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

Kehua Li
University of Wollongong, kl751@uowmail.edu.au

Rebecca Yang
Royal Melbourne Institute of Technology

Duane A. Robinson
University of Wollongong, duane@uow.edu.au

Jun Ma
University of Wollongong, jma@uow.edu.au

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

Follow this and additional works at: https://ro.uow.edu.au/eispapers1

Part of the Engineering Commons, and the Science and Technology Studies Commons
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

Abstract
This study presents 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. The proposed strategy takes the advantages of three dissimilarity measures (i.e. Euclidean distance, Pearson distance and Chebyshev distance) to calculate the difference between daily electricity usage profiles. Two-year hourly electricity usage data collected from two different university library buildings were employed to evaluate the performance of this strategy. It was shown that this strategy, which considered both magnitude dissimilarity and variation dissimilarity simultaneously, can identify more informative typical daily electricity usage profiles, in comparison with other twelve clustering-based strategies which used a single dissimilarity measure. Some interesting information related to building energy usage behaviours was also discovered with the help of visualisation techniques. Additional or hidden information discovered using this strategy can potentially be useful for fault detection and diagnosis and performance enhancement of library buildings.

Disciplines
Engineering | Science and Technology Studies

Publication Details

This journal article is available at Research Online: https://ro.uow.edu.au/eispapers1/2496
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

Kehua Li¹, Rebecca Jing Yang², Duane Robinson¹, Jun Ma³, Zhenjun Ma¹,*

¹Sustainable Buildings Research Centre, University of Wollongong, Australia
²School of Property, Construction and Project Management, RMIT University, Australia
³SMART Infrastructure Facility, University of Wollongong, Australia

* zhenjun@uow.edu.au

Abstract: This study presents 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. The proposed strategy takes the advantages of three dissimilarity measures (i.e. Euclidean distance, Pearson distance and Chebyshev distance) to calculate the difference between daily electricity usage profiles. Two-year hourly electricity usage data collected from two different university library buildings were employed to evaluate the performance of this strategy. It was shown that this strategy, which considered both magnitude dissimilarity and variation dissimilarity simultaneously, can identify more informative typical daily electricity usage profiles, in comparison with other twelve clustering-based strategies which used a single dissimilarity measure. Some interesting information related to building energy usage behaviours was also discovered with the help of visualisation techniques. Additional or hidden information discovered using this strategy can potentially be useful for fault detection and diagnosis and performance enhancement of library buildings.

Keywords: Building electricity usage; Shared Nearest Neighbours; multiple dissimilarity measures; cluster analysis
Nomenclature

cov  Covariance

dist  pairwise distance

l  numeric vector containing Clustering Validity Indices values

j  number of Clustering Validity Indices used for calculation of $V_k$

K  number of nearest neighbours in Shared Nearest Neighbours technique

k  number of clusters

$k_{opt}$  optimal number of clusters

$PCC$  Pearson’s correlation coefficient

r  parameter

sd  standard deviation

$SNND$  Shared-Nearest-Neighbours-based dissimilarity

$SNND_{com}$  combined pairwise dissimilarity based on $SNND$

$V_k$  global clustering quality index

$x, y$  Profiles

$x_i, y_i$  $i^{th}$ element of $x$ and $y$, respectively

$\bar{x}, \bar{y}$  average values of $x$ and $y$, respectively

Subscript

C  Chebyshev distance

E  Euclidean distance

P  Pearson distance
1. Introduction

Data mining, known as a set of efficient techniques that can reveal useful information from huge datasets [1], has been widely applied to improve the energy performance of buildings and building energy systems. In recent years, multiple data mining techniques have been used for building system fault detection and diagnosis (FDD), building load forecasting, occupancy behaviour prediction and energy profiling of a group of buildings. For instance, decision trees have been used for prediction of building electricity usage [2] and FDD of air handling units [3]. Artificial neural networks have been employed to simulate energy performance of air-conditioning systems [4, 5] and forecast the electricity load [6]. Association rule mining has been used to discover useful information hidden in the dataset collected from building automation systems [7].

