [50], Davies- Bouldin (DB) index [51] and Silhouette index [52]. A higher value of CH index, Dunn index and Silhouette index indicates a better clustering result, while a lower value of C-index and DB index means a better clustering result. The comparison of clustering results for the three buildings is shown in Tables A1–A3. The selected value of each CVI corresponding to the optimal cluster number is highlighted in bold. All of the CVIs except the Dunn index support only 2 clusters in most cases, indicating the other four CVIs may not precisely present the quality of the clustering result when using the proposed clustering method and the cases in this study.

Appendix B. Comparison of clustering methods

In this study, k-means and GMM clustering algorithm were selected to compare the performance of the proposed clustering method. The Dunn index was employed to evaluate clustering results for k-means and GMM clustering techniques. A higher Dunn index indicates a better clustering result.

Before clustering, the raw time series electricity consumption data were dealt with based on the same data pre-processing. Three buildings (A033, A155 and A169) in case study were all used for comparison. The optimal number of clusters for k-means and GMM clustering technique is selected ranging from 2 to 8. Fig. B1 and B2 show the clustering results of the Dunn index for k-means and GMM clustering techniques, respectively. It can be seen that k-means clustering suggests only 2 cluster numbers for the three buildings, while GMM clustering provides 2 clusters for A033 and 6 clusters for A155 and A169.

Table A1
Comparison of CVIs when using the proposed clustering strategy for A033.

Number of clusters	C-index	CH index	Dunn index	DB index	Silhouette
2	0.0495	715.0163	0.0828	**0.5840**	**0.5871**
3	0.0389	**821.2868**	0.0570	0.6853	0.5169
4	0.0320	701.5296	0.0656	0.8480	0.4174
5	0.0308	692.0458	**0.1022**	0.9661	0.3756
6	0.0293	624.2869	0.0932	1.0202	0.3432
7	**0.0286**	573.0234	0.0838	1.0609	0.3157
8	0.0311	539.3377	0.0875	1.1259	0.2988

Table A2
Comparison of CVIs when using the proposed clustering strategy for A155.

Number of clusters	C-index	CH index	Dunn index	DB index	Silhouette index
2	**0.0258**	**855.3203**	0.0952	**0.5667**	**0.60823**
3	0.0495	808.6681	0.0653	0.7140	0.4902
4	0.0282	799.5241	**0.0991**	0.8088	0.4466
5	0.0389	711.2439	0.0769	1.0014	0.3703
6	0.0308	659.0876	0.0872	1.0989	0.3400
7	0.0323	607.9064	0.0766	1.1518	0.3206
8	0.0277	576.9888	0.0766	1.0819	0.3308

Table A3
Comparison of CVIs when using the proposed clustering strategy for A169.

Number of clusters	C-index	CH index	Dunn index	DB index	Silhouette index
2	**0.0105**	703.6249	0.1065	**0.5440**	**0.6349**
3	0.0285	**727.1074**	0.0779	0.6683	0.5352
4	0.0161	699.1215	**0.1234**	0.7925	0.4796
5	0.0274	649.1215	0.0781	0.9211	0.4195
6	0.0284	573.9431	0.0918	1.0670	0.3631
7	0.0249	525.9075	0.0937	1.1926	0.3195
8	0.0250	485.4931	0.1012	1.1571	0.3095

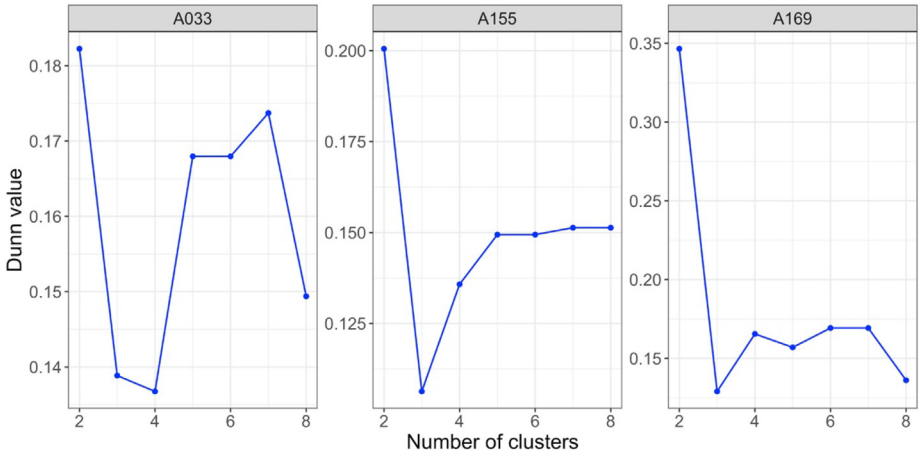


Fig. B1. Optimal number of clusters for k-means clustering based on raw DEPLs.

