Pattern-Preserved Normalization Enabled User Profiling

Fengchao Chen(✉), Lide Zhou, Junni Su, and Xin Zhang

Dongguan Power Supply Bureau, Guangdong Power Grid Corporation, Dongguan 523000, Guangdong, China
csgcfc@126.com

Abstract. The legacy power grid is evolving into a more intelligent grid, and the classical preventive control paradigm is also evolving into a more modern data-driven control paradigm. However, the massive data also poses challenges on the data-driven techniques. In this paper, we focus on the clustering problem in the residential energy sector based on long-term energy consumption data. We employ the classical k-means clustering algorithm and analyze the drawbacks of Min-Max normalization and the disadvantages of utilizing Euclidean distance. We further provide a potential solution, PP-normalization, to solve these issues to achieve a better performance in residential consumption data clustering.

Keywords: Clustering · data-driven · Consumer Analysis

1 Introduction

Data analytic methods are changing every aspect of our daily life. This is also true for the power grid. Thanks to the pervasive sensing technology, the system operator gathers vast amounts of data daily, enabling the smart grid's data-driven control paradigm. However, like many other application domains, the transition from the classical preventive control paradigm to a data-driven control paradigm in the electricity sector is not very smooth. The major drawback is that data-driven methods suffer from data pre-processing problems. In this paper, we take the clustering problem at the end-user level as an example, to explore pattern-preserved-normalization (PP-normalization) enabled clustering method based on long-term energy consumption habits.

1.1 Related Works

We identify two lines of research related to our work. The first is about residential energy consumption clustering in power system, and the other focuses on energy consumption data analysis.

Carmo *et al.* put forward the cluster analysis for residential heat load profiles in [6]. Cui *et al.* propose a clustering oriented pricing scheme based on consumer's

This work was supported in part by the science and technology study of China Southern Power Grid Cooperation (No. GDKJXM20200475).

H. Yang et al. (Eds.): SmartGIFT 2022, LNICST 483, pp. 331–341, 2023.
https://doi.org/10.1007/978-3-031-31733-0_28

load profile [5]. Maqsood *et al.* propose STFT cluster analysis for DC pulsed load monitoring in [10]. Wakeel *et al.* explore K-means based cluster analysis for residential smart meter measurements [2]. Wu *et al.* define a novel predictability matrix from an information theoretic perspective to cluster different kinds of loads [23]. References [3,11,26] give detailed reviews of the clustering approach to electricity load profile characterisation using consumption data. Wang *et al.* utilize the historical load profile data from existing users to conduct effective clustering, which contributes to the load forecasting for a new user in the power system [19]. Clustering for the electric load is one of the most classic tasks in power system, and thus is well investigated. Our paper focuses on the preprocessing analysis before clustering, and includes PP-normalization to improve the clustering performance.

Another research focuses on the energy consumption data processing methods. Pavlo *et al.* provides a review for the approaches to large-scale data analysis [13]. Vandijk reviews some statistical load data processing methods in [18]. Noussan *et al.* give the data analysis on district heating load patterns in [12]. Cui *et al.* examine how data quantity and data quality affect the online dispatch efficiency in [4]. Jin *et al.* put forward a load modeling by finding support vectors of load data from field measurements in [8]. Wu *et al.* propose a data mining approach for spatial modeling in small area load forecast in [24]. Afshar *et al.* make data analysis and short term load forecasting for Iran electricity market in [1]. This line of research pays more attention to general data processing techniques. However, in our paper, we seek to design efficient data processing method, i.e., PP-normalization, for load profile clustering.

1.2 Paper Organization

The remainder of the paper is organized as follows. Section 2 discusses the fundamental features of the data and the idea of preprocessing. We also explain the dropping majority of data and the primary process of clustering detection. Then after preprocessing, Sect. 3 introduces how to employ the classical K-means clustering algorithm to accomplish the task. Furthermore, Sect. 4 attempts to improve the performance of K-means clustering using PP normalization. Section 5 conducts a numerical study to highlight the proposed improved k-means clustering performance and the proposed method's limitations. Finally, concluding remarks are delivered in Sect. 6.

