Study on clustering analysis of building energy consumption data

Yehong Li¹,²*, Jiankun Yang¹ and Xiangyang Jiang¹

¹ Guangzhou Institute of Building Science Co. Ltd., Guangzhou, Guangdong, 510440, China
² Guangzhou Construction Engineering Co., Ltd., Guangzhou, Guangdong, 510030, China
* Corresponding author’s e-mail: yehongli_gibs@yeah.net

Abstract. A large amount of building energy consumption data has been accumulated nowadays. Such data can help to find problems in building energy efficiency and put forward suggestions for improvements. However, the present situation is that a big amount of data has been mothballed and its value has not been fully exploited. In this paper, clustering method is used to analyse the energy consumption data of 784 public buildings in Guangzhou. The method combines K-means algorithm and Euclidean distance for similarity measure, aiming to classify the time-series data of energy consumption. The analysis results identify well the energy consumption trends of these buildings. Three clusters of buildings with increasing energy consumption trends are taken as key target for energy conservation and four clusters of buildings with downward energy consumption trends are treated as example buildings for experience learning. A further analysis of the clusters finds the characteristics of each group, this helps understand better the patterns of building energy consumption. The clustering analysis facilitates more effective diagnosis in energy efficiency and also supports policy making regarding building energy conservation.

1. Introduction
An effective diagnosis of building energy efficiency must rely on data analysis. Nowadays, there is a trend to transfer the focus of building energy efficiency from energy saving products and technologies to data-driven energy supervision and management [1]. Data-driven energy supervision and management can clearly reveal the energy performance problems and make suggestions for energy performance improvements. For this reason, a complete energy supervision system for building energy consumption has been established in China, including energy monitoring platforms for public buildings in major cities, annual energy consumption statistics for civil buildings, and energy audit for key buildings [2]. As a result, a bulk of data regarding to building energy consumption have been collected. Such data has brought new opportunities to building energy conservation, but also accompanied by new challenges [3]. Currently, majority of the collected building energy consumption data are put on the shelf, and the value of such data has not been fully exploited.

The demand of building energy consumption data analysis is three-fold. Firstly, it needs to analyse the status of building energy consumption, focusing on annual, monthly or daily energy consumption indicators and helping understand better energy consumption situation. Secondly, it involves the cause analysis of building energy consumption, emphasising on influencing factors and their relationships. Thirdly, it is the predictive analysis of building energy consumption, aiming to forecast the future...
trend of energy consumption through existing data. Currently, there are researchers using different methods to analyse building energy consumption data for different purposes. For example, Xiao [4] established a prediction model of building energy consumption using RBF neural network. Li and Jiang [5] predicted the power consumption of public buildings using BP neural network. Lin et al. [6] proposed a building energy consumption analysis model based on K-means clustering and FP growth association rules. Su [7] used association rules to analyse the linkage relationship between subentry energy consumption in a library. Generally, there are many kinds of data mining methods; while it is not an easy job to select the appropriate method required for a specific analysis demand.

Clustering analysis is one of the most commonly used methods for data mining. When there is a large amount of samples needing to be concurrently classified into various groups regarding a specific attribute, clustering analysis is suitable. Wei [8] used clustering method of K-means algorithm for analysing power consumption data of a residential area in Shanghai and found that the results of clustering analysis is more accurate and reliable than that of logical reasoning analysis. In China, statistical data of building energy consumption are collected every year. For these data, clustering analysis can be used to classify buildings according to their numeric value or variation trend of energy consumption. Such analysis helps identifying building groups of high value and/or continuous growth of energy consumption, thus facilitating energy efficiency diagnosis. However, there are still few studies on clustering of energy consumption trends of a large number of building samples in China.

In this paper, energy consumption data of 784 public buildings in Guangzhou are collected. Clustering method combining Euclidean distance and K-means algorithm is selected for analysing the shape pattern (i.e. variation trend) of building energy consumption during the recent five years (2014 - 2018). The tool RapidMiner is selected for implementation of the clustering analysis. A further analysis of the classified groups is conducted in order to find the features of the building groups, so as to help energy efficiency diagnosis.

