District household electricity consumption pattern analysis based on auto-encoder algorithm

To cite this article: Yuan Jin et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 609 072028

View the article online for updates and enhancements.
District household electricity consumption pattern analysis based on auto-encoder algorithm

Yuan Jin1, Da Yan1,*, Xingxing Zhang2, Mengjie Han2, Xuyuan Kang1, Jingjing An1 and Hongsan Sun1

1 School of Architecture, Tsinghua University, Beijing 100084, China
2 School of Industrial Technology and Business Studies, Dalarna University, Falun 79188, Sweden
*yanda@tsinghua.edu.cn

Abstract. The energy shortage is one key issue for sustainable development, a potential solution of which is the integration with the renewable energy resources. However, the temporal sequential characteristic of renewable resources is different from traditional power grid. For the entire power grid, it is essential to match the energy generation side with the energy consumption side, so the load characteristic at the energy use side is crucial for renewable power integration. Better understanding of energy consumption pattern in buildings contributes to matching different source of energy generation. Under the background of integration of traditional and renewable energy, this research focuses on analysis of different household electricity consumption patterns in an urban scale. The original data is from measurement of daily energy consumption with smart meter in households. To avoid the dimension explosion phenomenon, the auto-encoder algorithm is introduced during the clustering analysis of daily electricity use data, which plays the role of principal component analysis. The clustering based on auto-encoder gives a clear insight into the urban electricity use patterns in household. During the data analysis, several feature variables are proposed, which include peak value, valley value and average value. The distinction analysis is also conducted to evaluate the analysis performance. The study takes households in Nanjing city, China as a case study, to conduct the clustering analysis on electricity consumption of residential buildings. The analysis results can be further applied, such as during the capacity design of district energy storage.

1. Introduction
Residential energy consumption accounts for around 20% to 30% electricity consumption in different countries and is of great potential for energy conservation[1]. In 2016, the household electricity consumption takes up 27.2% proportion of the world total electricity consumption[2]. With the development of smart grid and sustainable energy integration, the combination of renewable energy becomes more promising to save energy and to face the challenge of energy shortage. During the integration of the renewable energy source with the traditional electricity grid, because of the profile discrepancy, it is necessary to make the energy generation and end use side comply with each other. Based on the developing technology of energy consumption measurement, such as smart meter, it is easier to analyse the electricity use with different temporal resolutions[3, 4]. Since then, a lot of researches on the electricity profile and demand response have arisen, and the new smart grids also bring many challenges to the demand response research[5-8]. The electricity use data with different temporal resolution possesses different potentials for research and engineering application. The hourly data contributes to the load profile study in one day and instructs the strategy of demand response. While the daily data helps with the fuel supplies, maintenance operations[9] as well as energy storage design.

Many researchers have conducted analysis based on the daily electricity use data. Valor[10] conducted research on relationship between daily electricity consumption and air temperature, with the objective to simulate the response of energy consumption to weather conditions. The research showed that the energy consumption was strongly related to the outdoor air temperature, which indicated the
influence of temperature on the energy consumption of heating or cooling. The analysis divided the data from the weekdays and the weekends (including holidays). Compared with the weekends and holidays, the electricity use in weekdays presented stronger and clearer correlation. Beccali[11] proposed a combined approach based on unsupervised and supervised neural networks to forecast and predict the electric energy demand, correlating the electricity use with the weather data, including temperature, relative humidity and global solar radiation. Besides the regression and prediction work, many researchers focus on the clustering analysis. To instruct the building management for forecasting and detection of abnormal energy use, Li[12] made classification of the energy consumption in buildings with outlier detection methods. The features of the energy consumption were firstly extracted, used for the abnormal outlier detection as well as the classification work. Wen[13] proposed a shape-based clustering method to recognize the electricity use patterns in residential buildings. There are also many clustering methods about the electricity use data, such as Dirichlet process mixture model, k-means method, fuzzy clustering and decision trees[14-16]. Zhou[17] also used an improved fuzzy c-means method to cluster monthly electricity use data. Different algorithms meet the requirement for data with different scale and amount. In the process of clustering analysis, k-means clustering method shows advantage than other algorithms[13]. However, it should be noticed that the k-means method is based on the distance analysis among different observations. Once the dimension of the observation reaches to a large scale, the “dimension explosion” shall happen, which makes Euclidean distance meaningless and the clustering analysis may not work. In this study, the data contains the daily electricity use data for a whole year, with 365 values, and the dimension of the data is relatively high. Therefore, to deal with the clustering analysis of such high dimensional data, the dimension reduction should be introduced.

