A Two-Step Time-Series Data Clustering Method for Building-Level Load Profile

Preprint

Jiyu Wang, Xiangqi Zhu, and Barry Mather

National Renewable Energy Laboratory

Presented at the 2023 IEEE Power & Energy Society General Meeting
Orlando, Florida
July 16-20, 2023

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

Conference Paper
NREL/CP-5D00-84634
September 2023

Contract No. DE-AC36-08GO28308

A Two-Step Time-Series Data Clustering Method for Building-Level Load Profile

Preprint

Jiyu Wang, Xiangqi Zhu, and Barry Mather

National Renewable Energy Laboratory

Suggested Citation
Wang, Jiyu, Xiangqi Zhu, and Barry Mather. 2023. *A Two-Step Time-Series Data Clustering Method for Building-Level Load Profile: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-5D00-84634. https://www.nrel.gov/docs/fy23osti/84634.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5D00-84634
September 2023

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

A Two-Step Time-Series Data Clustering Method for Building-Level Load Profile

Jiyu Wang, Xiangqi Zhu, Barry Mather
National Renewable Energy Laboratory (NREL), Golden, Colorado.
Xiangqi.Zhu@nrel.gov

***Abstract*—Residential and commercial buildings have huge potential to contribute value to improve grid resilience by participating grid services. To reveal the significant value, it is critical to estimate the grid service capability from these buildings. Unlike the large-scale distributed energy resources such as wind and solar farms, those buildings need to participate grid services in aggregation, not by individual. Therefore, it is important to appropriately group buildings for aggregation. In this paper, we develop a load profile clustering method to classify the building-level load profiles for grid service capability estimation. In our two-step clustering approach, we first calculate the total load consumption for each building, clustering the load profiles based on energy consumption level. Then, we further cluster the load profiles in each energy cluster based on the load shape. The parameter selection for each clustering step is discussed. The proposed method is applied on actual building-level load profiles, and the results have proved the effectiveness of this method.**

***Index Terms*— Advanced metering infrastructure (AMI), energy consumption, load profile cluster, load shape.**

I. Introduction

The integration of distributed energy resources (DERs) in distribution systems has helped to reduce the emissions and pollutions from traditional power resources; however, the increasing penetrations of DERs might cause problems, such as power back-feeding, overvoltage, and large voltage ramps [1]–[2]. Traditional voltage regulation devices—such as load tap changers, voltage regulators, and switched capacitor banks—can help regulate the voltages to some extent [3]. Customer loads, however, which are often located far from the load tap changers and capacitor banks, could still experience voltage problems during the peak load or peak photovoltaic generation period because of the large voltage changes across the lines [4].

With the modernization of smart grids from electric utilities, demand response technologies started being used to provide different types of grid services to eliminate the problems from DER integration. Among the different types of demand resources, buildings consume approximately 75% of electricity in the United States, and thermal controllable load contributes to more than 30% of electricity consumption in buildings [5]. This characteristic provides buildings great potential to participate in demand response activities to provide grid services. To implement demand response in distribution system operation, it is critical for electric utilities to know the available demand response capability at different times of the day. In recent years, electric utilities in the United States have deployed advanced metering infrastructure (AMI) to further modernize the grid [6]. The deployment of AMI on the customer side can enable the collection of smart meter measurements from the grid edge and set a solid foundation for data-driven demand response capability estimation; On the other hand, those grid edge customers need participate grid services in aggregation instead of by individual. The load profiles from different buildings can vary widely because of different building functions, building sizes, and living patterns, making it hard for electric utilities to select appropriate buildings to aggregate together for participating grid services.

An effective solution to this problem is to cluster the load profiles into different groups by load patterns. The load profiles in the same group will have similar characteristics, leading to similar demand response capability. On the other hand, in many utilities, smart meters are not fully deployed to all customers. With the clustering and estimation results from areas with similar demographical information, the load profiles for customers without AMI measurements can be estimated based on the available building information.

