J. Mod. Power Syst. Clean Energy (2018) 6(2):306–316
https://doi.org/10.1007/s40565-017-0288-x

Long-term system load forecasting based on data-driven linear clustering method

Yiyan LI[1], Dong HAN[1], Zheng YAN[1]

Abstract In this paper, a data-driven linear clustering (DLC) method is proposed to solve the long-term system load forecasting problem caused by load fluctuation in some developed cities. A large substation load dataset with annual interval is utilized and firstly preprocessed by the proposed linear clustering method to prepare for modelling. Then optimal autoregressive integrated moving average (ARIMA) models are constructed for the sum series of each obtained cluster to forecast their respective future load. Finally, the system load forecasting result is obtained by summing up all the ARIMA forecasts. From error analysis and application results, it is both theoretically and practically proved that the proposed DLC method can reduce random forecasting errors while guaranteeing modelling accuracy, so that a more stable and precise system load forecasting result can be obtained.

Keywords Long-term system load forecasting, Data-driven, Linear clustering, Autoregressive integrated moving average (ARIMA), Error analysis

CrossCheck date: 12 January 2017

Received: 30 April 2016 / Accepted: 12 January 2017 / Published online: 17 May 2017

✉ Yiyan LI
mmaawwxx@163.com

Dong HAN
qbilial@163.com

Zheng YAN
yanz@sjtu.edu.cn

[1] Key Laboratory of Control of Power Transmission and Conversion, Ministry of Education, Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

1 Introduction

Power system load forecasting investigates the changing pattern of the power load, seeks intrinsic correlations between power load and the factors that influence it, and then forecasts the future load scientifically based on corresponding historical data [1]. Particularly, long-term system load forecasting plays an important role in power system planning.

Classical methods for long-term system load forecasting are mainly in three categories: time series models [2–5], correlation models [6–9] and artificial intelligence models [9–13]. Time series models forecast the future load based on the historical load data, so the underlying assumption is that the future load will follow the same trend as its past. Thus, the forecasting error will significantly increase once the trend changes. In short, the major problem of common time series methods is that they don’t adapt well to a changing environment. Meanwhile, because of the complicated set of factors influencing the power load and insufficient annual data, correlation models and artificial intelligence models sometimes cannot work well either.

For several reasons, long-term system load forecasting for a developed city becomes a problem. When a city enters the late stage of urbanization, economic growth slows down, industry is restructuring, and the population starts to saturate [14]. Under such circumstances, the load trend is changing from the fast growing stage to the saturating and fluctuating stage, which weakens the regularity and makes it difficult to conduct accurate forecasting work. Take Shanghai as an example. The average annual growth rate of power consumption in Shanghai from 2001 to 2010 was 8.83%. However, since 2011, the average rate sharply dropped to 1.42% over the next 4 years, with an unprecedented negative growth of −1.42% in 2014. The official value of the average annual growth rate used for planning for Shanghai is 3.2% from 2016 to 2020, which is also much lower than before [15], while contrasting with the experience of the previous 5-year period. In fact, there are quite a number of cities in this transitional period around the world, so it is timely to propose a corresponding effective forecasting method.

In recent years, two main trends have appeared in the development of long-term system load forecasting research. One is that hybrid models are gradually becoming the mainstream. Reference [16] proposes a hybrid model combining dynamic and fuzzy time series approaches to forecast the power consumption in household, commerce and industry respectively. Reference [17] utilizes an Ensemble Empirical Mode Decomposition method to extract the electricity consumption characteristics in multiple time scales, and then construct a relational model between these characteristics and the factors they affect to improve forecasting. Reference [18] constructs a semi-parametric model to investigate the uncertainties in mid-to-long-term forecasting and estimate the probability distribution of the future load, while a novel Kullback-Liebler divergence-based similarity measure strategy is combined to identify the significant impact factors. A Grey model optimized by the Ant Lion Optimizer and a regression model optimized by Improved Particle Swarm Optimization are proposed in [19, 20], respectively. Hybrid models incorporate advantages of different single models so that changing load patterns can be better described, and forecasting accuracy can be improved.

Another research trend in long-term system load forecasting field is that the “big data” concept is gaining increasing attention [21]. With the development of Smart Grids with Advanced Metering Infrastructure, massive power consumption data are available at different network levels, providing a new opportunity to understand the intrinsic characteristics of the power load and improve the forecasting accuracy. Reference [22] investigates how many lagged hourly temperatures and moving average temperatures are needed in a regression model based on a massive load and temperature dataset. In order to prepare for forecasting, [23–25] utilize high-resolution data at an hourly interval to analysis and recognize the load pattern. Particularly, clustering methods are widely used for load forecasting, especially in big-data analyses. In [26], customers are grouped according to consumption similarities, and system load forecasting is improved by combining the forecasting results of each group. References [27, 28] apply the hierarchical clustering method to put similar load curves into one cluster, and forecast the future load of each cluster respectively. Reference [29] introduces the Fuzzy Hopfield Neural Network to classify the hourly load curves based on the date information in order to weight different forecasting models. References [30, 31] extract and analyze the load pattern based on the clustering results, while spectral clustering and functional clustering are utilized to prepare for load forecasting in [32, 33]. In summary, by putting similar objects into one cluster, clustering methods can help recognizing load patterns, weighting different models, simplifying calculation and preparing for model construction in the load forecasting field.

Based on the big data idea, this paper proposes a data-driven linear clustering (DLC) method to improve the stability and accuracy of long-term system load forecasting. A large substation load dataset is utilized to investigate the composition of the system load and to reveal its changing pattern.

Two major contributions of this paper are:

1) The data-driven forecasting idea is introduced to address the forecasting difficulties caused by load fluctuation in developed cities. Based on an autoregressive integrated moving average (ARIMA) model, it is theoretically proved in this paper that the data-driven method is effective in reducing the random forecasting errors so that it can better adapt to the changing environment.
2) A novel linear clustering method is proposed to put complementary substation load curves into the same cluster. After clustering, more accurate ARIMA models can be constructed so that the forecasting error can be further reduced.

The rest of the paper is organized as follows. In Sect. 2, the proposed DLC method is introduced. In Sect. 3, the forecasting error is analyzed, while the results of applying the method to load data are demonstrated in Sect. 4. Finally, Sect. 5 concludes this paper.

2 Data-driven linear clustering method

Among all load forecasting problems, long-term system load forecasting has its own characteristics. In terms of time scale, it usually forecasts the annual power load in the next few years or even decades based on annual historical data, which means a low data quantity and resolution. In terms of spatial scale, system load forecasting focuses on a load system such as a city, a province or even the whole country, so that the load level is usually high and the load curve is relatively smooth.

