Energy Proceedings
2022
ISSN 2004-2965

# The Application Of Time Series Data Mining Algorithm In Building Energy Efficiency[#]

Ying Zeng[1], Qiang Gong[1*], Xiaodong Liu[2], Ebu Adu[1]

1 College of Architecture and Civil Engineering, Xiamen University of Technology, Xiamen 361024, China (*Corresponding Author)

2 College of Architecture and Urban Planning, Tongji University, Shanghai 200092, China

**ABSTRACT**

The purpose of this study is to conduct data mining research on building time series energy consumption data set. Research done so far mainly focused on cluster analysis. In this research, python language is used as the carrier, K-shape algorithm is adopted to perform cluster analysis on the time series energy consumption data from the perspective of time series. By analyzing the characteristics of its time series changes, the corresponding energy consumption patterns are obtained, and then corresponding energy saving measures are proposed to complete the construction of the entire method system. The final results show that the energy consumption curve obtained by clustering algorithm can effectively reflect the operation characteristics of the building. At the same time, based on the principle of peak cutting and valley filling, the energy consumption curve can be adjusted according to the characteristics of different modes, so that the building can effectively achieve the effect of energy saving.

**Keywords:** Cluster analysis, Energy consumption data, Time series; Energy profile

**NONMENCLATURE**

| | |
|---|---|
| *Abbreviations* | |
| EP | Energy Proceedings |
| *Symbols* | |
| n | Year |

## 1. INTRODUCTION

With the development of society, a great deal of attention is now being paid to environmental problems. Particularly after the release of carbon peaking and carbon neutrality goals, China has promoted the strategic goal of ecological civilization construction to a new height. In the past few decades, people's research on energy conservation has mainly focused on the structure of the building itself. However, with the emergence and maturity of artificial intelligence technology, new methods emerge one after another, providing new ideas for energy conservation research. In particular, during the operational phase of a building, an estimated 16% of building operational energy consumption is wasted due to widespread inappropriate energy consumption behavior and control strategies. (Park & Nagy, 2018) In addition, two common data mining methods based on clustering algorithm and association algorithm have been developed and widely used in other fields. (Zhao, et al., 2019) Li et al. (Li, et al., 2018) and Pan et al. (Song, et al., 2017) adopted clustering algorithms to obtain the daily building electricity consumption pattern from the hourly operation data. Moreover, experiments by Rhodes et al. through clustering algorithm revealed the daily electricity consumption patterns of residential buildings (Rhodes, et al., 2014) Xue et al. used association algorithms to extract the operating mode of the district heating system from historical operating data. (Xue, et al., 2017)

The purpose of this research is to use the clustering algorithm to analyze the time series energy consumption data in the benchmark energy consumption model.

## 2. DATABASE AND METHODS

### *2.1 Research Process*

The specific process of this research is shown in Figure 1, which is mainly composed of four important stages, including data collection, data mining, data analysis and conclusion. Firstly, gathering time series energy consumption data by using EnergyPlus software, which aims to simulate the building model provided by the US Department of Energy and build a csv-format data set. Subsequently, preprocessing collecting data in the second step, which involves converting units of energy consumption into kWh. After that, it commenced to

# This is a paper for the 14th International Conference on Applied Energy - ICAE2022, Aug. 8-11, 2022, Bochum, Germany.

complete the program code of K-shaped clustering for data mining of time series energy consumption database based on Python language. Lastly, analyzing the data mining results of the subentry consumption and total energy consumption, and then the relevant energy consumption patterns have been obtained.