2 Overview of the Dataset and Preprocessing

In this section, we first introduce some basic features of the dataset, which could be used to suggest how to conduct valid data preprocessing. In particular, given the data of 1-minute granularity, we check the monthly and seasonal patterns inherent in the data. We believe such information will help us choose the strategy to deal with missing values, select or drop out samples, and determine the efficient period for modeling and analysis.

2.1 Dataset Overview

The dataset used in this paper is the Pecan Street energy consumption dataset [14]. We regard continuous "0" observations (say more than two days) and negative observations as ineffective. To this end, only 182 out of 342 samples have complete effective data (the number standard is effective observations should be larger than 170000). Another 62 samples contain effective observations of 3 months. For the remaining 98 samples, four total samples do not contain any observation (all 0's) and are thus being removed from the analysis. For the remaining data, we regard the samples containing continuous "0" observations for more than seven days. We will propose a strategy to deal with these missing values based on the other fundamental features of the dataset.

2.2 Seasonality Check

We conduct the daily and weekly seasonality using the QS-test, a variant of the Ljung-Box test that computes the test statistic based on seasonal lags of the series. The QS-test requires that the samples are independently distributed and can be modeled using the SARIMA model with a period of l. Therefore, the partial auto-correlation between samples with lag l should be rather large. That is, the null hypothesis of non-seasonality of l will be rejected under a sufficiently large QS value.

Note that if the data sequence has significant daily seasonality, it should also have significant weekly seasonality since a week is an integer multiple of days. Therefore, time series decomposition with daily seasonality is required to capture the actual weekly seasonality. All the samples are tested with all the effective observation data. The empirical results are consistent with common sense: Most samples show significant daily and weekly seasonality. A few samples fail to pass the daily seasonality test and are removed from further analysis.

Most of the samples with daily seasonality will also appear to have weekly seasonality after removing the daily seasonality term. However, the weekly seasonality property is less robust. The weekly seasonality may fail if the QS test is only applied on a partition of the effective observations - say, two continuous months. Therefore, in this study, we will mainly focus on the day seasonality, and the word seasonality will be used instead for convenience.

Since clustering discussed in this study is mainly for providing support to long-term abnormal energy consumer detection. This relatively robust daily seasonality property will allow us to use the aggregate daily consumption profile to represent the consumption profile of the consumer.

2.3 Monthly Trend

Many other factors will affect the energy consumption profile, e.g., outside temperature. To form an effective strategy to fill in the missing data, it will be better if sample trends are straightforward. To detect the monthly trend of the samples, we adopt simple linear regression models for each household and each month.

We can also test whether this linear regression model is a good candidate. A simple approach is to treat the parameter estimation as constant (the actual parameter value) and then apply the t-test. From the empirical test result on the dataset, we observe that the monthly trends are different even for the same household for other months. Therefore, using the sample of the current month to extrapolate data from other months may be unreasonable.

2.4 Preprocessing Strategy

Based on the features discussed above, we choose not to fill in the missing data due to the different trends between months, which means the observations of the current month may not be a good estimation start point for another month. Also, samples without daily seasonality are removed from analysis due to the relatively small sample size. It will be hard to get a reliable result based on about ten samples. Instead, observations from 1-31 18:01 to 3-04 01:18 are selected to be used in this study since almost all the samples have continuous and complete observations over this time interval. Then by kicking out another four samples due to the lack of data over this period, we finally selected 317 samples, each with more than 40,000 observations. Though we remove many observations from the analysis, the remaining observations are still sufficient to construct a robust analysis. However, since we only do clustering over generally one month (February), the difference between the consumption profile of months should not be neglected.

The aggregate data of each day is used to simulate the consumption profile of consumers. Moreover, due to the limitation of computational power, we select only to use the aggregate time on each hour. Therefore, we begin with using a vector $v \in \mathbb{R}^{24}$ to capture the consumption profile of each sample, and we will use the word feature vector to represent this vector.