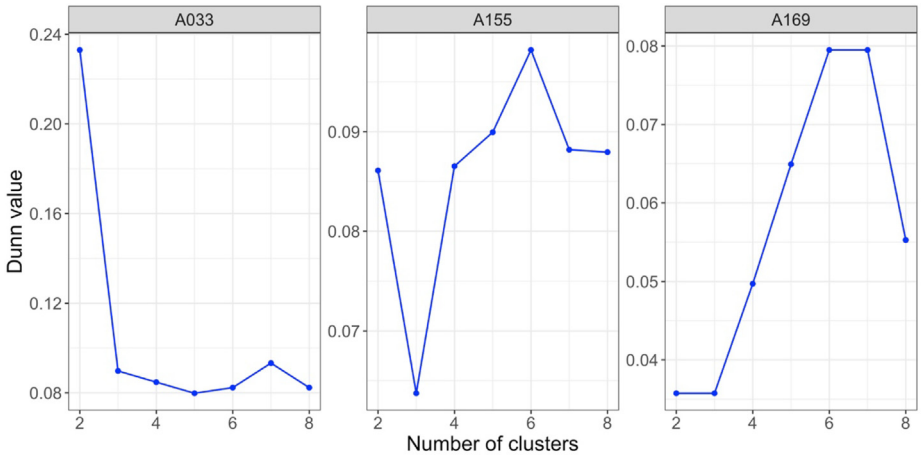


Fig. B2. Optimal number of clusters for GMM clustering based on raw DEPLs.

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.enbuild.2020.110601.