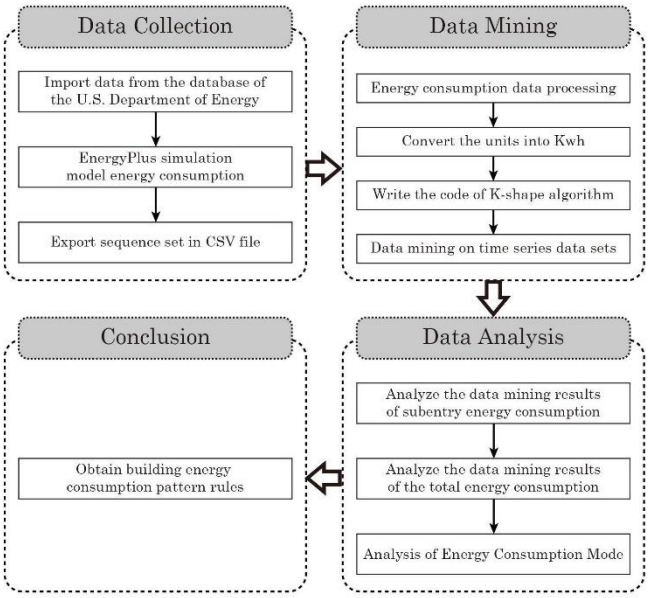


Figure 1 Outline of the research methodology workflow

### *2.2 Database sources*

There are two main sources of the database, one is actual monitoring, which is constructed by the energy consumption data of the actual operation of the building collected by the professional energy consumption monitoring data collector and the energy consumption monitoring system. These kinds of databases often have a huge amount of information, and can more accurately reflect the energy consumption patterns of the buildings being monitored. For example, Jang et al. used WEB technology to establish an online monitoring system for building energy consumption, which has the functions of energy consumption statistics and analysis. (Won-Suk Jang, et al., 2008) However, the disadvantage of this type of database is that due to the defects of the monitoring system, there will inevitably be some abnormal energy consumption data in the monitoring data. Therefore, some scholars have combined the statistical outlier detection method with clustering and successfully applied it to the abnormal data mining of energy consumption. (Seem JE, 2007)

The other type is software simulation, which is a database constructed from data obtained from modeling and simulation by professional energy simulation software. In the process of energy consumption simulation, such databases can flexibly adjust and set the variables according to their needs, so as to simulate different types of sub-item energy consumption and obtain more accurate and reliable energy consumption data. At the same time, the corresponding simulation models are representative, which can reflect the energy consumption pattern of this type of building, and have high application value.

The database of this study is based on the commercial building simulation database provided by the U.S. Department of Energy, which was published by the National Renewable Energy Laboratory in 2011, a report entitled The National Building Stock Commercial Reference Building Model of the U.S. Department of Energy, detailing the most common commercial buildings. The models in the database cover 16 building types, which have been monitored and designed by developers for a long time, with high accuracy and represent the energy characteristics of most commercial buildings. In this study, the primary school building type was selected as the research object among the 16 building types, and the Houston area of the United States was used as the input condition of the climate zone.

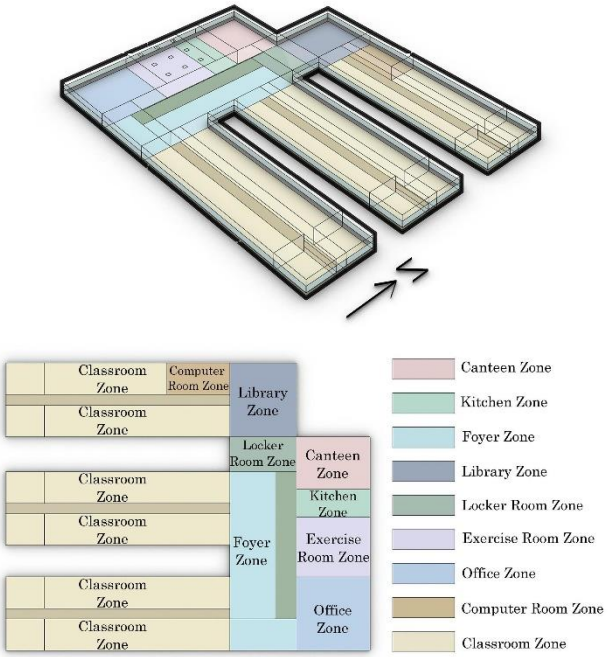


Figure 2 Schematic diagram of the primary school building energy consumption simulation model

