A novel decompose-cluster-feedback algorithm for load forecasting with hierarchical structure

Yang, Yang, Zhou, Hu, Wu, Jinran, Liu, Chan-Juan and Wang, You-Gan

Yang, Y., Zhou, H., Wu, J., Liu, C.-J. and Wang, Y.-G. (2022). A novel decompose-cluster-feedback algorithm for load forecasting with hierarchical structure. *International Journal of Electrical Power and Energy Systems, 142*(Part A), p. Article 108249. https://doi.org/10.1016/j.ijepes.2022.108249

A novel decompose-cluster-feedback algorithm for load forecasting with hierarchical structure

Yang Yang[a], Hu Zhou[a], Jinran Wu[b,*], Chan-Juan Liu[c], You-Gan Wang[b]

[a]College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, P. R. China
[b]School of Mathematical Sciences, Queensland University of Technology, Brisbane 4001, Australia
[c]School of Business Administration and Customs Affairs, Shanghai Customs College, Shanghai 201204, P. R. China

Abstract

In load forecasting fields, electricity demand with hierarchical structure is very popular where there are some differences among investigated load series because of geography or customers' habits. Common methods usually ignore their differences and introduce some complex models to improve forecasting performance. Therefore, appropriately dealing with the diverged series is necessary to achieve accurate predictions in hierarchical load forecasting. In this paper, we propose an iterative decompose-cluster-feedback algorithm, which is modified from CLC method, to further improve the performance of forecasts at the total level of hierarchy. Compared with CLC, this algorithm applies empirical mode decomposition (EMD) to decompose load series into sub-series with various amplitude-frequency characteristics, which can avoid directly operating on load series. Specifically, the divergence can have detrimental effects on forecasts if ignored. Finally, we test the proposed algorithm with three real tasks of load forecasting with hierarchical structure, and the experimental results show that the performance of our algorithm is at least 43% better than a SVR-BU method, 52% better than a TD-MLP and a TD-LSTM-SDE method, and 32% better than several methods belonging to middle-out method.

Keywords: Hierarchical time series, load forecasting, clustering, decomposition

1. Introduction

Time series with hierarchical structure are common in the field of load forecasting [1]. The total electricity consumption in a region can be dis-aggregated with its difference in sub-region of sale, price, weather and degree of electricity consumption [2, 3]. For example, in Buzna et al. [4], total load in a region can be obtained from the total load of different sub-regions. While in load forecasting, it is problematic to simply aggregate different time series because of divergence of load series [5]. Therefore, load forecasting considering hierarchical structure (HLF) has attracted attention of many scholars and has become an important research area of energy forecasting. In brief, hierarchical forecasting is an

*Corresponding author
Email addresses: j73.wu@qut.edu.au, orcid: 0000-0002-2388-3614 (Jinran Wu), you-gan.wang@qut.edu.au, orcid:0000-0003-0901-4671 (You-Gan Wang)

approach for modelling time series which have hierarchical structure. Here, a simple 3-level hierarchical structure is shown in Figure 1.

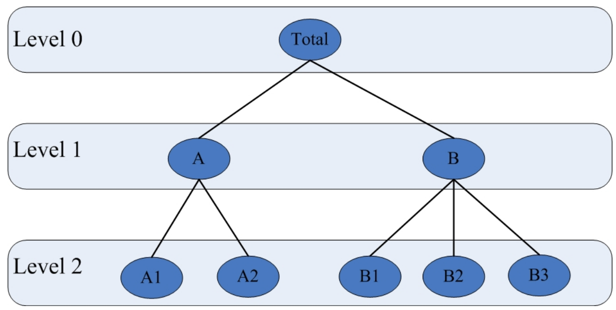

Figure 1: A simple 3-level hierarchical structure: each node represents one load time series, and the nodes such as A1 and B1 in level 2 represent the most basic load series units. The high-level series in hierarchy are the aggregation of the series from lower levels.

A common objective of HLF methods is to optimize the forecasting performance at a total level in the hierarchy. The existing approaches on HLF can be mainly divided into four groups: Top-Down approach, Bottom-Up approach, Middle-Out Approach and reconciling method [6, 7, 8, 9]. Firstly, the Top-Down approach is one of the most common methods in HLF. It follows a process of aggregation and then prediction, and the details can be found in Fliedner [10] and Chen et al. [11]. The second approach is the Bottom-Up which generates forecasts of individual series in hierarchy, and some examples can be seem in Zheng et al. [12], Ye et al. [13], Goehry et al. [14] and Black and Henson [15]. Contrary to the Top-Down method, this approach follows a process of first forecasting separately and then aggregating [16]. Thirdly, as for the Middle-Out approach, it follows a process which starts at an intermediate level in the hierarchy. Specifically, the Middle-Out approach can be divided into two categories: cluster-based method and ensemble-based method [7]. The cluster-based method is to cluster series with similar characteristics to reduce variance [17, 18], and the specific clustering criterion can be divided into density based, centroid-based and distribution-based methods [19]. While there is a dilemma that the optimal clustering is difficult to obtain by common clustering methods. The ensemble-based method can solve this problem to some extents for the centroid-based methods. The idea of this method is to make multiple groups of experiments with several choices of initial clustering number, and the results are obtained by weighted average approach. For this method, the shortcomings caused by the defects of clustering methods can be reduced. The last method is reconciling method which reconciles the independent individual forecasts in all levels of a hierarchy. Specifically, this method utilizes the hierarchical relationships to improve the predictions [20]. To conclude, two queries can be extracted in the existing approaches:

a. These existing approaches directly deal with the load time series in hierarchy. However, according to the existing research in Taieb et al. [21], different load series in hierarchy can contain very different patterns. Therefore, the forecasts generated at different levels of a hierarchy would not be coherent, and the forecasting results may occur significant errors. Therefore, we raise a question that directly dealing with complex load series is an appropriate operation? Whether this operation will ignore the different information sets available and generation conditions of load series?

b. The objective of clustering is not consistent with minimizing the forecasting errors at total level. In other words, the objective of clustering is to minimize the dissimilarity among load time series, which is independent of minimizing the forecasting errors [7].

Recently, a closed-loop clustering (CLC) algorithm has been proposed by Zhang and Li [7], which has a unique feedback mechanism to solve the problem of mismatch between clustering and minimizing forecasting errors. It should be noticed that this CLC algorithm still pay no attention to the first query though it appropriately verifies and solves the second query.

As for the first query, we have known that a time series may be affected by various factors and different time series in a hierarchy may exist divergence caused by extrinsic factors like geographical position and customers' habits. As pointed out by Zhang et al. [22], the sub-series decomposed by empirical mode decomposition (EMD) [23] method has unique physical meanings which can affect the generation of time series. Therefore, we hold an idea that decomposing complex load series in hierarchy into relatively simple sub-series by EMD method is a way to verify and explain the first query. As a result, we can operate on these decomposed sub-series instead of original complex load series. Therefore, we introduce EMD to this study. It can decompose non-stationary signals as sums of zero-mean amplitude modulation components, which are denoted as intrinsic mode functions (IMFs) and residuals [24, 25, 26]. In brief, EMD is an appropriate way to decompose complex series into sub-series with simple patterns.

From the perspective of physical meaning, directly operating on load series is unsuitable because this operation ignores the different generation conditions of load series in hierarchy. Therefore, operating on decomposed sub-series may be an appropriate method because some sub-series of different load series have similar patterns. As a result, this conjecture is demonstrated in the experiments in this study.

To sum up, hierarchical load forecasting (HLF) is an important task for power system dispatching, electricity consumption and planning for management departments. Zhang et al. [22] improved the forecasting performance at total level in hierarchy by clustering all load series and they proposed a feedback mechanism to optimize the clusters. Their method can commendably improve forecasting performance of HLF. However, a more appropriate action is clustering features of load series instead of original series because the inner features of different load series exist divergence, mainly due to the different generating conditions like geographical position and customers' habits among all load time series. Therefore, we introduce EMD algorithm to decompose original load series into sub-series containing different amplitude-frequency characteristic. Then we cluster these sub-series instead of load series to improve the forecasting performance. The experimental results suggest the proposed method can improve the forecasting performance of HLF. This proposed algorithm can appropriately solve the problem that the load series cannot be directly aggregated and can count the clusters tactfully and minimize the prediction errors.

Here, the main contributions in our paper can be summarized as:

a. This paper proposes an iterative decompose-cluster-feedback algorithm, which is composed with a decomposition part, clustering part and a feedback mechanism. The proposed algorithm introduces EMD algorithm to decompose the load series into sub-series with simply patterns. Clustering on these sub-series can avoid the problems caused by clustering original load series generated with various conditions. For HLF, the proposed algorithm can generate hierarchical predictions and improve the forecasting accuracy at total level of hierarchical structure of load series. The feedback mechanism in our algorithm can unify the two goals to iteratively reduce the prediction errors at total level of hierarchy though clustering and minimizing prediction errors are two different objectives.

b. In the proposed iterative decompose-cluster-feedback algorithm, we apply a zero-crossing rate criteria to classify the clustered series into three levels (low frequency, medium frequency and high frequency), and then, the RR, ELM and SVR methods are applied to model low, medium and high frequency series. For the clustered series with lower frequency or simpler patterns, we use RR and ELM to generate appropriate forecasting models and the SVR is applied to model series with high frequency or complex patterns.

c. In this study, we apply the proposed algorithm to generate predictions at a total level for three load forecasting tasks. The final experiment results demonstrate the best forecasting performance of the proposed algorithm among the comparison methods.

The remaining contents are organized as follows: Section 2 introduces the approaches involved in the proposed algorithm. Section 3 shows the details of the proposed algorithm which is based on decomposition, clustering, and feedback mechanism. Section 4 shows the experimental results of two different tasks and empirical analysis of the proposed algorithm and other benchmark models. Section 5 concludes our study.