2. Sources of building energy consumption data
The data source of this study comes from the statistical data of building energy consumption in Guangzhou. Such data mainly focuses on electricity consumption of buildings and the basic information of buildings such as building type, function, area, completion time, cooling type and so on. In total, 784 building samples are collected, which have energy consumption data within five consecutive years from 2014 to 2018. Figure 1 shows the proportion of building samples with different functions. It can be seen that commercial offices are the most among the building samples, accounting for 31%, followed by government offices (18%), shopping malls (15%), hotels (14%) and building complexes (10%). Stadiums, telecommunication buildings, cultural venues and transportation buildings are the least.

![Figure 1. Proportion of building samples with different functions.](image-url)
Figure 2 shows the proportion of building samples with different completion years. Buildings completed between 2000 and 2010 are the most, followed by buildings built between 1990 and 2000. This is related to the rapid economic development of China in these years. Buildings completed after 2010 account for a big proportion for stadiums and cultural venues. Old buildings built before 1990 are mainly hospitals, hotels, educational buildings, cultural venues as well as government offices.

Figure 3 shows the number of building samples with different cooling types. Buildings using centralised cooling have the largest number of 500 and mainly include commercial offices, shopping malls, government offices, hotels and building complexes. There are 177 buildings using split cooling, mainly government offices and commercial offices.

3. **Demand definition of building energy consumption data analysis**

Before analysing the data, it is essential to make clear the goal of data analysis, that is, the target problems to be solved and the expected results to be obtained. Regarding to clustering analysis of the collected data, it can be carried out according to numerical patterns or shape patterns of building energy consumption. This paper focuses on the clustering of shape pattern (variation trend) of energy consumption, which is a clustering for time-series data. In addition, clustering can be divided into Q-type and R-type. Q-type clustering is to classify samples and R-type clustering is to classify variables. This study uses Q-type clustering to classify building samples according to their energy consumption trend from 2014 to 2018, so as to identify buildings with continuous growth of energy consumption and to find buildings with continuous decline of energy consumption. This is aimed to help energy efficiency diagnosis, find energy consumption patterns, and support follow-up policy making regarding to energy conservation.

4. **Clustering analysis method for building energy consumption data**

4.1. **Data preprocessing**

Since the energy consumption of each building sample is quite different in value, the five-year energy consumption of each building sample is normalised before the analysis. Common normalisation methods include range method and z-score method. Range method is a linear transformation of the original data. Firstly, the range is calculated using the minimum and maximum values, and then the object is mapped between 0-1. Z-score method is based on the mean and standard deviation of the original data. The processed data conform to the standard normal distribution, that is, the mean value
is 0 and the standard deviation is 1. In this study, since each sample only has five time-series data points, so the 0-1 range method is selected.

4.2. Selection of similarity measures
Clustering analysis is essentially a process of similarity measure. Therefore, appropriate similarity measure index is especially important. It can help to attain better clustering results. Similarity measures for time-series data can be divided into two categories, i.e. lock-step measures and elastic measures [9]. Lock-step measures use one-point-to-one-point comparison between two time series; while elastic measures allows one-point-to-many-point comparison between two time series. The commonly used lock-step measures include Euclidean distance, Minkowski distance, Mahattan distance, cosine similarity, Pearson correlation coefficient and so on. Meanwhile, common elastic measures include edit distance and DTW (dynamic time warping) distance. Elastic measures such as DTW is based on dynamic programming algorithm, which considers the distortion of time series and is mainly used to align two series with different length.

In this study, all of the time series have the same length, so lock-step measures are applicable. Through comparing the clustering results measured by Euclidean distance, cosine similarity and Pearson correlation coefficient, it has been found that Euclidean distance can obtain better clustering results. The calculation formula for Euclidean distance is as follows:

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

(1)

Where \( \text{dist}(x, y) \) is the distance from point \( x \) to point \( y \), and \( i \) is the dimension of the point.

