**Abstract:**

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.

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.

The framework integrates multiple data mining techniques and mainly consists of three phases: **data preparation**, **identification of TELPs** and **knowledge discovery** in the patterns.

**Load profiling:**

Refers to the process of grouping temporal subsequences of measured electricity data to identify typical electricity consumption patterns of a building

A new **clustering method** with a two-step clustering analysis is proposed to identify the

TELPs at the individual building level.(The first one is density based which aims to detect outliers of DELPs, The second aims to group similar DELPs by means of the k-means algorithm to extract TELPs).

The effectiveness of the proposed framework(two step clustering is a framework, starting with DBSCAN and K-means clustering) was demonstrated with two single step clustering techniques.

**Preprocessing:**

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.

**Methodology:**

The proposed framework consists of three phases.

1. **Data preparation phase:**

*There are mainly three tasks in terms of data preprocessing, data segmentation and data normalization.*

1. **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.*

1. *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)*
2. Finally, a potential application in anomaly detection is demonstrated.

## **DATA PREPARATION:**

The three main steps of Data preparation:

1. **Data Preprocessing:** The main task of this step is *detecting and removing missing values* and *outliers* in the raw time series dataset. The missing values are identified (Q3 + 1.5IQR) and labeled as a missing value. Then the missing values were filled using the moving average method of window size 3.
2. **Data Segmentation:** means reshaping the raw time series data (i.e., annual hourly electricity consumption data) into DELPs.
3. **Data Normalization:** 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 p\_norm(t) = p(t)/ P\_max(t)
4. **Feature definition:** Time series have high dimension which is a problem in calculating distance in clustering. For this reason, the daily profile is divided into four segments according to working-schedule of the buildings, 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. Then the men of each segment are calculated. 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.
5. **Removing outliers from DELPs:** 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. (DBSCAN is capable of recognizing outliers in the low density areas)
6. **Extraction of TELPs :** 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. **Dunn index** was selected to evaluate the clustering results based on a comparison with the other four CVIs, as presented in Appendix A.

1. **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. And for this Decision Tree(CART) was used because of its interpretability compared to other methods.

The target variable (i.e, TELPs) is categorical and classification trees can be used to determine the expected pattern based on given input predictors.

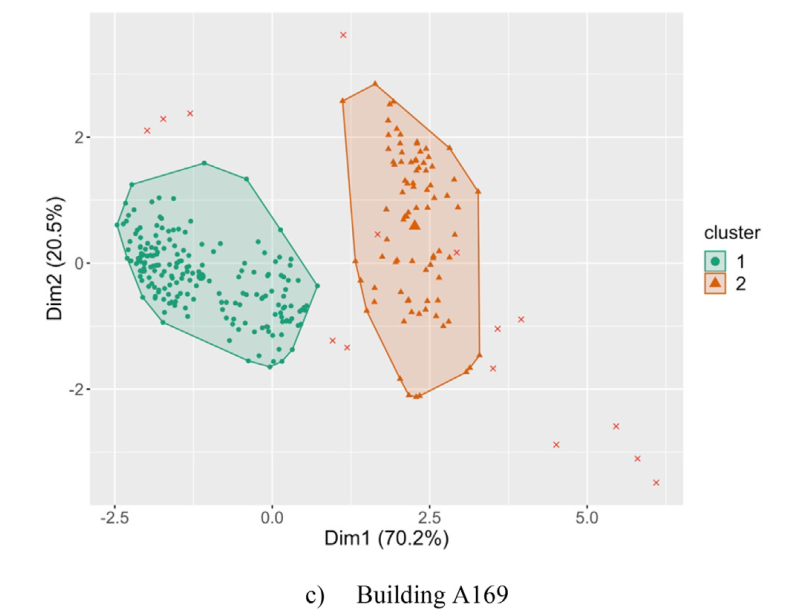
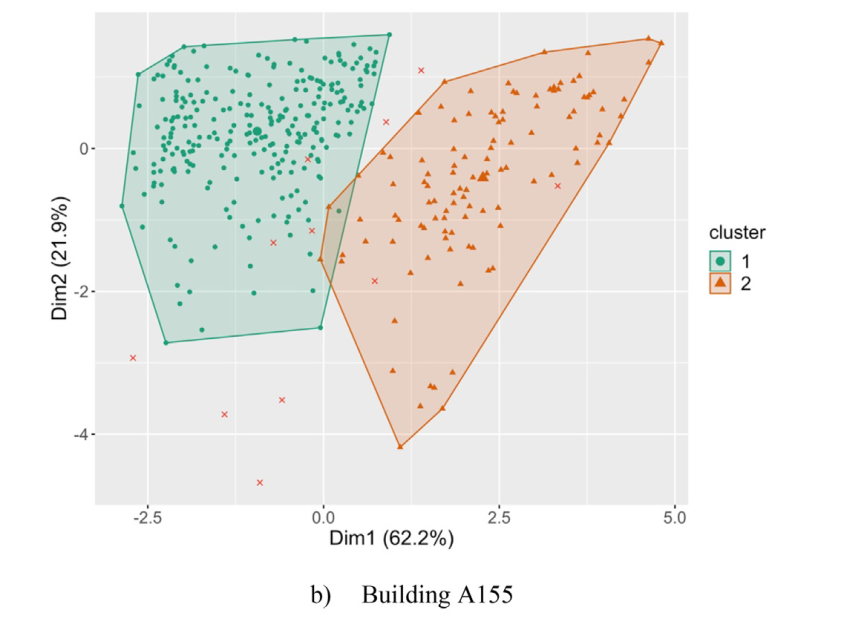
## **Case study applications**

