A New Similarity Measurement Method for the Power Load Curves Analysis

Xin Ning[1], Ke Zhu[1], Yuanshi Deng[1], Rui Zhang[1], Qi Chen[2(✉)], and Zhong Li[2]

[1] State Grid Sichuan Electric Power Research Institute, Chengdu 610072, Sichuan, China
[2] North China Electric Power University, Baoding 071003, Hebei, China

Abstract. In order to improve the quality of the power load curves similarity measurement, a new similarity measurement method based on Euclidean distance is proposed in this paper . Among the commonly used similarity measurement methods, Euclidean distance is not sensitive to the fluctuation of the load curves, which results in the lack of shape measurement capability. For the numerical distribution on the timeline is not concerned, the dynamic time warping (DTW) distance is not accord with the requirement of the power system load analysis. Focus on those issues, the proposed method introduced a correction factor that contains the dynamic characteristics of the numerical difference between two power load curves without compromising time warping. The advantages and performance of the proposed method are evaluated by similarity computing and clustering analysis. As shown in the experimental results of similarity computing, the proposed method performs as same as ED and DTW, but the calculating time is less than DTW. In the clustering analysis, it also decreases the calculating time from 3.9 s to 0.595 s compared with DTW and shows better clustering effect that make the Davies-Bouldin index from 0.438 for ED and 0.325 for DTW to 0.249.

Keywords: Power load curves · Similarity measurement · Euclidean distance · Cluster analysis

1 Introduction

A large number of valuable information contained in the power load curves has produced a great interest in power load analysis in recent years. The power load curves are always adopted to describe the variation of power load with time and reflect the characteristics and rules of users' electricity consumption. Generally, similarity measurement of the power load curves is popular to be used in load analysis [1]. As the basic of load forecasting and load pattern recognition [2], similarity measurement had various applications. Francisco et al. [3] used pattern sequence similarity for energy time series forecasting. Zhou and Li et al. [4] proposed a source-load-storage coordinated optimization model with source-load similarity. Singh [5] proposed a work on energy time series for users' behavioral analytics and energy consumption forecasting. Nagi, et al. [6] using support vector machines to detected nontechnical loss for metered customers in power utility.

M. Cheng et al. (Eds.): SmartGIFT 2020, LNICST 373, pp. 3–13, 2021.
https://doi.org/10.1007/978-3-030-73562-3_1

There are a lot of methods have been designed and proposed [7–14] to increase the quality of similarity measurement. Euclidean distance (ED) [7] and dynamic time warping distance (DTW) [8] are two typical examples. However, those two methods shown no applicable to varying degrees. Specifically, existing ED cannot recognize the change of curve shape. Besides, the calculating process of existing DTW made the timeline not aligned and had low efficiency, which is not suitable for the power system analysis. In addition to these two methods, Li and Yuan [9] considered the characteristic of vector difference and proposed a method for similarity estimate but only tested in graphic data set. Jia et al. [10] used imporved clustering method to evaluated the shape of the load curves. Yu et al. [11] presented a general guideline to find a better distance measure for similarity estimation based on statistical analysis of distribution models. Teeraratkul et al. proposed a shape-based approach to household electric load curves clustering and prediction based on DTW [12]. Although above methods [10–12] have achieved some results in load analysis, they all have complex calculation procedure that can't meet the requirement of load analysis for efficiency. Then a new similarity distance calculation method is urgently needed to be put forward.

This paper enumerates two classical distance-based similarity measurement methods, i.e. Euclidean distance and dynamic time warping distance, firstly. Then the comparison between the applicable scope, advantages and disadvantages are shown in detail. Aiming at the deficiency of similarity measurement and the characteristics of power load curves, this paper proposed a similarity measurement method, i.e. modified Euclidean distance (MED). Then the influence of the proposed improved algorithm in power load curve similarity measurement and its application effect in power load curve clustering are analyzed in the design experiment. The results are compared with Euclidean distance and dynamic time warping distance, respectively. The proposed MED method ensures accurate similarity calculation and increases the clustering results effectively.

