Analytical Methodology of Monthly Residential Building Electricity Consumption Based on Data Mining Models

Analytical Methodology of Monthly Residential Building Electricity Consumption Based on Data Mining Models

Xuyuan Kang, Da Yan*, Hongsan Sun, Yuan Jin

Department of Building Science, School of Architecture, Tsinghua University, Beijing, China.

*Corresponding author email: yanda@tsinghua.edu.cn

Abstract. With the development of urbanization, the energy consumption of residential buildings is growing rapidly. Previous researches mainly focus on the daily profile of residential electricity consumption, while monthly electricity consumption profile is essential for the design and optimization of power system and energy storage system. With the advancement of smart meters as well as mass data storage and transfer techniques, the availability of building energy data is growing rapidly. Thus, data mining methods are becoming possible and effective in the analysis of monthly electricity consumption in residential buildings. This paper first reviews the current research of monthly energy consumption profiles in residential buildings, then proposes a new stochastic distribution method based on data-driven models. A case study of Jiangsu residential building electricity consumption is conducted to illustrate and verify the application of this method. Finally, this paper discussed the future promotion of this method and possible perspective of monthly energy profiles.

1. Introduction

Building represents more than 1/5 of the total energy consumption in China and is believed to be a continuously growing sector for China as well as countries all around the world in future years. Of all sectors, the energy consumption of residential buildings takes up a large portion of total building energy consumption, and is growing rapidly.

Most recent researches mainly focus on the daily profile of electricity consumption in residential buildings. However, the study of monthly profile is of equal importance. The monthly electricity consumption shows strong characteristics among different seasons due to the different patterns of summer AC electricity use and winter heating electricity use. In order to cope with the control of the power system and the design of the energy storage system, a model that reflects the stochastic characteristics of monthly electricity consumption is required.

Traditionally, physical models are being used in the analysis of energy consumptions of buildings with the assistance of utility software such as DeST, EnergyPlus and eQuest. However, with the advancement of sensor technology, wireless data transmission protocol, network data transfer technology, cloud calculation and smart mobile devices, data-driven models are becoming more popular and efficient. Data-driven models learn and draw characteristics from the large amount of

historical data to acquire building energy-related knowledge. They are "data to data" methods, which do not require much detailed information of all aspects. They use machine learning techniques to extract relations, usually mathematical ones, between variables, and reproduce and regenerate them in a data-oriented way. The data required for this method is usually available from BMS system, making the data-driven model more convenient for energy analysis.

In the analysis of monthly electricity profiles in residential buildings, multiple researches have investigated into the methodology and application of data-driven models. K Zhou et al (2017) collected the data of electricity consumption of 1200 households in Jiangsu Province during the year of 2014 and defined 6 typical electricity use patterns based on fuzzy clustering method. This paper offered a description of energy use patterns of residential buildings, but the stochastic use pattern was not revealed in this research.

To W M. et al (2017) proposed a regression model based on historical monthly energy consumption. The researchers took monitored residential building energy consumption from 1970 to 2014 in Hong Kong as case study, and decomposed the energy consumption into the trend part and temperature-related part, using non-linear exponential regression to setup the prediction model. This method was very efficient in time-series monthly prediction of total energy consumption of residential buildings, yet failed to recognize the electricity use patterns of different households.

Besides, Apadula et al (2012) also used multiple linear regression for monthly electricity load prediction in residential buildings. Current researches mainly look into the monthly profile within two perspectives: time-series load prediction and clustering-based use pattern extraction. The characteristics of monthly electricity profile in residential buildings and stochastic distribution of household energy use are rarely given specific considerations.

As a result, this paper proposes a stochastic distribution method based on data-driven models to better represent and illustrate the monthly profile of residential electricity consumption. We use the monthly energy data from Jiangsu residential buildings as case study to illustrate and verify the method, and eventually discuss the future application and optimization of this method.

2. Methodology

2.1. Establishment of model

The purpose of stochastic distribution model is to extract features and give a description of the distribution for each feature. Thus, this method comprises the following steps:

1. Data cleaning. Clustering methods are used for data cleaning. The original dataset is clustered into multiple groups and we identify the abnormal data sample by identifying abnormal cluster centroids. All samples that belong to abnormal clusters will be removed from the original dataset.