For this case study 3 three non-residential buildings were selected. 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. Weather data of that area were also collected. (The following table shows that).

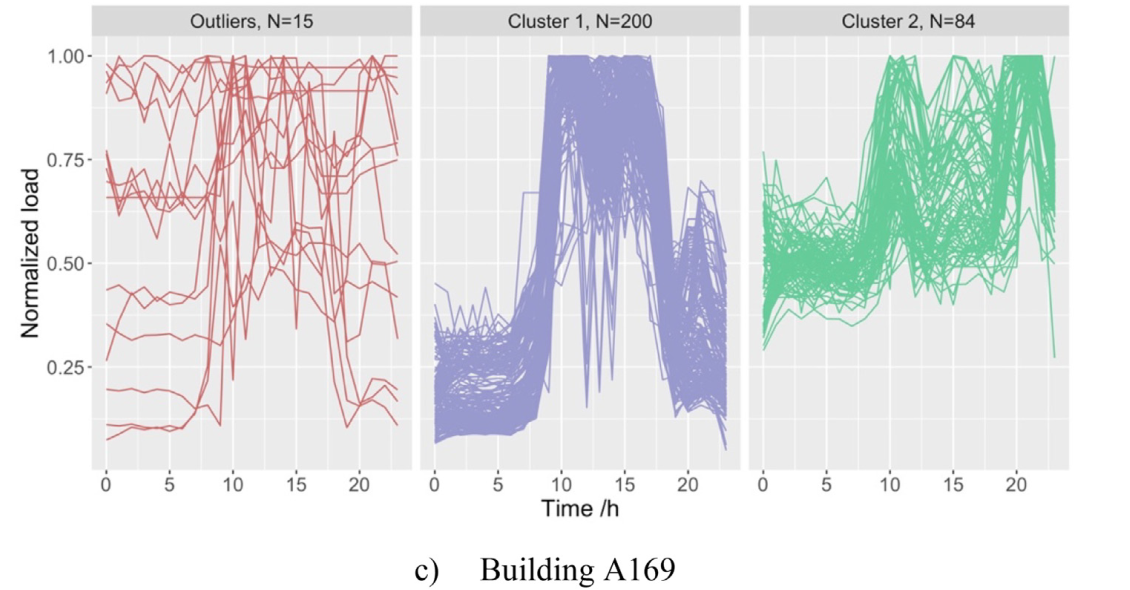
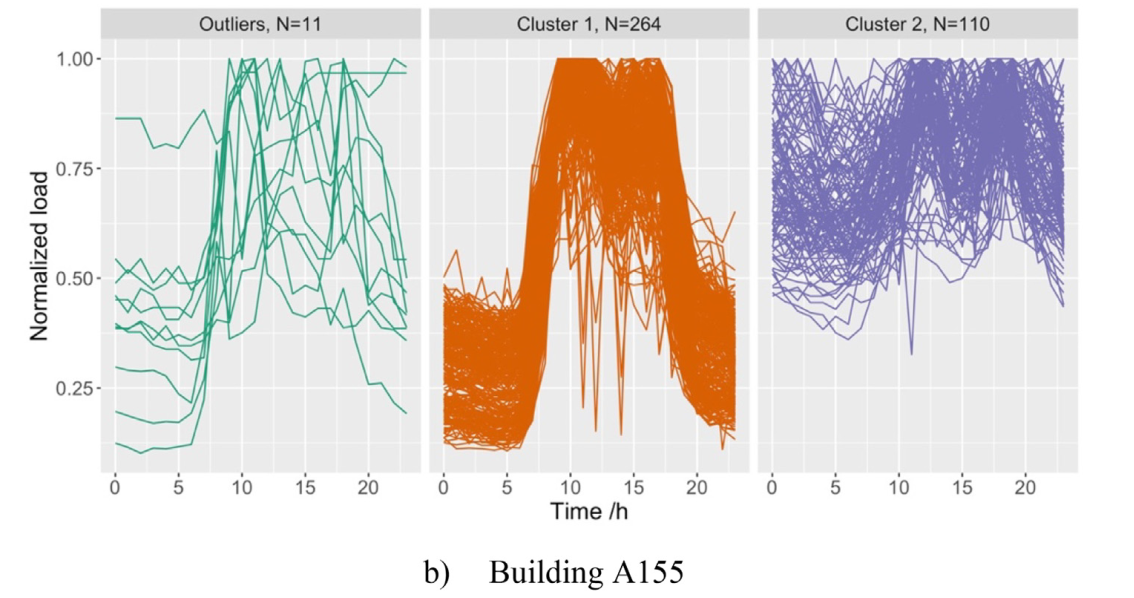
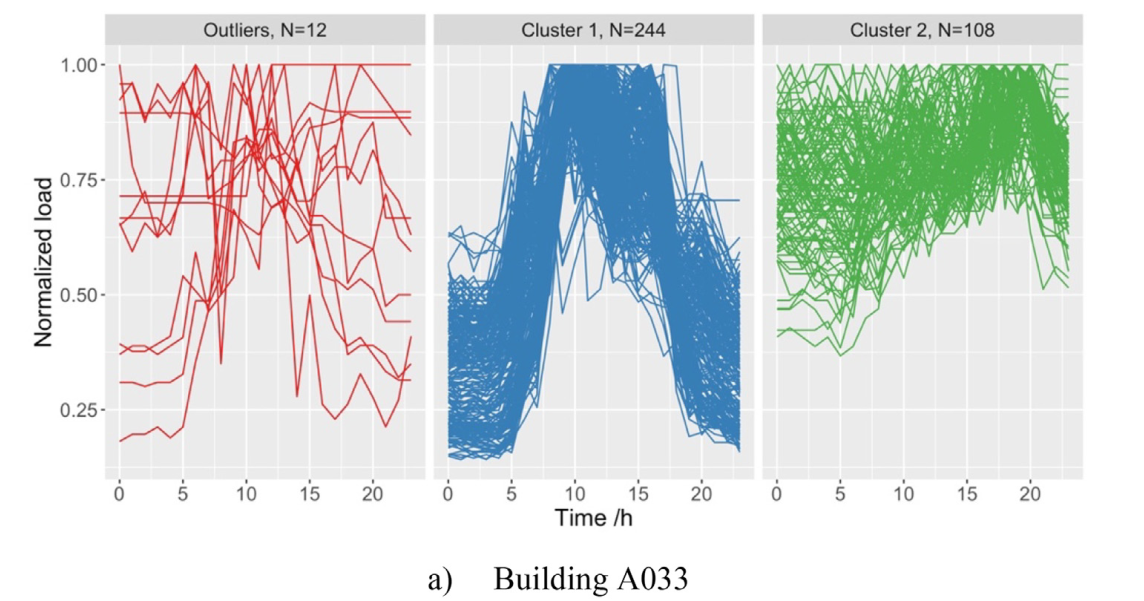