3 Classical K-Means Clustering

To cluster the consumption profiles, the measurement of similarity should be declared. Distance-based measurements could be reasonable approaches since we expect similar consumption profiles to simultaneously have minor differences between the consumption volumes. A considerable distance between feature vectors generally means significant differences exist in each element. Though this argument may not be valid sometimes, since a considerable distance may also cause a significant difference in some specific element pairs, distance measurement can provide insight into determining the similarity. Thus, we choose to use Euclidean distance as the starting point, and it will be natural first to try K-means, one of the popular distance-based clustering methods.

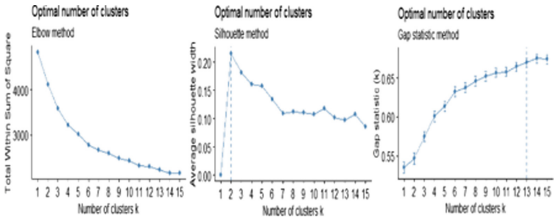


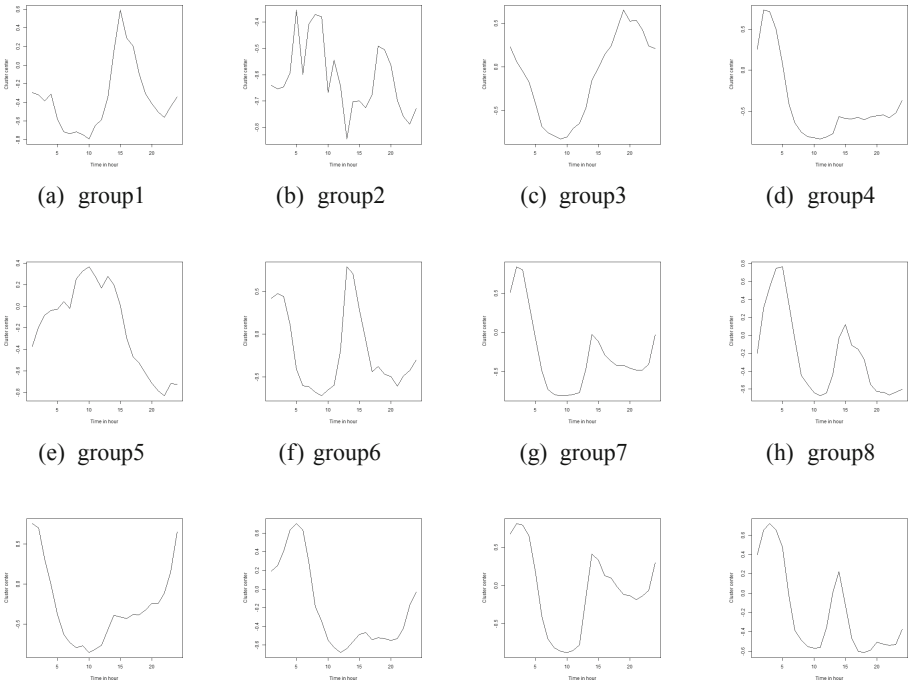
Fig. 1. Cluster number selection base on three different methods.

3.1 Initial Attempts

Recall the goal that we want to form a clustering model to help improve long-term abnormal energy consumer detection. This drives us to do clustering only based on the pattern of sample consumption profile but not the consumption volume of the samples. Therefore, a suitable normalization method should be applied to preserve the pattern and scale the samples in the same range. We will first try Min-Max normalization in the range 0 to 1. This is generally a promising approach for normalization tasks but still has some disadvantages that we will discuss later. At this point, we will directly try Min-Max normalization and then turn to K-means++ clustering, which is just a modified K-means that spreads out the k initial points as much as possible to obtain a more robust clustering outcome compared with the traditional K-means algorithm. In this study, K-means is the same as K-means++ for convenience of report writing.

Since K-means clustering requires an initial assignment of the number of clusters, the Elbow, Silhouette, and gap statistic methods are applied and serve as a reference to the final cluster number selection (see Fig. 1 for a comparison). The reason for applying these three methods is because the Silhouette method focuses more on the distance between clusters but not the difference between clusters; the gap statistic method cares more about the in-cluster performance, and the Elbow method generally performs better on considering both in-cluster and out-cluster performance. We can depend on all the results of these methods to get a better decision on the number of clusters.