However, the forecasting methods based on annual system load data are gradually incapable to grasp the load trend in the transitional period mentioned above. So, in order to improve the forecasting accuracy, we need more detailed information about the load system to better understand its structure and inherent regularity. In this case, we propose the DLC method to conduct long-term system load forecasting based on a large substation load dataset. Two main parts are included in this method: the linear clustering preprocessing part and the optimal ARIMA modelling and forecasting part. The large substation load dataset, comprising load time series with annual interval, is firstly clustered by the proposed linear clustering method. Then the time series of summed load of each cluster are modeled and forecasted using optimal ARIMA model. Finally, the system load forecasting results are obtained by summing up all the ARIMA forecasting results.

The flow chart of the DLC method is shown in Fig. 1.

2.1 Linear clustering preprocessing

Suppose $y_t(t = 1, 2, \ldots, T)$ is the time series of the system power load, and y_t is composed of subsequences $y_{k,t}(k = 1, 2, \ldots, N)$, where N is the number of subsequences and T is the number of time samples, usually at annual intervals. $y_{k,t}$ could be a substation load series, or the load series of a district load, and so on. Then we have:

$$\sum_{k=1}^{N} y_{k,t} = y_t \tag{1}$$

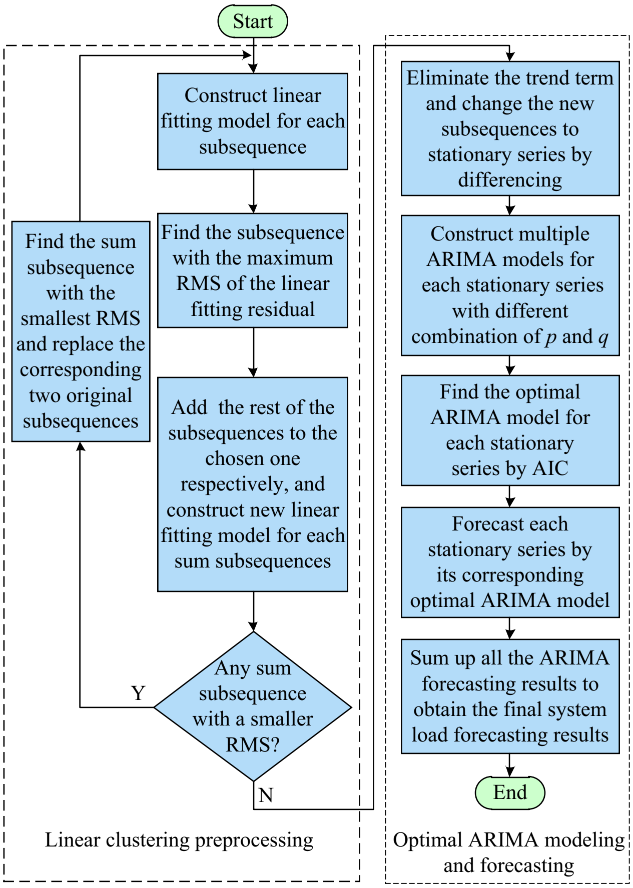

Start
Construct linear fitting model for each subsequence
Find the subsequence with the maximum RMS of the linear fitting residual
Add the rest of the subsequences to the chosen one respectively, and construct new linear fitting model for each sum subsequences
Any sum subsequence with a smaller RMS?
Y
N
Find the sum subsequence with the smallest RMS and replace the corresponding two original subsequences
Linear clustering preprocessing
Eliminate the trend term and change the new subsequences to stationary series by differencing
Construct multiple ARIMA models for each stationary series with different combination of p and q
Find the optimal ARIMA model for each stationary series by AIC
Forecast each stationary series by its corresponding optimal ARIMA model
Sum up all the ARIMA forecasting results to obtain the final system load forecasting results
End
Optimal ARIMA modeling and forecasting

Fig. 1 Flow chart of proposed DLC method

The proposed linear clustering preprocessing method aims to smooth the multiple substation or district load series in such a way that the modelling accuracy is improved. Linear clustering here refers to the clustering criteria. Traditional clustering methods usually classify objects into classes according to a measure of similarity. In order to prepare for modelling and forecasting, we cluster subsequences into classes such that the sum of subsequences in a class has a better linear property than the sum of all subsequences in the dataset. A better linear property means more obvious regularity so that modelling accuracy could be better.

Therefore, the proposed linear clustering is indeed an optimization problem to find the optimal clustering that provides the best global linearity, which can be described by:

$$\min \sum_{i=1}^{M} f_{\mathrm{RMS}}(S_{i,t} - S_{i,t}^{*}) \tag{2}$$

$$\text{s.t.} \begin{cases} S_{i,t} = \sum y_{k,t}, \quad i = 1, 2, \ldots, M \\ \sum\limits_{i=1}^{M} S_{i,t} = \sum\limits_{k=1}^{N} y_{k,t} = y_t \end{cases}$$

where $S_{i,t}$ is the sum of the obtained cluster i, $i = 1, 2, \ldots, M$, $M \le N$; $S_{i,t}^{*}$ is the corresponding linear fitting series; f_{RMS} is the root mean square (RMS) calculation:

$$f_{\mathrm{RMS}}(x) = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \ldots + x_n^2)} \tag{3}$$

where x is an n-dimension vector and $x = (x_1, x_2, \ldots, x_n)$.

In order to solve this problem, the following iterative algorithm is described:

Step 1: Construct least-squares linear fitting model for each subsequence $y_{k,t}$ and calculate the corresponding RMS value of the linear fitting residual, denoted u_k, k = 1,2,…,N, as a linearity measurement for each original subsequence.

Step 2: Find the subsequence with the maximum RMS value $u_{k\max}$ from Step 1 and mark it as $y_{k\max,t}$. Then $y_{k\max,t}$ is the subsequence with the most obvious fluctuation and usually the most difficult one to construct an accurate model for. Therefore, $y_{k\max,t}$ is our major optimization target for this iteration.

Step 3: Construct new linear fitting models for sum series $Y_{j,t} = y_{k\max,t} + y_{j,t}, j = 1, 2, \ldots, N, j \neq k$, and calculate the corresponding RMS values of the fitting residual marked as U_j. This step is to see whether there is any other subsequence that can be summed with $y_{k\max,t}$ to improve the linear fit.