1. **OUTLIER DETECTION:**

DBSCAN algorithm was adopted to identify the outliers of DELPs based on the defined features of DELPs. Dim1 and Dim2 are the two PCA components. 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. The following are the two visualizations of the outliers in general clusters of DELPs for three buildings.



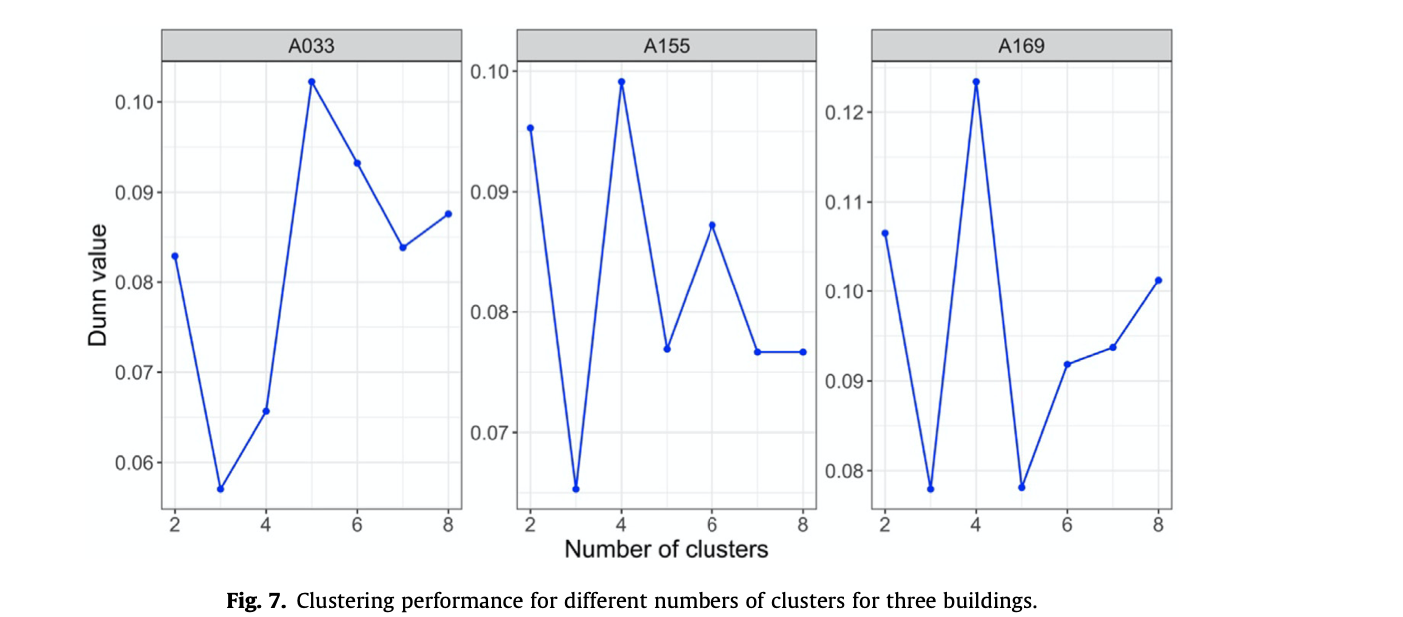
In Fig. 6(below), the outliers and general clusters of the daily profiles are displayed. The irregular curves are identified in the cluster called ‘‘outliers”. Note that there are only two electricity load patterns identified by DBSCAN algorithm in addition to the cluster of outliers for the three buildings, 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. The irregular curves are the outliers.