Cluster analysis is one of commonly used unsupervised data mining techniques to group objects so that an object is similar with other objects in the same group, but different from the objects in other groups [8]. A number of strategies based on cluster analysis techniques have been developed and employed to identify building typical energy usage profiles for various applications [9, 10]. For instance, Partitioning Around Medoids (PAM) was used in a strategy to identify building typical daily energy consumption profiles [11]. The time series data of electricity consumption and heating load from 19 university buildings were used to test the performance of this PAM-based strategy. The result showed that this strategy can provide useful information including the variation of daily energy consumption and the daily high energy consumption period. Zhou et al. [12] used a fuzzy c-means clustering algorithm to group 1,312 low-voltage residential buildings using one-month daily electricity usage data. Different cluster numbers were tested during the clustering process to discover the feature of building groups with the highest and lowest electricity usage, respectively. Luo et al. [13] developed a method to compare the energy performance of small-sized and medium-sized commercial buildings.
using representative daily electricity consumption profiles. After clustering daily electricity consumption profiles of more than 2,000 buildings, the typical patterns representing the electricity consumption of each building group were determined using a $k$-means based strategy. Gouveia et al. [14] analysed building daily electricity consumption profiles to discover the users’ behaviour on space heating and cooling, in which a cluster analysis using hierarchical clustering and squared Euclidean distance was implemented to group the households based on their response about space cooling and heating in different temperatures. Yang et al. [15] used a $k$-shape clustering algorithm with a shape-based dissimilarity measure to identify the shape patterns of the energy consumption of ten institutional buildings. The discovered patterns were then used to improve the forecasting of building energy usage using Support Vector Regression. A comparison showed that this strategy outperformed another similar strategy using Dynamic Time Warping as the dissimilarity measure. Capozzoli et al. [16] presented a strategy using Symbolic Aggregate approXimation technique, $k$-means and regression trees to facilitate the flexibility and robustness of identifying abnormal building energy usage patterns. Gianniou et al. [17] presented a strategy to discover the useful information from the daily heating consumption data of residential buildings. A $k$-means algorithm was used to group buildings based on the load intensity and representative patterns. The correlations between energy intensity and the characteristics of buildings and users, load profiles, and variation of consumption behaviours were also examined.

Calculation of the pairwise dissimilarity among the profiles is an important step when applying clustering algorithms to identify building typical energy usage profiles [18]. Several dissimilarity measures, as summarised in Table 1, have been mainly used to calculate the pairwise dissimilarity to facilitate cluster analysis. Each measure has its strength and weakness. To evaluate the performance of different dissimilarity measures in identifying typical daily electricity usage profiles (TDEUPs), Iglesias and Kastner [19] used a fuzzy c-means clustering
method with Euclidean distance (ED), Pearson distance (PD), Mahalanobis distance and Dynamic Time Warping (DTW) respectively to identify TDEUPs from five university buildings. The result showed that the ED outperformed the other three dissimilarity measures. Ma et al. [11] compared the performance of ED with PD for identifying TDEUPs of multiple buildings and concluded that PD performed better than ED in terms of identifying profiles with similar variation for the case buildings studied. These studies implied that it is less possible to identify building typical energy usage profiles with all useful features by using a single dissimilarity measure [15]. Accordingly, it is worthwhile to develop a strategy using multiple dissimilarity measures to identify building energy usage profiles and make use of the advantages and avoid the drawbacks of each dissimilarity measure.

Table 1 Main dissimilarity measures used in energy profiling studies.

<table>
<thead>
<tr>
<th>Dissimilarity measure</th>
<th>Key characteristics in cluster analysis</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean distance (ED)</td>
<td>Tend to group the energy usage profiles with similar magnitudes [11, 19].</td>
<td>[11-14, 19-30]</td>
</tr>
<tr>
<td>Pearson distance (PD)</td>
<td>Can identify the energy usage profiles with similar variations [11] but is not good at identifying the profiles with similar magnitudes [19].</td>
<td>[11, 19]</td>
</tr>
<tr>
<td>Chebyshev distance (CD)</td>
<td>Good at distinguishing profiles which have a large difference in a few dimensions [31].</td>
<td>[31, 32]</td>
</tr>
<tr>
<td>Dynamic Time Warping (DTW)</td>
<td>Similar to Euclidean distance but can better identify the clusters in which the profiles are highly correlated [19].</td>
<td>[15, 19]</td>
</tr>
<tr>
<td>Mahalanobis distance (MD)</td>
<td>Less effective with dataset which has fewer objects and high dimensionality [19].</td>
<td>[19]</td>
</tr>
</tbody>
</table>

In this paper, a clustering-based strategy to identify TDEUPs of buildings with complex electricity usage behaviours was presented. The novelty of this paper is to use multiple dissimilarity measures (i.e. ED, PD and CD) and Shared Nearest Neighbours (SNN) technique to enhance the clustering results. By using three different dissimilarity measures, the advantages of each dissimilarity measure can be utilised, and their limits can be avoided. The performance of this strategy was evaluated using two-year hourly electricity usage data retrieved from two
university library buildings in different climates. A comparison of this strategy with other twelve clustering-based strategies which used a single dissimilarity measure was also performed. It is worthwhile to note that, in this study, the characteristics of date, temperatures and other factors affecting building electricity consumption were not considered during the calculation of dissimilarities between daily electricity usage profiles (DEUPs).

2. Development of the agglomerative hierarchical clustering-based strategy

2.1 Outline of the proposed strategy

The clustering-based strategy proposed in this study to identify the TDEUPs of multifunction educational buildings is shown in Fig. 1. The strategy consists of five steps, which are data collection, data pre-processing, calculation of pairwise distance, cluster analysis and the results visualization and interpretation.