Step 4: Find the minimum value of U_j from Step 3 and mark it as $U_{j\min}$. If

$$U_{j\min} < u_{k\max} \tag{4}$$

it means that there exists a subsequence $y_{j\min,t}$ that can be summed with $y_{k\max,t}$ to improve the linear fit. In this case, we replace $y_{j\min,t}$ and $y_{k\max,t}$ by their sum $Y_{j\min,t}$ and go back to Step 1. The iteration stops when $U_{j\min} \geq u_{k\max}$, which means the subsequences cannot be smoothed any further by summation.

After such linear clustering preprocessing, the smoothness of the subsequences is improved while their number is reduced, which are better conditions for modelling and forecasting.

2.2 Optimal ARIMA modelling and forecasting

The ARIMA model proposed by Box and Jenkins in 1970s has a good performance when describing and forecasting a time series [34]. Therefore, we use it to forecast the summed load of each cluster and to analyze the load forecasting error. The ARIMA (p,d,q) model can be described by:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{5}$$

where ε_t is white noise; φ and θ are the coefficients.We can see that there are two parts contained in the ARIMA model: the autoregressive (AR) part

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t \tag{6}$$

and the moving average (MA) part

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{7}$$

The AR part describes the remembered characteristics of past system states, while the MA part reflects the influence of noise on the current system state. p and q are the corresponding orders of the two parts. Because ARIMA model is only suitable for stationary series, differencing preprocessing is required if the series is not stationary [35], with d representing the differencing order.

We construct optimal ARIMA models for $S_{i,t}$, $i = 1, 2, \ldots, M$, forecast their future values respectively, and sum them up to obtain the final system load forecast. The corresponding algorithm steps are as follows.

Step 1: The Unit Root Test [36] is firstly applied to the preprocessed series $S_{i,t}$ to check whether they are stationary or not. Any non-stationary series will be converted into a stationary one by differencing.

Step 2: Construct multiple ARIMA (p,d,q) models for each stationary series with different combination of parameters p and q. Due to limited series length, we limit p and q to a relatively low order in order to avoid overfitting [37]: p = 0,1,2; q = 0,1.

Step 3: Among all the ARIMA models constructed in Step 2, find the optimal one for each stationary series by the Akaike information criterion (AIC). This is a criterion to measure the modelling effect considering both the fitting accuracy and the complexity of the constructed model [38]:

$$AIC = 2n + T \ln(f_{\mathrm{RSS}}/T) \tag{8}$$

where n is the number of parameters in the constructed model; T is the length of the series; f_{RSS} is the residual sum of squared differences which reflects the modelling accuracy. Generally, the model with the smallest AIC value is the optimal one, so the mathematical description of choosing the optimal ARIMA model for $S_{i,t}$ is:

$$\begin{aligned} &\min \ AIC = 2n + T \ln(f_{\mathrm{RSS}}^2/T) \\ &\text{s.t.} \begin{cases} n = p + q \\ p = 0, 1, 2, \ q = 0, 1 \\ q \leq p \\ f_{\mathrm{RSS}} = \sqrt{\sum_{t=1}^{T} (S_{i,t} - S_{i,t}^*)^2} \end{cases} \end{aligned} \tag{9}$$

where $S_{i,t}^*$ is the ARIMA (p,d,q) fitting value of $S_{i,t}$.

Step 4: Forecast the future value of each preprocessed series $S_{i,t}$ based on its corresponding optimal ARIMA model selected in Step 3. The forecasting results are denoted by $S_{i,t+\tau}$, $\tau = 1, 2, \ldots, \Delta T$, in which ΔT is the forecasting period. Note that ΔT cannot be too big because of a limitation of the ARIMA model [39].

Step 5: Sum up all the ARIMA forecasting results to obtain the final system load forecasting results:

$$S_{t+\tau} = \sum_{i=1}^{M} S_{i,t+\tau} \tag{10}$$

3 Forecasting error analysis

When we forecast the power load, the forecasting error mainly consists of two parts: the modelling error and the random error [40]. The modelling error refers to the difference between the model fitting value and the true value. Generally, the smoother the load curve, the smaller the modelling error, so that the constructed model can better fit the pattern of changing load. The random error refers to the forecasting error caused by some random and unpredictable factors that change the original load changing pattern. Thus, in order to improve the forecasting accuracy, we should both improve the modelling accuracy and try to limit the random error.

Here, we analyze the forecasting error of different forecasting methods based on ARIMA model. For simplicity, we make two assumptions in advance [40]:

1) Because the ARIMA forecasting results mainly depend on the AR part, we assume that the time series of the power load follows the first term of the AR model in (6), denoted the AR(1) pattern:

$$y_t = \varphi_1 y_{t-1} + \varepsilon_t,\ y_{k,t} = \varphi_{k,1} y_{k,t-1} + \varepsilon_{k,t} \tag{11}$$

2) We assume that the white noise in the time series of power load is White Gaussian Noise (WGN), and that the standard deviation of the noise is proportional to the load level:

$$\varepsilon_t \sim N(0, \sigma^2 y_t^2),\ \varepsilon_{k,t} \sim N(0, \sigma^2 y_{k,t}^2) \tag{12}$$

where $\sigma > 0$ is the proportionality coefficient.

Suppose the modelling error in time t is v_t^m, and the random error is v_t^r. Then v_t^m depends on the AR(1) part, while v_t^r is related to ε_t according to the analysis above. Then the total forecasting error can be described by:

$$v_t = v_t^m + v_t^r \tag{13}$$

3.1 Modelling error

Based on (1), suppose the ARIMA modelling result for time series y_t is:

$$y_t^* = \varphi_1^* y_{t-1} + \varepsilon_t^* \tag{14}$$

From (2) we know that ε_t is WGN, so $\varepsilon_t^* = 0$. Then (14) becomes:

$$y_t^* = \varphi_1^* y_{t-1} \tag{15}$$

From (11), the actual value of φ_1 is:

$$\varphi_1 = \frac{y_{t-1} - \varepsilon_{t-1}}{y_{t-2}} \tag{16}$$

Therefore the parameter estimation error for φ_1 is:

$$\Delta\varphi_1 = \varphi_1 - \varphi_1^* = \frac{\varepsilon_{t-1}}{y_{t-2}} \tag{17}$$

where $\Delta\varphi_1$ is the source of the modelling error, which is proportional to the WGN ε_t and inversely proportional to the load level y_t according to (17). If we model and forecast the system load directly (called the direct method in this paper), the modelling error will be small because the load level y_t is high and the standard deviation of the noise ε_t is low due to the smoothness of the system load curve. On the other hand, if we model and forecast the subsequences of the system load and then sum them up to obtain the system load forecasting result (called the data-driven method in this paper), the modelling error for each subsequence will be more significant. The proposed DLC method constructs a forecasting model based on the smoothed sum series so that the modelling accuracy can be guaranteed to a certain extent, theoretically inferior to the direct method but better than the data-driven method.