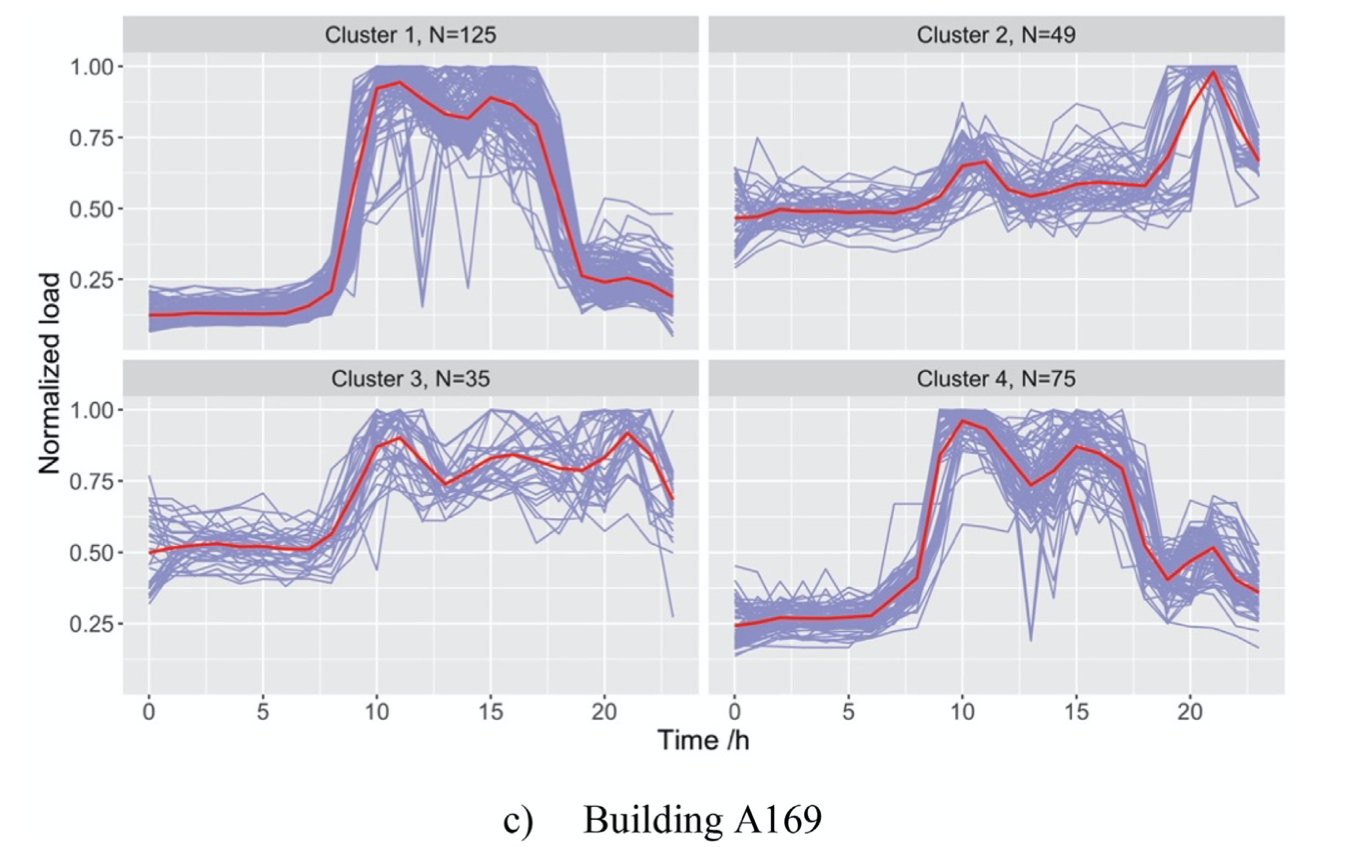
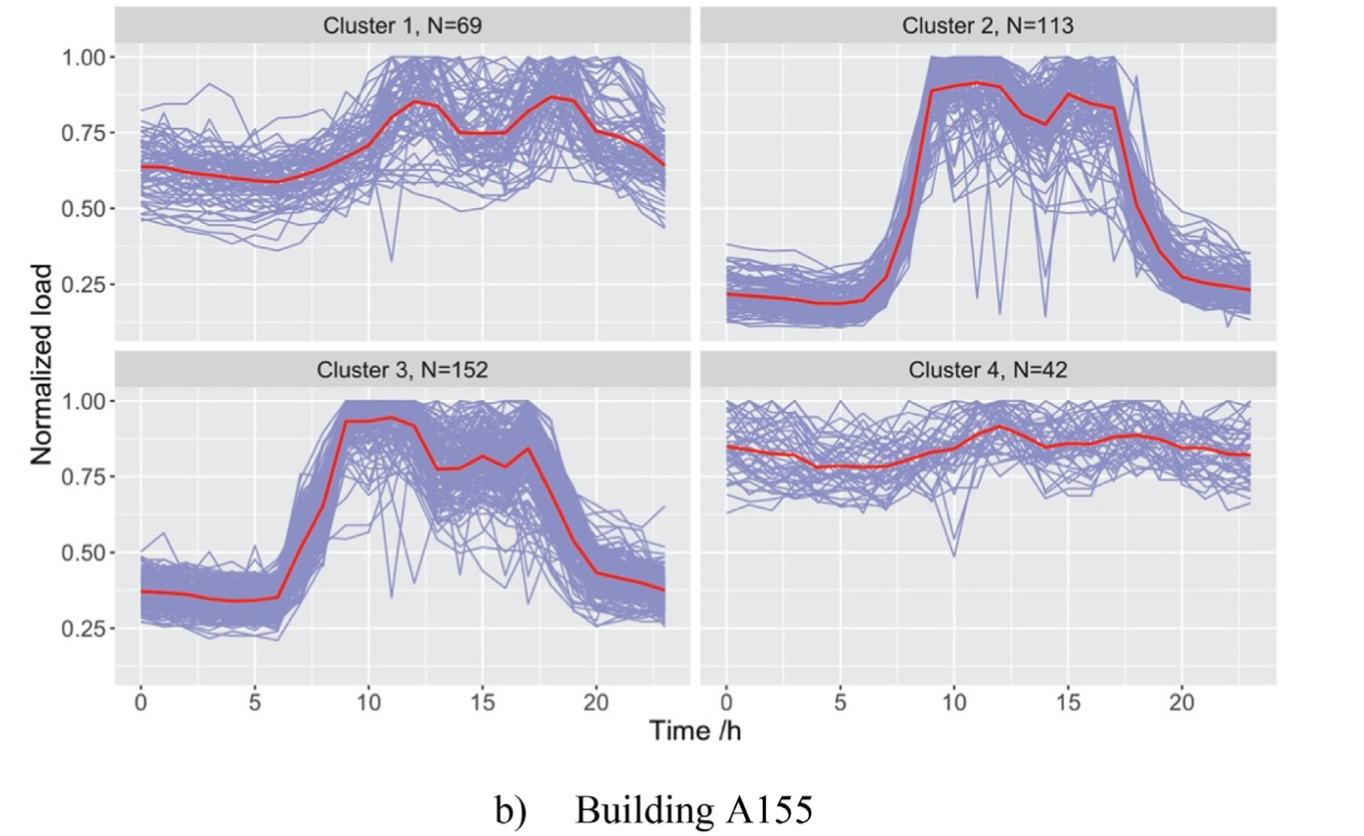