In the literature, researchers have developed a variety of methods to cluster the time-series data. The authors in [7] used hierarchical clustering and k-means to cluster the monitored load profiles into groups; however, the monitored load profiles are from the substation level instead of the building level. A clustering method based on ant colony optimization was proposed in [8] to group the load shapes. A load profile clustering method was proposed for load data classification approximation and spectral clustering. Because the shapes used in [8] and [9] are the normalized load profiles, the energy consumption level, which is a critical factor in demand response estimation, was not considered. In [10], a data-driven approach was proposed to cluster the seasonal load profiles with homeowner survey data. This study used the averaged seasonal load profile for clustering instead of the actual daily load consumption. To the best of our knowledge, there are not effective methods in the state-of-the-art to cluster grid edge load profiles.

In this paper, we propose a two-step load profile clustering method to group the load profiles with similar characteristics and patterns. The load profiles are clustered based on their energy consumption level in the first step, then the profiles in each energy group are further clustered based on the load shape. The k-means method is used for the initial clustering, and a multilayer perceptron (MLP) is used to classify the new incoming load profile. The contributions of this paper are two-fold:

- The proposed clustering method consider two critical factors in demand response capability estimation which are energy consumption level and consumption pattern.
- We develop approaches to test and select the k-means and MLP parameters for maximized clustering performance.

The rest of this paper is organized as follows. Section II describes the detailed load profile clustering method. Section III discusses the parameter selections for the proposed clustering method. Section IV presents the results generated from the proposed clustering method. Section V summarizes this paper and discusses the potential future work.

II. Load Profile Clustering Method

This section presents the detailed load profile clustering method we propose.

A. *Load Profile Clustering Steps*

The flowchart of the proposed load profile clustering algorithm is shown in Fig. 1. The raw AMI measurements from smart meters are first preprocessed to exclude the data with missing or bad measurements. This process also extracts the data from the desired period that we want to cluster. Then the extracted load profiles are clustered into *M* groups based on the energy consumption level. After that, the profiles in each energy group are further clustered into N_i clusters. These clusters are the final clustering results, and the total number of clusters can be calculated by:

$$N_{tot} = \sum_{i=1}^{M} N_i \qquad (1)$$

where N_{tot} is the total number of clusters, M is the total number of energy groups, and N_i is the total number of clusters in each energy group.

B. *K-Means Clustering*

The k-means clustering algorithm is one of the most widely used classification methods due to its versatility and applicability to large data sets [11]. It is an unsupervised learning algorithm, so there are no labeled data for this clustering. It performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster [12]. The k-means clustering algorithm aims to divide the time-series data into different clusters with maximized distance among clusters and minimized distances among profiles inside one cluster. In this study, k-means is used for both energy consumption-level clustering and load profile clustering. The Euclidean distance is selected as the reference for clustering, which can be calculated by:

$$d(p,q) = \sqrt{(p-q)^2} \qquad (2)$$

where d is defined as the Euclidean distance between profile p and q.

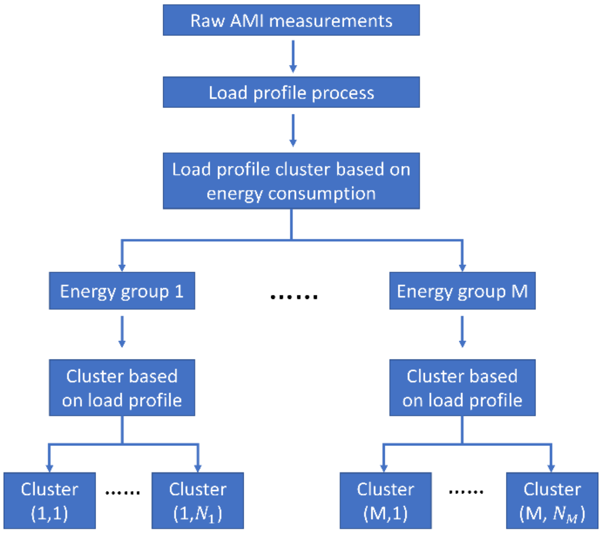


Fig. 1. Flowchart of proposed load profile clustering algorithm

C. *Silhouette Index*

In this study, the Silhouette index is selected to evaluate the performance of the clustering results [13]. The silhouette refers to a method of interpretation and validation of consistency within clusters of data [14]. The silhouette index is a measure of how similar a profile is to its own cluster compared to other clusters, so it serves as an indicator of the clustering performance, including whether the number of clusters is well chosen or whether the appropriate profiles are clustered into one group. The Silhouette index we used in this paper is defined as:

$$s(ii) = \frac{b(ii)-a(ii)}{max\{a(ii),b(ii)\}}, \qquad (3)$$

$$a(ii) = \frac{1}{n_{C_i}-1}\sum_{j\in C_i, ii\neq jj} d(ii,jj), \qquad (4)$$

$$b(ii) = min\left(\frac{1}{n_{C_j}}\sum_{j\in C_j} d(ii,jj)\right), j \neq i, j \in (1,\dots,k) \qquad (5)$$