The modelling error of the forecasting results can be evaluated by:

$$v_t^{*m} = \frac{1}{T}\sum_{t=1}^{T}\frac{\left|y_t - y_t^*\right|}{y_t} \times 100\% \tag{18}$$

3.2 Random error

From (12) we know that the WGN of a subsequence $\varepsilon_{k,t} \sim N(0, \sigma^2 y_{k,t}^2)$. Because of the mutual independence property of WGN, we have:

$$\sum_{k=1}^{N}\varepsilon_{k,t} \sim N\left(0, \sum_{k=1}^{N}\sigma^2 y_{k,t}^2\right) \tag{19}$$

Because $\sigma > 0$, $y_{k,t} > 0$, and $y_{k,t}$ are not all equal for different k, we have:

$$\sum_{k=1}^{N}\sigma^2 y_{k,t}^2 < \sigma^2\left(\sum_{k=1}^{N} y_{k,t}\right)^2 = \sigma^2 y_t^2 \tag{20}$$

Equation (20) is the theoretical basis of the data-driven method: the variance of the WGN is smaller than for the direct method. In this way, the forecasting random error can be limited and a more stable system load forecasting result can be obtained by the data-driven method. And this is exactly the value of using a large quantity of substation load data. Similarly, the proposed DLC method also takes advantage of the large dataset, so that its random forecasting error will be smaller than that of the direct method.

The forecasting error can be evaluated by:

$$v_t^* = \frac{1}{\Delta T}\sum_{\tau=1}^{\Delta T}\frac{\left|y_{T+\tau} - S_{T+\tau}\right|}{y_{T+\tau}} \times 100\% \tag{21}$$

According to (13), the random forecasting error can be evaluated as:

$$v_t^{*r} = v_t^* - v_t^{*m} \tag{22}$$

In short, the direct method usually performs well in modelling, but will probably gain an uncontrollable random error when forecasting. On the contrary, the data-driven method can limit the forecasting random error, but makes it harder to construct a precise model for each subsequence. As a combination of the above two methods, the proposed DLC method can reduce the random forecasting error while guaranteeing modelling accuracy, providing improved forecasting results.

4 Application results

Peak load data from Shanghai are used to test the effectiveness of the proposed DLC method [41]. The annual peak loads from 2001 to 2015 are shown in Fig. 2a. The system load is composed of 83 substation loads at 220 kV ($N = 83$), of which the corresponding peak load curves are plotted in Fig. 2b.

Here, we construct a model based on the load data from 2001 to 2012, and conduct virtual forecasting from 2013 to 2015 to test its effectiveness. For comparison, four different forecasting schemes are applied:

1) Direct method: construct an optimal ARIMA model for the system load data in Fig. 2a directly and forecast the system peak load.
2) Data-driven method: construct optimal ARIMA model for each original subsequence in Fig. 2b and forecast each one, then sum up all the forecasting results to obtain the system load forecasting results.
3) DLC method: based on the subsequence data in Fig. 2b, using the forecasting algorithm proposed in Sect. 2 to obtain the system load forecasting results.
4) Classical methods: apply some classical forecasting methods, such as the scrolling GM(1,1) model, the elasticity coefficient model and a regression model to forecast the system peak load.

Additionally, we apply the proposed DLC method to another four cities to test its adaptability. Finally, we forecast the future load in Shanghai from 2016 to 2020 based on DLC method.

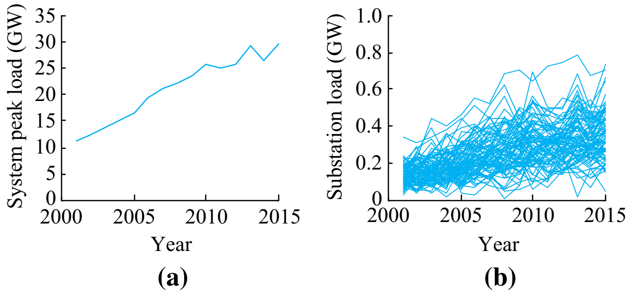


Fig. 2 Annual peak load and corresponding annual peak load of 83 substations at 220 kV in Shanghai

4.1 Direct method

The optimal ARIMA model was applied directly forecast the system load. The model fitting and forecasting results are shown in Fig. 3.

The modelling error of the direct method is 2.18% and the average forecasting error is 10.02%, so the random error is 7.84%. From Fig. 3 we can see that the annual system load curve in Shanghai is relatively smooth, which is advantageous for modelling and leads to a high modelling accuracy. However, the forecasting results are not desirable. This is mainly due to the changing pattern of the load growth. Shanghai has a high urbanization level and is under industrial restructuring, in which backward production facilities are closed down and the development of tertiary industry is accelerated. Meanwhile, the population in Shanghai is becoming saturated. Under such circumstances, the pattern of load growth is changing, having shown significant fluctuation since 2009. This makes it difficult for the direct method to work well. In order to obtain a better forecasting result, more information about the load system is required to explore the internal regularity of the fluctuating system load.

4.2 Data-driven method

Optimal ARIMA modelling and forecasting was conducted for each 220 kV substation load in Fig. 2b, and the results are shown in Fig. 4.

The average modelling and forecasting error in Fig. 2a are 20.32 and 26.46% respectively, so the average random error is 6.13%. The significant increase of the modelling error is due to the low load level and the high fluctuation of substation loads, which has been discussed in Sect. 3. And this is also the main reason for the large forecasting error for individual substation loads.