2. Background

2.1. Empirical mode decomposition

The Empirical mode decomposition (EMD) was proposed by Huang et al. [23] to adaptively decompose non-stationary signals as sums of zero-mean amplitude modulation frequency modulation components [27]. Compared with wavelet decomposition approach, the most remarkable feature of EMD approach is that it does not need any defined function as the base but can adaptively generate intrinsic mode functions according to the non-stationary signal itself.

For a given non-stationary signal $x(t)$, the mathematical expression of the decomposition result is written as follows:

$$x(t) = m_K(t) + \sum_{k=1}^{K} d_k(t),$$

where K is the defined initial clustering number, $m_K(t)$ represents the residual and the $d_k(t)$ are constrained to be zero-mean amplitude modulation and frequency modulation waveforms, which represent the original signal can be expressed in different frequency bandwidth scales.

In this paper, EMD approach is applied to decompose load time series in hierarchy. According to the characteristics of EMD method, the decomposed IMFs of each load time series can be arranged from small to large in frequency.

2.2. K-means clustering algorithm

K-means clustering algorithm is the most popular unsupervised learning clustering algorithm and has been widely used in various fields like label-free data grouping problem [28] and data dimension reduction problem [29]. In this paper, this algorithm is applied for the initial time series clustering for its advantages of simplicity and fast clustering speed, which is only for the initial clustering, while the number of clusters would be adjusted in our new proposed algorithm.

Given a group of time series $x_i, i = 1, \cdots, N$, its idea is randomly determining L cluster centers. Then these clustering centers are iterated with the objective of minimizing the following cost function:

$$J = \sum_{j=1}^{L} \sum_{i=1}^{n_j} |x_i^j - C_j|,$$

where x_i^j represents the time series which belongs to the cluster n_j and C_j is the corresponding cluster center. In the next iterative process, cluster centers are searched by the following formula:

$$C_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i.$$

In general, there exists several weaknesses in the practical application of K-means algorithm, mainly including the problem of uncertain cluster number L and poor initialization of cluster centres. However, K-means algorithm is still used in our algorithm. The main reasons are that these weaknesses can be remedied by the feedback mechanism because our algorithm can iteratively adjust and determine the best clustering results, regardless of the initial clustering situation.

2.3. Ridge regression

Ridge regression (RR) algorithm [30] is an objective for ordinary linear regression models

$$y = \omega^T x,$$

which has weight parameters ω and inputs matrix x. Specifically, it has the following objective function,

$$\min_{\omega}((y - \hat{y})^2 + \lambda \sum_{i=1}^{p} \omega_i^2),$$

where y and $\hat{y}$ are real observations and predictions respectively, λ is a hyperparameter and p is the number of weight parameters.

The weight parameter ω trained by RR function can be limited relatively small, which can reduce the sensitivity to the noise of input variables. Typically, the larger the hyperparameter λ is, the smaller the weight ω will become.

2.4. Extreme learning machine

ELM is a commonly used feed-forward neural network, which is proposed by Huang et al. [31]. The advantage of ELM is that the weight parameters between input layer and hidden layer and bias in hidden layer are randomly selected. In addition, the weight between hidden layer and output layer are obtained by calculating a least square solution:

$$\hat{\omega}_o = H^{\dagger}T,$$

where T represents the real observations and $H^{\dagger}$ represents the generalized inverses matrix of output H in hidden layer.

In brief, ELM has the characteristics of fast training speed and good generalization performance.

2.5. Support vector regression

SVR is a branch of SVM, which is a classic method for electricity demand forecasting [32, 33]. For linear SVR, the model is optimized by maximizing the width of the interval band and the total loss. Specify formula is as follows:

$$\begin{aligned} &\min \frac{1}{2}||\omega||^2 \\ &s.t. \quad \begin{cases} |y_i - (\omega x_i + b)| \leq \epsilon \quad \forall i, \end{cases} \end{aligned}$$

where ω and b are the parameters of linear models, ϵ controls the interval boundary.

For non-linear problem, SVR applies a "kernal trick" to map the features of the original samples to higher dimensions. This method can help SVR solve the problems that cannot get a good regression result by using the original sample features. The commonly used kernel functions include linear kernel, polynomial kernel, Gaussian kernel, and Laplace kernel.

3. Iterative decompose-cluster-feedback stratagem

This paper proposed an iterative decompose-cluster-feedback algorithm for HLF, which can iteratively determine the optimal clustering of the decomposed load time series and effectively improve the forecasting accuracy of HLF. The detailed processes of the proposed stratagem are as follows.

3.1. Decomposition and clustering

Given N load time series $X_i, i = 1, \cdots, N$, the first measure is to decompose these time series by EMD approach. We can obtain a group of IMFs after decomposing and denote them as a set of sub-series

$x = \{x_k | k = 1, 2, \cdots, h\}$. For all sub-series in x, we use K-means algorithm to obtain L initial clusters $n_j(0)$:

$$n_j(0) = \{n_0(0), n_1(0), n_2(0), \cdots, n_L(0)\},$$

where L, j satisfy $1 \leq L \leq N$, $j = 1, 2, \cdots, L$, respectively. Then the decomposed sub-series set can be specifically denoted as

$$x = \{x_k^{(j)} | k \in n_j, j = 1, \cdots, L\},$$

where the constant h is assumed to be the total number of sub-series after decomposing. After that, we can give all sub-series an initial membership C representing the cluster they belong to

$$M(0) = \{M_1(0), M_2(0), M_3(0), \cdots, M_h(0)\}_{h\times 1}^T, \tag{1}$$

where $M(0)$ is a $h \times 1$ vector and represents the corresponding clusters of sub-series after initial clustering measure.

Remark 1. It would be better to choose a proper larger number for the initial clusters L (but must be smaller than N), because the prior information of clusters of the decomposed sub-series is unknown. In our proposed algorithm, we combine all sub-series from a cluster together for training a model for the cluster. Apparently, if the initial cluster number is smaller, some sub-series that should be in different clusters would be clusters, which would decrease the variance of parameters in our forecasting model. Thus, in practice, we empirically set a proper large initial cluster number. A large L can minimize this risk as much as possible because finding the best clustering can be realized by the iterative feedback mechanism in our algorithm. According to updating clustering membership in this proposed algorithm, some redundant clusters would be eliminated during the modelling procedure, and the computing burden would reduce along the training.

Remark 2. The k-means algorithm is used to obtain the initial clusters. More advanced clustering algorithms can be employed for the initialization. The shortcoming of the k-means method is to determine the number of clusters. However, according to the feedback mechanism, the membership of each cluster and the number of the cluster are adaptively adopted in our training procedure. In other words, the cluster is updated based on each sub-series forecasting error minimization criteria, then all the sub-series in the same cluster would model again to generate a new forecasting model. The number of cluster and the membership for each cluster are tuned during the iterative procedure. In practice, generally, a better initial clustering method would optimize the algorithm flow and reduce our computational costs.

3.2. Classification for clusters

For the sub-series belong to a same cluster, we can assume that they have similar amplitude-frequency characteristic and can aggregate them as a single series $S_j(0)$. Then we can obtain a set $S(0) = \{S_j | j = 1, \cdots, L\}$, where each element in $S(0)$ represents a group of series belonging to a same cluster.

Since the decomposed IMFs by EMD have two traits: a) The number of extreme points and zero crossings must be equal or differ by no more than one; b) The upper envelope and the lower envelope are locally symmetrical with respect to the time axis. Therefore, we introduce a definition rule to evaluate the frequency levels of series [34]:

$$Z_n = \frac{n_{zero}}{M}, \tag{2}$$

where Z_n represents the frequency standard of series, n_{zero} is the number of zero crossings for series and M represents the size of samples.

In our algorithm, we divide the elements in $S(0)$ into three types, i.e. low frequency level, medium frequency level and high frequency level. In addition, we apply Ridge Regression (RR), Extreme Learning Machine (ELM) and Support Vector Regression (SVR) to forecast the low frequency, medium frequency, and high frequency respectively.

3.3. Feedback mechanism

Once we have the trained models for each cluster in $n_j(0)$ and they are noted as

$$F(0) = \{F_{n_1(0)}(S_1(0)), F_{n_2(0)}(S_1(0)), \cdots, F_{n_L(0)}(S_1(0))\},$$

the sub-series $x_k, h = 1, \cdots, h$ are tested for the trained forecasting models. The fitness for sub-series x_k on $F_{n_j(0)}$ is denoted as $\sigma_{n_j(0)}$, which is specifically defined as

$$\sigma_k^{n_j(0)} = \sum_{q=1}^{Q} |y_{n_j(0),q} - \hat{y}_{n_j(0),q}|$$

where Q represents the size of forecasting samples, y_q and $\hat{y}_q$ are the real values of sub-series and the predictions respectively. In fact, the fitness function is equivalent to MAE criterion. The fitness matrix can be written as

$$E = \begin{bmatrix} \sigma_1^{n_1(0)} & \sigma_1^{n_2(0)} & \cdots & \sigma_1^{n_L(0)} \\ \sigma_2^{n_1(0)} & \sigma_2^{n_2(0)} & \cdots & \sigma_1^{n_L(0)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_h^{n_1(0)} & \sigma_h^{n_2(0)} & \cdots & \sigma_h^{n_L(0)} \end{bmatrix}_{h \times L}. \tag{3}$$

For each row in E (i.e., a sub-series of x_k), it returns the index of the minimum $\sigma_k^{n_j(0)}$ in the process of feedback mechanism. Meanwhile, the new membership of this sub-series is updated by the index returned, which means the cluster this sub-series belong to is also updated. The membership vector is updated as follows

$$M(1) = \{M_1(1), M_2(1), M_3(1), \cdots, M_h(1)\}, \tag{4}$$

and the new clustering results can be expressed as

$$n_j(1) = \{n_0(1), n_1(1), n_2(1), \cdots, n_{L^*}(1)\}, \tag{5}$$

where $L^* \leq L$, and L^* can gradually decrease until it converges to a constant. This means the cluster is trimmed and merged. In this process, each sub-series is adjusted to minimize the forecasting errors of this sub-series, which can achieve the objective of unifying clustering and minimizing the overall prediction errors.