Here, $\sum_{j\in C_i, ii\neq jj} d(ii,jj)$ represents the sum of the distances between the data point ii and other data points in the same cluster, C_i, and $\sum_{j\in C_j} d(ii,jj)$ represents the sum of the distances between the data point ii and the data points in another cluster, C_j. In this study, each data point represents a load profile in a cluster. The silhouette index ranges from -1 to +1, where a high value indicates that the load profile is well matched to its own cluster and poorly matched to neighboring clusters. While testing different numbers of clusters in k-means, the one with a higher silhouette index and an appropriate number of clusters will be selected.

D. *MLP Classifier*

MLP is a supervised learning algorithm that learns to classify by training on a data set. It is a feed-forward artificial neural network model that maps sets of input data to a set of appropriate outputs [15]. An MLP consists of multiple layers. The nodes of the layers are neurons, with nonlinear activation functions, except for the nodes of the input layer [16]. It relies on this underlying neural network to perform the task of classification. The structure of an MLP is shown in Fig. 2. The number of hidden layers and the number of cells in each layer is predefined. An MLP can handle large amounts of input data and make quick predictions after training. In this study, after the initial k-means clustering, the MLP will be trained using the k-means clustering algorithm to classify the new incoming load profiles.

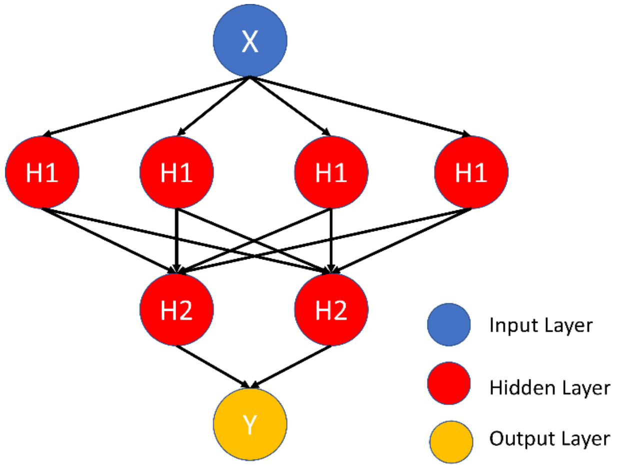


Fig. 2. Structure of the MLP

III. Algorithm Parameter Selection

This section introduces the data used in this study and the tests for different parameters of the k-means and MLP in the load profile clustering method.

A. *Data Preparation*

In this study, realistic daily building-level load profiles are used to test and select parameters of the proposed algorithm. The data resolution is one. hour, so there are 24 data points in each profile. Some example building-level profiles are shown in Fig. 3. It can be observed that this data set contains different types of load shapes with different energy consumption levels.

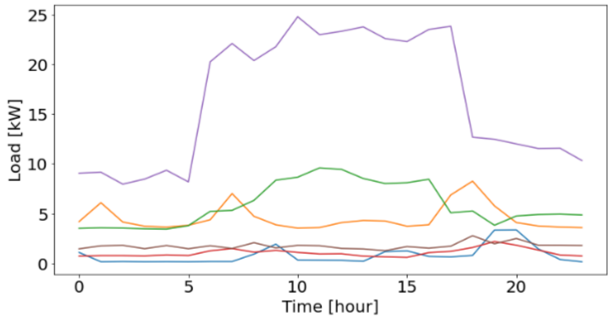


Fig. 3. Example building-level load profiles

B. *K-Means Parameter Selection*

First, the parameters for the k-means are tested for the energy consumption-level clustering. Two thousand daily load profiles are clustered by the k-means clustering algorithm, and different numbers of predefined clusters are tested. Because the large number of profiles and their daily energy ranges from 0 to 5000 kWh, the number of clusters are tested from 8 to 15. For each case, the k-means clustering algorithm is conducted, and the silhouette score is calculated. In addition to the silhouette score, the size of each cluster is considered as an evaluation criterion for the results. We define the clusters with less than 5 profiles as a small cluster, and clusters with more than 50 profiles as a large cluster. Ideally, the results should have more large clusters and less small clusters so each cluster can represent a typical case. Fig. 4 shows that the silhouette score is stable when the number of clusters is between 8 and 12. After comparing the number of small clusters and large clusters, we select the case when there are 9 clusters to be the best result.