Depicted in Figure 2 is the primary school building model adopted in EnergyPlus software. Illustrated in Table 1 and Table 2 is the basic information of the small hotel and the envelope condition for the whole building, which are fed into a simulation tool to simulate building energy consumption. All these parameters are fed into the simulation tool to simulate the energy consumption of the building. After the EnergyPlus simulation is completed, the relevant data can be directly stored in a CSV file, and the data can be recorded according to the

hourly time resolution and different energy consumption types.

Table 1 The parameters of a primary school building in Houston.

| Building Type | Floor Area | | Aspect Ratio | No. of Floors | Floor-to Floor Height | | Floor-to Ceiling Height | | Glazing Fraction |
|---|---|---|---|---|---|---|---|---|---|
| | ft² | m² | | | ft | m | ft | m | |
| Primary School | 73,960 | 6,871 | E-Shape | 1 | 13 | 3.96 | 13 | 3.96 | 0.35 |

Table 2 Building envelope related parameters.

| Roof U-Values (Btu/h·ft² ºF) | Steel Frame Wall U-Values (W/M² K) | Mass Wall U-Values (W/M² K) | Window Overall U-Value (Btu/h·ft² ºF) | Window Solar Heat Gain Coefficient |
|---|---|---|---|---|
| 0.063 | 0.704 | 3.293 | 1.22 | 0.25 |

### *2.3 Algorithm types*

#### *2.3.1 Python language and algorithm*

The Python language is a powerful interpreted programming language, it has high-efficiency high-level data structure, simple and effective implementation of object-oriented programming. Its concise syntax and support for dynamic typing, combined with the nature of an interpreted language, make it an ideal scripting language in many areas on most platforms. (Kuhlman Dave, 2012)

In the field of computing, an algorithm is defined as a well-defined finite number of instructions and computational steps to accomplish a job or solve a problem, which take one or a set of values as input, and process one or a set of values as output. When using a computer to solve practical problems, it is often necessary to design an algorithm first, and then use a programming language to implement the algorithm.

#### *2.3.2 Clustering algorithm*

Clustering is a method of unsupervised learning and is a commonly used statistical data analysis technique in many fields. The clustering algorithm is to divide the data set into multiple categories according to the intrinsic similarity of the data for a large number of unknown labeled data sets, so that the similarity of the data within the category is large and the similarity of the data between the categories is small. Commonly used clustering algorithms are: K-means algorithm, DBSCAN clustering algorithm, mean shift clustering algorithm, etc.

#### *2.3.3 Time series clustering algorithm*

The time series clustering algorithm, as the name suggests, is a clustering algorithm based on time series. At present, time series clustering algorithms can be divided into three types of clustering based on statistics, shape and deep learning.

The first statistical-based clustering method refers to clustering based on statistical feature information in time series data and some higher-order features, such as mean, standard deviation, coefficients of ARIMA model, fractal metrics and other indicators, or divide the windows, calculate these statistical features in each window, and then aggregate them. The second shape-based clustering method starts from the perspective of time-series image changes, clustering by finding the changing relationship and shape of similar graphics, and ignoring the differences in the amplitude and time scale of the images. Currently available algorithms are K-shape algorithm, DTW algorithm, and Shapelets algorithm. The third is the deep learning-based clustering method. This method mainly reduces the dimension of time series data based on the encoder model. However, at present, this method has shortcomings and lacks a general method to capture time series characteristics.