![Fig. 1. Outline of the proposed clustering-based strategy.](image)

The first step is to collect building hourly electricity usage data which can be generally retrieved from building management systems. In the data pre-processing step, the time series data of building electricity usage were first divided by the building floor area and then segmented into DEUPs. The DEUPs which have missing data were not considered in the following analysis.
In the third step, the pairwise distances of all DEUPs were first calculated using three different dissimilarity measures (i.e. ED, PD and CD), respectively. The reason for using these three different dissimilarity measures is presented in Section 2.2. After the transformation and combination of three pairwise distances using Shared Nearest Neighbours (SNN), a clustering analysis using agglomerative hierarchical clustering (AHC) algorithm was conducted to group the DEUPs according to the calculated combined distance matrix. The DEUPs which have a high dissimilarity with any other DEUPs were identified as outliers and ignored in the later analysis to avoid their impact on the TDEUP identification. The TDEUPs were then determined by averaging all DEUPs in each cluster. In the last step, the identified TDEUPs were visualized, evaluated and interpreted.

2.2 Dissimilarity measures

As mentioned before, different dissimilarity measures have different characteristics. In this study, three dissimilarity measures, including ED, PD and CD, which have been used in previous studies and are complementary with each other (to be illustrated in Section 2.2.4), were used to generate a new dissimilarity measure by using the SNN technique for cluster analysis.

2.2.1 Euclidean distance

ED is the most widely used dissimilarity measure in identifying building energy usage profiles and is good at identifying profiles with different magnitudes [11]. ED is defined in Eq. (1) [8].

\[
\text{dist}_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

where \(x\) and \(y\) represent two profiles which have the same dimensionality, and the subscript \(i\) stands for the \(i^{th}\) dimension of the profile.

2.2.2 Pearson distance
Pearson distance is a dissimilarity measure developed based on Pearson’s correlation coefficient, which can reflect the linear correlation between the two profiles [8]. Pearson distance is defined in Eq. (2) [11].

\[
\text{dist}_p(x, y) = 1 - PCC(x, y)
\]

\[
PCC(x, y) = \frac{\text{cov}(x, y)}{\text{sd}(x) \times \text{sd}(y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}
\]

where \(\text{cov}\) and \(\text{sd}\) denote the covariance and standard deviation, respectively.

### 2.2.3 Chebyshev distance

Chebyshev distance is the maximum difference between the two profiles in any dimension. Chebyshev distance is good at distinguishing profiles that have large differences in a few dimensions [33]. For the time series data (e.g. DEUPs), the Chebyshev distance between two profiles becomes larger when there is a difference between the two time series profiles in a specific time [32]. Chebyshev distance is defined in Eq. (4) [8].

\[
\text{dist}_c(x, y) = \lim_{r \to \infty} \left( \sum_{i=1}^{n} |x_i - y_i|^r \right)^{\frac{1}{r}} = \max_i (|x_i - y_i|)
\]

where \(r\) is a parameter.

### 2.2.4 Comparison among ED, PD and CD

A comparison among ED, PD and CD was conducted to illustrate their difference in terms of distinguishing DEUPs with different characteristics. Fig. 2 presents four DEUPs with different characteristics which were used for the comparison. The data used were given only for illustration purpose. It can be seen that these DEUPs were different in terms of magnitude and variation, or different at a specific time. For instance, the electricity usage of Profile 1 was considerably lower than that of Profile 2 at 10:00 am, the magnitude of Profile 3 was obviously higher than the other profiles, and the variation of Profile 4 was different from the other three profiles. These differences should be considered simultaneously during the identification of
TDEUPs. Fig. 3 presents the similarity matrix among these four DEUPs when using ED, PD and CD to calculate the pairwise distance. The colour of the blocks in each plot indicated the distance between each pair of the profiles (e.g. a lighter colour means a larger distance). The difference between the two profiles may not be likely identified during the clustering process if the distance between them is too small. As shown in Fig. 3a, ED was less effective to identify the difference among Profile 1, Profile 2 and Profile 4 as these three profiles have similar magnitudes but it is very effective in identifying the difference of Profile 3 from Profiles 1, 2 and 4. As shown in Fig. 3b, PD cannot clearly distinguish Profile 1, Profile 2 and Profile 3 from each other due to their similar variations but identified their difference with Profile 4. In Fig. 3c, CD was sensitive to the difference between Profile 1 and Profile 2 as it was effectively identified the large difference at 10:00 am. CD also clearly identified the difference between Profile 1 and Profile 3 due to their large difference at each hour. As CD is very sensitive to the difference between the two profiles at a specific time, in some scenarios, such difference may be an outlier. The combination of CD with ED can avoid an undesired clustering result. From the above results, it can be seen that the characteristics of the three dissimilarity measures were functionally complementary to each other but they were not overlapped with each other.

![Illustration of four different DEUPs.](image)

Fig. 2. Illustration of four different DEUPs.
Fig. 3. Comparison among three different dissimilarity measures.