3.4. Iterative process

Once we have the re-assignment of the clusters and membership for all decomposed sub-series, the sub-series in each cluster can be aggregated as a series individual repeatably, which can be adjusted as $S(1) = \{S_j | j = 1, \cdots, L^*\}$. Then, the formula (2) is also repeatably used to determine the levels of series. With this process, the corresponding models for the series with different levels can be established as $F(1) = \{F_{n_1(1)}(S_1(1)), F_{n_2(1)}(S_2(1)), \cdots, F_{n_{L^*}(1)}(S_{L^*}(1))\}$.

The fitness function will be calculated again for all sub-series. Specifically, fitness values of all sub-series are combined into a new fitness function matrix, which is in the form of formula (3). Then the formula (4) and (5) can be updated by this matrix. In this way, our proposed algorithm can form a closed-loop process and constantly look for the best clusters from the perspective of minimizing forecasting errors. An alternative understanding can be interpreted as that the objectives of clustering and minimizing forecasting errors of sub-series at the total level in hierarchy are unified as a mutual objective. Furthermore, the proposed algorithm can adaptively obtain the optimal cluster number K^* and the corresponding clustering members for each cluster.

Remark 3. Clustering is to group objects according to the information found in data describing objects and their relationships. The purpose is that the objects in a group are like each other (related), while the objects in different groups are different (unrelated). However, the goal of minimizing prediction error is not related to the goal of clustering, and the two goals belong to two types of tasks. In our study, we use clustering algorithm to generate initial clusters because we want to aggregate the sub-sequences which belong to a same model (i.e., have minimal error) and we assume the sub-sequences belong to a same model may have similar patterns. As a result, we apply k-means algorithm to generate initial clusters. In fact, in the feedback mechanism after the initial clustering, the goal of our algorithm is changed as minimizing forecasting errors. This is consistent with the main objective of our algorithm, i.e., minimizing the final prediction errors. Meanwhile, an important reason for using k-means algorithm for initial clustering is to reduce the time cost. Apparently, removing the initial clustering link is more in line with the process of finding the optimal solution.

Now, the flowchart of the proposed algorithm is shown in Figure 2, and corresponding training procedure is listed as follows:

Step 1: Decompose all load time series $X_i, i = 1, \cdots, N$ into a set of sub-series $x = \{x_k | k = 1, 2, \cdots, h\}$.

Step 2: Obtain initial clusters for the decomposed sub-series by K-means algorithm.

Step 3: Aggregate the sub-series in same cluster and determine the levels of all aggregated series.

Step 4: Model different levels of series by using Ridge Regression, ELM and SVR methods.

Step 5: Calculate fitness function matrix (3) for all sub-series, according to all established models.

Step 6: Updated the membership vector (4) and clusters (5) based on the fitness function matrix.

Step 7: Repeat Steps 3-6 until the forecasting errors at total level converge or the clusters remain unchanged.

Remark 4. Our proposed algorithm is modified from the closed-loop clustering algorithm by Zhang et al. [22] where they state, theoretically the number of cluster change is less than 1 and the maximum number of iterations is infinity where there would be no switches among clusters when the process terminated and convergence is guaranteed, and they point out that the number of cluster change and the maximum number can be adjusted by the user according to the data size and computation limits.

4. Case studies

In this paper, we applied the proposed iterative decompose-cluster-feedback algorithm for three real forecasting tasks of HLF. The first task is making load forecasts of next 7 days for 12 regions in USA. This task aims to provide prediction information at national level for management departments' decisions. The second one is making predictions for 180 distribution zone substations in Australia. This task belongs to Short Term Load Forecasting (STLF) and is used to make short predictions for a region which contains hundreds of distribution substation or customers. The last task is generating predictions for 370 customers' electricity consumption at a total level. This task is to test and verify the performance of the proposed method in terms of obtaining an overall prediction for entire customers in a region.

4.1. Evaluation criterion

To comprehensively evaluate the performance of experimental results, three evaluation criteria are applied in our study:

$$\text{RMSE} = \sqrt{\frac{1}{Q}\sum_{i=1}^{Q}(y_i - \hat{y}_i)^2},$$

$$\text{MAE} = \frac{1}{Q}\sum_{i=1}^{Q}|y_i - \hat{y}_i|,$$

and

$$\text{MAPE} = \frac{100}{Q}\sum_{i=1}^{Q}\frac{|y_i - \hat{y}_i|}{y_i},$$

where Q represents the size of samples, y_i and $\hat{y}_i$ represent the observations and predictions, respectively. The RMSE and MAE criteria are used to count the average errors between observations and predictions and the MAPE criterion is used to evaluate the excellence of the model.

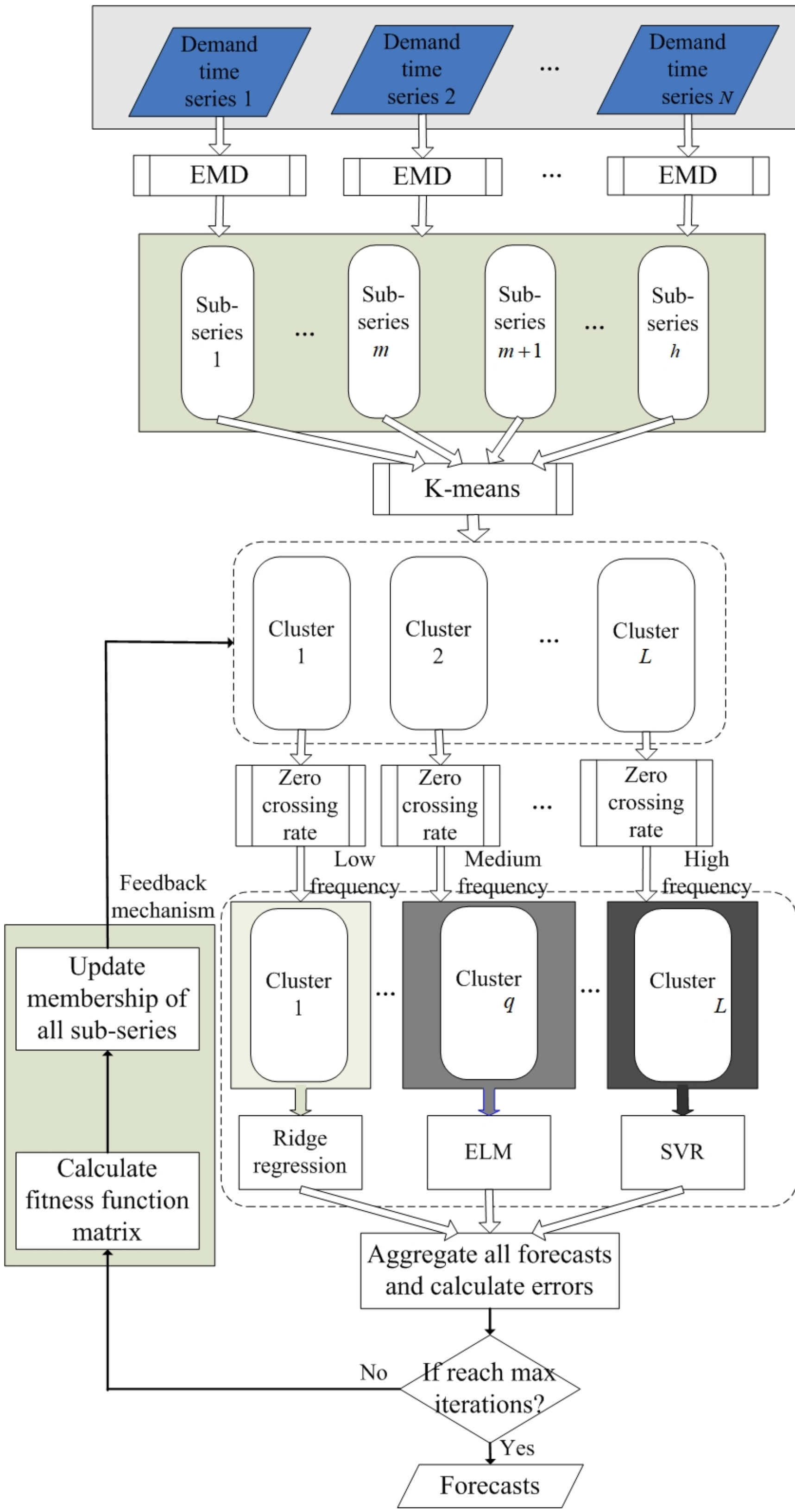

Figure 2: The flowchart of the proposed algorithms.

In addition to above three indicators, a FuzzyEn criterion in the reference of Chen et al. [35] is applied to our study where FuzzyEn introduces fuzzy membership function and is used to measure the probability of new patterns. The greater the measure value, the greater the probability of new patterns, that is, the greater the complexity of series. This criterion is applied to evaluate the degree of complexity of decomposed sub-series.