For this simple attempt, we can find these methods provide different suggestions on the cluster number selection. Moreover, though the default function result suggests 13 clusters for the gap statistic curve, there is no significant global maximum with $k \leq 15$. The result of the gap statistic method remains doubtful. For the Silhouette method, which suggests 2 clusters are the best, its result is counter-intuition. Since it is hard for the result of these three methods to reach an agreement, we decide to try from $k = 13$. The idea of this decision is based on the intuition that if the K-means algorithm cannot perform well on a large number of k, then it will even perform worse when the k is small. Thus, choosing a relatively large k at the beginning is reasonable. This simple attempt's cluster centers are shown in Fig. 2.



(a) group1 (b) group2 (c) group3 (d) group4

(e) group5 (f) group6 (g) group7 (h) group8

(i) group9 (j) group10 (k) group11 (l) group12

Fig. 2. Cluster centers of Min-Max + K-means approach. Cannot find significantly different between center of group 5 and center of group 10.

3.2 Drawbacks of Classical K-Means Clustering

By looking into the cluster centers shown in Fig. 2, it can be found that the center of group 5 is close to the center of group 9. The Euclidean distance between these two centers is even less than some of the distance between samples in groups 5 and 9 and their centers. We doubt whether these two groups show no significant difference from each other. However, if we carefully check the in-cluster samples, more than one possible pattern has been grouped into the same cluster. This phenomenon suggests increasing the number of clusters. However, if we change the number of clusters from k = 13 to k = 17, we can still find samples with possible different patterns within the same cluster, but some clusters will only contain a few samples (less than or equal to 3) at this time. Therefore, this phenomenon is not caused by the excessive or insufficient number of clusters.

As mentioned in the introduction paragraph of Sect. 3.1, the value of Euclidean distance cannot guarantee the number of significant different elements between two vectors. In other words, the relatively small Euclidean distance does not mean the two feature vectors are the same but only close to each other in high dimensional space. For example, consider a consumption pattern only with one peak at 19:00–20:00. If there is another consumption pattern whose con-

sumption peaks appear at 7:00–8:00 and 19:00–20:00, and these two consumption patterns are almost the same except 7:00–8:00. By intuition, samples with these two consumption patterns should not be grouped. However, the difference in Euclidean distance caused by the "morning consumption peak" may not be significant enough to separate these two samples.

4 Proposed Improvements

One possible solution to the problem is to make a dimensional reduction. Unfortunately, traditional dimensional reduction methods like PCA and T-SNE do not help much. We want to improve the in-cluster performance from other approaches. Another problem that needs to be concern is the Min-Max normalization. Though Min-Max normalization is an excellent approach to maintaining the consumption pattern of samples, it cannot remove noises in the feature vector. The noises discussed here are small random fluctuations within a relatively constant consumption period. These noises may cause extra distance between two series with the same pattern. Besides, Min-Max normalization probably will enlarge the within-sample difference of relatively flat samples since it will always normalize the curve to a pre-defined region. To avoid the drawbacks of Min-Max normalization, we constructed a modified Min-Max method so-called Pattern-Preserved normalization (PP-normalization).

4.1 PP-Normalization

The basic idea behind the PP-normalization is that: We can consider the PP-normalization as a mapping function f,

$$f(\mathbf{x}) = \mathbf{y} \tag{1}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and each element in $\mathbf{y}$ is bounded by $[L, U]$ which is determined by input vector $\mathbf{x}$.