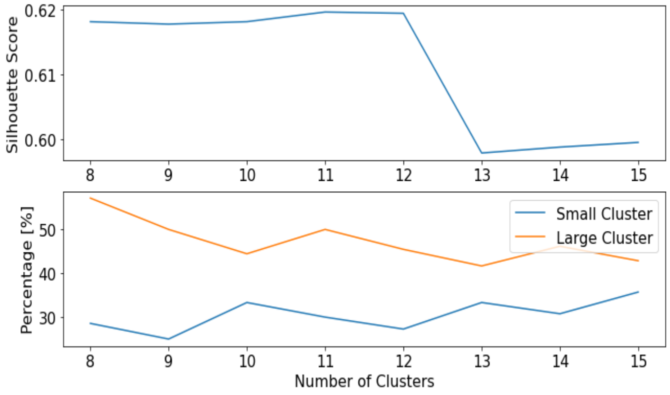


Fig. 4. Parameter test for k-means clustering on energy consumption level

Second, the parameters for the k-means are tested for the load shape clustering. One cluster from the energy consumption-level clustering result is selected as an example. This cluster has 126 load profiles, and the number of clusters are tested from 3 to 10. We define the clusters with less than 3 profiles as a small cluster and clusters with more than 15 profiles as a large cluster. Fig. 5 shows that the best result happens when there are 4 clusters. This case has the highest silhouette score with more large clusters and less small clusters. In the clustering process, this procedure should be repeated for all energy groups to select the best number of clusters for all cases.

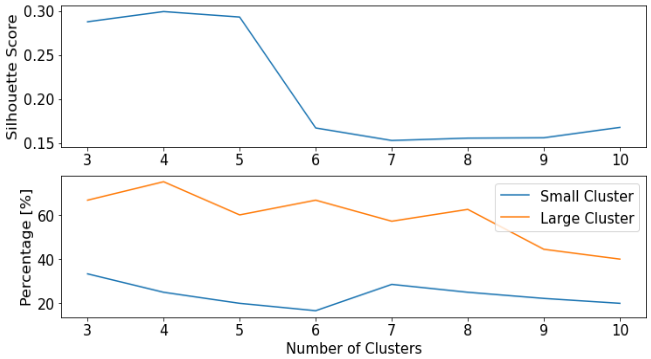


Fig. 5. Parameter test for k-means clustering on load shape

C. MLP Parameter Selection

After the initial clustering results are generated by the k-means algorithm, they can be used as a training data set to train the MLP neural network; therefore, we do not need to use k-means for the clustering again for the new incoming profiles because they can be classified by using MLP. While training the MLP classifier, the number of hidden layers and the number of cells in each layer need to be predefined. For the energy consumption-level MLP classification model, we tested the performance with different parameters, and the results are presented in Table 1. The training data set for the MLP is the result of 2,000 clustered profiles from k-means clustering, and the testing data set contains 5,000 load profiles excluding the training data set. The results show that with different combinations, the highest silhouette score is near 0.61. Considering the computational time, we selected 3 hidden layers and 6 cells in each layer to be the best result.

The MLP parameters for the load shape clustering were also tested. The results are shown in Table 2, and the highest silhouette score happened when there are 5 hidden layers and 4 cells in each layer. Similar to the process in the k-means parameter selection, this procedure can be used to select the best parameters for all cases.