4.2. Comparison of methods

In order to show the performance of the algorithm proposed more clearly, some advanced comparison methods are investigated:

a. Bottom-Up method: This method follows the idea of firstly forecasting solely and then aggregating the predictions. For this type of method, we introduce a model named SVR-BU algorithm [36] as a comparative model.

b. Top-Down method: Contrary to Bottom-Up method, this method follows a process of aggregating firstly and then make predictions. For the aggregated load time series, we applied two methods to construct contrastive models [8, 3]. The first is Multilayer Perceptron (MLP) regression and the second is a novel method which consists of Long Short-Term Memory (LSTM) and Stochastic differential equation (SDE) driven by Brownian motion. We denote them as TD-MLP and TD-LSTM-SDE, respectively.

c. Cluster-based method: We use K-means algorithm to respectively cluster the load time series and decomposed sub-series $x_k, k = 1, \cdots, h$ in this method. The two classes in this method are denoted as cluster and cluster-EMD. Next, the formula (2) is used to determine the levels of clusters. Then each cluster is modeled according to its level after aggregating the sub-series belong to this cluster. The final predictions are obtained by aggregating the predictions of all clusters.

d. Ensemble-based method: Following the measure of the study of Wang et al. [37], we conduct groups of predictions with different initial clustering number of k-means approach. Then the total forecasts are obtained by aggregating all groups of forecasts with weighted average method. Because this method is based on cluster-based method, this method can be similarly divided into two categories corresponding to cluster-based method, which are denoted as ensemble and ensemble-EMD.

e. In the study of Zhang and Li [7], they proposed a closed-loop cluster algorithm for HLF problem which introduce a feedback mechanism. Therefore, we also compared our algorithm with their method.

f. Brégère and Huard [20] proposed a novel three-stage meta algorithm, which is conposed of three steps (Generation, Aggregation and Projection) for hierarchical series. This algorithm is one of the reconciling methods. We apply this algorithm to our experiments, and for convenience, we denote this method as TSM algorithm in experiments.

g. In Nystrup et al. [38]'s study, the authors proposed an algorithm for temporal hierarchies with autocorrelation for load forecasting, which is a kind of reconciling method. We denote this algorithm as THWA and compared it with our algorithm.

4.3. Task 1: Predictions for load series in USA

4.3.1. Dataset

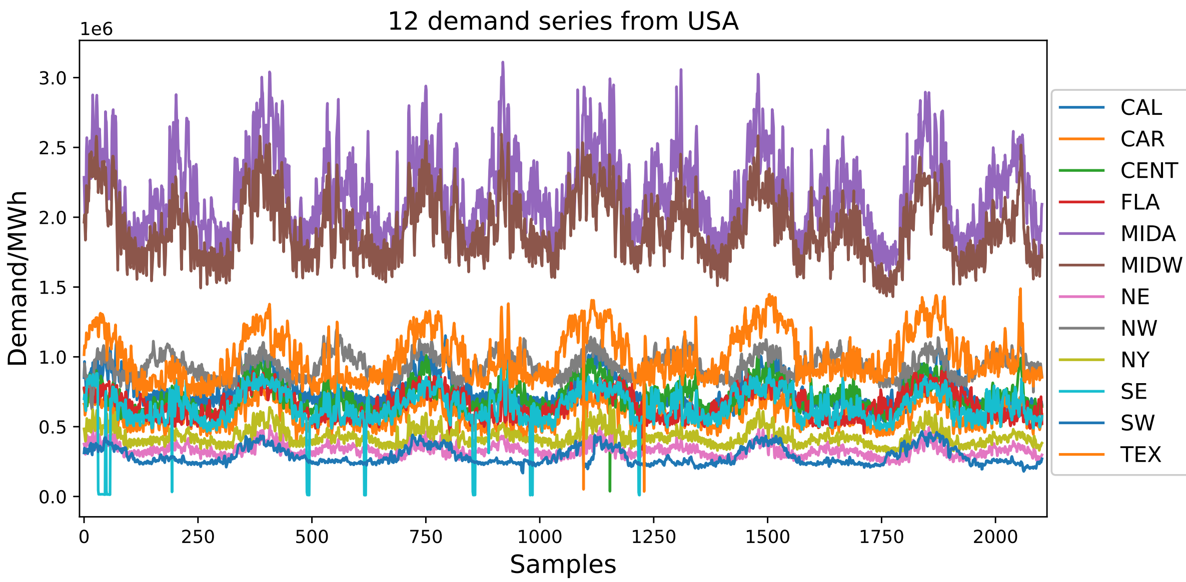

Figure 3: The figures for 12 load series from USA.

In this group of experiments, the electrical load series from 12 regions in United States are applied to make predictions at total level. Specifically, these time series are from July 1, 2015, to April 2, 2021, with a total of 2103 samples and all series have same time scale. Each sample is the sum of the electricity load of corresponding region in one day. Each time series is divided into training set (70%) and test set (30%). This group of load series are shown in Figure 3. Specific information can be obtained from Energy Information Administration (EIA) and its URL is https://www.eia.gov/beta/.

4.3.2. Experimental configuration

The goal of the experiment is to obtain the total load forecast for the next 7 days based on the total level of 12 load time series in the United States. As for the experiments, we aim to obtain the load predictions of sum of the following 7 days at a total level of the 12 load time series in USA. The prediction results obtained in this paper can provide guidance for the rational planning and application of national electric energy. The most outstanding model in this task can provide better predictions for management departments to make better decisions.

Details of this task are set as follows: Firstly, the initial number of clusters of the proposed algorithm is set to 12, which is a moderate number in this task. Secondly, after trials for all sub-series, we set the

Table 1: Experimental results on USA datasets

Method	MAE(10^4)	RMSE(10^4)	MAPE(%)	Improve(%)[1]
TD-MLP	238.16	297.83	3.3	69.7
TD-LSTM-SDE	244.63	320.67	3.4	70.6
SVR-BU	237.92	294.09	3.3	69.7
Cluster	233.84	287.92	3.3	69.7
Cluster-EMD	156.97	206.64	2.1	52.4
Ensemble	234.96	288.97	3.3	69.7
Ensemble-EMD	131.68	172.75	1.8	44.4
CLC	252.05	312.87	3.5	71.4
TSM	211.75	270.58	2.9	65.5
THWA	230.84	286.30	3.2	68.8
The proposed	73.23	92.14	1.0	*None*

[1] The improvement by the proposed method in terms of MAPE.

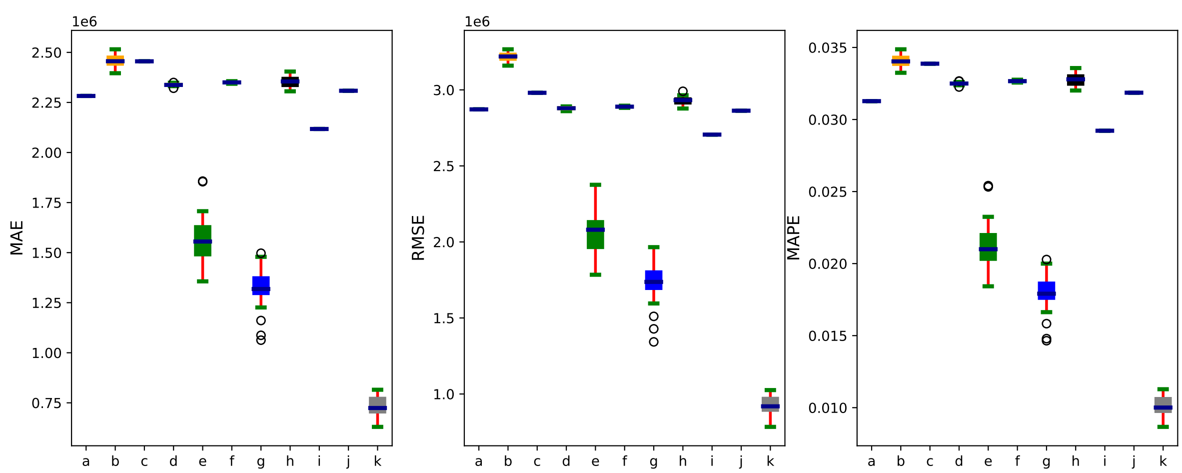

Figure 4: The boxplot for results on USA data set. The corresponding models of these boxes are a: TD-MLP, b: TD-LSTM-SDE, c: SVR-BU, d: Cluster, e: Cluster-EMD, f: Ensemble, g: Ensemble-EMD, h: CLC, i: TSM, j: THWA and k: The proposed.

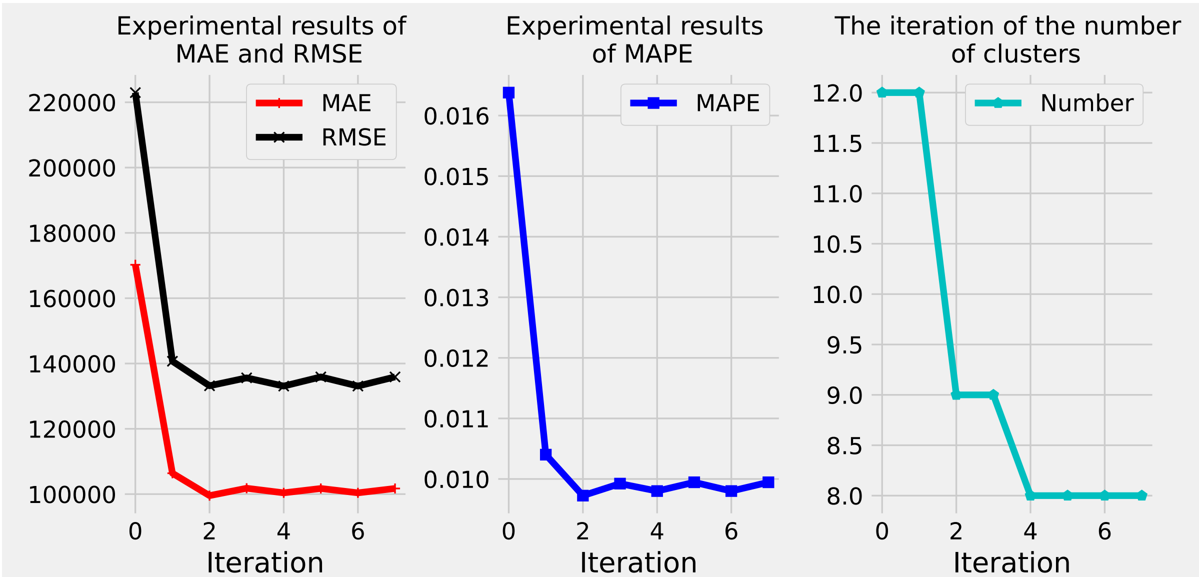

Figure 5: The one-day ahead experimental results on first group of datasets of the proposed algorithm.