2.3 Shared Nearest Neighbours-based dissimilarity

Since the above three dissimilarity measures focus on different features of the DEUPs, it was meaningless to compare the values of the three dissimilarity measures. It is also not meaningful to calculate the average value or total value of the three dissimilarity measures to estimate the overall difference among DEUPs. In this study, the pairwise distances calculated using ED, PD and CD were transformed into an SNN similarity so that the pairwise distances can be compared and combined (Fig. 4). The SNN similarity is a secondary similarity measure based on the primary dissimilarity measures (i.e. ED, PD and CD in this study). The basic idea of SNN is that the more similar two objects are, the more nearest neighbours they share, in which the nearest neighbours stand for the most similar objects to a specific object [34].
Fig. 4. Flowchart of the SNN-based transformation of pairwise distances.

To transform a primary distance between two profiles into an SNN similarity, a positive integer $K$ which is smaller than the total number of the profiles in the dataset needs to be first chosen. The number of the same data points in the $K$ nearest neighbours of the two data points is then considered as the SNN similarity between the two data points. Since the AHC algorithm used in this study for cluster analysis requires dissimilarity as the input, the SNN similarity was transformed into a dissimilarity distance using Eq. (5) [8] before performing the cluster analysis.

$$SNND(x, y) = 1 - \frac{SNN(x, y)}{K}$$

(5)

where $SNN$ is the SNN similarity and the $SNND$ is the dissimilarity which was transformed from the SNN similarity.

In this study, the distance matrices calculated using each individual dissimilarity measure were first transformed into $SNND$ (as shown in Fig. 4) in which the $K$ was set to 50, and then combined using Eq. (6) to distinguish two profiles which have a difference in any aspect of variation or magnitude, or a difference at a specific time.
\[ SNND_{com}(x, y) = \max(SNND_E(x, y), SNND_P(x, y), SNND_C(x, y)) \] (6)

### 2.4 Agglomerative hierarchical clustering

AHC technique was used to cluster the DEUPs of each building. To conduct AHC in a dataset, each object in the dataset is first considered as an individual cluster and then successively merge the two closest clusters until all objects have been merged in a single cluster [8]. An important component of a hierarchical clustering algorithm is linkage criterion which determines whether the two clusters should be merged. In this study, Ward’s method was employed as the linkage criterion. In Ward’s method, two clusters should be merged if the merge can minimise the increase in the sum of the squared error [35]. The advantages of the hierarchical clustering include that the number of clusters is not required to be determined before the clustering and the overall process can be interpreted with a dendrogram [11].

### 3. Performance test and evaluation of the proposed strategy

R language [36] was used to implement the proposed strategy, and the R packages `dbscan` [37] and `cluster` [38] were used to conduct SNN and AHC, respectively. The R package `ggplot2` [39] was used to plot the majority of the figures presented.

#### 3.1 Introduction to the case study university library buildings

The performance of the proposed strategy was tested and evaluated using the electricity usage data collected from two university library buildings in Australia. Building A (Fig. 5a) is located at the RMIT University, Melbourne, Australia, which is in the mild temperature climate zone. The building was built in 1994 and has six levels with a total floor area of 10,200 m². The building consists of a library, a café shop, two auditoriums and several offices and studios. Building B (Fig. 5b), located at the University of Wollongong, Wollongong, Australia, which is in the warm temperature climate zone, was originally built in 1972 and retrofitted in 2008. This building has three levels with a total floor area of 14,800 m². Building B consists of a
library, a café shop and several computer labs and group study rooms. The hourly electricity usage data of the two library buildings from 1\textsuperscript{st} July 2014 to 30\textsuperscript{th} June 2016 were collected and used in this study.

![a) Building A](image)

![b) Building B](image)

Fig. 5. The two case study buildings.