After summing up the modelling and forecasting results in Fig. 4a to obtain the system results in Fig. 4b, the system modelling and forecasting errors are 2.35 and 3.65% respectively, with a random error 1.30%. We can see that

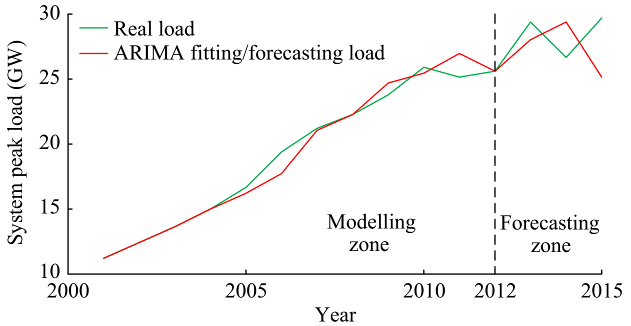
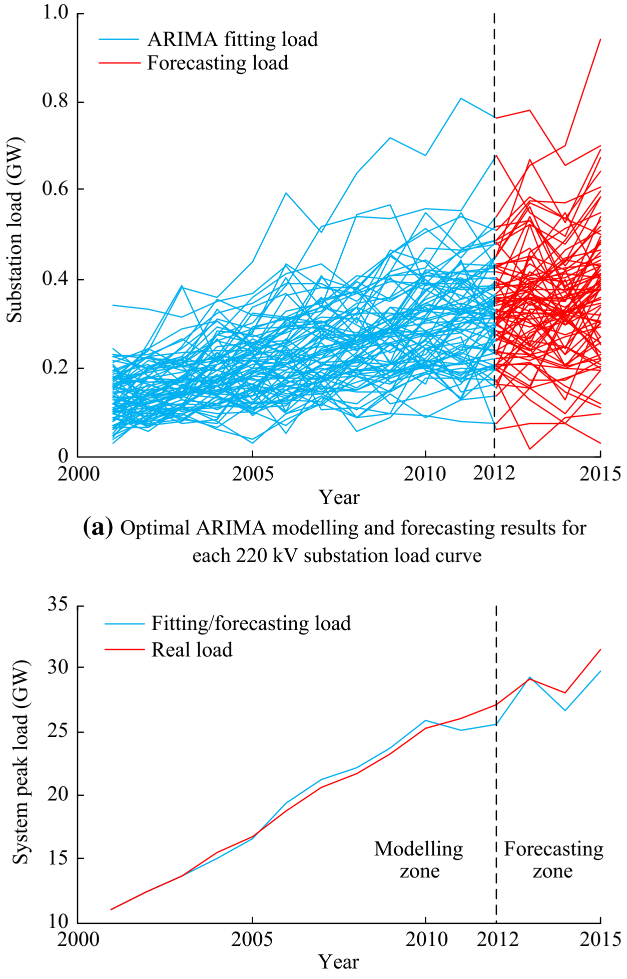


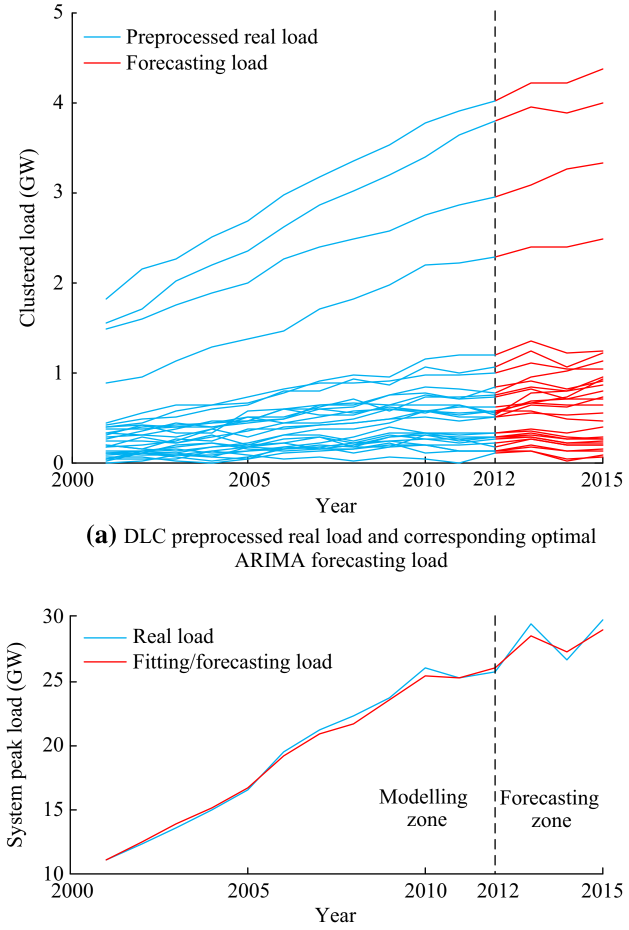
Fig. 3 Modelling and forecasting results by direct method



(a) Optimal ARIMA modelling and forecasting results for each 220 kV substation load curve

(b) Corresponding system load fitting and forecasting results

Fig. 4 Load curves with data-driven method



(a) DLC preprocessed real load and corresponding optimal ARIMA forecasting load

(b) Corresponding system load fitting and forecasting results

Fig. 5 Load curves with DLC method using clustering criterion 1

the random error has been effectively reduced from 7.84 to 1.30% compared with the direct method, therefore the corresponding forecasting error is reduced. This is the value of the data-driven method, which has also been discussed in Sect. 3. But on the other hand, the modelling error is 2.35% and becomes the major part of the forecasting error.

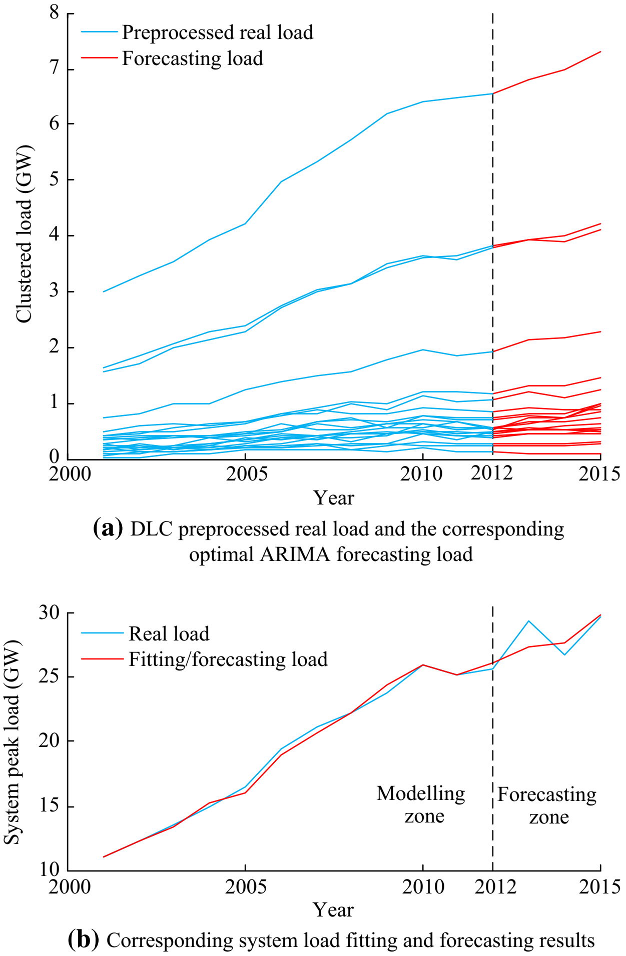
4.3 DLC method

In order to improve the modelling accuracy of the data-driven method, the substation load data were preprocessed using the proposed linear clustering method. The modelling and forecasting results are shown in Fig. 5.