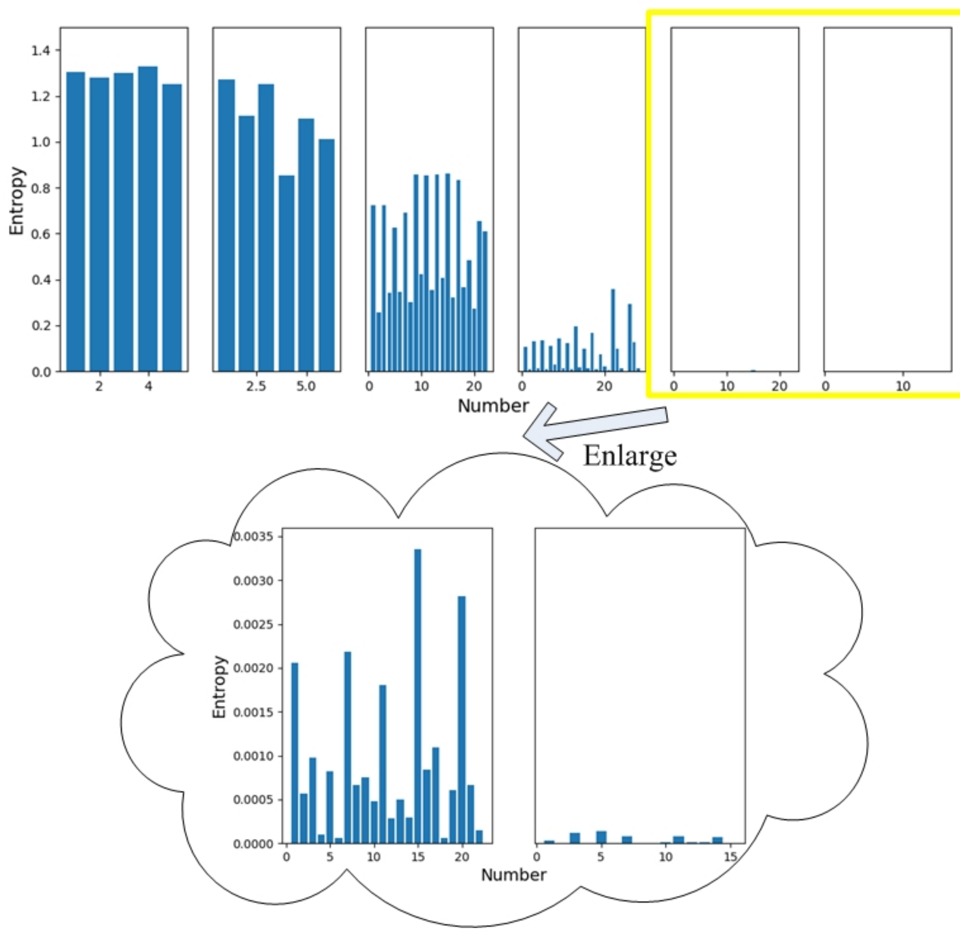

Figure 6: The clustering results on entropy values of the proposed algorithm on USA data set.

Table 2: The constructed models and time consume of benchmark models in training process.

Method	RR	ELM	SVR	Constructed models	Time(s)
TD-MLP	0	0	7	7	3
TD-LSTM-SDE	-	-	-	7	369
SVR-BU	0	0	84	84	31.68
Cluster	0	35	28	63	122
Cluster-EMD	77	0	7	84	3
Ensemble	7	147	154	308	503
Ensemble-EMD	392	21	112	525	26
CLC	5	47	39	91	41
TSM	-	-	-	-	55
THWA	-	-	-	-	19
The proposed	124	7	103	234	510
modeling time[1]	0.0006	0.16	0.17	-	-

[1] The average modeling time for the RR, ELM and SVR model.

boundaries of zero-crossing rate as 50 and 150 separately for series between low frequency and medium frequency and the series between medium frequency and high frequency. Finally, the maximum number of iterations is set to 8 after a group of trials. As the number of initial clusters increases, the number of iterations also increases.

4.3.3. Empirical analysis

In this section, we will compare and analyse the algorithm proposed in this paper and other methods from the following aspects.

a. Comparisons among the models using EMD or not using EMD: In our study, we introduce the EMD method to decompose the original 12 load time series and we hold the idea that it is better to process the decomposed sub-series than to directly process the original load series because load time series cannot be deal with directly for the different conditions for the production of different load series. Table 1 shows the experimental results (i.e., the values of MAE, RMSE and MAPE criterion) of the proposed algorithm and other comparative models. In all benchmark models, EMD method is used in Cluster-EMD, Ensemble-EMD methods and the proposed algorithm. We can find that the results of models using EMD are stably better than the same models without EMD in terms of all the three criteria. According to the last column in Table 1, the Cluster-EMD and Ensemble-EMD methods are about 32% better than their corresponding methods without EMD (i.e., Cluster and Ensemble).

   To evaluate the information complexity of sub-series, we apply FuzzyEn criterion in this task. We calculate the values of FuzzyEn for all sub-series and show the final clustering results in Figure 6. We can find that the sub-series whose FuzzyEn values are in a similar range are selected as a cluster and the range of all clusters are arranged according to the rank of fuzzy entropy. In brief, our idea of decomposing load series into sub-series with simple patterns and clustering sub-series with similar patterns can be elaborated in this part.

b. Comparisons among the proposed algorithm and other methods: We have obtained the conclusion that the models using EMD (i.e., Cluster-EMD, Ensemble-EMD and the proposed algorithm) perform better than the models without EMD (i.e., BU, TD-MLP, TD-LSTM-SDE, Cluster, Ensemble, CLC, TSM and THWA). According to the fourth, sixth and tenth rows in Table 1, the algorithm proposed in this paper is the best method among all models using EMD. This result displays the feedback mechanism in the proposed algorithm works and can improve the accuracy of predictions. As shown in the fifth column of Table 1, the algorithm proposed in our study is more than 44.4% better than all other methods, which is mainly improved by EMD and the feedback mechanism.

c. Analysis of the proposed algorithm: As shown in Figure 5, we can find the MAE, RMSE and MAPE indexes gradually decrease with the increase of the number of feedback mechanisms. In the two figures on the left side of Figure 5, these three indexes finally converge. Similarly, the number of the clusters in Figure 5 also converges at similar time and the final clustering results is shown in Figure 6. Therefore, we can conclude that the feedback mechanism can gradually adjust the clusters of the decomposed sub-series. This represents that the decomposed sub-series with different amplitude-frequency characteristics of each load time series converge to a constant situation.

d. Analysis of the time cost of all methods: as shown in Table 2, we can find that the Ensemble and the proposed method take the most time (503s and 510s). The constructed models constructed by the method with EMD is much more than those method without EMD, which should make the methods with EMD cost much more time in theoretically. However, the Cluster-EMD and Ensemble-EMD methods cost less time than the Cluster and Ensemble methods, which is mainly due to establishing an ELM and a SVR model need more time than establishing a RR model. As for the proposed method, it needs to construct less models and half of them are RR models. The reason for the most time cost of the proposed method is mainly due to the computing other than the modeling part, which is a potential improving direction for the proposed method in the future.

4.4. Task 2: Predictions for load series in Australia

In this task, we do short term load predictions for 180 time series from Australian distribution substations. This task can provide load forecasting at total level for an area and can satisfy the power companies' demand for accurate load forecasting. Here we set up the same comparison models and experiments as in Task 1.

4.4.1. Dataset

The dataset in this task is from distribution zone substations in Australia. These time series have same time scale, from June 25, 2019, to July 14, 2019, with total of 1920 samples. Each sample is the

Table 3: Experimental results on Australia dataset

Method	MAE(10^6)	RMSE(10^6)	MAPE(‰)	Improve(%)[1]
TD-MLP	24.13	33.33	8.7	73.6
TD-LSTM-SDE	16.52	23.94	6.0	61.7
SVR-BU	17.93	23.44	4.2	42.9
Cluster	9.5	14.1	3.5	34.3
Cluster-EMD	12.0	17.5	4.3	46.5
Ensemble	9.5	13.95	3.4	32.4
Ensemble-EMD	10.52	15.44	3.8	39.5
CLC	9.5	14.0	3.4	32.4
TSM	14.09	18.71	5.2	55.8
THWA	10.76	15.87	3.8	39.4
The proposed	6.6	11.45	2.3	$None$

[1] The improvement by the proposed method in terms of MAPE.

sum of load in 15 minutes. Each time series is divided into training set (70%) and test set (30%). Specific information can be obtained from AUSGRID (https://www.ausgrid.com.au/).

4.4.2. Experimental configuration

The detailed goal of this task is to obtain single-step predictions of load series (i.e., 15 minutes in this task) at the total level.

Details of all benchmark models are set as follows: Firstly, the initial number of clusters of the proposed algorithm is set to 25, which is a moderate number in this task. Secondly, we also set the boundaries of zero-crossing rate between low frequency and medium frequency and between medium frequency and high frequency are 50 and 150 respectively, which is an appropriate setting for modelling different sub-series. Finally, the maximum number of iterations is set to 15, of which the proposed algorithm can achieve convergence. As the number of initial clusters increases, the number of iterations also should increase.