2. Feature extraction. The data series of each household contains 12 data points which represent the electricity use of 12 months. In order to reflect the characteristics of electricity use of every single household, 3 features are extracted for description:

- Monthly-averaged electricity consumption μ: energy use level of a household
- Coefficient of Variance of monthly electricity consumption CV=σ/μ: the fluctuation of the energy uses of a household
- Electricity use patterns ($a_1, \dots, a_{12}$): behavioural characteristics of the energy use of a household

Z-score standardization method is being used to extract the 3 features. The z-score standardization is expressed as follows:

$$a_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

Where x_i represents the total electricity consumption of the ith month (kWh), μ represents the monthly-averaged electricity consumption (kWh), σ represents the standard deviation of monthly electricity consumption (kWh), i=1…12

3. Cluster analysis: K-means method is used to run clustering analysis. K-means clustering requires the number of clusters as an input, so we select the number of clusters based on Calinski-Harabaz (C-H) principle. C-H index is defined as ratio of within-cluster dispersion and the between-cluster dispersion. It can be denoted as:

$$S(k) = \frac{tr(B_k) * (m-k)}{tr(W_k) * (k-1)} \quad (2)$$

Where *m* is the number of samples, *k* is the number of clusters, B_k refers to the between-cluster dispersion matrix, W_k refers to the within-cluster dispersion matrix, *tr* refers to the trace of the matrix. *S(k)* is the Calinski-Harabaz index.

In order to determine the k value (the number of clusters), we iterate k within a specific range and set the k with maximum C-H index.

Then we use K-means method to cluster the original dataset by features and formed multiple sub-clusters for further analysis.

4. Distribution fitting: 3 features are extracted for each household; different forms of probability distribution are utilized under different scenarios. In this case, we use logarithm normal distribution to fit μ and CV, and use normal distribution to fit energy use patterns $(a_1, \dots, a_{12})$

With the four steps mentioned above, the stochastic distribution model is established for the case of residential building monthly electricity consumption. Using this model, a random electricity use curve can be simulated.

2.2. Simulation of random electricity use curves

The simulation process is more likely to be a process of "unzip", or "regeneration". Based on the fitted distribution, we can randomly generate 3 electricity use features, and then restore the electricity use curve of with the following equation:

$$x_i = a_i * \sigma + \mu = (a_i * CV + 1) * \mu \quad (3)$$

In which x_i refers to the electricity consumption of the ith month (kWh), μ refers to the monthly averaged electricity consumption (kWh), σ refers to the standard deviation of monthly electricity consumption (kWh), CV refers to the Coefficient of Variance, a_i refers to the pattern parameter of the ith month. i = 1…12

In this way the electricity curves of a sub-cluster can be fully regenerated from distribution-based random parameters using the methods shown above.

2.3. Validation of the models

We use 3 methods to validate the effectiveness of the model: parametric validation, simulation validation and testing set validation.

First method of validation is the parametric validation. T-test is used to verify the effectiveness of distribution fitting. Under the confidence level of 95%, the t-test is passed if $p > 0.05$.

Second method of validation is simulation validation. We randomly generate a specific number of samples of electricity curves based on the model, then pick same number of samples randomly from the original dataset. The above-mentioned procedure is defined as "one simulation". The simulation repeats 100 times to see the relative difference statistically.

Third method of validation is testing set validation. The whole model is set based on the data from training set. The model is used to simulate the testing set and compare the total electricity consumption with that of the real data.

3. Case Study

In this part, energy profiles of Jiangsu Province are taken as the case study to analyse the monthly electricity consumption residential buildings. The data is oriented from the monitored data of 86,672 households from six cities in Jiangsu Province – Lianyungang (LYG), Nanjing (NJ), Nantong (NT), Wuxi (WX), Xuzhou (XZ) and Taizhou (TZ). The time span of the monitored data covers the whole year of 2014. After data cleaning, the number of households is reduced to 83,243. The distribution of these households in six cities are shown in **Table 1**. For the validation of the model, samples of TZ (13,458 households) are assigned as the testing set and the samples of the other five cities (69,785 households) as the training set. The following parts of the setup of the model are all based on the data from the training set.