We can see from Fig. 5a that the preprocessed data obtained by linear clustering method are much smoother than the original data in Fig. 4a, making them more suitable for time series modelling. The average modelling error has been reduced to 10.71%. Also, the number of subsequences is reduced from 83 to 30 ($M = 30$) so the computational work is reduced. More importantly, the forecasting results for substation load clusters are more stable: the average forecasting error is reduced to 18.54%, with a random error 7.76%.

After summing up the modelling and forecasting results in Fig. 5a to obtain the system results in Fig. 5b, the system modelling error is 1.40%, the forecasting error 2.67%, and the random error 1.27%. The more accurate results prove that the proposed DLC forecasting algorithm takes advantage of clustering to limit the random error while guaranteeing the modelling accuracy.

In the proposed linear clustering preprocessing method, the clustering criterion in Step 4 in Sect. 2.1 is crucial. Different clustering criteria will result in different clustering results, thus leading to different modelling and forecasting effects. In the DLC method presented above, the clustering criterion is shown in (4), denoted “criterion 1”. Consider relaxing it to (23), denoted “criterion 2”:



(a) DLC preprocessed real load and the corresponding optimal ARIMA forecasting load

(b) Corresponding system load fitting and forecasting results

Fig. 6 Modelling and forecasting results using clustering criterion 2

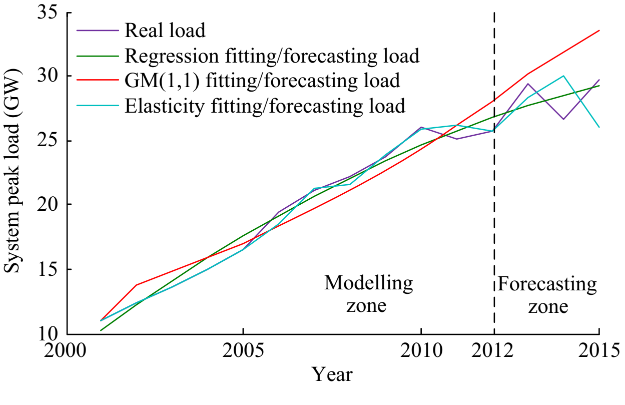


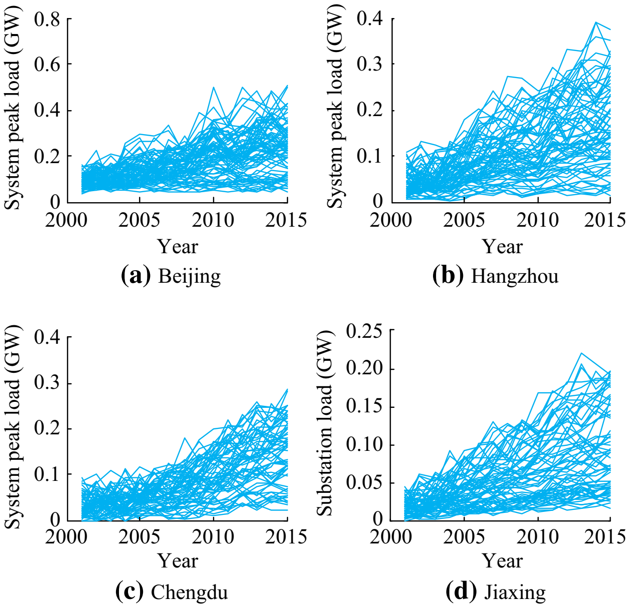
Fig. 7 Modelling and forecasting results by GM(1,1) model, regression model and elasticity coefficient model

$$U_{j\min} < \sqrt{u_{k\max}^2 + u_j^2} \tag{23}$$

where u_j is the RMS value of the linear fitting residual for $y_{j\min,t}$. The new modelling and forecasting results are shown in Fig. 6.

Table 1 Results summary

Method	Modelling error (%)	Random error (%)	Forecasting error (%)
Direct	2.18	7.84	10.02
Data-driven	2.35	1.30	3.65
DLC-criterion 1	1.40	1.27	2.67
DLC-criterion 2	1.34	2.13	3.47
Regression	3.51	1.26	4.76
Scrolling GM (1,1)	5.97	5.28	11.25
Elasticity coefficient	1.22	8.21	9.43