4.4.3. Empirical analysis

In this part, we would analyze the results of this group of experiments from similar aspects to the first group of experiments.

a. Comparisons among the models using EMD or not using EMD: According to Table 3, different from the results in the first group of experiments, we cannot find obvious improvement by EMD in two pairs of methods: Cluster, Cluster-EMD and Ensemble, Ensemble-EMD. Considering the case where the number of decomposed sub-series are large and the weakness of K-means method, the main reason is the terrible clustering performance of K-means algorithm. Therefore, we can obtain an inference that single EMD approach seems cannot improve the performance of forecasts in some cases. However, this inference contradicts the conclusion in Task 1, where the EMD has been proved to improve the forecasting performance at the total level in complex hierarchical load forecasting. Observing the good results of the proposed algorithm in Table 3 and Figure 7, we can

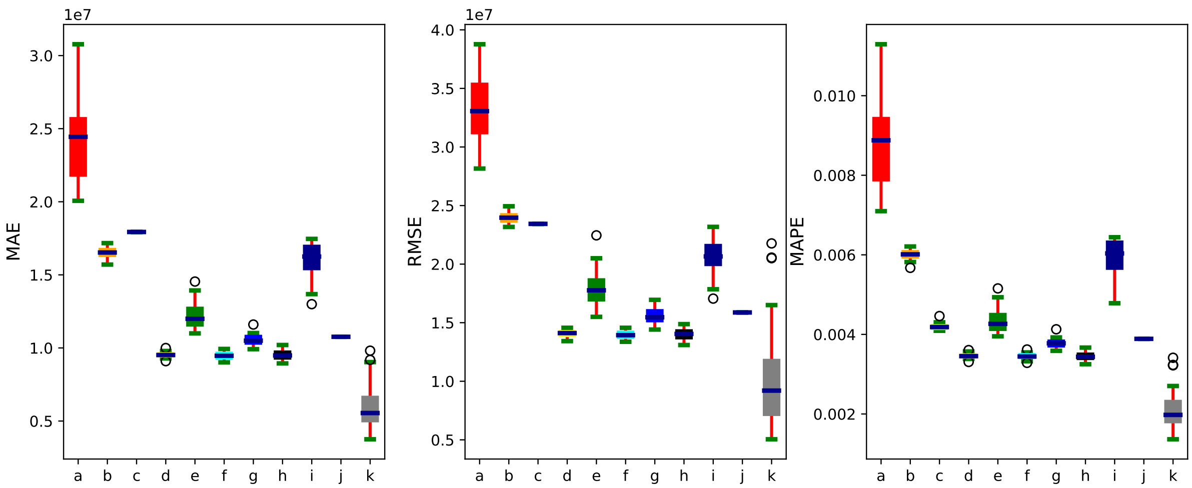

Figure 7: The boxplot for results on Australia data set. The corresponding models of these boxes are a: TD-MLP, b: TD-LSTM-SDE, c: SVR-BU, d: Cluster, e: Cluster-EMD, f: Ensemble, g: Ensemble-EMD, h: CLC, i: TSM, j: THWA and k: The proposed.

demonstrate the efficiency of the feedback mechanism for terrible clusters. This phenomenon also proves that the feedback mechanism can solve the clustering problem caused by K-means method when the number of decomposed sub-series is large and adjust clusters better.

As with task 1, we apply FuzzyEn criterion to evaluate the sub-series in this task. As shown in Figure 10, we can find each cluster has a rough range of FuzzyEn values and the ranges of all clusters are arranged. Moreover, considering the best performance of the proposed algorithm, we can confirm the effects of applying EMD method.

b. Comparisons among the proposed algorithm and other methods: In this task, we can get an obvious conclusion that the two TD related methods (TD-MLP and TD-LSTM-SDE) have the worst forecasting accuracy according to Table 3. The Ensemble and CLC method can improve the performance slightly compared with the Cluster method. As shown in Figure 8, we can find that the criterion of forecasts (MAE, RMSE and MAPE) only get a little better though the number of clusters decreases gradually. The main reasons are that directly cluster the original load time series can easily reach the limit of optimization. On the contrary, we can find that our algorithm proposed in this study which use EMD method can break this limit and have better forecasts.

c. Analysis of the proposed algorithm: As shown in Figure 9, we can find that the performance (MAE, RMSE and MAPE) of our proposed algorithm can decrease gradually and finally converge which is same as the results in the first group of experiments. In both Figures 9 and 5, the criteria (including MAE, RMSE and MAPE) exist fluctuation but the overall trend of them is declining. Furthermore, the clustering performance can be seen in Figure 10 and the good clustering results can demonstrate the effectiveness of our algorithm in terms of adjusting clusters. In brief, we can

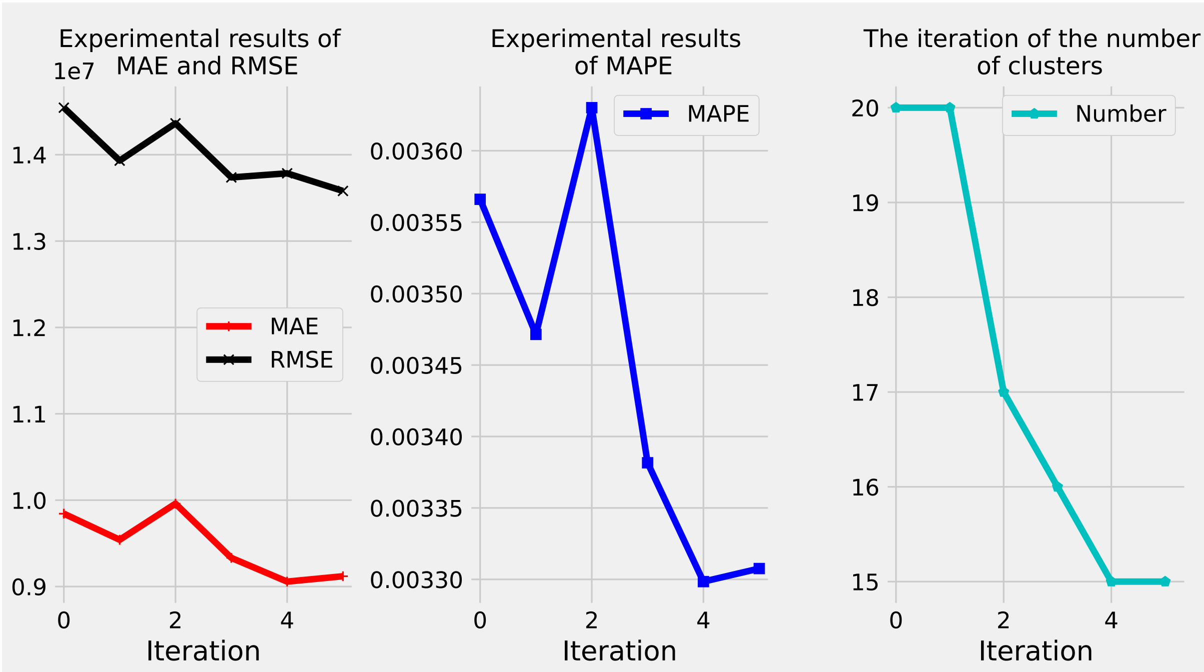

Figure 8: The expermental results on second group of datasets of CLC approach.

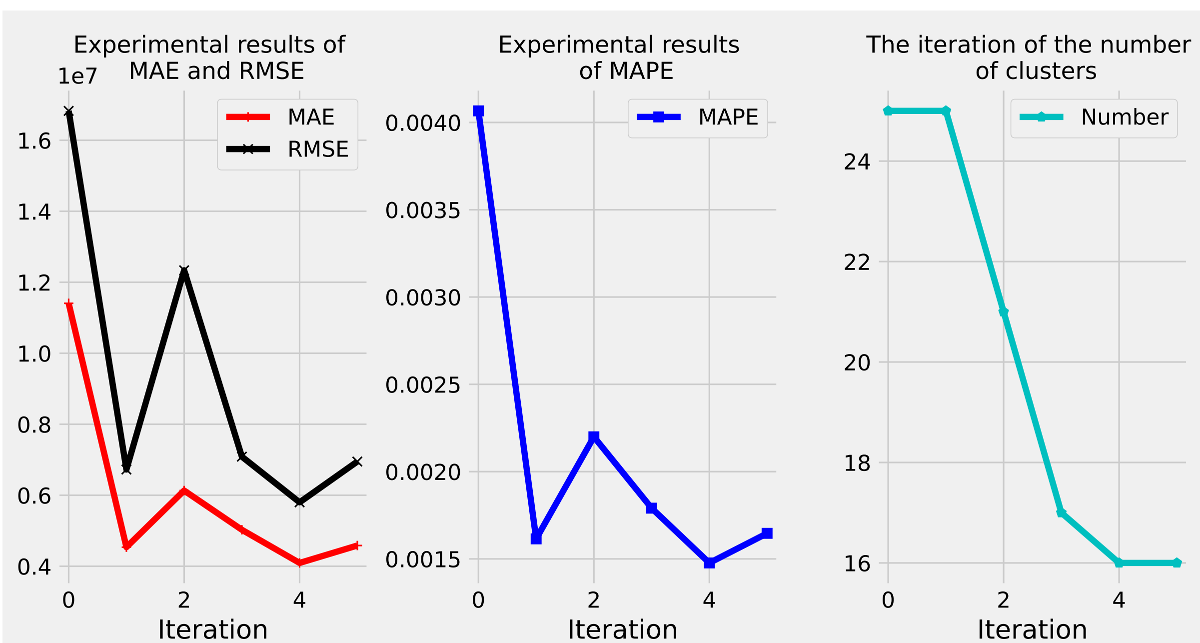

Figure 9: The expermental results on second group of datasets of the proposed algorithm.