**3.2 Clustering results using the combined dissimilarity measures**

The collected time series data of hourly electricity usage of each building were first divided by the total building floor area and then segmented into the DEUPs. After removing the DEUPs with missing data, a total of 712 and 731 DEUPs remained for Building A and Building B, respectively. The DEUPs of the two buildings are shown in Fig. 6, in which each curve denoted a DEUP. It can be seen that both buildings had a high diversity of electricity usage and complexity in the shapes of the DEUPs.
The pairwise distances of all DEUPs of each building were calculated using ED, PD and CD, respectively. After transformed into $SNND$, the three sets of the pairwise distances were combined into one set of the pairwise distance ($SNND_{com}$), and then clustered using the AHC. The similarity matrix of $SNND_{com}$ of Building A is shown in Fig. 7. The rows and columns of the similarity matrix were reordered by the dendrogram generated by the AHC so that the profiles which were similar to each other will be gathered in the same area. A total of eleven blue blocks can be roughly identified in the plot. There was a small dissimilarity between the members within each block, while there was a large dissimilarity between the members in different blocks, which indicated that each block can be seen as a cluster [8]. Accordingly, eleven clusters were visually identified. Since there was not a clear block shown on the top right area of the similarity matrix, this indicated that the profiles in this area were not similar to any other profiles, and the DEUPs in this area were therefore identified as the outliers.
The median of each cluster was then identified as the TDEUP and is shown as the black lines in Fig. 8. The variation of all DEUPs in each cluster was shown by the boxplot at each hour and the height of a box presented the significance of the variance at the corresponding hour. Table 2 summarises the main information of the identified TDEUPs. It can be seen that some TDEUPs identified had a less difference with other TDEUPs in terms of magnitude or variation. For instance, TDEUP 8 and TDEUP 9 were similar to each other in terms of magnitude and had a high possibility to be grouped into one cluster by using a clustering-based strategy with ED as the dissimilarity measure. TDEUP 3 and TDEUP 4 had a similar variation and could be grouped into the same cluster by using a clustering-based strategy with PD as the dissimilarity measure. However, these TDEUPs were all separately identified by using the proposed strategy since it considered the dissimilarity in terms of both magnitude and variation simultaneously. Moreover, TDEUP 1 and TDEUP 6 were highly similar to each other in terms of both magnitude and variation, but they were also distinguished by the proposed strategy. This was due to the fact that CD was sensitive to the difference in a few dimensions.
The distribution of the TDEUPs of Building A is shown in Fig. 9 in a calendar view to provide a better understanding of the temporal distribution of the TDEUPs identified. The uncoloured blocks in Fig. 9 represented the profiles which contained the missing data and had been removed during the data pre-processing. Interestingly, the TDEUPs of Building A during the first twelve months were quite different from that during the second twelve months. Such difference is an interesting point that is worthwhile to investigate for potential operational issues. It was shown that, from July 2014 to the middle of 2015, the majority of the profiles on Sunday was represented by TDEUP 8 and TDEUP 9. TDEUP 6 and TDEUP 7 mainly represented the profiles on Saturday and the session break. TDEUP 1 and TDEUP 2 mainly represented the profiles of the weekdays during the session time. During the second year, TDEUP 10 and TDEUP 11 represented the profiles on Sunday while the majority of the profiles on Friday and Saturday were represented by TDEUP 5. TDEUP 3 and TDEUP 4 represented the profiles of the weekdays during the session time.

Fig. 8. Clustering result of Building A using the proposed strategy.
# Table 2 Main information of the identified TDEUPs of Building A.

<table>
<thead>
<tr>
<th>TDEUP No</th>
<th>Number of DEUPs in this cluster</th>
<th>Estimated high electricity usage period</th>
<th>Maximum hourly electricity demand (kWh/m²)</th>
<th>Main characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>109</td>
<td>04:00–19:00</td>
<td>0.038</td>
<td>The electricity usage increased with fluctuations from 03:00 and reached a local peak at 06:00. The maximum demand occurred during 10:00-15:00 and started to drop quickly at 19:00 and then became stable at 21:00.</td>
</tr>
<tr>
<td>2</td>
<td>77</td>
<td>05:00–20:00</td>
<td>0.045</td>
<td>The electricity usage increased significantly from 03:00 to 10:00 and then kept stable until 16:00. It then started to drop quickly at 20:00 and then became stable at 22:00.</td>
</tr>
<tr>
<td>3</td>
<td>96</td>
<td>08:00–20:00</td>
<td>0.029</td>
<td>The electricity usage jumped at 07:00 and then increased slightly. After the peak during 11:00-15:00, it then reduced and became stable at 21:00.</td>
</tr>
<tr>
<td>4</td>
<td>95</td>
<td>08:00–20:00</td>
<td>0.038</td>
<td>Have a similar variation to TDEUP 3 but the peak was higher than TDEUP 3.</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>08:00–17:00</td>
<td>0.028</td>
<td>The electricity demand from 08:00 to 17:00 was significantly higher than the rest of the day while there was not a clear peak.</td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>05:00–20:00</td>
<td>0.036</td>
<td>Have a similar variation to TDEUP 2 but the demand was lower than TDEUP 2 during the high demand period.</td>
</tr>
<tr>
<td>7</td>
<td>42</td>
<td>04:00–19:00</td>
<td>0.031</td>
<td>Have a similar variation to TDEUP 1 but the demand was lower than TDEUP 1 during the high demand period.</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
<td>06:00–19:00</td>
<td>0.018</td>
<td>The electricity demand from 06:00 to 19:00 was slightly higher than the rest of the day. A sharp peak and a smooth peak occurred at 06:00 and 15:00, respectively.</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>07:00–20:00</td>
<td>0.019</td>
<td>The electricity demand from 07:00 to 20:00 was slightly higher than the rest of the day. A sharp peak and a smooth peak occurred at 07:00 and 14:00, respectively.</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>06:00–17:00</td>
<td>0.017</td>
<td>Have a similar variation to TDEUP 5 but the high demand period started earlier at 06:00.</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
<td>06:00–17:00</td>
<td>0.026</td>
<td>Have a similar variation to TDEUP 10 but the peak was higher than TDEUP 10.</td>
</tr>
</tbody>
</table>
Fig. 9. Distribution of the identified TDEUPs of Building A.