4.3. Selection of clustering algorithm
After selecting the similarity measure, it also important to select proper clustering algorithm. There are many algorithms for clustering, such as K-means and hierarchical algorithm. K-means algorithm is based on distance calculation, while hierarchical algorithm is based on similarity calculation. Therefore, this paper chooses K-means algorithm, combined with Euclidean distance for clustering. K-means is an iterative clustering algorithm. It firstly selects \( k \) objects randomly as the initial clustering centres, then calculates the distance between each object and each clustering centre, and assign the object to the nearest clustering centre. Each time an object is assigned, the cluster centre will be recalculated. This process will be repeated until a termination condition is met. This termination condition can be that no objects are reassigned to different clusters, no cluster centres change again and/or that the sum of squares of errors is minimum. When Euclidean distance is used, the criterion function of K-means is to minimise the sum of squares of the distances between the object and its cluster centre, the calculation formula is as follows:

\[
\min \sum_{\ell=1}^{k} \sum_{x \in c_{\ell}} \text{dist}(c_{\ell}, x)^2
\]  

(2)

Where \( k \) is the number of clusters, \( c_{\ell} \) is the centre of the \( \ell \)th cluster, \( \text{dist}(c_{\ell}, x) \) is the distance from \( x \) to \( c_{\ell} \).

4.4. Selection of clustering analysis tools
The mainstream data mining tools include Python, R and MATLAB. Although such tools are powerful, they need programming skills. In order to simplify the process of analysis, the tool RapidMiner is used for this study. RapidMiner is a user-friendly tool, which does not need any programming skills.

5. Cluster analysis results of building energy consumption data
Using the above selected method and tool, energy consumption data of the 784 building samples were clustered. Through setting different K values, finally 15 categories were determined since each cluster can well reflect its shape pattern. The results show that three clusters of the samples have an overall upward trend, in total 191 buildings, accounting for 24.36% of the samples. Four clusters show an overall downward trend, in total 254 buildings, accounting for 32.40% of the samples. Eight clusters
have fluctuation trends. Figure 4 (a) and (b) show two of the clusters that have upward trends, (c) and (d) present two of the clusters that have downward trends, (e) and (f) show two of the clusters that have fluctuating trends. It can be seen clearly that different clusters show different curve shapes, reflecting different energy consumption trends.

![Figure 4](image-url)

Figure 4. Different curve shapes in clustering results of energy consumption trends.

A further analysis of the characteristics of different clusters was conducted. Table 1 shows the proportion of building samples with different energy consumption trends and functions. It can be seen that shopping malls and building complexes have much bigger proportion in downward trends than upward trends. While hospitals have more proportion in upward trends than downward trends, which is related to the growing demand for medical and health care. For hotels, government offices and commercial offices, the building samples with downward trends of energy consumption are only slightly less than those with upward trends.
Table 1. Proportion of building samples with different energy consumption trends and functions.

|                      | Hotels | Government offices | Educational buildings | Shopping malls | Commercial offices | Hospitals Building complexes |
|----------------------|--------|--------------------|-----------------------|----------------|--------------------|-------------------------------|
| **Upward trend**     | 21.30% | 25.17%             | 25.00%                | 20.83%         | 26.42%             | 36.67%                        |
| **Downward trend**   | 29.63% | 27.97%             | 21.43%                | 45.83%         | 29.27%             | 13.33%                        |
| **Fluctuating trend**| 49.07% | 46.85%             | 53.57%                | 33.33%         | 44.31%             | 50.00%                        |

Table 2 shows the proportion of building samples with different energy consumption trends and cooling types. It can be seen that the building samples with downward trends of energy consumption are more than those with upward trends, especially for the buildings which use centralised cooling.

Table 2. Proportion of building samples with different energy consumption trends and cooling types.

|                      | Split cooling | Centralised cooling | Mixed cooling |
|----------------------|---------------|--------------------|--------------|
| **Upward trend**     | 26.55%        | 23.60%             | 18.18%       |
| **Downward trend**   | 29.38%        | 34.00%             | 22.73%       |
| **Fluctuating trend**| 44.07%        | 42.40%             | 59.09%       |

Table 3 shows the proportion of building samples with different energy consumption trends and completion years. For the buildings built before 1980, they have equivalent proportion in upward and downward trends of energy consumption, and majority of them have fluctuating trends. For the buildings built between 1990 and 2010, they have more samples with downward trends than those with upward trends. For the buildings built after 2010, they have more samples with upward trends than those with downward trends.