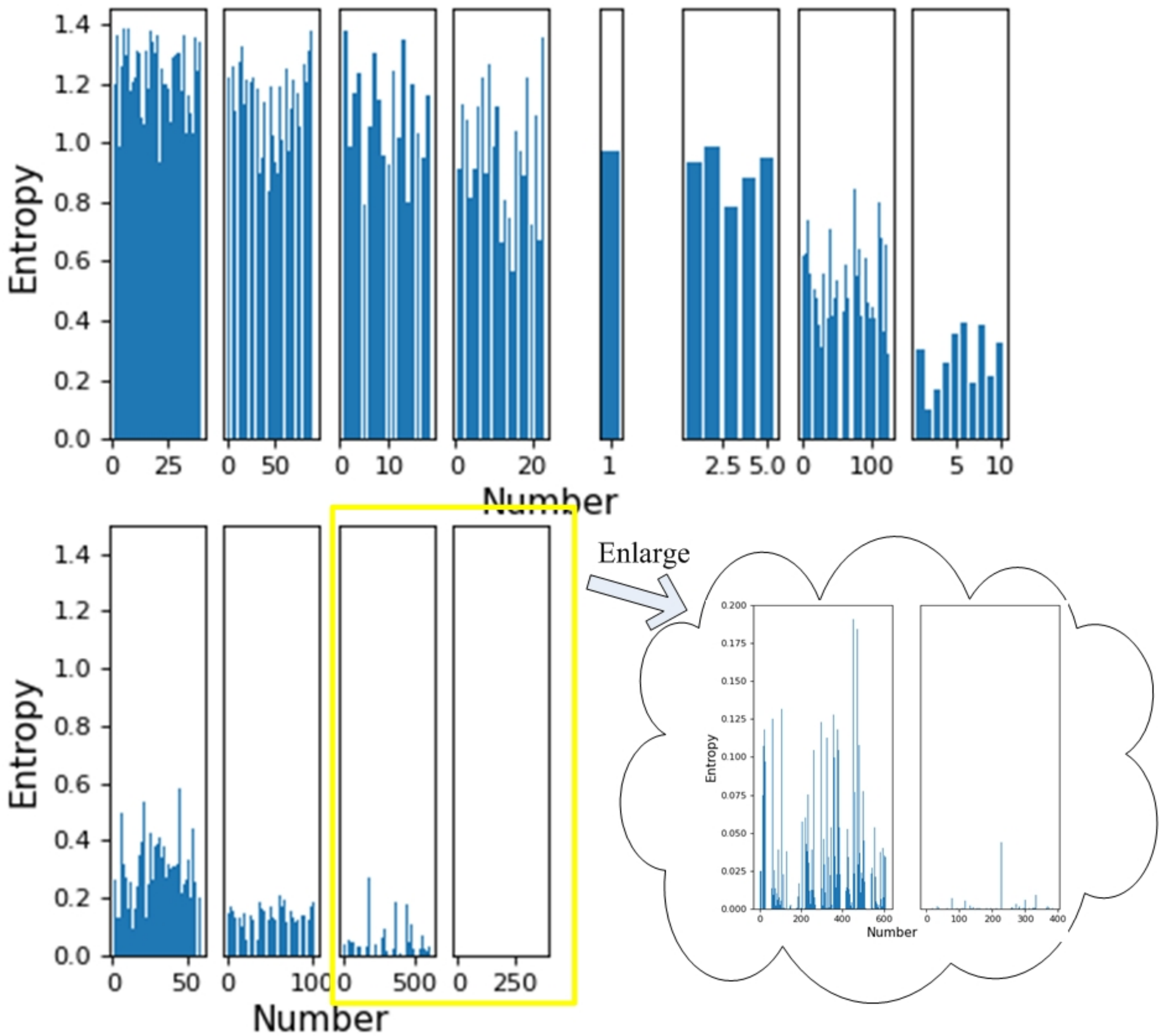

Figure 10: The clustering results on entropy values of the proposed algorithm on Australia data set.

Table 4: The constructed models and time consume of benchmark models in training process.

Method	RR	ELM	SVR	Constructed models	Time(s)
TD-MLP	0	0	1	1	6
TD-LSTM-SDE	-	-	-	1	45
SVR-BU	0	0	180	180	29
Cluster	0	14	0	14	6
Cluster-EMD	615	674	180	1469	105
Ensemble	0	70	0	70	12
Ensemble-EMD	61	49	0	110	33
CLC	9	129	0	138	391
TSM	-	-	-	180	215
THWA	-	-	-	180	37
The proposed	34	38	58	130	4947
modeling time[1]	0.0006	0.09	0.11	-	-

[1] The average modeling time for the RR, ELM and SVR model.

declare that our proposed algorithm can effectively improve the forecasting performance for HLF with simple or complex hierarchy.

d. Analysis of the time cost of all methods: The results of all methods are shown in Table 4. we can find that the CLC and the proposed method take the most time (391s and 4947s), which is mainly due to the increase of time series in hierarchy. The proposed method takes the most time, which is the same as Section 4.3 (Task 1) though the proposed method need not to construct too many models. The main reason is mainly due to the computing in addition to modeling part.

4.5. Task 3: Predictions for electricity consumption of 370 customers

4.5.1. Dataset

In this task, we mainly focus on the forecasting performance and make electricity predictions at a total level for a public electricity dataset with electricity consumption of 370 customers. Each time series is from January 1, 2014, to December 31, 2014. We process the original minutely time series into hourly ones. The training set is from January 1, 2014, to December 17, 2014, and the test set is from December 17, 2014, to December 31, 2014. The specific dataset can be obtained in https//archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams 20112014.

4.5.2. Experimental configuration

The goal of this task is to obtain single-step predictions at total level for 370 electricity consumption time series. The initial cluster of the proposed method is set to 25. The boundaries of zero-crossing rate between low frequency and medium frequency and between medium frequency and high frequency are 20 and 200 after trials and errors, respectively. The number of iterations of the proposed method is set to 16.

4.5.3. Empirical analysis

Table 5: Experimental results on the electricity consumption of 370 customers

Method	MAE(10^4)	RMSE(10^4)	MAPE(%)	Improve(%)[1]
TD-MLP	5.55	7.00	7.9	67.1
TD-LSTM-SDE	4.17	5.78	5.4	51.9
SVR-BU	2.84	4.88	4.9	47.3
Cluster	2.64	4.52	4.5	42.2
Cluster-EMD	3.02	6.09	5.6	53.7
Ensemble	2.63	4.49	4.4	41.9
Ensemble-EMD	3.03	6.12	5.7	64.7
CLC	2.73	4.70	4.7	44.6
TSM	3.17	4.77	5.2	50.3
THWA	2.69	4.05	4.3	39.4
The proposed	1.63	2.17	2.6	*None*

[1] The improvement by the proposed method in terms of MAPE.

According to Table 5, we can find the TD-MLP and TD-LSTM-SDE method can only achieve the worst performance in terms of the three criterion (MAE, RMSE and MAPE). For the three methods using EMD, i.e., Cluster-EMD, Ensemble-EMD and the proposed method, the front two methods perform worse than corresponding methods without EMD. However, the proposed method can achieve the best performance among all methods. This phenomenon, which is the same as that in task 2, can suggest that directly using EMD may not achieve a better prediction for a hierarchy with massive time-series. This can be mainly attributed to the terrible initial clustering results of K-means and the mismatching between clustering and minimizing prediction errors. However, the proposed method, which use the EMD and feedback mechanism, can achieve a better forecasting performance. This proves the adjustment ability of the feedback mechanism in adjusting the clustering results. Comparing the intervallic performance between the proposed algorithm and Cluster-EMD and Ensemble-EMD, we can conclude that the terrible initial clustering results are improved by the feedback mechanism. Meanwhile, the result of the proposed algorithm achieving best performance in this task is consistent with that in task 1 and task 2.

4.6. General discussion

In this work, we focus on the application in hierarchical load forecasting, and we show our algorithm can provide a highly accurate predictions compared with all considered benchmark models based on experiments on three datasets. Furthermore, we experimentally validate the convergence of the proposed algorithm. According to results from Figures 5 and 9 of the two tasks in our study, in terms of error index, our algorithm finally converges. In terms of the final number of clusters, our algorithm also converges. According to the final clusters of two tasks in Figures 6 and 10, we can conclude that the sub-sequences which have similar information are clustered as a set. Overall, we can see that our proposed algorithm converges according to our experiments.

5. Conclusion

In this paper, we have proposed a new decompose-cluster-feedback algorithm for hierarchical load forecasting. An EMD method has been developed to decompose load time series in hierarchy into sub-series with different amplitude-frequency characteristics in our algorithm, which can effectively improve the final forecasting accuracy compared with directly clustering load time series in hierarchy. This idea has been proved to be an effective approach to improve the performance of forecasts for HLF according to the three tasks of making predictions for three groups of load series. Specifically, as mentioned before, the clustering and minimizing forecasting errors at the total level are two different objectives which is difficult for commonly used clustering methods to obtain good clustering results that help minimize the prediction error at the same time. The results of three investigated tasks have proved that the feedback mechanism in our algorithm can solve this problem above and iteratively determine the optimal clusters of forecasts.

One of the limitations for the proposed algorithm is mainly the increasing time cost because of the EMD algorithm and the feedback mechanism in the proposed algorithm. In the future, a novel clustering algorithm that can generate better initial clusters can be applied to this study because better clusters can reduce the time cost of the proposed algorithm. Meanwhile, the three modeling approaches in this study can be replaced by novel and better approaches that can generate more accurate forecasts with less time cost. This study can be extended to other fields, like supermarket revenue forecasting and wind turbine generation forecasting, etc. Moreover, the proposed algorithm focuses on electricity demand forecasting with hierarchical structure. We consider the most general condition without bad or missing data points. In practice, the presence of bad or missing data points would reduce the forecasting performance of the proposed framework. For handling time series with outliers, some model-based, density-based/dissimilarity-based, histogramming methods [39] can be effectively incorporated in our algorithm to deal with outliers. In addition, some robust loss functions, like lncosh loss [40, 41] and general loss [42], can be employed to train the algorithm. As for datasets with missing data points, some modelling strategies can be used to impute the missing points; some advanced machine learning and statistical modelling methods can be found in references of Che et al. [43] and Chao et al. [44]. Meanwhile, according to the type of missing data mechanism, like missing completely at random (when the probability that an individual value will be missing is dependent of observations), missing at random (when the probability that an individual value will be missing is independent of missed observations), and missing nonignorable (when the probability that an individual value will be missing depends on observations), some corresponding modifications are potential in improving the forecasting performance [45]. Last, a mathematical proof of the convergence of the procedure is very crucial but challenging. In our procedure, the convergence of clustering can be guaranteed as proof by Pollard [46]. In our procedure, we define the data error from the forecasting model for each cluster as the distance for clustering in the training procedure. Our proposed procedure is like the EM algorithm, and it would be potential to explore the convergence by following the work of Wu [47].