Table 1. The distribution of households in six cities in Jiangsu

City	*LYG*	*NJ*	*NT*	*WX*	*XZ*	*TZ*
Households	13,697	10,870	16,093	15,235	13,890	13,458

3.1. Data Cleaning

The original dataset is initially clustered into 30 groups. Examples of abnormal clusters are shown below. After the data cleaning, the number of the samples reduced from 86,672 households to 83,243 households.

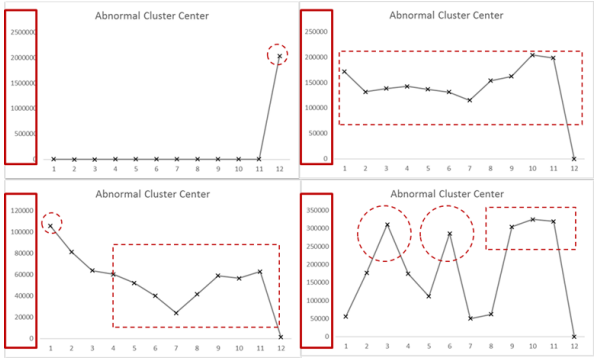

Figure 1. Examples of abnormal cluster centres

3.2. Feature Extraction

For each sample among all 83,243 households, we extract the monthly-averaged electricity consumption, the coefficient of variance and the electricity use patterns as features. The monthly electricity consumption of a single household can be described with 3 features as described above.

3.3. Cluster Analysis

K-means method is utilized in clustering analysis. Using Calinski-Harabaz principle, we iterate k value within [2, 20] and eventually set k=4.

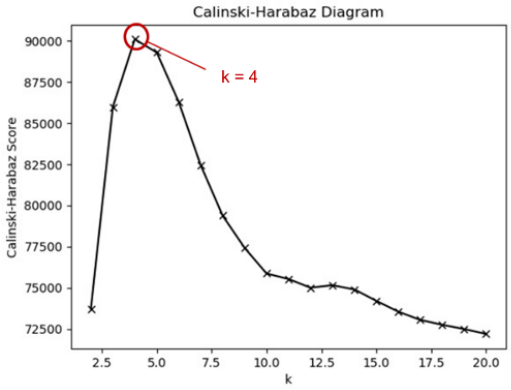

Figure 2. Calinski-Harabaz diagram to select the best number of clusters

Using K-means algorithm and C-H principle, all samples are firstly divided into 4 clusters by (μ, σ), which represents 4 levels of total energy consumption. Then we run clustering by electricity use patterns $(a_1, \ldots, a_{12})$ to acquire 4 sub-clusters within every cluster, thus getting a total of 16 sub-clusters. Details of sub-clusters are shown below.

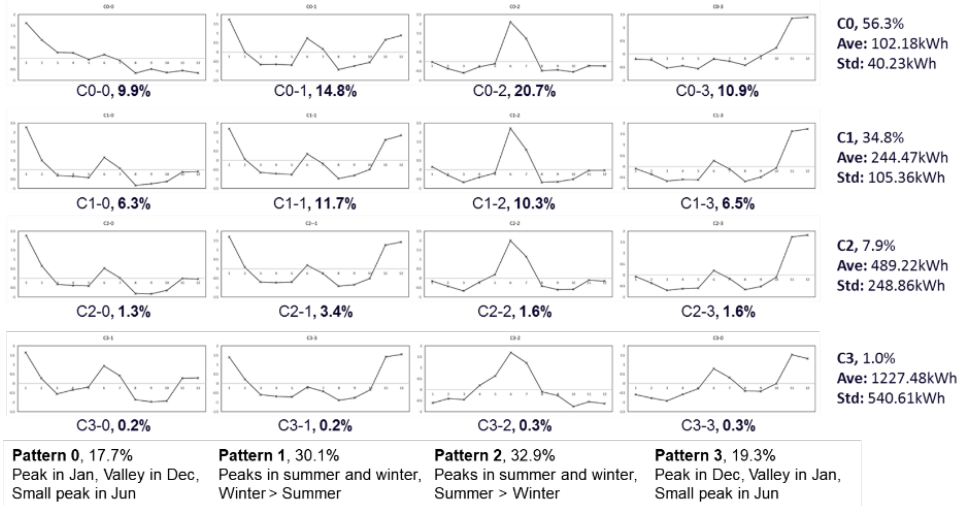

Figure 3. Distributions and statistics of 16 sub-clusters

3.4. Distribution fitting of features

Here logarithm normal distribution is used to fit μ and CV, and normal distribution is used to fit energy use patterns $(a_1, \ldots, a_{12})$. Results of sub-cluster C2-1 are shown in **Figure 4** to illustrate the outcome of distribution fitting.