(a) Beijing

(b) Hangzhou

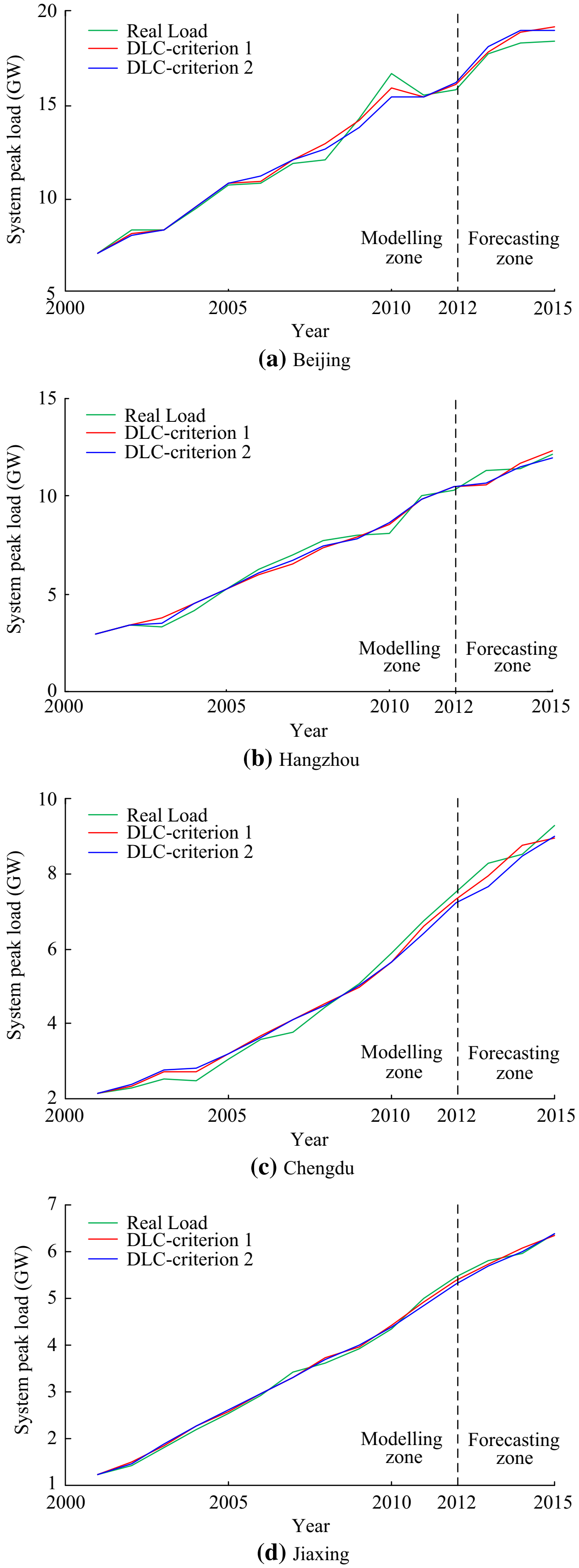
(c) Chengdu

(d) Jiaxing

Fig. 8 Corresponding substation load data

We can see that after relaxing the clustering criterion, the number of the subsequences has further reduced to 21 ($M = 21$), and each of them is smoother. The average modelling error is 7.64%, forecasting error 14.96%, and random error 7.32%. After summing them up to obtain the system results in Fig. 6b, the modelling error is 1.34%, the forecasting error 3.47%, and the random error 2.13%.

Generally, a relaxed criterion will result in smoother load curves with a lower number of clusters, which is advantageous to the modelling accuracy but disadvantageous to reducing the random error. A stricter criterion will lead to an opposite effect. Therefore, an ideal clustering criterion should be a proper compromise between the number of clusters and smoothness of load curves, so that the forecasting accuracy can be optimized. In order to obtain such an optimal criterion, characteristics of the load should be considered, and clustering results with different criteria should be analyzed and compared.



(a) Beijing

(b) Hangzhou

(c) Chengdu

(d) Jiaxing

Fig. 9 DLC modelling and forecasting results

4.4 Classical methods

For further comparison, the classical scrolling GM(1,1) model, the elasticity coefficient model and a regression model are constructed to forecast the system power load directly [42]. The modelling and forecasting results are illustrated in Fig. 7.

The results show that all the three classical models have an adequate modelling accuracy: 5.973.51 and 1.22% respectively. However, they have common difficulty in capturing the changing load pattern in the forecasting zone, with the forecasting errors 11.25, 4.76 and 9.43% respectively.

The final system forecasting results of forecasting methods applied are summarized in Table 1.

In order to demonstrate the adaptability of the proposed DLC method, we collected substation load data from four different cities, which are shown in Fig. 8. Note that the four cities are in different stage of urbanization. The modelling and forecasting results of the DLC method with both criteria for each city are also plotted in Fig. 9, and the forecasting errors are shown in Table 2.

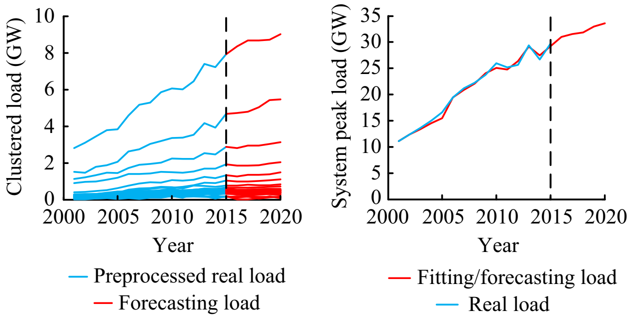
The comparison of results in Table 1 proves the effectiveness of the proposed DLC forecasting method. Firstly, the three methods based on the data-driven methodology all show a successful reduction of random error when compared with the other methods. Secondly, the DLC method can provide modelling accuracy at almost the same level as the direct method, so that the forecasting accuracy is improved. The DLC method also performs better than the three classical models mentioned above. Additionally, the forecasting results for four different cities shown in

Table 2 DLC forecasting error in four different cities

Method	DLC forecasting error (%)				
	Beijing	Hangzhou	Chengdu	Jiaxing	Shanghai
DLC-criterion 1	2.33	3.25	3.20	1.50	2.67
DLC-criterion 2	2.67	2.74	3.81	0.87	3.47

Table 3 DLC forecasting results in Shanghai from 2016 to 2020

Year	Forecasting value (GW)	Annual growth rate (%)
2016	30.99	6.14
2017	31.54	1.79
2018	31.84	0.94
2019	32.98	3.57
2020	33.59	1.85



(a) DLC preprocessed real load and the corresponding optimal ARIMA forecasting load

(b) System load fitting and forecasting results in Shanghai

Fig. 10 Load curves forecast to 2020 using the proposed DLC method

Table 2 indicate the adaptability and stability of the DLC method.

Finally, we forecast the peak load in Shanghai from 2016 to 2020, based on the proposed DLC method with criterion 1 as an example in Table 3. Figure 10 shows load growth and ARIMA forecasting results for each cluster and the overall peak load growth.

The modelling error is 1.92%, and the average annual growth rate from 2016 to 2020 for peak load in Shanghai is 2.86%.

5 Conclusion

In this paper, we propose a data-driven linear clustering method to solve the long-term system load forecasting problem caused by load fluctuations in some developed cities. In order to grasp the internal structure of the system load and improve the forecasting accuracy, we introduce a data-driven method to conduct modelling and forecasting based on a large quantity of substation load data. We have theoretically proved that this data-driven method is effective in reducing the forecasting random error so that a more stable result can be obtained. However, the data-driven method can result in modelling difficulty, which is disadvantageous for forecasting accuracy. For this problem, we propose a linear clustering method to preprocess the substation load data, making it more smooth and thereby reducing the modelling error. When applied to load data from Shanghai the proposed DLC method is shown to be effective in both reducing the forecasting random error and guaranteeing the modelling accuracy, so that a more stable and accurate system load forecasting result can be obtained. Furthermore, applying the same method to load data from another four cities indicates that the proposed DLC method is adaptable and stable.

Future work could theoretically investigate the optimal clustering criterion and level to further improve the forecasting stability and accuracy. Meanwhile, substation load curves in the same cluster have a linear complementarity property, which provides an opportunity to conduct correlation analysis of urbanization characteristics such as industrial structure, population and land utilization. This would help to understand and quantify the structural influences behind the changing peak load.

Acknowledgement This work was supported by the National Energy (Shanghai) Smart Grid Research Center and the National Natural Science Foundation of China (No. 51377103).