In this research, an auto-encoder based method is proposed to solve the high-dimension problem during the clustering analysis of electricity use data. As for the data pre-processing, the k-nearest neighbour analysis is used to conduct the outlier analysis. One case study with daily electricity data for one year is conducted to illustrate the proposed clustering analysis method. Based on the clustering analysis, the typical daily and monthly electricity pattern in the urban scale are obtained. Also, the peak and valley values of the typical electricity pattern are analysed. The primary results will help with the capacity design of district energy power station. Based on this research, further the model for district electricity use will be established, which will be used for analysis of power integration with renewable resource.

2. Methodology
This study focuses on the clustering analysis of electricity use data, which possesses high-dimensional characteristic. The data is firstly processed with data cleaning, dimension reduction and outlier analysis. Here, as for the dimension reduction, the proposed improved method of auto-encoder algorithm to extract principle component with much lower dimension. Then the electricity data is clustered according to the principle components. Finally, the evaluation and illustration of each cluster is performed. The clustering results are further used for energy management and demand response analysis. The overall methodology illustration is as shown in Figure 1.

2.1. Data pre-processing
The data pre-processing is divided into three parts, (1) data cleaning, (2) dimension reduction and (3) outlier detection.

Firstly, data cleaning is required. Because of the errors of measurement and recording, the electricity use data may not be complete. In this study, if the data is less than the 90% length of the full observation, this observation will be deleted from the whole dataset. As for the remaining observations, the blank electricity use will be filled with the average daily energy consumption in this month to represent the average electricity use.

Secondly, this research focus on daily electricity consumption data for a certain time, which is considered as high-dimensional. The traditional clustering methods are distance-based or density-based, among which the k-means algorithm[21] presents superiority of clustering performance. However, the high-dimensional data may bring about the so-called “curse of dimensionality” to clustering analysis, which means that with the increase of observation’s dimension, the Euclidean distance may be invalid.
to depict the relationship among observations. Many studies have been conducted to deal with this problem. Principal component analysis and shape-based clustering method are proposed by Wen to solve this problem[13]. The similar methods are listed as wavelet packet decomposition method, discrete Fourier transform, piecewise linear regression, etc[22-24]. In this study, the cluster method is improved by the introduction of auto-encoder[25], which plays the role of dimension reduction.

**Figure 1.** Methodology illustration of auto-encoder based clustering analysis.

The auto-encoder is an unsupervised neural network, including three layer-structures, the input layer, output layer and the hidden layer(s). The particularity is that the input layer and the output layer are exactly the same during the model training (Figure 2). Through well training process, normally with backward propagation[26], the model can well represent nearly all the information from input layer to output layer. In this way, the useful information is extracted and preserved in the neurons of hidden layers, the information in hidden layers is considered to be adequate for the representation of the almost all the details of input layers. The structure from input layer to hidden layer(s) is called the encoding process, and the structure from hidden layer(s) to output layer is called the decoding process. If the dimension of the hidden layer is less than the dimension of the input observation, then the high-dimensional data is reduced to low-dimensional data in hidden layer.

According to the dimension scale of the original data, the stacked auto-encoder is introduced when the dimension of the data is rather high. The whole neural network is symmetrical, with stacked auto-encoder networks shown in Figure 2, which indicates multi-time dimension reduction during the model training. Hence, the stacked auto-encoder network reduces the dimension of the data step by step, and the middle of hidden layer is the results of the dimension reduction, as the core hidden layer. The detailed stacked auto-encoder network is further illustrated in section 3.

Thirdly, due to the abnormal electricity use may occur no matter related to the malfunction of the smart meter or the consumption boundary confusion, outlier detection is necessary during the data pre-processing. Here, the consumption boundary confusion means that the region classified as residential however the building used as commercial or office functions, and this may occur in the case study. In this study, the households amount is also at a large scale, so an efficient outlier detection for raw data with large amount is introduced[18].

The outlier detection, also called anomaly detection, is used to detect the abnormal or deviant observations in a dataset. Mainly there are two procedures of the outlier detection, outlier ranking and determination. Normally, an index or score is defined to represent the abnormal and anomalous level of one observation. In this study, the outlier detection combines k-nearest neighbours and local outlier factors to conduct the outlier detection[19]. Considering the dataset with large amount observations, the k-nearest neighbours will firstly conduct to help with the rough category and reduce the computation cost[20]. Then in the scope of their k-nearest neighbours, the index of local outlier factor (LOF) is
estimated to represent the “outlier level” of each observation. The “outlier level” depends on how isolated the object is compared with the surrounding neighbours[19]. It can be understood as the average of the ratio of the local reachability density of a sample and those of its k-nearest neighbours. The inlier observation’s LOF tends to be close to 1, while outliers tend to have a larger LOF value. By this means, the outlier observations of the electricity dataset are recognized and removed.