The rest of the paper is organized in the following way. Existing ED method and dynamic time warping distance (DTW) are shown in Section II. In Section III, the proposed method, which is optimized by improvement factor is introduced with detailed analysis. Experimental validation of the proposed algorithm is given in Section IV to compare with existing ED and DTW. In Section V, conclusion is given to summarize the advantages of the proposed method.

2 Background

2.1 The Similarity of the Power Load Curves

For the power load curves, there are two factors influence the degree of similarity. The one factor is numerical similarity, which is reflected by the distance of two power load curves on the same sampling instant. The other factor, shape similarity, is the consistency degree of the dynamic changes of two load curves during the whole sampling time. As shown in Fig. 1, the distance between point a and point b reflect the power load curves numerical similarity. The shape similarity can be measured by comparing the shapes and variation trend of each curves.

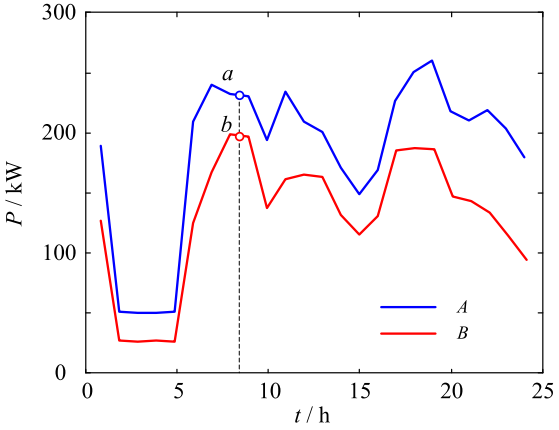

Fig. 1. Illustration diagram of the power load curves similarity.

2.2 Classic Similarity Measurement Methods

Among all the similarity measurement methods, Euclidean distance and dynamic time warping (DTW) distance are used commonly.

Euclidean Distance (ED). Suppose there are two load curves with length n, X ($X = x_1$, x_2, …, x_i, …, x_n) and Y ($Y = y_1, y_2, \ldots, y_i, \ldots, y_n$). To compare the similarity between X and Y, the Minkowski distance is used as:

$$D(X, Y) = \sqrt[r]{\sum_{i=1}^{n} (x_i - y_i)^r} \tag{1}$$

where r ($r \geq 1$) is a distance coefficient of Minkowski distance. Different values of r result in different similarity measurement methods. Euclidean distance is a special Minkowski distance when $r = 2$. So, the Euclidean distance between X and Y is:

$$D_{ED}(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

The process of Euclidean distance works is a strictly point-to-point calculation and simple to implement. But on account of using the square of the difference between x_i and y_i, Euclidean distance ignores the sign of $(x_i - y_i)$. Therefore, Euclidean distance is not sensitive to the fluctuation of the load curves, which results in the lack of shape measurement capability.

Figure 2 shows three daily power load curves A, B and C of different users on the same date, where each curve has 24 sampling points (take samples once an hour). Calculating the Euclidean distance between A and B, A and C, the results are: D_{ED}(A, B) = 610.6907, D_{ED}(A, C) = 965.0453, D_{ED}(A, B) < D_{ED}(A, C).

The calculations results indicate that the load curve A is more similar with the load curve B. But load curve A is more similar with load curve C when comparing the shape and variation trend of three load curves as shown in Fig. 2.

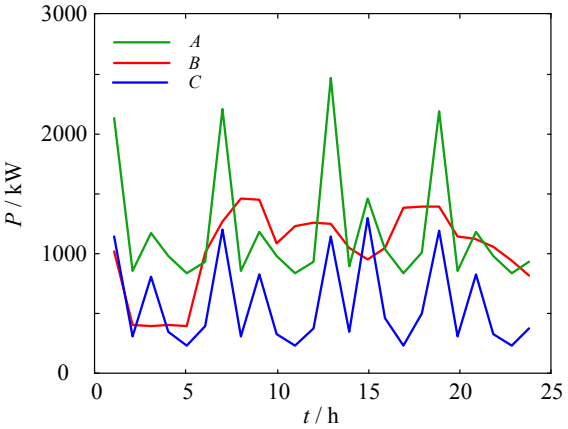

Fig. 2. Comparison of different load curves using Euclidean diatance.