Table 1. Parameter test on MLP model for energy-level classification

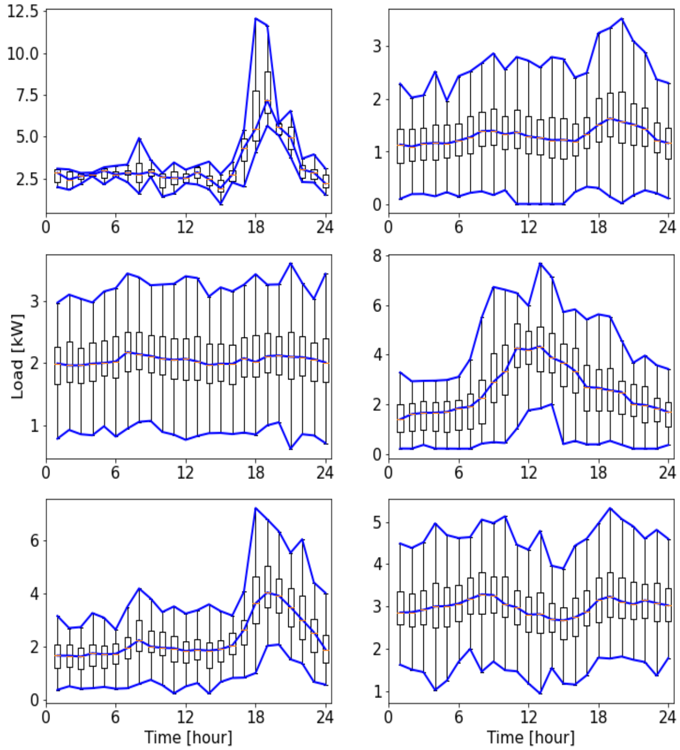
Silhouette score		**Number of hidden layers**			
		3	**5**	**7**	**9**
Number of cells in each layer	**4**	0.49	0.49	0.56	0.58
	5	0.56	0.48	0.56	0.53
	6	0.61	0.56	0.61	0.60
	7	0.60	0.60	0.60	0.57
	8	0.60	0.61	0.59	0.61
	9	0.61	0.61	0.61	0.60

Table 2. Parameter test on MLP model for load shape classification

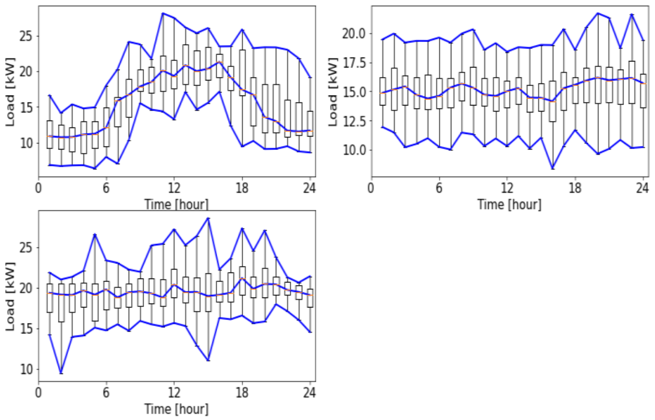
Silhouette score		**Number of hidden layers**			
		3	**5**	**7**	**9**
Number of cells in each layer	**4**	0.36	0.59	0.34	0.58
	5	0.16	0.25	0.25	0.26
	6	0.01	0.25	0.22	0.23
	7	0.23	0.24	0.15	0.19
	8	0.15	0.21	0.17	0.24
	9	0. 11	0.25	0.15	0.28

IV. Clustering Results

This section presents the results from the proposed load profile clustering method. The boxplot of load profiles in some example results are presented in Fig. 6. It can be observed that load profiles with different characteristics are classified into different clusters. For example, the first subplot in Fig. 6 (a) represents load shapes with a giant evening peak load, the fourth subplot is the cluster for load profiles with a higher consumption in daytime, and the fifth subplot contains load profiles with an evening peak. On the other hand, the fourth subplot in Fig. 6 (a) and the first subplot in Fig. 6 (b) have similar shapes. However, they belong to two different clusters because their energy consumption levels are different.



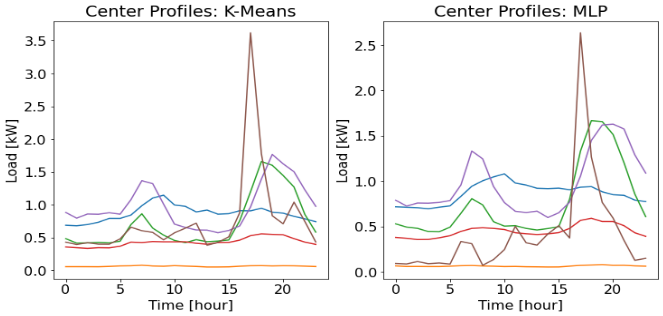
(a) Cluster results from energy group 1



(b) Cluster results from energy group 2

Fig. 6. Clustering results from two example energy group

The center profiles for each cluster in each energy group when using the k-means and MLP are shown in Fig. 7. It can be observed that the center profile of each cluster is very similar for the two methods, which means that the trained MLP model can capture the results from the k-means well and classify the new incoming load profiles.