2.2. Clustering analysis
After dimension reduction processing, the results of lower dimension are used to make clustering analysis. In this study, k-means algorithm is recommended. To decide the proper cluster amount, the Calinski-Harabaz index is introduced, under the circumstance that the true cluster label cannot be obtained. The higher index value indicates that the analysis possesses better defined clusters.

2.3. Typical electricity use pattern
After the clustering analysis of the dataset, the results for principle components also the original electricity use pattern can be illustrated and further analysed with different temporal resolution. For each cluster, the average daily electricity use data of the households in this cluster can be calculated and analysed. As for the daily electricity consumption profile, the feature indicators like peak value, valley value and average value are analysed. For the monthly electricity consumption data, the peak value and valley value can also be analysed. From the analysis, the relationship between electricity consumption and the season and temperature can be inferred.

3. Case study
In this study, the residential electricity use data from Nanjing city, Jiangsu province, in China is used as the case study. The buildings include apartment buildings and detached buildings. The daily consumption data is measured by smart meter from January 1, 2014 to December 31, 2014, lasting for one year in total. The data is recorded in the temporal resolution of 1 hour. This dataset is used to depict the clustering analysis based on auto-encoder algorithm, which shows potential of daily electricity consumption data analysis to district power storage capacity design and integration with renewable energy.

3.1 Data pre-processing
For the first step of data pre-processing, the valid days of electricity consumption data are analysed and used for data cleaning. The household with amount of valid days less than 90% will be deleted from the
dataset. Figure 3 shows the histogram of the statistical results of the amount of the valid days for all the households in the dataset. After the data cleaning, there are 17,031 valid households in the dataset. As for the vacant data, the monthly average electricity consumption will be filled into the blanks.

**Figure 3.** Histogram of amount of valid days of all the households in the original dataset.

The auto-encoder network is established based on package called “Scikit-learn” in python[29], with the symmetrical structure and same output layer with input layer, however. The stacked auto-encoder is trained through the neural network training, and the detailed parameter settings and network structure is as shown in Table 1. The middle hidden-layer compromise 3 neurons, which is the reduced dimension of the electricity use data.

| Solver   | α   | Activation function | Network size of hidden layers |
|----------|-----|---------------------|-------------------------------|
| ADAM[30] | 0.001 | f(x) = max(0, x) | (180,60,12,3,12,60,180) |

The outlier detection is conducted to avoid the abnormal data due to malfunction of the measurement equipment. The k-nearest neighbours are introduced, and the k is set to be 1,000. The local outlier factor (LOF) for each household is computed according to the 1,000 nearest neighbours fold data. The larger LOF value refers to more anomaly. The histogram of LOF values is illustrated in Figure 4. There are 5.9% households with LOF larger than 1.5, and this figure shows the results of LOF under 1.5. After the outlier detection, the inlier households of the electricity consumption data are 14,478 households.

**Figure 4.** Histogram of Local Outlier Factors during the outlier detection process.

3.2 Clustering analysis

With the training of auto-encoder network, the output layer is close to the input layer, which has already been standardized, and the 3 neurons is considered as the dimension-reduced factors with almost all the key feature of the daily electricity consumption data lasting for one year. The information of middle hidden-layer is named as core layer for easy understanding. The clustering analysis is performed on the core data. Figure 5 shows the results for the clustering of the core data. The meaning of each value in
the core data is not specific, however, the information from the core layer is adequate to represent the characteristic of the input data. According to the cluster labels for each household, the average daily electricity consumption data is also analysed, which is shown in Figure 6. The proportion and values of each cluster centre are shown in Table 2.

![Figure 5. Clustering results of core data.](image)

![Figure 6. Clustering results of average daily electricity consumption data.](image)