Dynamic Time Warping (DTW) Distance. DTW based on the thinking of dynamic programming. This method searches for a shortest path through time stretching or warping, which makes the distance between two load curves is minimum. For time series: $X = \{x_1, x_2, \ldots, x_i\}$and $Y = \{y_1, y_2, \ldots, y_j\}$. The definition of DTW is:

$$D_{\mathrm{DTW}}(X, Y) = \begin{cases} 0, i = j = 0; \\ \infty, i = 0 \text{ or } j = 0; \\ D_b(x_1, y_1) + \min \begin{cases} D_{\mathrm{DTW}}(\mathrm{Re}st(X), \mathrm{Re}st(Y)), \\ D_{\mathrm{DTW}}(\mathrm{Re}st(X), Y), \\ D_{\mathrm{DTW}}(X, \mathrm{Re}st(Y)) \end{cases} \end{cases} \tag{3}$$

where i and j are the length of the load curves X and Y respectively, Rest(X) = $\{x_2, \ldots, x_i\}$, Rest(Y) = $\{y_2, \ldots, y_j\}$. D_b is basic distance between x_i and y_j, usually be used in Euclidean distance.

Two points connected by a black line in Fig. 3 are the similarities of the load curves A and B. When several points correspond to one point, the time warping will occur. The sum of the distances of all the similarities is DTW distance.

In spite of DTW could recognize the load curves' shape features to some extent from Fig. 3, it's not a good choice for the power system load analysis. Because the timeline of the load curves is not fully aligned, DTW neglected the numerical distribution on the time line. This measure may cause over warping sometimes and not reflect the real change of the power load. In addition, DTW also have the shortcoming of high complexity.

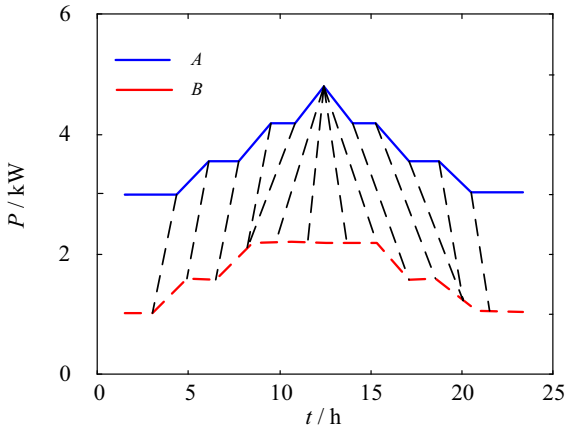

Fig. 3. Illustration diagram of DTW.

3 A New Similarity Measurement Method for the Power Load Curves

Generally, the similarity measurement method should be designed with specific problems. In the power system load analysis, both numerical similarity and shape similarity of the power load curves need to be concerned by similarity measurement method. Even though Euclidean distance and DTW are competitive for many fields, they have some shortcomings as a method to measure the similarity of the power load curves. Each of them could not consider numerical and shape similarity at the same time and the calculation of DTW has no advantage in speed.

Focus on these issues, a modified Euclidean distance (MED) for the power load curves is proposed as the new similarity measurement method. The MED between the load curves X ($X = x_1, x_2, \ldots, x_i, \ldots, x_n$) and Y ($Y = y_1, y_2, \ldots, y_i, \ldots, y_n$) is defined as:

$$D_{\mathrm{MED}}(X, Y) = D_{\mathrm{ED}}(X, Y) + \delta \tag{4}$$

where δ is a shape correction factor defined as:

$$\delta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i - \mu)^2} \tag{5}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i) \tag{6}$$

where $(x_i - y_i)$ is the difference between the corresponding point of two load curves on the same sampling instant.