The similarity matrix of Building B is shown in Fig. 10, in which ten clusters were identified visually through the number of the blocks on diagonal. The profiles corresponding to the top right area of the plot were also identified as the outliers as there was not a clear block. The identified TDEUPs of Building B are shown in Fig. 11. The main information of the identified TDEUPs is summarised in Table 3. Similar to Building A, some TDEUPs identified had a less difference to each other in terms of magnitude or variation. These TDEUPs were all separately identified in the clustering result. The advantages of using CD were also illustrated since TDEUP 2 and TDEUP 6 were highly similar to each other in terms of both magnitude and variation but they were distinguished by the proposed strategy. The grey spots in Fig. 11 indicated the extreme electricity usage. It is noted that a number of extreme values were abnormally higher than the electricity usage of other DEUPs and some of them were identified as the outliers. Further investigation of these extreme values might be useful to identify potential operational issues.
Fig. 10. Similarity matrix and dendrogram of Building B.

Fig. 11. Clustering results of Building B using the proposed strategy.

Fig. 12 shows the distribution of the TDEUPs of Building B. It can be seen that, different from Building A, the majority of Saturday and Sunday tended to have the same TDEUPs, which were represented by the TDEUPs 2, 6, 7 and 9. The rest of the TDEUPs represented the daily
profiles on weekdays. Some interesting periodical patterns can also be observed in this plot. For instance, the weekdays of the first four weeks and the five weeks of each year before Christmas were mainly represented by TDEUP 1, while TDEUP 10 only existed during the first three weeks in February each year. The difference between the electricity usage of the two buildings may be resulted by factors such as weather conditions, building functions, building construction details, and operation & maintenance schedules as well as occupant schedules.

Fig. 12. Distribution of the identified TDEUPs of Building B.
Table 3 Main information of the identified TDEUPs of Building B.

<table>
<thead>
<tr>
<th>TDEUP No</th>
<th>Number of DEUPs in the cluster</th>
<th>High electricity demand period</th>
<th>Max. hourly electricity demand (kWh/m²)</th>
<th>Main characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>105</td>
<td>04:00–17:00</td>
<td>0.029</td>
<td>The electricity demand started to rise at 03:00. After a flat period during 08:00 and 15:00, the demand started to decrease and became stable at 18:00.</td>
</tr>
<tr>
<td>2</td>
<td>95</td>
<td>09:00–17:00</td>
<td>0.023</td>
<td>The electricity demand from 10:00 to 17:00 was significantly higher than the rest of the day while there was not a clear peak.</td>
</tr>
<tr>
<td>3</td>
<td>67</td>
<td>04:00–21:00</td>
<td>0.024</td>
<td>The electricity demand increased at 03:00 and kept almost stable from 08:00 to 21:00. It then started to drop and became stable at 22:00.</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
<td>04:00–21:00</td>
<td>0.031</td>
<td>Have a similar variation to TDEUP 3 but a smooth peak occurred at around 12:00.</td>
</tr>
<tr>
<td>5</td>
<td>73</td>
<td>05:00–21:00</td>
<td>0.024</td>
<td>Have a similar variation to TDEUP 3 but started to increase at 04:00.</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
<td>09:00–16:00</td>
<td>0.023</td>
<td>Have a similar variation to TDEUP 2 but the decrease of the electricity demand occurred earlier.</td>
</tr>
<tr>
<td>7</td>
<td>39</td>
<td>Not clear</td>
<td>0.006</td>
<td>No significant variation during the day.</td>
</tr>
<tr>
<td>8</td>
<td>86</td>
<td>04:00–21:00</td>
<td>0.036</td>
<td>Have a similar variation to TDEUP 4 but the peak demand was slightly higher than TDEUP 4.</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
<td>09:00–21:00</td>
<td>0.022</td>
<td>Have a similar variation to TDEUP 2 but the decrease of the electricity demand occurred at 21:00.</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>10:00–19:00</td>
<td>0.034</td>
<td>The electricity demand rose at midnight then kept stable until there was a gradual increase at 05:00. After a flat from 08:00 to 18:00, the demand started to decrease until 20:00.</td>
</tr>
</tbody>
</table>