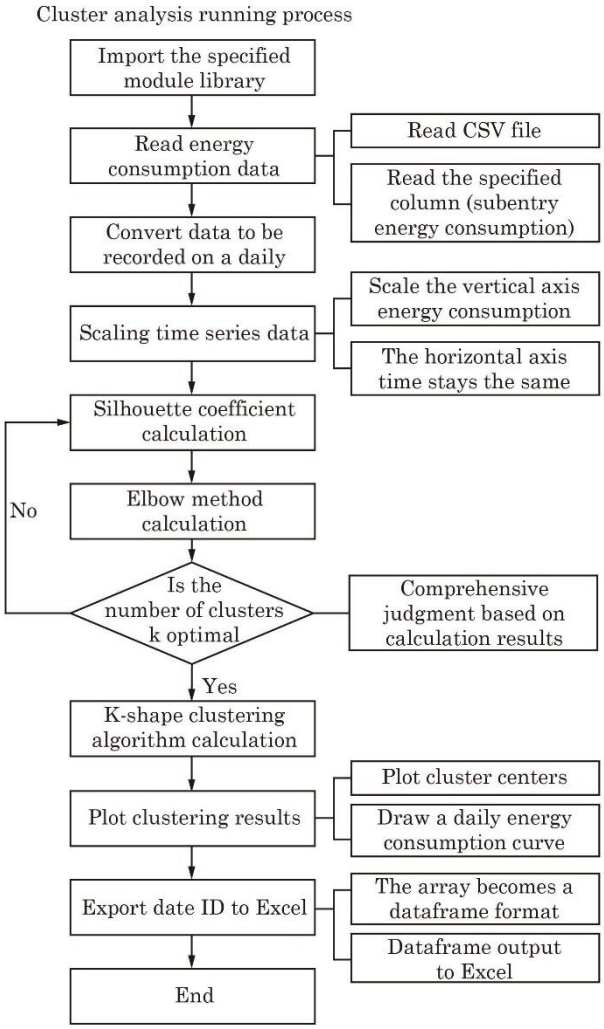


Figure 3 Logic of the clustering algorithm

This study adopts the second shape-based time series clustering method, and Figure 3 shows the overall flow of the analysis of the clustering algorithm.

### *2.4 Data circle path*

Depicted in Figure 4 is the relationship between data cycles in the entire research process. Overall, there are 365 days in a year, 24 hours a day, and one data point per hour, for a total of 8,760 energy consumption data in a year, which together make up the data set for the study. In the cluster analysis, the daily data constitutes a transaction set, that is, there are 365 transaction sets in total, and each set has 24 data points. The cluster analysis is based on the 365 transaction sets in this year.

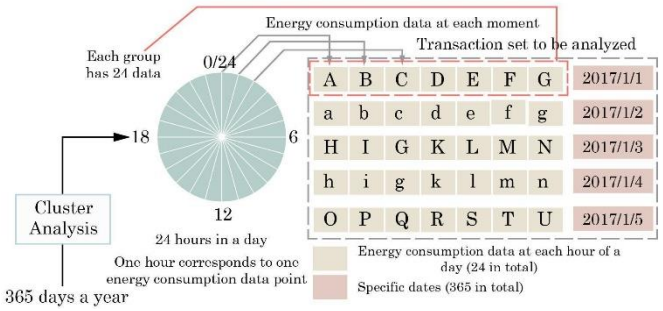


Figure 4 Relationship between data cycles in the entire research process

## 3. CALCULATION AND ANALYSIS

### *3.1 Perform indicator—— Silhouette coefficient*

Silhouette coefficient uses two basic concepts of cohesion and separation to measure the clustering effect. Cohesion measures the similarity between objects and the clusters to which they belong, while separation compares the similarity between different clusters. The similarity measure is finally performed using the silhouette coefficient value, which ranges from -1 to 1. The number k of cluster types corresponding to the larger silhouette coefficient value is the search target.