### Table 2. Proportion of each cluster of electricity consumption data.

| Cluster | Proportion | Core layer value |
|---------|------------|------------------|
|         |            | 1    | 2    | 3    |
| 1       | 21.9%      | -34.6| 34.9 | 87.1 |
| 2       | 36.6%      | -8.0 | 6.2  | 19.2 |
| 3       | 41.5%      | -19.9| 17.0 | 50.0 |
3.3 Typical electricity use pattern
Based on the cluster analysis results, the feature indicators of electricity consumption in each cluster are analysed. Figure 6 has already illustrated the average electricity use pattern for each cluster. For further application of integration with renewable energy and traditional energy resource, the detailed pattern of the electricity use shall be explored. In this study, the peak and valley value of daily and monthly electricity use data are analysed. Figure 7 illustrates the peak and valley values of average daily electricity use data for one year of three clusters. Ten households of load profile from each cluster are picked out randomly from the raw dataset and depicted in Figure 8, which shows difference between the single household electricity use pattern and the average pattern in one district. The randomness and variety should be considered during the future analysis. For example, the histogram of peak values of all the households in each cluster is illustrated in Figure 9. The peak value of single household is higher than the peak value of average electricity use, which should be carefully considered during further electricity simulation work. In Figure 10, the monthly electricity use data is illustrated. The energy consumption varies with different season, which infers that the electricity use is strongly affected by the heating and cooling consumption in the case study district. This characteristic is promising to be considered for electricity use model establishment. Finally, the summary of typical electricity use pattern are shown in Table 3.

![Figure 7. Peak value and valley value of average electricity use data for each cluster.](image-url)
Figure 8. Electricity use data of random ten households for each cluster.

Figure 9. Histogram of peak value of households in each cluster.

Figure 10. Average monthly electricity use for each cluster.
Table 3. Summary of typical electricity use pattern for each cluster.

| Cluster | Proportion | Average daily electricity consumption (kWh) | Daily peak value (kWh) | Daily valley value (kWh) | Monthly peak value (kWh) | Monthly valley value (kWh) |
|---------|------------|---------------------------------------------|------------------------|--------------------------|--------------------------|---------------------------|
| 1       | 22%        | 10.9                                        | 25.3                   | 6.1                      | 488.3                    | 213.0                     |
| 2       | 37%        | 2.6                                         | 6.7                    | 1.9                      | 110.3                    | 62.1                      |
| 3       | 41%        | 6.4                                         | 16.8                   | 4.2                      | 274.0                    | 142.7                     |

4 Discussion

In this study, the clustering analysis is conducted on urban electricity use data. To depict the relationship among the different clusters, one of analysis of variance, called Duncan’s test[27, 28] is performed to the cluster results to analyse the correlation among each cluster centre, which shows the significance of different cluster centres and relationship among them. Normally, the variables are listed in the sequence of the average value. Through Duncan’s test, the significance and relationship among different clusters can be analysed, normally according to the average and variance values. Table 4 shows the results for Duncan’s test. The three category under the 0.05 α value shows significant distinction, which verifies the performance of clustering analysis.

Table 4. Table of Duncan’s test results.

| Cluster | N   | α = 0.05 category |
|---------|-----|-------------------|
|         |     | 1    | 2    | 3    |
| 2       | 365 | 2.55 |      |      |
| 3       | 365 |      | 6.41 |      |
| 1       | 365 |      |      | 10.92|
| Significance | 1.000 | 1.000 | 1.000 |

This study is promising to provide a novel insight of the electricity consumption data analysis. From the temporal resolution of month, the seasonal characteristic can also be analysed. Because the clustering is conducted through daily use data, the seasonal features may depict more relation between the daily use and the seasonal characteristic. More correlation work between the electricity use and weather condition can be studied, and detailed energy use model in residential buildings in urban scale is going be established.

Furthermore, there still exists challenging work to improve. The outlier detection is limited when the outlier data accounts for a larger proportion, so the outlier recognition should be improved. Meanwhile, the cluster analysis should also be combined with more specific and detailed information including the household information and weather condition, so that the analysis will be more solid.

5 Conclusion

This study proposes a novel clustering analysis method based on auto-encoder algorithm, to deal with the analysis of daily electricity consumption data. Firstly, the data is pre-processed using a novel method of outlier detection based on k nearest neighbours and local outlier factor. To avoid the “curse of dimensionality” during the clustering analysis, the auto-encoder algorithm is combined with the clustering analysis, which plays a role as dimension reduction. By this means, the electricity data can be analysed and clustered in a more refined temporal resolution, which includes more detailed information. The results of this study are promising for further integration between the residential grid and renewable energy resources, such as the power storage capacity design and evaluation of the solar panels for residential buildings.

Acknowledgement

This study was supported by National Natural Science Foundation of China (grant number 51778321): Research on the quantitative description and simulation methodology of occupant behavior in buildings.
References

[1] Teeraratkul, T., D. O’Neill, and S. Lall, Shape-based approach to household electric load curve clustering and prediction. IEEE Transactions on Smart Grid, 2018. 9(5): p. 5196-5206.

[2] IEA, Electricity. 2018.