The definition of MED shows that Euclidean distance can be made having shape measurement capability by introducing a correction factor and not compromising time warping. The correction factor δ is standard deviation of $(x_i - y_i)$, which represents the

fluctuation degree of $(x_i - y_i)$. The more similar the two load curves are, the smaller the δ is. δ offsets the shortcoming of Euclidean distance that it couldn't consider the sign of $(x_i - y_i)$. The algorithm of MED is shown as follows, it can be easily implemented.

The algorithm of MED
Input: Daily load curves L_1 and L_2.
Output: Calculated distance.
Step 1: The mean value of the difference between L_1 and L_2 is calculated as (6).
Step 2: Calculating the distance of L_1 and L_2 by ED as ED (L_1, L_2).
Step 3: The parameter δ, which is the factor used to show the change trend of load curve is calculated as (5) ;
Step 4: Output the value combing ED (L_1, L_2) and Sig together as (4);

4 Experiment

To verify the feasibility and effectiveness of the proposed method, three described methods, i.e. ED, DTW and MED are applied to the load curves with similarity measurement and clustering. Then the comparisons between them are shown with detailed analysis. In this part, the experiment results of similarity measurement are shown at first. Then the clustering results of K-means are shown with analysis of Davies-Bouldin index (DBI), calculating time and iterations. Compile software: MatlabR2016a, operating system: Windows10, CPU: Inter(R) Core(TM) i3-8100, dominant frequency: 3.6GHz, internal storage: 16G, hard drive capacity:1T.

4.1 Experiment Data

Experiment data in this part is commercial and residential hourly load profiles for all TMY3 locations in the United States from Open EI [15]. This data set contains hourly load profile data for 16 commercial building types (based off the DOE commercial reference building models) and residential buildings (based off the Building America House Simulation Protocols). This data set also uses the residential energy consumption survey (RECS) for statistical references of building types by location.

4.2 Similarity Measurement

There are four load curves selected from the above data set adopted here to compare different similarity measurements, where the x-coordinate of the curves is the time point and the y-coordinate is the load value. The calculating results are shown in Table 1.

As shown in Table 1, the calculating results show that the calculation accuracy of MED is same as existing methods ED and DTW. Among all adopted methods, curve L_3 and L_4 are the closest, which conforms to the result of manual interpretation. And the second highest similarity is L_1 and L_4. For all methods, L_3 and L_4 owns the farthest distance. In terms of calculating time, as shown in Table 2, the calculating time of MED is little more than ED, but less than DTW (Fig. 4).

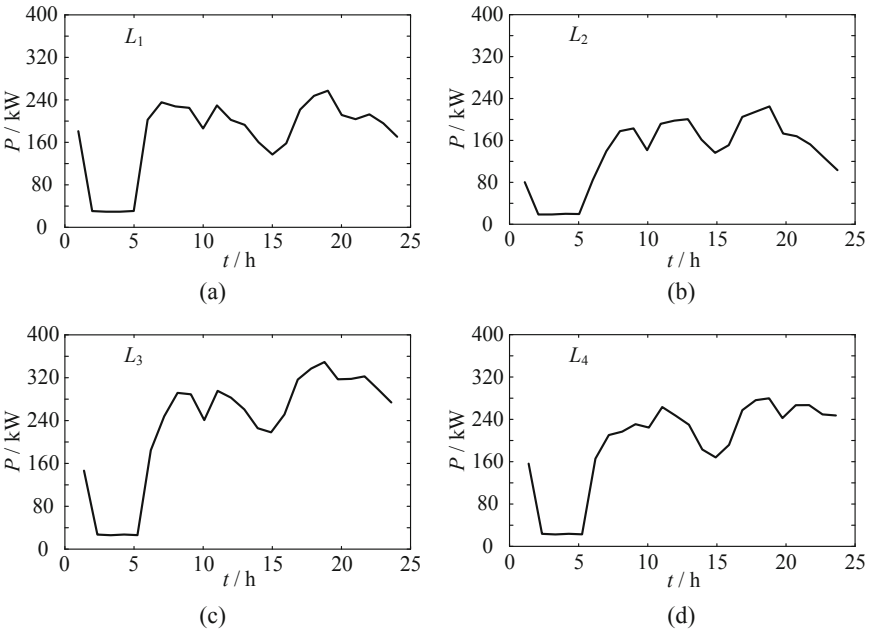

Fig. 4. The daily load curves selected for similarity calculation.