To determine the in-sample fluctuation level of the data $\mathbf{x}$, we set uniform distribution as reference. Let $\mathbf{x}_i, i = 1, 2, \cdots, n$ be the original series data and $\mathbf{x}_{(i)}$ is the ordered data and $\mathbf{x}'_{(i)}$ be series of ordered data from uniform distribution. Assume $\mathbf{x}_{(i)}$ follows uniform distribution, then $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(n)}$ can be estimated as:

$$\hat{\mathbf{x}}_{(1)} = q_{0.5} - (q_{0.75} - q_{0.25}) \tag{2}$$

$$\hat{\mathbf{x}}_{(n)} = q_{0.5} + (q_{0.75} - q_{0.25}) \tag{3}$$

where q_p is p quantile of series data $\mathbf{x}_{(i)}$. If the series has significant consumption peak time and consumption valley time, the whole series will be more compact to the median compare to series $\mathbf{x}'_{(i)}$. Therefore, pick control parameter as θ, general range parameter α, β, L, U are defined as:

$$L = \begin{cases} \alpha, & \min(\mathbf{x}_{(1)}, \theta\hat{\mathbf{x}}_{(1)}) = \mathbf{x}_{(1)} \\ MinMax(\mathbf{x}_{(i)}, \alpha, \beta), & \min(\mathbf{x}_{(1)}, \theta\hat{\mathbf{x}}_{(1)}) = \theta\hat{\mathbf{x}}_{(1)} \end{cases} \tag{4}$$

$$U = \begin{cases} \beta, & \min(\mathbf{x}_{(n)}, \theta\hat{\mathbf{x}}_{(n)}) = \mathbf{x}_{(n)} \\ MinMax(\mathbf{x}_{(n)}, \alpha, \beta), & \min(\mathbf{x}_{(n)}, \theta\hat{\mathbf{x}}_{(n)}) = \theta\hat{\mathbf{x}}_{(n)} \end{cases} \tag{5}$$

4.2 Noise Removing

Let $\bar{x}$ to be the average of the sample $\mathbf{x}$. After choosing an reference example denoted as $\tilde{x}$, we want to remove the sample noise by applying the following idea,

$$\mathbf{x} = \begin{cases} MinMax(\tilde{x}, \alpha, \beta), & |\mathbf{x} - \tilde{x}| \leq \gamma\bar{x} \\ MinMax(\mathbf{x}, \alpha, \beta), & Otherwise \end{cases} \tag{6}$$

That means, if the difference between the real sample and the reference sample is less than the predefined threshold $\gamma\bar{x}$, we will utilize the reference value to substitute the original value. Otherwise, we will keep the original value. Clearly, through this processing, it is easily to remove the useless noisy information, and will contribute to the subsequent clustering.

5 Numerical Studies

Select the number of clusters to be 12 and apply K-means clustering directly. The examples of normalization are shown in Fig. 3.

Indeed, the result of K-means based on PP-normalization is only slightly improved from the simple attempt. Compared to the K-means based on Min-Max normalization, PP-normalization improves the in-sample performance of K-means. By looking into each cluster, the number of samples without obvious different patterns decreases, though such a phenomenon still exists. However, PP-normalization cannot solve the problem that the cluster center of Group 4 and Group 9 are two close to each other. However, decreasing the number of clusters will result in a worse performance.

Moreover, it can be found obviously in the sample plot of Group 0 and Group 10 that one of the samples with a consumption peak more than 5 h apart from another one has been grouped into the same cluster. Intuitively speaking, if the difference in consumption peaks of two samples is only 1 h, their consumption pattern is probably the same. However, if their consumption peak is 5 h apart, it will be hopeless that they hold the same consumption pattern.

This miss-clustering result is also due to the Euclidean distance-based similarity measurement. No matter what kind of distance and L_p norm for any P, this problem will still exist since the traditional distance calculation only focuses on the difference between two corresponding elements in the vector but does not care about the difference between a specific element and its front and back.