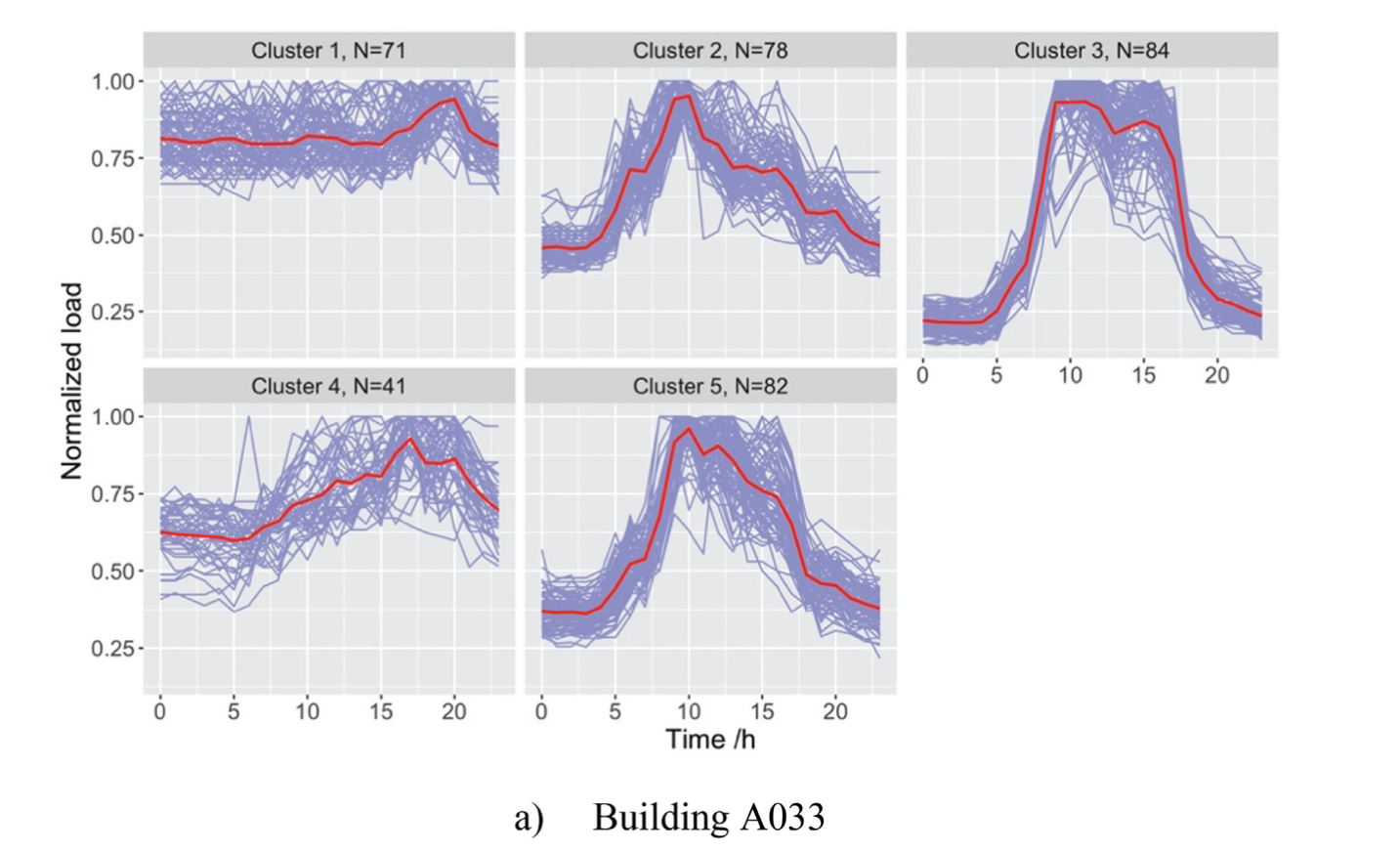
1. **Identification of typical electricity load patterns:**