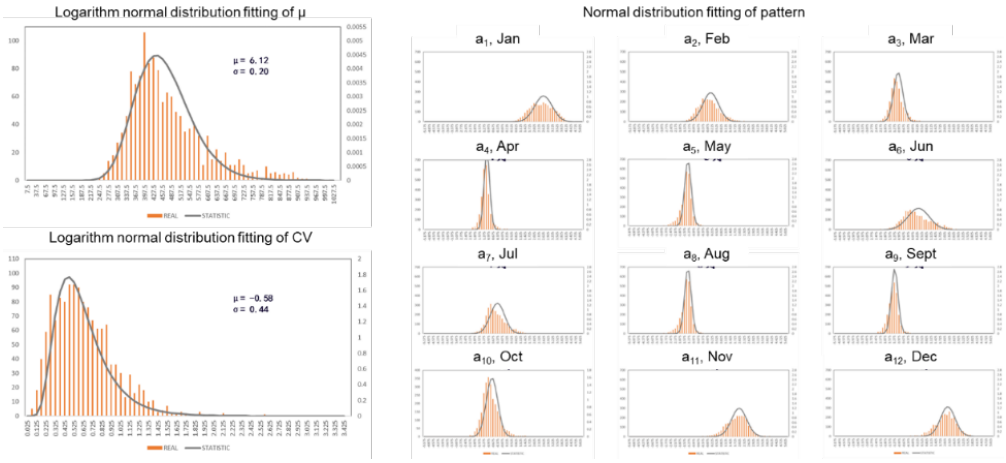

Figure 4. Distribution fitting of μ, CV and patterns

3.5. Simulation of random electricity use curves

Using the data from training set, the stochastic distribution model is fully established for monthly electricity consumption in Jiangsu residential buildings. This model is then used to simulate the testing set. The ratio of each sub-clusters is identified within the testing set, then corresponding number of samples are randomly generated to simulate the testing set.

3.6. Validation of the model

First, parametric validation. We take sub-cluster C2-1 as the example. **Figure 5** illustrates the p-value of 14 parameters (μ, CV and 12 pattern parameters) of sub-cluster C2-1. The p-values are all above 0.05, which suggest that t-test is passed.

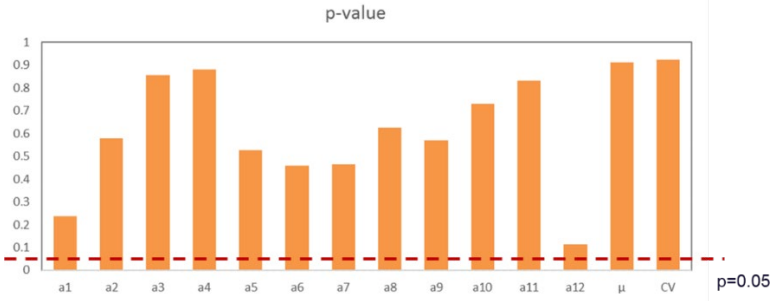

Figure 5. The p-value of 14 parameters in the t-test of sub-cluster C2-1

Second, simulation validation. We randomly generate 1,000 samples of electricity curves based on the model, meanwhile picking 1000 samples randomly from the original dataset in "one simulation". The simulation repeats 100 times to see the relative difference statistically. It can be seen that the maximum relative difference never exceeds 8%, and averages 3.77%, suggesting a high accuracy.