The specific calculation steps of silhouette coefficient are as follows:

1) Calculate the average distance a(i) between a sample i and the others within the same cluster. Corresponding sample should be classified into the cluster with smallest distance. The average value of all samples in a cluster a(i) is called the similarity degree of the cluster.
2) Compute the average distance b(i) between samples in different clusters, and define b(i)=min{bi1, bi2, …… , bik}. b(i) measures the degree of dissimilarity between clusters, which means that the sample is required to be labelled to another cluster if b(i) is large.
3) In line with the two parameters defined above, the formula for calculating the silhouette coefficient is as follows:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Equation 1

$$S(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, a(i) < b(i) \\ 0, a(i) = b(i) \\ \dfrac{a(i)}{b(i)} - 1, a(i) > b(i) \end{cases}$$

Equation 2

The above two formulas Equation 1 and Equation 2 are the equivalent calculation methods of the silhouette coefficient s(i) and the second formula is the first deformation.

4) Judge the clustering effect according to the calculation results. The closer s(i) is to 1, the more reasonable for the clustering number k is set.

### *3.2 Energy consumption curve shape analysis*

The time series energy consumption will show various shapes of energy consumption curves after clustering, including inverted U-shaped, M-shaped, inverted V-shaped and inverted V-shaped, which can reflect the changing characteristics of building energy consumption in a specified time period. For example, the inverted U shape shows the characteristics of high in the middle and low on both sides, reflecting that the building consumes more energy during the day and less at night.

Cluster the different sub-energy consumption and total energy consumption, analyze the shape of these energy consumption curves, find the highest point, the lowest point, and the characteristics of shape change, etc., and then we can get the characteristics of the change of building energy consumption throughout the year. According to these characteristics of energy consumption adjustment can better achieve the effect of energy saving.

### *3.3 Corresponding Seasonal Date Analysis*

Each energy consumption curve corresponds to a time period, so when analyzing energy consumption, it is necessary to combine the curve and the corresponding time season for analysis. For a building, no matter where it is located, its energy consumption will show a periodic change with the local climate change. This cyclical change is related to the local climate, and with the curve trend of energy consumption, the period with greater energy saving potential can be identified, thus providing a more scientific basis for the formulation of energy saving measures.

## 4. RESULTS

In general, the whole algorithm needs to import data first, then determine the optimal number of clusters k value by the elbow rule and the silhouette coefficient, and finally, use the optimal k value to run the algorithm to get the final analysis result. Table 3 shows the silhouette coefficients corresponding to different k values, and Figure 5 shows the final results of different sub-items energy consumption after cluster analysis.

Table 3 Silhouette coefficients corresponding to different k values

| Number of clusters | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| Cooling | 0.122 | 0.133 | 0.078 | 0.041 | 0.031 | 0.115 | 0.012 | 0.038 |
| Water systems | 0.397 | 0.386 | 0.373 | 3.134 | -0.057 | -0.055 | -0.003 | -0.011 |
| Pump | 0.277 | 0.229 | 0.220 | 0.182 | 0.122 | 0.215 | 0.203 | 0.213 |
| Heating | 0.110 | 0.162 | 0.101 | 0.114 | 0.106 | 0.061 | 0.148 | 0.075 |
| Refrigeration | 0.332 | 0.276 | 0.229 | 0.305 | 0.235 | 0.165 | 0.137 | 0.221 |
| Sum All | 0.219 | 0.318 | 0.076 | 0.005 | 0.184 | 0.034 | 0.240 | 0.102 |

It can be seen from the final results that the energy consumption patterns of the primary school building have different characteristics in winter and summer, whether it is the energy consumption by item or the total energy consumption.

For example, the overall cooling electricity consumption curve shows a steady state, which is due to the hot climate in Houston, requiring refrigeration equipment to operate almost year-round. From 9:00-17:00, as students attend classes, the overall energy level is maintained at a high level, while at other times it is possible to further reduce energy consumption as it is not class time. Therefore, the two time periods of 6:00-8:00 and 17:00-21:00 should be paid attention to reducing energy consumption and achieve the purpose of energy saving in schools.