After the outliers of DELPs were removed, TELPs for three buildings were identified through the k-means clustering method. The optimal number of clusters was selected to be from

2 to 8, and the results were evaluated by the **Dunn index(**Higher Dunn index represents better clustering results**)**.



The following graph shows the identified TELPs by K-means clustering.

****

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.

**Analysis of cluster**

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

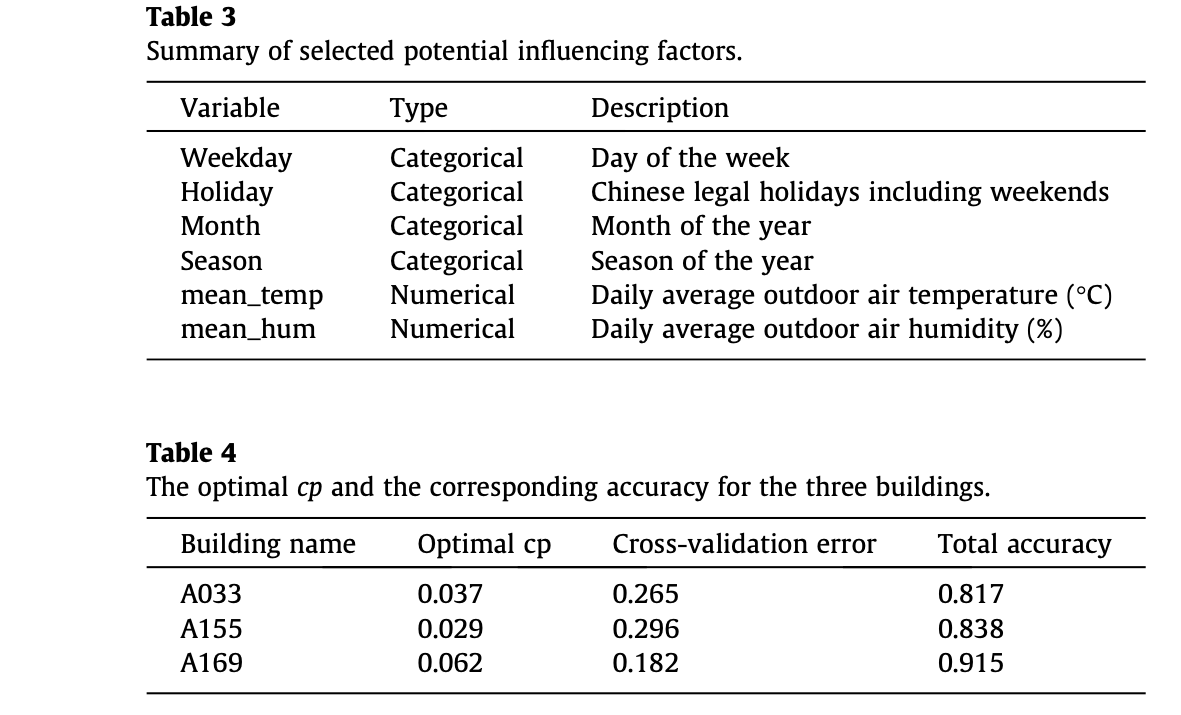
working is likely to occur in the evening in this cluster. For cluster 2, the peak electricity loads appear at 10:00 am and 21:00 pm. Different from the patterns of other three clusters, cluster 3 pattern remains stable during the daytime and evening except for two peaks at 10:00 am and 21:00 pm.

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.

The following graph shows the electricity usage intensity (EUI) reflecting electricity consump- tion levels of each cluster for three buildings, where the EUI was calculated by the ratio of hourly electricity consumption to total floor area.