4. Comparison of the proposed strategy with other clustering-based strategies using a single dissimilarity measure

In this section, the performance of the proposed strategy was compared with other twelve clustering-based strategies which used a single dissimilarity measure. The twelve strategies were developed based on the combination of four clustering algorithms (i.e. AHC using Ward’s method, k-means, PAM and Self-organizing maps (SOM) methods) and three dissimilarity measure (i.e. ED, PD and CD). The differences among these strategies and the proposed strategy are summarised in Table 4.
Table 4 Details and differences among twelve clustering-based strategies used a single dissimilarity measure and the proposed strategy.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>No.</th>
<th>Clustering algorithm</th>
<th>Dissimilarity measure</th>
<th>Clustering validity method</th>
<th>Search range of cluster numbers</th>
<th>Removal of outlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed strategy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#1</td>
<td>AHC</td>
<td>A combined dissimilarity measure</td>
<td>Similarity matrix and block-diagonal</td>
<td>-</td>
<td>Profiles which were not within any clear blocks</td>
</tr>
<tr>
<td></td>
<td>#2</td>
<td>AHC</td>
<td>ED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#3</td>
<td>AHC</td>
<td>PD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#4</td>
<td>AHC</td>
<td>CD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#5</td>
<td>k-means</td>
<td>ED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#6</td>
<td>k-means</td>
<td>PD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#7</td>
<td>k-means</td>
<td>CD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#8</td>
<td>PAM</td>
<td>ED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#9</td>
<td>PAM</td>
<td>PD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#10</td>
<td>PAM</td>
<td>CD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#11</td>
<td>SOM</td>
<td>ED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#12</td>
<td>SOM</td>
<td>PD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#13</td>
<td>SOM</td>
<td>CD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The input data and data pre-processing methods used for these other twelve clustering-based strategies were the same as those used in the proposed strategy. For the twelve clustering-based strategies which used a single dissimilarity measure, the raw time series data of the electricity usage per unit floor area of each building were segmented into 24-hour DEUPs. The only difference among them is that they used different combinations of the clustering method and dissimilarity measure. The linkage criterion used in the AHC-based strategy was Ward's method. In the SOM-based strategy, a grid in rectangular topology was adopted [8]. It is noted that the method used to determine the optimal number of clusters in the proposed strategy is less suitable in the three AHC-based strategies that used a single dissimilarity measure. This was because the similarity matrices, as shown Fig. 13a-c) and Fig.14a-c), did not show a clear block-diagonal pattern. Moreover, the method used to determine the optimal number of clusters in the proposed strategy is also less suitable in the strategies based on k-means, PAM and SOM since the number of clusters in these strategies should be determined before the plotting of the similarity matrix. Therefore, Clustering Validity Indices (CVIs) were employed to determine
the optimal number of clusters for these twelve strategies in order to achieve reasonable clustering results. Since there was not a single CVI that can work well in all conditions [21], five commonly used CVIs, including Silhouette index [40], Dunn index [41], Calinski-Harabasz index [42], Davies-Bouldin index [43] and C-Index [44], were used together to determine the optimal number of the clusters. The details of these five CVIs can be found in [45]. To determine the optimal number of the clusters for each strategy, the five CVIs were first calculated for each clustering result when $k$ was in the range of 2-20. For Davies-Bouldin index and C-Index, a smaller value indicates a better clustering result, which is opposite to the rest of the three indices. Therefore, the calculated values of these two indices were converted so that a larger value indicates a better clustering result by multiplied -1 before further processing. Then, the calculated value of each index was rescaled to 0-1 based on Eq. (7) [21].

$$V_k = \sum_j \frac{l_{j,k} - \min(l_j)}{\max(l_j) - \min(l_j)}$$

(7)

where $V_k$ is a global quality index calculated based on the five CVIs, $j$ indicates the $j^{th}$ CVI, $l_j$ is a numeric vector containing the value of $j^{th}$ CVI calculated when the number of clusters was set to 2-20, and $l_{j,k}$ is a member of $l_j$ indicating the CVI value when the number of clusters was set to $k$. The optimal number of the clusters was eventually determined using Eq. (8) [21]. Once $k_{opt}$ was determined, the clusters which had less than 5% of the number of all DEUPs in each building were identified as the outliers and removed before the identification of TDEUPs in order to improve the clustering result.

$$k_{opt} = \arg \max_k (V_k)$$

(8)

where $k_{opt}$ is the optimal number of the clusters.

Fig. 13 demonstrates the similarity matrices of Building A calculated using the twelve strategies with different dissimilarity measures. The block-diagonals were not clear as that shown in Fig. 7, which indicated that the profiles which had a strong similarity cannot be
successfully grouped by the strategies using a single dissimilarity measure, when compared with the proposed strategy. Similar results can also be seen in the similarity matrices of Building B (Fig. 14). The optimal numbers of the clusters identified for each strategy were summarised in Table 5. The cluster numbers using the AHC-based strategies were determined through dendrogram and CVIs. It can be seen that the clustering-based strategies using a single dissimilarity measure identified considerably fewer clusters than that of the proposed strategy, which means that many features cannot be identified.
Fig. 13. Similarity matrices of Building A calculated using different strategies used a single dissimilarity measure.

Fig. 14. Similarity matrices of Building B calculated using different strategies used a single dissimilarity measure.
Table 5 Optimal number of clusters identified for different strategies with individual dissimilarity measures.