Table 1. The distance results for similarity calculation.

Distance	ED	DTW	MED
$D(L_1, L_2)$	221.4492	681.3785	253.6778
$D(L_1, L_3)$	301.1149	1140.8	340.2479
$D(L_1, L_4)$	193.2995	672.4120	222.0406
$D(L_2, L_3)$	*445.5682*	*1600*	*490.4716*
$D(L_2, L_4)$	342.0599	1275.8	379.8231
$D(L_3, L_4)$	**135.7468**	**427.7579**	**154.5377**

Table 2. The calculating time results for similarity calculation.

Methods	ED	DTW	MED
Time	**0.007456**	*0.671990*	0.007468

4.3 Clustering Analysis

K-means clustering method is applied in this part to analysis. Additionally, the K values adopted in different methods are all determined by ED for those methods. The reason

for adopting it for three methods is to ensure the fairness of comparison. DBI is always used to confirm the value K. And the DBI is introduced as following.

A dispersion value S_i is defined in DBI to represent the dispersions of clusters i as:

$$S_i = \left\{ \frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^q \right\}^{\frac{1}{q}} \tag{7}$$

where T_i and A_i is the number of vectors and the centroid in cluster i, respectively. X_j represents the data point in clusters j. When q is taken as 1, it means the mean value of the distance from each point to the center. When q is taken as 2, it means the standard deviation of the distance from each point to the center. Both of them can be used to measure the degree of dispersion.

M_{ij} is the distance between vectors which are chosen as characteristic of cluster i and j and be calculated as:

$$M_{ij} = \left\{ \sum_{k=1}^{N} |a_{ki} - a_{kj}|^p \right\}^{\frac{1}{p}} \tag{8}$$

where a_{ki} is the k_{th} component of the n-dimensional vector a_{i}, which is the centroid of cluster i. Then R_{ij} is conducted to show the similarity between cluster i and j as:

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \tag{9}$$

where S_i and S_j are the dispersions of clusters i and j, respectively.

Then, DBI index is calculated as the mean value of R_i, which is the maximum value of R_{ij}.

$$\overline{R} = \frac{1}{N} \sum_{i=1}^{N} R_i \tag{10}$$

In this part, 640 daily load curves of 16 kinds commercial and residential customers of 4 different cities are chosen from adopted data set. As show in Fig. 5, K is better chosen 5 to realize clustering. Then the clustering results of three different similarity measurements are produced.

Clustering Results. The clustering results of three similarity measurement methods are shown in Fig. 6, 7 and 8, respectively, where (a) is the clustering results and (b) is the calculated clustering center.

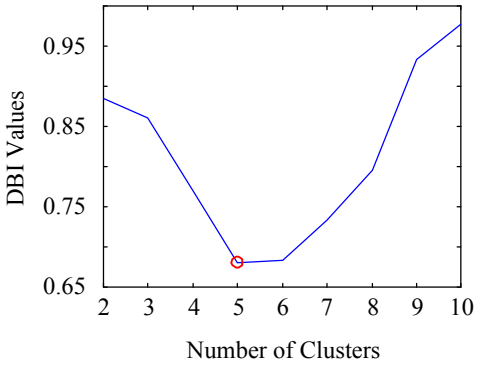

Fig. 5. DBI index during various clusters number.

As shown in Fig. 6, 7 and 8, all three methods realize clustering efficiently. But the clustering results are a little bit different. The clustering results are shown in Table 3, MED decreases the DBI index compared with ED and DTW, which results in a more accurate clustering result. Besides, the clustering time of MED is the shortest, while the calculating time of DTW is much longer.