1. **Knowledge Discovery by CART**

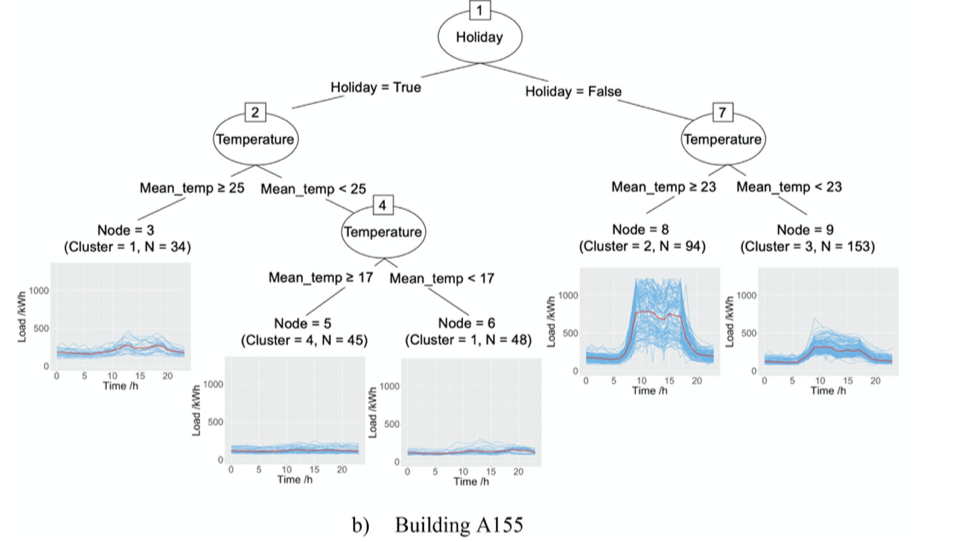
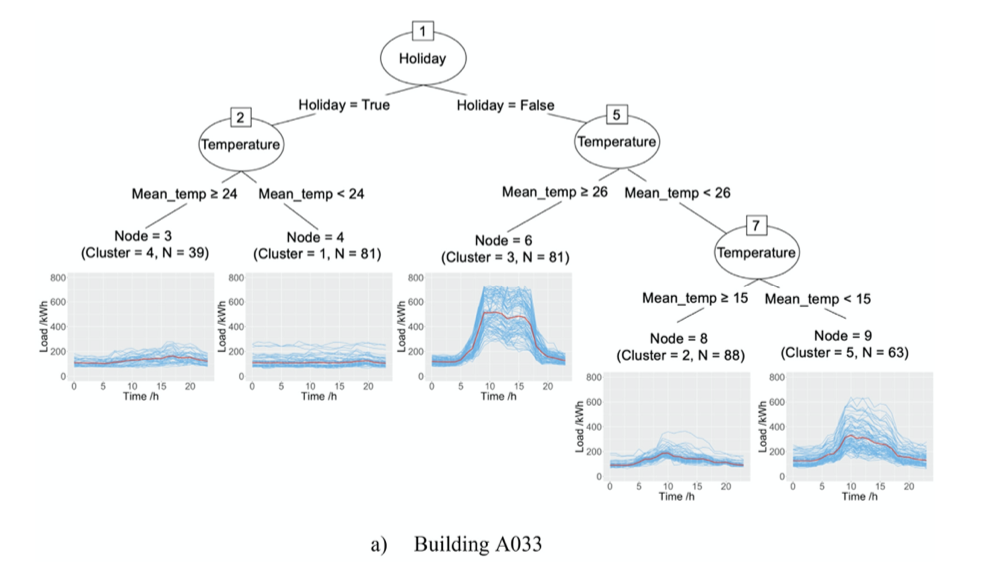
To further understand the reasons leading to different identified TELPs based on the proposed clustering analysis, i.e to improve the interpretability of the clustering results, a classification model was developed using CART algorithm to explore the underlying relations between the identified patterns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

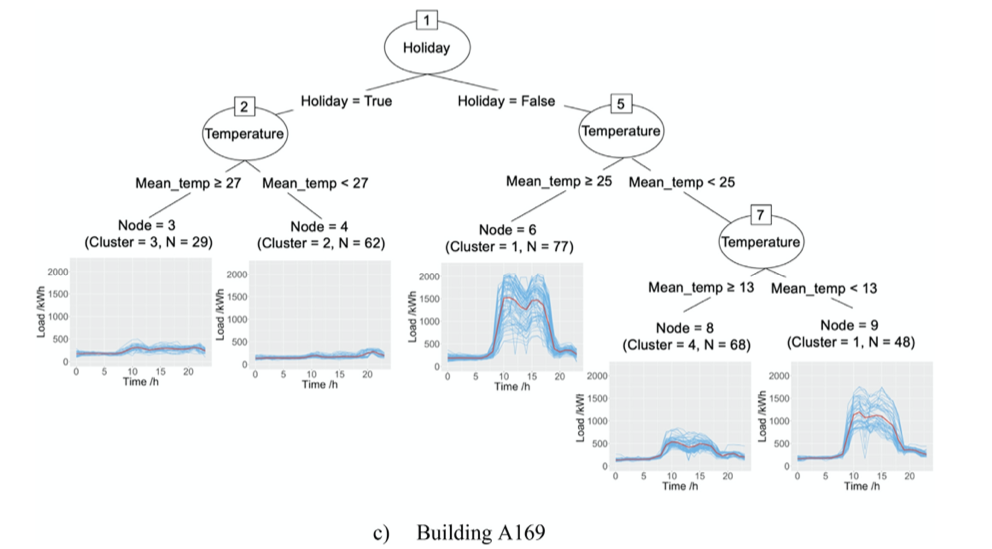


Note : Note that the original electricity load profiles without data normalization are provided in this figure in order to directly display the electricity consumption levels of each class

and help elucidate the characteristics of TELPs.

Result discussion.





Holiday and mean\_temp were the two main variables that contributed to the different TELPs identified for each building.