Center Profiles: K-Means

Center Profiles: MLP

Fig. 7. Center profiles for each cluster

V. Conclusion and Future Work

This paper presents a two-step time-series data clustering method to cluster the building-level load profiles. Real world building-level AMI measurements are used in this study. The load profiles are first clustered into energy groups based on their energy consumption level. Then the load profiles in each energy group are further clustered based on the actual load shape. The k-means algorithm is used to generate the initial clustering results. After that, an MLP model was trained to classify the new incoming data. The algorithm was tested with different parameters, and the one with the best result was selected. The clustering results demonstrate the effectiveness of the proposed method. As part of future work, we will develop methods to quantify the demand response capability in each load profile cluster.

Acknowledgments

This work is authorized in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

[1] F. Ding, H. Padullaparti, M. Baggu, S. Veda, and S. Meor Danial, “Data enhanced hierarchical control to improve distribution voltage with extremely high PV penetration,” *Power & Energy Society General Meeting (PESGM)*, IEEE, 2019.

[2] X. Zhu, J. Wang, N. Lu, N. Samaan, R. Huang, and X. Ke, “A hierarchical VLSM-based demand response strategy for coordinative voltage control between transmission and distribution systems,” *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 4838-4847, Sept. 2019.

[3] M. Chamana and B. H. Chowdhury, “Optimal voltage regulation of distribution networks with cascaded voltage regulators in the presence of high PV penetration,” *IEEE Transactions on Sustainable Energy*, vol. 9, no. 3, pp. 1427-1436, 2018.

[4] H. V. Padullaparti, M. Lwin, and S. Santoso, “Optimal placement of edge-of-grid low-voltage SVCs in real-world distribution circuits,” in *2017 IEEE Workshop on Power Elec. and Power Quality Applications*.

[5] U.S. Energy Information Administration, “Annual energy review,” Washington, D.C., 2010.

[6] U.S. Department of Energy, “Advanced metering infrastructure and customer systems,” Washington, D.C., Tech Rep., Sep. 2016.

[7] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, “Development of low voltage network templates—Part I: Substation clustering and classification," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3036-3044, Nov. 2015, doi: 10.1109/TPWRS.2014.2371474.

[8] G. Chicco, O. Ionel, and R. Porumb, “Electrical load pattern grouping based on centroid model with ant colony clustering,” *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1706-1715, May 2013, doi: 10.1109/TPWRS.2012.2220159.

[9] S. Lin, F. Li, E. Tian, Y. Fu, and D. Li, “Clustering load profiles for demand response applications," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1599-1607, March 2019, doi: 10.1109/TSG.2017.2773573.

[10] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, “Clustering analysis of residential electricity demand profiles,” *Applied Energy*, vol. 135, pp. 461–471, Dec. 2014.

[11] S. Yilmaz, J. Chambers, and M. K. Patel, “Comparison of clustering approaches for domestic electricity load profile characterisation-Implications for demand side management,” *Energy*, vol. 180, 665-677, 2019.

[12] R. Xu and D. Wunsch, *Clustering*. New York: Wiley, 2008.

[13] A. Dudek, “Silhouette index as clustering evaluation tool,” In *Conference of the Section on Classification and Data Analysis of the Polish Statistical Association*. Springer, Cham, 2019.

[14] S. Chaimontree, K. Atkinson, and F. Coenen, “Best clustering configuration metrics: Towards multiagent based clustering,” in *International Conference on Advanced Data Mining and Applications*. Springer, Berlin, Heidelberg, 2010.

[15] K. Sabancı and M. Köklü, “The classification of eye state by using kNN and MLP classification models according to the EEG signals,” 2015.

[16] M. W. Gardner and S. R. Dorling, “Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences,” *Atmospheric Environment*, vol. 32, no. 14-15, 2627-2636, 1998.