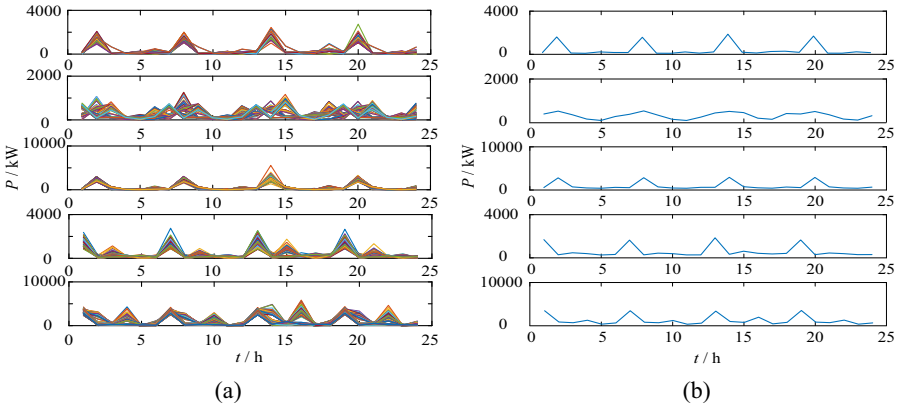

Fig. 6. Clustering results for ED.

4.4 Analysis of Experimental Results

It can be seen that the proposed MED method realized the similarity measurement effectively. And the proposed MED method having the same calculation result with ED, the calculating time of the proposed MED method is between the time of ED and DTW. It can be seen from the clustering results that the MED method increases the clustering results, which results in a lower DBI index. With the decrease of DBI, the proposed algorithm increases the iterations compared with ED. However, the iterations and calculating time of MED are all less than DTW.

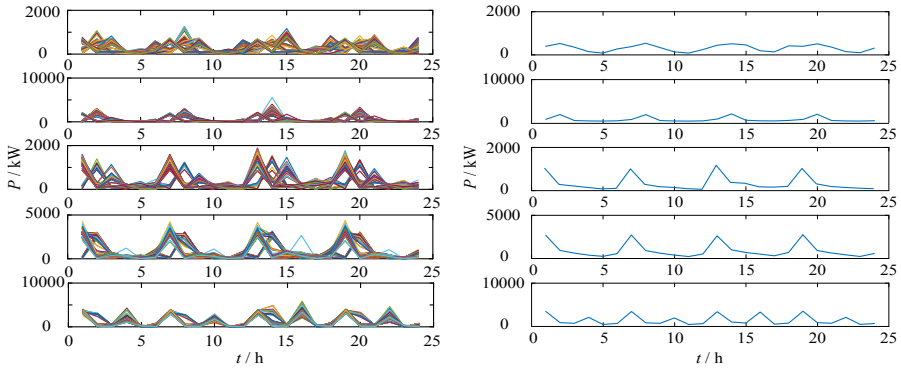

Fig. 7. Clustering results for DTW.

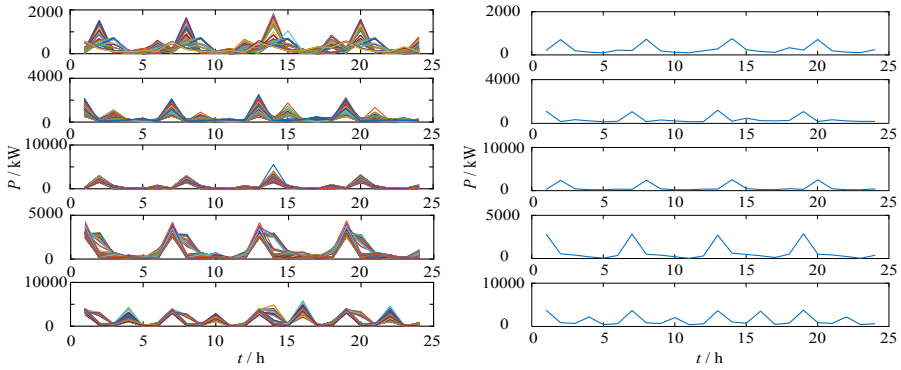

Fig. 8. Clustering results for MED.