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution Statement

Yang Yang: Project administration. Hu Zhou: Software, Visualization, Formal analysis, Writing-original draft. Jinran Wu: Supervision, Investigation, Formal analysis, Writing-original draft, Writing-review & editing. Chanjuan Liu: Writing-review & editing. You-Gan Wang: Supervision, Project administration, Investigation, Writing-review & editing.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant 61873130 and Grant 61833011, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191377, in part by the 1311 Talent Project of Nanjing University of Posts and Telecommunications, in part by Natural Science Foundation of Nanjing University of Posts and Telecommunications under Grant NY220194 and under Grant NY221082, by Australian Research Council (ARC) Discovery Project (DP160104292) and Industrial Transformation Training Centres (IC190100020).

References

[1] Tao Hong, Jingrui Xie, and Jonathan Black. Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, 35(4):1389–1399, 2019.

[2] Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

[3] Yuchao Qin, Haochen Hua, and Junwei Cao. Stochastic optimal control scheme for battery lifetime extension in islanded microgrid via a novel modeling approach. *IEEE Transactions on Smart Grid*, 10(4):4467–4475, 2018.

[4] Luboš Buzna, Pasquale De Falco, Gabriella Ferruzzi, Shahab Khormali, Daniela Proto, Nazir Refa, Milan Straka, and Gijs van der Poel. An ensemble methodology for hierarchical probabilistic electric vehicle load forecasting at regular charging stations. *Applied Energy*, 283:116337, 2021.

[5] Rob J Hyndman. Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, 4(4):43–46, 2006.

[6] George Athanasopoulos, Roman A Ahmed, and Rob J Hyndman. Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, 25(1):146–166, 2009.

[7] Chi Zhang and Ran Li. A novel closed-loop clustering method for hierarchical load forecasting. *IEEE Transactions on Smart Grid*, 12(1):432–441, 2021.

[8] Haochen Hua, Yuchao Qin, Chuantong Hao, and Junwei Cao. Stochastic optimal control for energy internet: A bottom-up energy management approach. *IEEE Transactions on Industrial Informatics*, 15(3):1788–1797, 2018.

[9] Rob J Hyndman, Alan J Lee, and Earo Wang. Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, 97:16–32, 2016.

[10] Gene Fliedner. An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers & Operations Research*, 26(10-11):1133–1149, 1999.

[11] Kunjin Chen, Kunlong Chen, Qin Wang, Ziyu He, Jun Hu, and Jinliang He. Short-term load forecasting with deep residual networks. *IEEE Transactions on Smart Grid*, 10(4):3943–3952, 2018.

[12] Zhuang Zheng, Hainan Chen, and Xiaowei Luo. A kalman filter-based bottom-up approach for household short-term load forecast. *Applied Energy*, 250:882–894, 2019.

[13] Chengjin Ye, Yi Ding, Peng Wang, and Zhenzhi Lin. A data-driven bottom-up approach for spatial and temporal electric load forecasting. *IEEE Transactions on Power Systems*, 34(3):1966–1979, 2019.

[14] Benjamin Goehry, Yannig Goude, Pascal Massart, and Jean-Michel Poggi. Aggregation of multi-scale experts for bottom-up load forecasting. *IEEE Transactions on Smart Grid*, 11(3):1895–1904, 2019.

[15] Jonathan D Black and William LW Henson. Hierarchical load hindcasting using reanalysis weather. *IEEE Transactions on Smart Grid*, 5(1):447–455, 2013.

[16] Hjörleifur G Bergsteinsson, Jan Kloppenborg Møller, Peter Nystrup, Ólafur Pétur Pálsson, Daniela Guericke, and Henrik Madsen. Heat load forecasting using adaptive temporal hierarchies. *Applied Energy*, 292:116872, 2021.

[17] Rui Xu and Donald C Wunsch. Clustering algorithms in biomedical research: a review. *IEEE Reviews in Biomedical Engineering*, 3:120–154, 2010.

[18] Peter Nystrup, Henrik Madsen, Emma MV Blomgren, and Giulia de Zotti. Clustering commercial and industrial load patterns for long-term energy planning. *Smart Energy*, 2:100010, 2021.

[19] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

[20] Margaux Brégère and Malo Huard. Online hierarchical forecasting for power consumption data. *International Journal of Forecasting*, 38(1):339–351, 2022.

[21] Souhaib Ben Taieb, James W Taylor, and Rob J Hyndman. Coherent probabilistic forecasts for hierarchical time series. In *International Conference on Machine Learning*, pages 3348–3357. PMLR, 2017.

[22] Shaotong Zhang, Jinran Wu, Yonggang Jia, You-Gan Wang, Yaqi Zhang, and Qibin Duan. A temporal lasso regression model for the emergency forecasting of the suspended sediment concentrations in coastal oceans: Accuracy and interpretability. *Engineering Applications of Artificial Intelligence*, 100:104206, 2021.

[23] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998.

[24] Jyotirmayee Naik, Prachitara Satapathy, and PK Dash. Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression. *Applied Soft Computing*, 70:1167–1188, 2018.

[25] Zhenhai Guo, Weigang Zhao, Haiyan Lu, and Jianzhou Wang. Multi-step forecasting for wind speed using a modified emd-based artificial neural network model. *Renewable Energy*, 37(1):241–249, 2012.

[26] Farhan Riaz, Ali Hassan, Saad Rehman, Imran Khan Niazi, and Kim Dremstrup. Emd-based temporal and spectral features for the classification of eeg signals using supervised learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(1):28–35, 2015.

[27] Patrick Flandrin, Gabriel Rilling, and Paulo Goncalves. Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 11(2):112–114, 2004.

[28] Imran Khan, Zongwei Luo, Joshua Zhexue Huang, and Waseem Shahzad. Variable weighting in fuzzy k-means clustering to determine the number of clusters. *IEEE Transactions on Knowledge and Data Engineering*, 32(9):1838–1853, 2019.

[29] Christos Boutsidis, Anastasios Zouzias, Michael W Mahoney, and Petros Drineas. Randomized dimensionality reduction for k-means clustering. *IEEE Transactions on Information Theory*, 61(2): 1045–1062, 2014.

[30] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[31] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[32] Yongbao Chen, Peng Xu, Yiyi Chu, Weilin Li, Yuntao Wu, Lizhou Ni, Yi Bao, and Kun Wang. Short-term electrical load forecasting using the support vector regression (svr) model to calculate the demand response baseline for office buildings. *Applied Energy*, 195:659–670, 2017.

[33] Youlong Yang, Jinxing Che, Chengzhi Deng, and Li Li. Sequential grid approach based support vector regression for short-term electric load forecasting. *Applied Energy*, 238:1010–1021, 2019.

[34] Jian Li, Daiyu Deng, Junbo Zhao, Dongsheng Cai, Weihao Hu, Man Zhang, and Qi Huang. A novel hybrid short-term load forecasting method of smart grid using mlr and lstm neural network. *IEEE Transactions on Industrial Informatics*, 17(4):2443–2452, 2020.

[35] Weiting Chen, Jun Zhuang, Wangxin Yu, and Zhizhong Wang. Measuring complexity using fuzzyen, apen, and sampen. *Medical Engineering & Physics*, 31(1):61–68, 2009.

[36] Juan Pablo Karmy and Sebastián Maldonado. Hierarchical time series forecasting via support vector regression in the european travel retail industry. *Expert Systems with Applications*, 137:59–73, 2019.

[37] Yi Wang, Qixin Chen, Mingyang Sun, Chongqing Kang, and Qing Xia. An ensemble forecasting method for the aggregated load with subprofiles. *IEEE Transactions on Smart Grid*, 9(4):3906–3908, 2018.

[38] Peter Nystrup, Erik Lindström, Pierre Pinson, and Henrik Madsen. Temporal hierarchies with autocorrelation for load forecasting. *European Journal of Operational Research*, 280(3):876–888, 2020.

[39] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3):1–33, 2021.

[40] Omer Karal. Maximum likelihood optimal and robust support vector regression with lncosh loss function. *Neural Networks*, 94:1–12, 2017.

[41] Yang Yang, Hu Zhou, Yuchao Gao, Jinran Wu, You-Gan Wang, and Liya Fu. Robust penalized extreme learning machine regression with applications in wind speed forecasting. *Neural Computing and Applications*, 34(1):391–407, 2022.

[42] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4331–4339, 2019.

[43] Zhengping Che, Sanjay Purushotham, Guangyu Li, Bo Jiang, and Yan Liu. Hierarchical deep generative models for multi-rate multivariate time series. In *International Conference on Machine Learning*, pages 784–793. PMLR, 2018.

[44] Fengqing Chao, Patrick Gerland, Alex R Cook, and Leontine Alkema. Global estimation and scenario-based projections of sex ratio at birth and missing female births using a bayesian hierarchical time series mixture model. *The Annals of Applied Statistics*, 15(3):1499–1528, 2021.

[45] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[46] David Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, pages 135–140, 1981.

[47] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, pages 95–103, 1983.