In addition, Figure 5 also shows the clustering results of the daily total energy consumption curve of the primary school building. It can be seen that the main differences of the three curve types of total energy consumption are concentrated on the seasons. In general, type 1 is similar to the inverted U-shaped cooling electricity consumption curve pattern, type 2 is similar to the heating gas consumption curve in winter, and type 3 presents a transitional season energy consumption pattern. Therefore, the overall mode of the total energy consumption is similar to the sub-item energy consumption form, and can be adjusted according to different mode characteristics, and then the corresponding energy-saving strategy can be obtained.

## 5. DISCUSSION

During the operation of the building, data related to energy consumption are continuously generated, including water energy, electric energy and natural gas,

among which, electric energy is the main form of building energy consumption. However, the electric consumption is recorded in the form of time series. In the past, the research on power consumption mainly used instantaneous data such as total energy consumption or sub-item energy consumption, and there was less research on the data of time series energy consumption. With the emergence of new methods, the study of building time series energy consumption data has become possible, avoiding a large amount of waste of related data.

In addition, with the advent of the era of big data, various data mining algorithms emerge in an endless stream. These algorithms perform computational analysis on a large number of data sets according to different mathematical principles, in order to find various laws hidden in the data. The emergence of data mining algorithms also provides new research ideas for the field of building energy conservation. Due to the lack of effective analysis methods, the large amount of energy consumption data recorded in the past is often inevitably discarded and cannot form an effective research data set. Now data mining methods can provide new possibilities for the research of building energy efficiency based on big data sets.

## 6. CONCLUSION

To sum up, in terms of building energy consumption data, most of the current research mainly focuses on data analysis using traditional methods without considering advanced artificial intelligence algorithms in the computer field. The method proposed in this study uses advanced artificial intelligence algorithms to mine building time series energy consumption data from the perspective of time series, successfully obtains energy information hidden behind a large amount of data, and discovers building energy consumption patterns. The final results show that the energy consumption patterns in winter and summer have different characteristics whether it is the subentry energy consumption or the total energy consumption. At the same time, the energy consumption patterns of several time series are grouped according to the shape of the curve, which clearly explains the potential energy saving time in a day, and proposes corresponding energy saving strategies based on this, which can effectively make the building achieve the effect of energy saving.