References

[1] International Energy Agency (IEA), Energy Technology Perspectives Scenarios and Strategies to 2050, IEA, Paris, 2012, p. 2012.
[2] C. Fan, F. Xiao, Z. Li, J. Wang, Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: a review, Energy Build. 159 (2018) 296–308, https://doi.org/10.1016/j.enbuild.2017.11.008.
[3] C. Miller, Z. Nagy, A. Schlueter, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, Renew. Sustain. Energy Rev. 81 (2018) 1365–1377, https://doi.org/10.1016/j.rser.2017.05.124.
[4] EIA. How many smart meters are installed in the United States, and who has them?; 2020 https://www.eia.gov/tools/faqs/faq.php?id=108&t=3 (Accessed on 11 July 2020).
[5] N. Wei, W. Yong, S. Yan, D. Zhongcheng, Government management and implementation of national real-time energy monitoring system for China large-scale public building, Energy Policy. 37 (2009) 2087–2091, https://doi.org/10.1016/j.enpol.2008.12.032.
[6] J. Hou, Y. Liu, Y. Wu, N. Zhou, W. Feng, Comparative study of commercial building energy-efficiency retrofit policies in four pilot cities in China, Energy Policy 88 (2016) 204–215, https://doi.org/10.1016/j.enpol.2015.10.016.
[7] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, J. Li, A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis, Energy Built Environ. 1 (2020) 149–164, https://doi.org/10.1016/j.enbenv.2019.11.003.
[8] K. Li, Z. Ma, D. Robinson, J. Ma, Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering, Appl. Energy 231 (2018) 331–342, https://doi.org/10.1016/j.apenergy.2018.09.050.
[9] A. Rajabi, M. Eskandari, M.J. Ghadi, L. Li, J. Zhang, P. Siano, A comparative study of clustering techniques for electrical load pattern segmentation, Renew. Sustain. Energy Rev. 120 (2020), https://doi.org/10.1016/j.rser.2019.109628.
[10] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renew. Sustain. Energy Rev. 82 (2018) 1027–1047, https://doi.org/10.1016/j.rser.2017.09.108.
[11] S. Aghabozorgi, A. Seyed Shirkhorshidi, T. Ying Wah, Time-series clustering – A decade review, Inf. Syst. 53 (2015) 16–38, https://doi.org/10.1016/j.is.2015.04.007.
[12] 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, https://doi.org/10.1016/j.energy.2017.05.191.
[13] K. Li, R.J. Yang, D. Robinson, J. Ma, Z. Ma, An agglomerative hierarchical clustering-based strategy using Shared Nearest Neighbours and multiple dissimilarity measures to identify typical daily electricity usage profiles of university library buildings, Energy 174 (2019) 735–748, https://doi.org/10.1016/j.energy.2019.03.003.
[14] Y. Wang, Q. Chen, C. Kang, Q. Xia, Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications, (2017). https://doi.org/10.1109/TSG.2016.2548565.
[15] Y. Lu, Z. Tian, P. Peng, J. Niu, W. Li, H. Zhang, GMM clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system, Energy Build. 190 (2019) 49–60, https://doi.org/10.1016/j.enbuild.2019.02.014.
[16] J.Y. Park, X. Yang, C. Miller, P. Arjunan, Z. Nagy, Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset, Appl. Energy 236 (2019) 1280–1295, https://doi.org/10.1016/j.apenergy.2018.12.025.
[17] L. Wen, K. Zhou, S. Yang, A shape-based clustering method for pattern recognition of residential electricity consumption, J. Clean. Prod. 212 (2019) 475–488, https://doi.org/10.1016/j.jclepro.2018.12.067.
[18] C.M.R. Do Carmo, T.H. Christensen, Cluster analysis of residential heat load profiles and the role of technical and household characteristics, Energy Build. 125 (2016) 171–180, https://doi.org/10.1016/j.enbuild.2016.04.079.
[19] M. Verleysen, D. François, The curse of dimensionality in data mining and time series pre- diction, In: Proceedings of International Work-Conference on Artificial Neural Networks (IWANN 2005), Heidelberg, Berlin, Springer, 2005. pp. 758–770. 2005. https://doi.org/10.1007/11494669_93.
[20] S. Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Trans. Acoust Speech Singal Process. 1 (1978) 159–165.
[21] A. Sard, Comparing Time-Series Clustering Algorithms in R Using the dtwclust Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package, (2019).
[22] X. Luo, T. Hong, Y. Chen, M.A. Piette, Electric load shape benchmarking for small- and medium-sized commercial buildings, Appl. Energy. 204 (2017) 715–725, https://doi.org/10.1016/j.apenergy.2017.07.108.
[23] T. Räsänen, M. Kolehmainen, Feature-Based Clustering for Electricity Use Time Feature-Based Clustering for Electricity Use, In: Proceedings of international conference on adaptive and natural computing algorithms (LNCS 5495). Berlin, Germany: Springer-Verlag; 2009. pp. 401–412. 2009. https://doi.org/10.1007/978-3-642-04921-7.
[24] S. Haben, C. Singleton, P. Grindrod, Analysis and clustering of residential customers energy behavioral demand using smart meter data, IEEE Trans. Smart Grid. 1 (2016) 136–144, https://doi.org/10.1109/TSG.2015.2409786.
[25] C. Fan, F. Xiao, C. Yan, A framework for knowledge discovery in massive building automation data and its application in building diagnostics, Autom. Constr. 50 (2015) 81–90, https://doi.org/10.1016/j.autcon.2014.12.006.