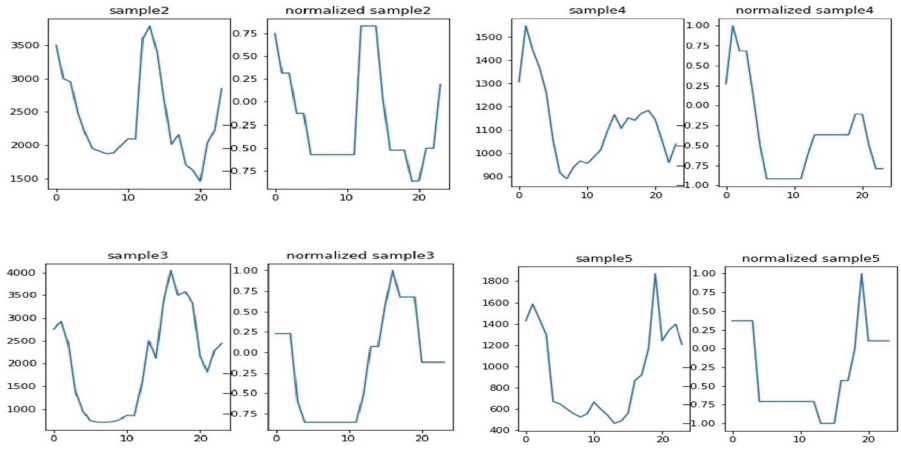
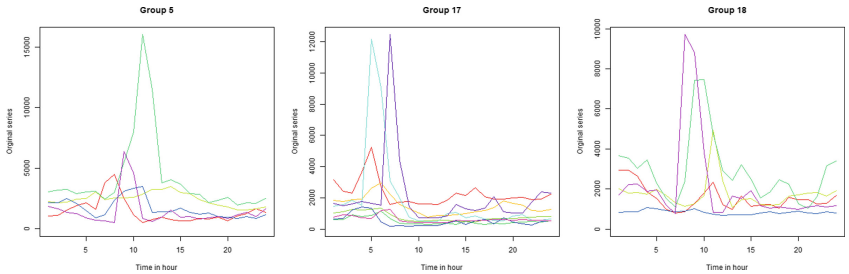


Fig. 3. PP-normalization result.

We are solving the problem caused by using Euclidean distance as a similarity measurement. We could use the alignment distance for K-mean clustering. Through numerical studies, we found that alignment distance provides a solution to cluster two samples with small peak shifts together. Now a method to separate samples with large peak shifts is needed. For simplicity, we will start by trying directly give a penalty on the large peak shift based on the Euclidean distance. By the intuition of patterns of the samples, no more than two significant consumption peaks occur in each sample. Also, no more than two consumption valleys for each sample. Let p_1, p_2 denote the index (time) of the peaks, and v_1, v_2 denote the index of valleys. The feature vector v is modified to a new feature vector $v' \in \mathbb{R}^{28}$. Since $v \in [-1, 1]$, it automatically assigns a high penalty on different peak and valley times. By applying K-means clustering on the new feature vector v', the result of several distinct groups discussed above are shown as follows:

By applying K-means based on the new feature vector, the gap statistic method and Elbow method both suggest a large number of clusters. After parameter selection, we finally choose k = 25 to obtain a relatively robust result. Different from the previous attempt, increasing the number of clusters will cause several clusters only to contain 1 sample. By selecting k = 25, we do not observe this phenomenon for K-means based on the new feature vector. For the in-cluster performance, K-means clustering based on new feature vectors is sensitive to the peak shift of samples. As shown in Fig. 4, cluster 0 discussed in Sect. 3.3 is generally separated into three clusters with close peaks.



(a) group5 (b) group17 (c) group18

Fig. 4. Cluster 0 in PP-normalization + K-means and its corresponding clusters based on new feature vector (cluster 5, 17, and 18).

6 Conclusions

This paper takes the clustering task for long-term energy consumption data as an example to explore the preprocessing techniques in data-driven methods. Through analyzing the disadvantages of classic Min-Max normalization using Euclidean distance in consumption clustering, we propose a novel PP-normalization idea to improve the clustering performance. Numerical studies shows that our PP-normalization is effective in energy consumption clustering task. We believe PP-normalization is a very promising data pre-processing technique, which could enable many other data-driven tasks, such as LMP prediction [7], convex hull pricing analysis [15,16], demand response program design [21,22], and storage control [9,17,20,25].