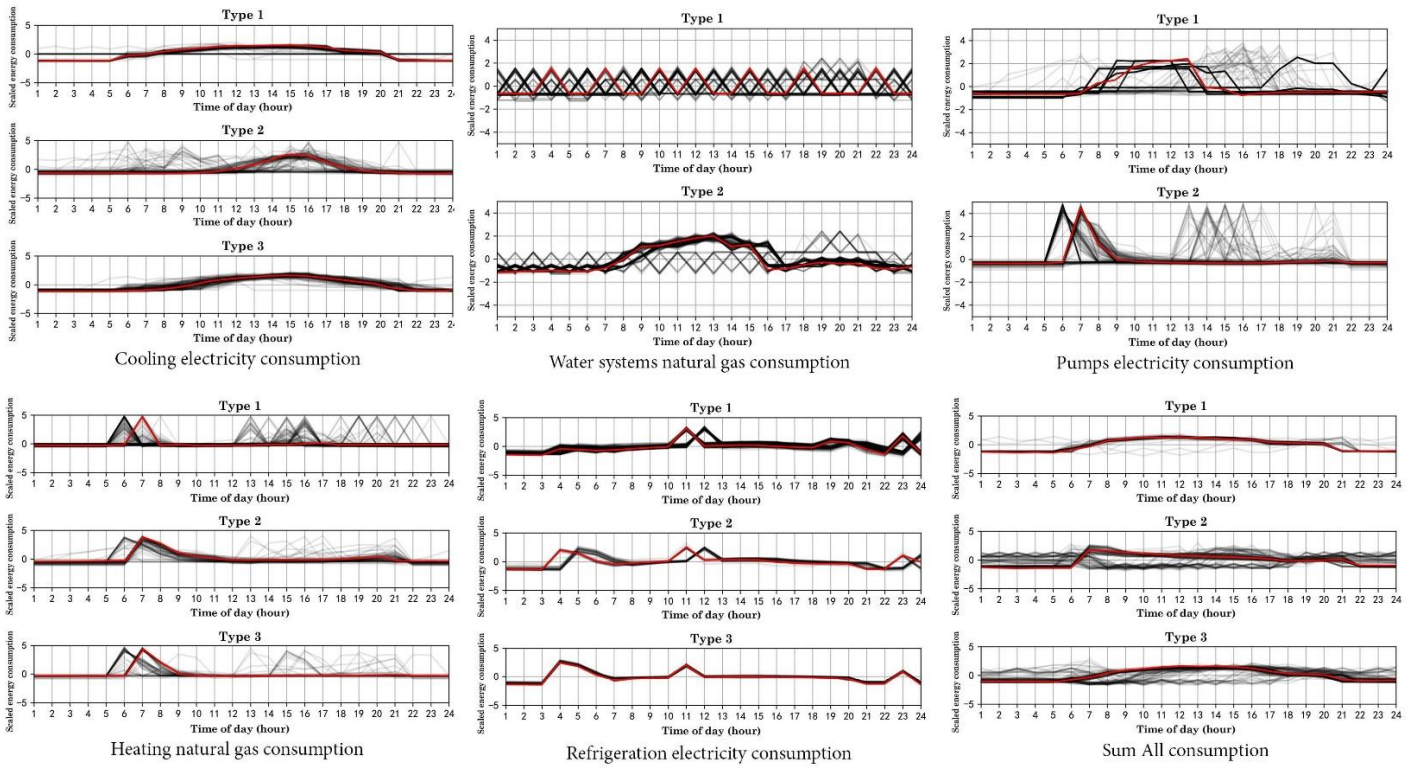


Figure 5 Results of different sub-items energy consumption after cluster analysis.

**REFERENCE**


[1] K. D., 2012. *A Python Book: Beginning Python, Advanced Python, and Python Exercises.* [Online] Available at: http://www.worldcolleges.info/sites/default/files/py3 [Accessed 01 05 2022].

[2] Li, K., Ma, Z., Robinson, D. & Ma, J., 2018. Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering. *Applied Energy*, 231(1), p. 331–342.

[3] Pan, Y., 2013. *Handbook of Practical Building Energy Simulation.* Beijing: China Architecture & Building Press.

[4] Park, J. Y. & Nagy, Z., 2018. Comprehensive analysis of the relationship between thermal comfort and building control research - A data-driven literature review. *Renewable and Sustainable Energy Reviews*, 82(3), p. 2664–2679.

[5] Rhodes, J. D. et al., 2014. Clustering analysis of residential electricity demand profiles. *Appl. Energy*, 135(15), p. 461–471.

[6] Seem JE, 2007. Using intelligent data analysis to detect abnormal energy consumption in buildings. *Energy and Buildings*, 39(1), pp. 52-58.

[7] Song, P. et al., 2017. Cluster analysis for occupant-behavior based electricity load patterns in buildings: a case study in Shanghai residences. *Building Simulation*, Volume 10, p. 889–898.

[8] Won-Suk Jang, W. M. H. & Miroslaw J. Skibniewski, 2008. Wireless sensor networks as part of a web-based building environmental monitoring system. *Automation in Construction*, 17(6), pp. 729-736.

[9] Xue, P. et al., 2017. Fault detection and operation optimization in district heating substations based on data mining techniques. *Appl. Energy,* Volume 205, p. 926–940.

[10] Zhao, Y. et al., 2019. A review of data mining technologies in building energy systems: load prediction, pattern identification, fault detection and diagnosis. *Energy and Built Environment*, 1(2), p. 149–164.