[26] F. Wang, K. Li, N. Dui, Z. Mi, B. Hodge, M. Sha, J.P.S. Catalão, Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns, 171 (2018) 839–854. https://doi.org/10.1016/j.enconman.2018.06.017.
[27] A. Capozzoli, M.S. Piscitelli, S. Brandi, D. Grassi, G. Chicco, Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings, Energy 157 (2018) 336–352, https://doi.org/10.1016/j.energy.2018.05.127.
[28] F. McLoughlin, A. Duffy, M. Conlon, A clustering approach to domestic electricity load profile characterisation using smart metering data, Appl. Energy 141 (2015) 190–199, https://doi.org/10.1016/j.apenergy.2014.12.039.
[29] Z. Ma, J. Song, J. Zhang, A real-time detection method of abnormal building energy consumption data coupled POD-LSE and FCD, Procedia Eng. 205 (2017) 1657–1664, https://doi.org/10.1016/j.proeng.2017.10.334.
[30] W. Zhao, S. Du, Spectral-spatial feature extraction for hyperspectral image classification: a dimension reduction and deep learning approach, IEEE Trans. Geosci. Remote Sens. 54 (2016) 4544–4554, https://doi.org/10.1109/TGRS.2016.2543748.
[31] S. Sarkar, A.K. Ghosh, On perfect clustering of high dimension, low sample size data, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1, https://doi.org/10.1109/tpami.2019.2912599.
[32] M. Ester, H. Kriegel, X. Xu, D.- Miinchen, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, In: Proceedings of the 2nd ACM SIGKDD, Portland, Oregon; 1996. pp. 226–231.
[33] M. Hahsler M. Piekenbrock S. Arya D. Mount R, Package 'dbscan' 2020 (Accessed on 11 July 2020).
[34] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (2005) 645–678, https://doi.org/10.1109/TNN.2005.845141.
[35] I. Benítez, A. Quijano, J.L. Díez, I. Delgado, Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers, Int. J. Electr. Power Energy Syst. 55 (2014) 437–448, https://doi.org/10.1016/j.ijepes.2013.09.022.
[36] B. Lkhagva, Y. Suzuki, K. Kawagoe, New Time Series Data Representation ESAX for Financial Applications New Time Series Data Representation ESAX for Financial Applications, 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 2006, pp. x115–x115, https://doi.org/10.1109/ICDEW.2006.99.
[37] G. Brock, V. Pihur, S. Datta, S. Datta, ClValid: An R package for cluster validation, J. Stat. Softw. 25 (2008) 1–22. https://doi.org/10.18637/jss.v025.i04.
[38] R. Agrawal, Mining Association Rules between Sets of Items in Large Databases, In: Proc of the 1993 ACM-SIGMOD international conference on management of data. pp. 207–216.
[39] T.M. Therneau, E.J. Atkinson, An Introduction to Recursive Partitioning Using the RPART Routines, 2019 https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf (Accessed on 11July 2020).
[40] P. Rousseeuw, A. Struyf, M. Hubert, M. Studer, P. Roudier, Package 'cluster', 2020 https://cran.r-project.org/web/packages/cluster/cluster.pdf (Accessed on 11 July 2020).
[41] T. Therneau, B. Atkinson, B. Ripley, M.B. Ripley, R Package 'rpart', 2020 http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf (Accessed on 11 July 2020).
[42] X. Li, R. Yao, Q. Li, Y. Ding, B. Li, An object-oriented energy benchmark for the evaluation of the office building stock, Util. Policy 51 (2018) 1–11, https://doi.org/10.1016/j.jup.2018.01.008.
[43] L. Zhao, J.L. Zhang, R.B. Liang, Development of an energy monitoring system for large public buildings, Energy Build. 66 (2013) 41–48, https://doi.org/10.1016/j.enbuild.2013.07.007.
[44] C. Miller, F. Meggers, The building data genome project: an open, public data set from non-residential building electrical meters, Energy Procedia 122 (2017) 439–444, https://doi.org/10.1016/j.egypro.2017.07.400.
[45] Z. Wang, T. Hong, Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN), Energy Build. 224 (2020) 110299, https://doi.org/10.1016/j.enbuild.2020.110299.
[46] K. Sun, D. Yan, T. Hong, S. Guo, Stochastic modeling of overtime occupancy and its application in building energy simulation and calibration, Build. Environ. 79 (2014) 1–12, https://doi.org/10.1016/j.buildenv.2014.04.030.

[47] C. Biernacki, G. Celeux, G. Govaert, Assessing a mixture model for clustering with the integrated completed likelihood, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 719–725, https://doi.org/10.1109/34.865189.
[48] L. Hubert, J. Schultz, Quadratic assignment as a general data analysis strategy, Br. J. Math. Stat. Psychol. 29 (1976) 190–241, https://doi.org/10.1111/j.2044-8317.1976.tb00714.x.
[49] T. Caliński, J. Harabasz, A dendrite method foe cluster analysis, Commun. Stat. 3 (1974) 1–27, https://doi.org/10.1080/03610927408827101.
[50] J.C. Dunn, Well-separated clusters and optimal fuzzy partitions, J. Cybern. 4 (1974) 95–104, https://doi.org/10.1080/01969727408546059.
[51] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1979) 224–227, https://doi.org/10.1109/TPAMI.1979.4766909.
[52] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.