References

1. Afshar, K., Bigdeli, N.: Data analysis and short term load forecasting in Iran electricity market using singular spectral analysis (SSA). Energy **36**(5), 2620–2627 (2011)
2. Al-Wakeel, A., Wu, J.: K-means based cluster analysis of residential smart meter measurements. Energy Proc. **88**, 754–760 (2016)
3. Chicco, G.: Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy **42**(1), 68–80 (2012)
4. Cui, J., Gu, N., Wu, C.: Quantity or quality? Data enabled online energy dispatch. In: Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers, pp. 606–612 (2021)
5. Cui, J., Wang, H., Wu, C., Yu, Y.: Vulnerability analysis for data driven pricing schemes. In: 2020 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5. IEEE (2020)
6. Do Carmo, C.M.R., Christensen, T.H.: Cluster analysis of residential heat load profiles and the role of technical and household characteristics. Energy Build. **125**, 171–180 (2016)

7. Jiang, W.J., Cao, S., Wu, C.: LMP prediction with incomplete information. In: 2022 IEEE Power & Energy Society General Meeting, pp. 1–5. IEEE (2022)
8. Jin, M., Renmu, H., Hill, D.J.: Load modeling by finding support vectors of load data from field measurements. IEEE Trans. Power Syst. **21**(2), 726–735 (2006)
9. Kalathil, D., Wu, C., Poolla, K., Varaiya, P.: The sharing economy for the electricity storage. IEEE Trans. Smart Grid **10**(1), 556–567 (2017)
10. Maqsood, A., Oslebo, D., Corzine, K., Parsa, L., Ma, Y.: STFT cluster analysis for DC pulsed load monitoring and fault detection on naval shipboard power systems. IEEE Trans. Transp. Electrif. **6**(2), 821–831 (2020)
11. McLoughlin, F., Duffy, A., Conlon, M.: A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl. Energy **141**, 190–199 (2015)
12. Noussan, M., Jarre, M., Poggio, A.: Real operation data analysis on district heating load patterns. Energy **129**, 70–78 (2017)
13. Pavlo, A., et al.: A comparison of approaches to large-scale data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, pp. 165–178 (2009)
14. Street, P.: Pecan street dataport. Website (2016). https://dataport.pecanstreet.org
15. Sun, J., Gu, N., Wu, C.: Market power in convex hull pricing. In: Proceedings of the Eleventh ACM International Conference on Future Energy Systems, pp. 398–400 (2020)
16. Sun, J., Wu, C.: Temporal vulnerability assessment for convex hull pricing. In: Proceedings of the Twelfth ACM International Conference on Future Energy Systems, pp. 124–136 (2021)
17. Sun, J., Zhang, Y., Yu, Y., Wu, C.: Storage control for carbon emission reduction: opportunities and challenges. In: 2020 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5. IEEE (2020)
18. Vandijk, G.: Statistical load data processing. NASA. Langley Res. Center Advanced Approaches to Fatigue Evaluation (1972)
19. Wang, Q., Chen, Z., Wu, C.: Clustering enabled few-shot load forecasting. In: 2021 IEEE Sustainable Power and Energy Conference (iSPEC), pp. 2417–2424. IEEE (2021)
20. Wu, C., Kalathil, D., Poolla, K., Varaiya, P.: Sharing electricity storage. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 813–820. IEEE (2016)
21. Wu, C., Mohsenian-Rad, H., Huang, J.: Vehicle-to-grid systems: ancillary services and communications. José MR Gonçalves **129** (2012)
22. Wu, C., Mohsenian-Rad, H., Huang, J., Wang, A.Y.: Demand side management for wind power integration in microgrid using dynamic potential game theory. In: 2011 IEEE GLOBECOM Workshops (GC Wkshps), pp. 1199–1204. IEEE (2011)
23. Wu, C., Tang, W., Poolla, K., Rajagopal, R.: Predictability, constancy and contingency in electric load profiles. In: 2016 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 662–667. IEEE (2016)
24. Wu, H.C., Lu, C.N.: A data mining approach for spatial modeling in small area load forecast. IEEE Trans. Power Syst. **17**(2), 516–521 (2002)
25. Wu, J., Wang, Z., Wu, C., Wang, K., Yu, Y.: A data-driven storage control framework for dynamic pricing. IEEE Trans. Smart Grid **12**(1), 737–750 (2020)
26. Yang, S.L., Shen, C., et al.: A review of electric load classification in smart grid environment. Renew. Sustain. Energy Rev. **24**, 103–110 (2013)