Table 3. Proportion of building samples with different energy consumption trends and completion years.

|                      | Before 1980 | 1980-1990 | 1990-2000 | 2000-2010 | After 2010 |
|----------------------|-------------|-----------|-----------|-----------|------------|
| **Upward trend**     | 20.00%      | 28.05%    | 24.78%    | 21.10%    | 37.74%     |
| **Downward trend**   | 20.00%      | 26.83%    | 32.61%    | 35.07%    | 30.19%     |
| **Fluctuating trend**| 60.00%      | 45.12%    | 42.61%    | 43.84%    | 32.08%     |

6. Conclusions

This paper explores the use of clustering method to analyse the energy consumption data of 784 public buildings in Guangzhou. By combining the Euclidean distance and K-means algorithm, the energy consumption trend of each building sample in recent five years was classified. The clusters obtained by using this method well reflect different curve shapes and help to find buildings with continuous growth trends of energy consumption and buildings with continuous decline trends of energy consumption. Through further analysis of different clusters, the proportion of buildings with different energy consumption trends in different types of buildings was identified, so as to find some regular patterns of building energy consumption. For example, hospitals with sustained growth in energy consumption account for a large proportion. Since energy consumption level of hospitals is relatively higher than other functional buildings. Therefore, effective policies or technical measures could be taken to slow down the growth trends of energy consumption in hospitals.

Using the clustering method, it can quickly locate the groups of buildings with increasing energy consumption that need further improvement in energy conservation and also identify benchmark buildings with continuous decreasing energy consumption for experience learning. This facilitates the implementation of energy efficiency measures in individual buildings but also the formulation of more effective policies from the government level, so as to improve the current status of building energy consumption. For example, energy consumption quota of individual buildings could be established based on their energy consumption in the previous year, that is, energy consumption is required to decrease by a specified percentage on the basis of the previous year. If using this energy quota policy, the clustering method in this paper will be very practical to determine if a single building meets the
quota requirements or not, thus helping to diagnose building energy efficiency and to take effective energy conservation measures.

**Acknowledgments**

This paper is one of the phased achievements of the following funded project: China Postdoctoral Science Foundation project (2018M643044), the science and technology progress fund project of Guangzhou Institute of Building Science Co., Ltd. (2019Y-KJ08) and the science and technology plan project of Guangzhou Construction Engineering Co., Ltd. (2019-KJ024).

**References**

[1] Wang, X., Wei, Q., Jiang, Y. (2010) Energy efficiency diagnosis and application to air conditioning systems based on energy consumption data and indicators. Heating Ventilating & Air Conditioning, 40(8): 22-24, 4.

[2] Ding, H., Liu, H., Yin, S. (2017) Research on the current situation and trend of public building energy conservation supervision platform construction in China. Construction Science and Technology, (23): 10-11.

[3] Mo, Z. (2015) New Strategy for Building Energy Savings in the Big Data Era. City & House, (8): 11-13.

[4] Xiao, D. (2012) Method Research and System Development of Energy Analysis for Public Buildings based on Data Mining. Chongqing University, Chongqing, China.

[5] Li, J., Jiang, Y. (2019) Research on Prediction of Electricity Consumption of Public Buildings Based on BP Neural Network. Software Guide, 18(7): 49-52.

[6] Lin, S., Hu, F., Hao, C., Li, D., Fu, Y. (2018) Study on the power consumption analysis model of building based on data mining. Electrical Measurement & Instrumentation, 55(20): 52-59.

[7] Jiang, D. (2013) Analysis of Energy-saving System Power Consumption Items Coupling Relationship based on Association Rule Mining for the Study and Implement. South China University of Technology, Guangzhou, China.

[8] Wei, Y. (2017) Data mining method and case analysis of building energy demand side management. Heating Ventilating & Air Conditioning, 47(09): 27-32, 54.

[9] Chen, H., Liu, C., Sun, B. (2017) Survey on similarity measurement of time series data mining. Control and Decision, 32(1): 1-11.