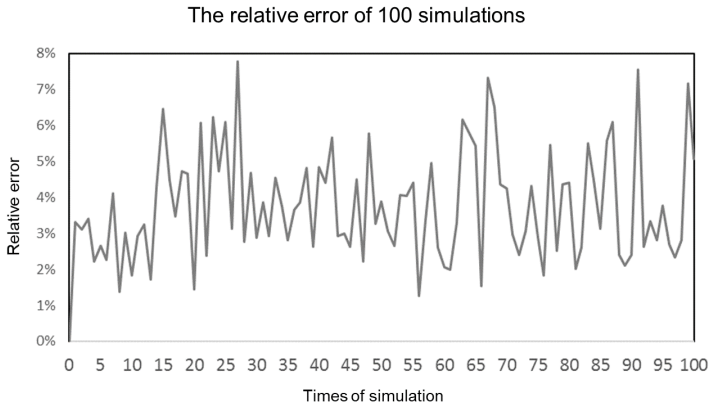

Figure 6. The relative error of 100 simulations

Third, testing set validation. We compared the simulated total electricity consumption of testing set with the real total electricity consumption. The result is shown in **Figure 7** with an average relative error of 1.6%, suggesting a great fit of simulated and real data.

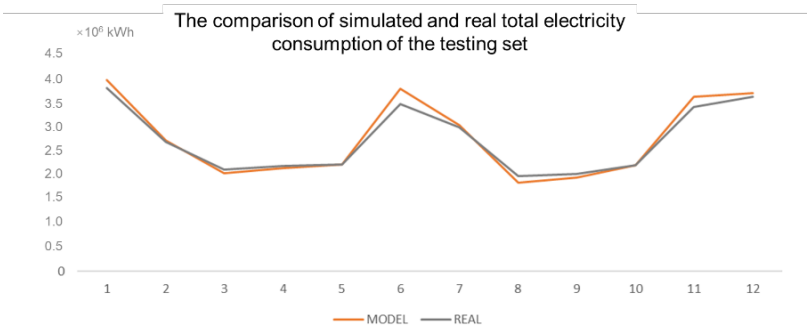

Figure 7. The relative error of 100 simulations

4. Conclusions

This paper proposed a stochastic distribution model for the description of monthly electricity profile for residential buildings. The example of Jiangsu residential buildings is introduced as the case study to better illustrate the application and validation of the models in various perspectives. This method serves for the description of energy use patterns for residential buildings and offers a way to estimate monthly electricity consumption profiles in different scales. However, this method doesn't represent the time-relevance of the electricity consumption of adjacent months. Future research may look into the optimization of time dependence factors in the establishment of the models.

References

[1] Building Energy Research Center. 2017. *China Building Energy Use 2017*. China Architecture & Building Press.
[2] Zhao H, and Frédéric M. 2012. "A review on the prediction of building energy consumption." *Renewable and Sustainable Energy Reviews* 16.6: 3586-3592.
[3] Amasyali K, El-Gohary N M. 2018. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews,* 81: 1192-1205.
[4] Wei, C., & Li, Y. 2011. Design of energy consumption monitoring and energy-saving

management system of intelligent building based on the Internet of things. *In Electronics, Communications and Control (ICECC), 2011 International Conference on (pp. 3650-3652). IEEE.*

[5] Ouf, M. M., Issa, M. H., Azzouz, A., Sadick, A. M. 2017. Effectiveness of using WiFi technologies to detect and predict building occupancy. *Sustainable Buildings*, 2, 7.

[6] Koseleva, N., & Ropaite, G. 2017. Big data in building energy efficiency: understanding of big data and main challenges. *Procedia Engineering*, 172, 544-549.

[7] To W M, Lee P K C, Lai T M. 2017. Modeling of Monthly Residential and Commercial Electricity Consumption Using Nonlinear Seasonal Models—The Case of Hong Kong[J]. *Energies*, 10(7):885.

[8] Zhou, K., Yang, S., & Shao, Z. 2017. Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study. *Journal of cleaner production,* 141, 900-908.

[9] Apadula, F., Bassini, A., Elli, A., & Scapin, S. 2012. Relationships between meteo-rological variables and monthly electricity demand. *Applied Energy*, 98, 346-356.

Acknowledgement
This research is supported by Tsinghua University Initiative Scientific Research Program, Grant No. 20171080282.