Table 3. The clustering results for three methods.

	DBI	Calculating time	Iterations
ED	*0.4382*	0.637700 s	**14**
DTW	0.3248	*3.902797 s*	*29*
MED	**0.2487**	**0.595371 s**	23

5 Conclusion

A modified similarity measurement method based on conventional Euclidean distance is designed and proposed in this paper. Focus on the issues for existing similarity measurement methods, proposed method reflected the change of curve shape without complex

calculation. Besides, the advantages of proposed are proved by similarity calculation and clustering analysis in experimental validation. Compared with Euclidean distance and dynamic time warping distance methods, the proposed method increases the clustering effect without increasing the calculating burden significantly. In future, proposed algorithm will be implied into other clustering algorithms and the parameter that reflects the change in shape will be improved to increase the clustering effect further.

Acknowledgments. This work is supported by the State Grid Corporation of China (52199719002M).

References

1. Shi, L., Zhou, R., et al.: New energy-load characteristic index based on time series similarity measurement. Electr. Power Autom. Equipment **39**(5), 75–81 (2019)
2. Lin, R., Wu, B., Su, Y.: An adaptive weighted parson similarity measurement method for load curve clustering. Energies **11**, 2466 (2018)
3. Alvarez, F.M., Troncoso, A., Riquelme, J.C., Ruiz, J.S.A.: Energy time series forecasting based on pattern sequence similarity. IEEE Trans. Knowl. Data Eng. **23**(8), 1230–1243 (2011). https://doi.org/10.1109/TKDE.2010.227
4. Zhou, R., et al.: Source-load-storage coordinated optimization model with source-load similarity and curve volatility constraints. Proc. CSEE **40**(13), 4092–4101 (2020)
5. Singh, S., Yassine, A.: Big data mining of energy time series for behavioral analytics and energy consumption forecasting. Energies **11**, 452 (2018)
6. Nagi, J., Yap, K.S., Tiong, S.K., et al.: Nontechnical loss detection for metered customers in power utility using support vector machines. IEEE Trans. Power Delivery **25**(2), 1162–1171 (2010)
7. Yu, K., Guo, G., et al.: Quantum algorithms for similarity measurement based on euclidean distance. Int. J. Theor. Phys. **59**, 3134–3144 (2020)
8. Mei, J., Liu, M., Wang, Y., Gao, H.: Learning a mahalanobis distance-based dynamic time warping measure for multivariate time deries vlassification. IEEE Trans. Cybern. **46**(6), 1363–1374 (2016)
9. Li, Z., Yuan, J.: An estimation similarity measure method based on the characteristic of vector difference. Int. J. Inf. **14**(3), 1067–1074 (2011)
10. Jia, H.M., He, G.Y., Fang, C.X., Li, K.W., Yao, Y.Z., Huang, M.M.: Load forecasting by multi-hierarchy clustering combining hierarchy clustering with approaching algorithm in two directions. Power Syst. Technol **31**, 33–36 (2007)
11. Yu, J., Amores, J., Sebe, N., Radeva, P., Tian, Q.: Distance learning for similarity estimation. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 451–462 (2008)
12. Teeraratkul, T., O'Neill, D., Lall, S.: Shape-based approach to household electric load curve clustering and prediction. IEEE Trans. Smart Grid **9**(5), 5196–5206 (2018)
13. Gao, M., Gong, T., Lin, R., et al.: A power load clustering method based on limited DTW algorithm. In: Information Technology, Networking, Electronic and Automation Control Conference, Chengdu, pp. 253–256. IEEE (2019)
14. Lin, R., Wu, B., Su, Y.: An adaptive weighted pearson similarity measurement method for load curve clustering. Energies **11**(9), 2466 (2018)
15. https://openei.org/datasets/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states