[3] Kong, W.C., et al., An Extensible Approach for Non-Intrusive Load Disaggregation With Smart Meter Data. IEEE Transactions on Smart Grid, 2018. 9(4): p. 3362-3372.

[4] Ponocko, J., J.V. Milanovic, and IEEC, Application of Data Analytics for Advanced Demand Profiling of Residential Load using Smart Meter Data. 2017 IEEC Manchester Powertech. 2017.

[5] Godina, R., et al., Model Predictive Control Home Energy Management and Optimization Strategy with Demand Response. Applied Sciences-Basel, 2018. 8(3): p. 19.

[6] Nan, S.B., M. Zhou, and G.Y. Li, Optimal residential community demand response scheduling in smart grid. Applied Energy, 2018. 210: p. 1280-1289.

[7] Setlhaolo, D., X.H. Xia, and J.F. Zhang, Optimal scheduling of household appliances for demand response. Electric Power Systems Research, 2014. 116: p. 24-28.

[8] Vázquez-Canteli, J.R. and Z. Nagy, Reinforcement learning for demand response: A review of algorithms and modeling techniques. Applied energy, 2019. 235: p. 1072-1089.

[9] Stoll, H., Least-cost electric utility planning. New York: John Wiley Sons Inc, 1989.

[10] Valor, E., V. Meneu, and V. Caselles, Daily Air Temperature and Electricity Load in Spain. Journal of Applied Meteorology, 2001. 40(8): p. 1413-1421.

[11] Beccali, M., et al., Forecasting daily urban electric load profiles using artificial neural networks. Energy Conversion and Management, 2004. 45(18-19): p. 2879-2900.

[12] Li, X.L., C.P. Bowers, and T. Schnier, Classification of Energy Consumption in Buildings With Outlier Detection. IEE Transactions on Industrial Electronics, 2010. 57(11): p. 3639-3644.

[13] Wen, L.L., K.L. Zhou, and S.L. Yang, A shape-based clustering method for pattern recognition of residential electricity consumption: Journal of Cleaner Production, 2019. 212: p. 475-488.

[14] Granell, R., C.J. Axon, and D.C.H. Wallom, Clustering disaggregated load profiles using a Dirichlet process mixture model. Energy Conversion and Management, 2015. 92: p. 507-516.

[15] Ryu, S., et al., Customer Load Pattern Analysis using Clustering Techniques. KEPCO Journal on Electric Power and Energy, 2016. 2(1): p. 61-69.

[16] Qi, Y., et al. Load pattern recognition method based on fuzzy clustering and decision tree. in 2017 IEEE Conference on Energy Internet and Energy System Integration (E2I). 2017. IEEE.

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

[18] Liu, H., et al., Efficient outlier detection for high-dimensional data. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017(99): p. 1-11.

[19] Breunig, M.M., et al. LOF: identifying density-based local outliers. in ACM sigmod record. 2000. ACM.

[20] Cover, T.M. and P.E. Hart, Nearest neighbor pattern classification. IEEE transactions on information theory. 1967. 13(1): p. 21-27.

[21] Hartigan, J.A. and M.A. Wong, Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. 28(1): p. 100-108.

[22] Khan, J., et al. PMU data analysis in smart grid using WPD. in IECON 2014-40th Annual Conference of the IEEE Industrial Electronics Society. 2014. IEEE.

[23] Torriti, J., A review of time use models of residential electricity demand. Renewable and Sustainable Energy Reviews, 2014. 37: p. 265-272.

[24] Eichinger, F., et al., A time-series compression technique and its application to the smart grid. The VLDB Journal—The International Journal on Very Large Data Bases, 2015. 24(2): p. 193-218.

[25] Li, C., et al., Building energy consumption prediction: An extreme deep learning approach. Energies, 2017. 10(10): p. 1525.
[26] Rummelhart, D.E., *Learning internal representations by error propagation*. Parallel Distributed Processing: I. Foundations, 1986: p. 318-362.

[27] Bernhardson, C.S., *Type I Error Rates When Multiple Comparison Procedures Follow a Significant F Test of ANOVA*. Biometrics, 1975. 31(1): p. 229-232.

[28] Harter, H.L., *Critical values for Duncan's new multiple range test*. 1960, AEROSPACE RESEARCH LABS WRIGHT-PATTERSON AFB OH.

[29] Pedregosa, F., et al., *Scikit-learn: Machine learning in Python*. Journal of machine learning research, 2011. 12(Oct): p. 2825-2830.

[30] Kingma, D. and L. Ba, *Adam: A Method for Stochastic Optimization*. 2015.