<table>
<thead>
<tr>
<th>Dissimilarity measure</th>
<th>Building A</th>
<th>Clustering algorithm</th>
<th>Building B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AHC</td>
<td>k-means</td>
<td>PAM</td>
</tr>
<tr>
<td>ED</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>PD</td>
<td>8</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>CD</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Figs. 15 and 16 show the TDEUPs identified by twelve clustering-based strategies used a single dissimilarity measure after the removal of outliers. It can be seen that the TDEUPs identified by using these twelve strategies for both Building A and Building B were less than that identified by the proposed strategy. Some interesting patterns discovered by the proposed strategy cannot be identified by any of these twelve strategies with a single dissimilarity measure. For example, there was a small peak in the morning in TDEUP 8 and TDEUP 9 of Building A identified by the proposed strategy were not identified by any strategy using a single dissimilarity measure. Similarly, the late increase and late decrease of the electricity demand identified in TDEUP 9 of Building B were also not identified by any of these twelve strategies. These results showed that the proposed strategy with multiple dissimilarity measures outperformed the clustering-based strategies using a single dissimilarity measure. A further comparison of the proposed strategy with other three clustering strategies used $k$-means, PAM and SOM with the three dissimilarity measures (i.e. ED, PD and CD) is presented in Appendix.

It is worthwhile to note that the combined pairwise distance based on ED, PD and CD used in the proposed strategy was not the only possible combination of the dissimilarity measures for identification of TDEUPs. However, it opens the opportunities to further improve the clustering results. As energy usage behaviour of each building is different and the purpose of
data analysis may also be different, there is not a one-fit solution that can meet all requirements. Therefore, the number of dissimilarity measures to be combined and which dissimilarity measure could be used may be different for different applications.

Fig. 15. TDEUPs of Building A identified by different strategies using a single dissimilarity measure.
Fig. 16. TDEUPs of Building B identified by different strategies using a single dissimilarity measure.

5. Conclusion

This paper presented a strategy based on Shared Nearest Neighbours (SNN) and agglomerative hierarchical clustering to identify TDEUPs of university library buildings with high diversity and complexity of daily electricity usage profiles. In this strategy, three dissimilarity measures, i.e. Euclidean distance, Pearson distance and Chebyshev distance, were
used to generate a new dissimilarity measure with the SNN technique. An agglomerative hierarchical clustering algorithm was then used to identify TDEUPs using the combined pairwise distance.

Two-year hourly electricity usage data collected from two university library buildings in Australia were used to evaluate the performance of this strategy. The results showed that this strategy considered dissimilarity in terms of both magnitude and variation simultaneously can discover more useful information of the electricity usage of the two university library buildings. Some hidden information on the building energy usage behaviours were also discovered with the help of the visualisation techniques. Compared with twelve clustering-based strategies which used ED, or PD or CD as the single dissimilarity measure, the proposed strategy can discover more informative TDEUPs. The results from this study can be potentially helpful in the fault detection and diagnosis of building energy systems to enhance building energy performance.

Appendix:

In this appendix, the clustering performance of the proposed strategy was further compared with the other three clustering strategies which used $k$-means, PAM. And SOM as the clustering algorithm, respectively. In this comparison, these three strategies used the same input data, the same data pre-processing and the same dissimilarity measures (i.e. the combination of ED, PD and CD) as those used in the proposed strategy. In the SOM-based strategy, a grid in rectangular topology was adopted. As stated in Section 4, the similarity matrix is less suitable to determine the optimal clustering numbers for $k$-means, PAM and SOM. In this comparison, the five CVIs used in Section 4 were also employed in these three strategies to determine the optimal clustering number.
Fig. A.1 presents the similarity matrices calculated by using $k$-means, PAM and SOM based on the optimal clustering number determined and time series data of Building A. The results of using AHC can be found in Fig. 7. It can be seen that the block-diagonals in each plot were similar to that in Fig. 7. However, the patterns showed in the top-right area of Fig. 7 were different from those in Fig. A.1, which illustrated that it was less possible to identify the outliers in the DEUPs by using $k$-means, PAM and SOM and the identified TDEUPs might be influenced by the unidentified outliers.

![Similarity matrices calculated using $k$-means, PAM and SOM with multiple dissimilarity measures.](image)

Fig. A.1 Similarity matrices calculated using $k$-means, PAM and SOM with multiple dissimilarity measures.

The TDEUPs identified by using $k$-means, PAM and SOM with three dissimilarity measures for Building A are presented in Fig. A.2. It is worthwhile to note that 20 was used as the upper limit to search for the optimal clustering number. It can be seen that the number of TDEUPs identified using $k$-means, PAM and SOM were all higher than that using AHC (Fig. 8) and the differences among some TDEUPs were insignificant, which may increase the complexity for interpretation of the clustering results. It was shown that AHC still outperformed the other three algorithms in terms of the outlier identification, and the information provided in each TDEUP as well as the complexity for the result interpretation.
Fig. A.2 TDEUPs of Building A identified using k-means, PAM and SOM with multiple dissimilarity measures.

Acknowledgements

The first author would like to acknowledge China Scholarship Council (CSC) for the provision of his PhD scholarship and support.

References