References

[1] Kang CQ, Xia Q, Zhang BM (2004) Review of power system load forecasting and its development. Autom Electric Power Syst 28(17):1–11

[2] Atsawathawichok P, Teekaput P, Ploysuwan T (2014) Long term peak load forecasting in Thailand using multiple kernel Gaussian Process. In: 11th International conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON), 2014, Nakhon Ratchasima, Thailand, 14–17 May 2014, pp 1–4

[3] Box GE, Jenkins GM, Reinsel GC et al (2015) Time series analysis: forecasting and control. Wiley, New York

[4] Zhang DQ (2009) Application of improved Gray Verhulst model in middle and long term load forecasting. Power Syst Technol 33(18):124–127

[5] Carpinone A, Langella R, Testa A, et al (2010) Very short-term probabilistic wind power forecasting based on Markov chain models. In: IEEE 11th international conference on probabilistic methods applied to power systems (PMAPS), Singapore, 14–17 June 2010, pp 107–112

[6] Sanstad AH, McMenamin S, Sukenik A et al (2014) Modelling an aggressive energy-efficiency scenario in long-range load forecasting for electric power transmission planning. Appl Energy 128:265–276

[7] Cohen J, Cohen P, West SG et al (2013) Applied multiple regression/correlation analysis for the behavioral sciences. Routledge, London

[8] Hyndman RJ, Fan S (2010) Density forecasting for long-term peak electricity demand. IEEE Trans Power Syst 25(2):1142–1153

[9] Bian H, Wang X (2014) Saturated load forecasting based on nonlinear system dynamics. In: Weigo WANG (ed) Proceedings of the second international conference on mechatronics and automatic control. Springer, Berlin, pp 353–362

[10] Ghelardoni L, Ghio A, Anguita D (2013) Energy load forecasting using empirical mode decomposition and support vector regression. IEEE Trans Smart Grid 4(1):549–556

[11] Torrini FC, Souza RC, Oliveira FLC et al (2016) Long term electricity consumption forecast in Brazil: a fuzzy logic approach. Socio-Econ Plan Sci 54:18–27

[12] Yang M, Lin Y, Zhu SM et al (2015) Multi-dimensional scenario forecast for generation of multiple wind farms. J Mod Power Syst Clean Energy 3(3):361–370. doi:10.1007/s40565-015-0110-6

[13] Cui MJ, Ke DP, Gan D et al (2015) Statistical scenarios forecasting method for wind power ramp events using modified neural networks. J Mod Power Syst Clean Energy 3(3):371–380. doi:10.1007/s40565-015-0138-7

[14] Shahid Y, Tony S (2008) China urbanizes: consequences, strategies, and policies. World Bank Publications, New York

[15] Li YY, Han D, Yan Z et al (2016) Saturated load forecasting model under complex urbanization characteristics. Power Syst Technol 40(9):2824–2830

[16] Lee WJ, Hong J (2015) A hybrid dynamic and fuzzy time series model for mid-term power load forecasting. Int J Electr Power Energy Syst 64:1057–1062

[17] Shao Z, Gao F, Yang SL et al (2015) A new semiparametric and EEMD based framework for mid-term electricity demand forecasting in China: hidden characteristic extraction and probability density prediction. Renew Sustain Energy Rev 52:876–889

[18] Shao Z, Gao F, Zhang Q et al (2015) Multivariate statistical and similarity measure based semiparametric modelling of the probability distribution: a novel approach to the case study of mid-long term electricity consumption forecasting in China. Appl Energy 156:502–518

[19] Zhao H, Guo S (2016) An optimized grey model for annual power load forecasting. Energy 107:272–286

[20] Ardakani FJ, Ardehali MM (2014) Long-term electrical energy consumption forecasting for developing and developed economies based on different optimized models and historical data types. Energy 65:452–461

[21] Zhijian Qu, Chen G (2015) Big data compression processing and verification based on Hive for smart substation. J Mod Power Syst Clean Energy 3(3):440–446. doi:10.1007/s40565-015-0144-9

[22] Wang P, Liu B, Hong T (2015) Electric load forecasting with recency effect: a big data approach. Hugo Steinhaus Center, Wroclaw University of Technology

[23] Hong T, Wilson J, Xie J (2014) Long term probabilistic load forecasting and normalization with hourly information. IEEE Trans Smart Grid 5(1):456–462

[24] Xie J, Hong T, Stroud J (2015) Long-term retail energy forecasting with consideration of residential customer attrition. IEEE Trans Smart Grid 6(5):2245–2252

[25] Azad HB, Mekhilef S, Ganapathy VG (2014) Long-term wind speed forecasting and general pattern recognition using neural networks. IEEE Trans Sustain Energy 5(2):546–553

[26] Quilumba FL, Lee WJ, Huang H et al (2015) Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities. IEEE Trans Smart Grid 6(2):911–918

[27] Shen C, Qin J, Sheng WX et al (2016) Study on short-term forecasting of distribution transformer load using wavelet and clustering method. Power Syst Technol 40(2):521–526

[28] Xiao B, Nie P, Mu G et al (2015) A spatial load forecasting method based on multilevel clustering analysis and support vector machine. Autom Electric Power Syst 39(12):56–61. doi:10.7500/AEPS20140520001

[29] Niu XD, Wei YN (2013) Short-term power load combinatorial forecast adaptively weighted by FHNN similar-day clustering. Autom Electric Power Syst 37(3):54–57

[30] Grzegorz D (2015) Pattern similarity-based methods for short-term load forecasting Part 2: models. Appl Soft Comput 36:422–441

[31] Kálmán T, Lóránt K, András O et al (2016) Classification for consumption data in smart grid based on forecasting time series. Electric Power Syst Res 141:191–201

[32] Liu D, Wang JL, Wang H (2015) Short-term wind speed forecasting based on spectral clustering and optimised echo state networks. Renew Energy 78:599–608

[33] Goia A, May C, Fusai G (2010) Functional clustering and linear regression for peak load forecasting. Int J Forecast 26(4):700–711

[34] Box GEP, Pierce DA (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc 65(332):1509–1526

[35] Pappas SS, Ekonomou L, Karamousantas DC et al (2008) Electricity demand loads modelling using Auto Regressive Moving Average (ARMA) models. Energy 33(9):1353–1360

[36] Pesaran MH (2007) A simple panel unit root test in the presence of cross-section dependence. J Appl Econom 22(2):265–312

[37] Zhang GP (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159–175

[38] Hu S (2007) Akaike information criterion. Center for Research in Scientific Computation, 2007

[39] Wei WWS (1994) Time series analysis. Addison-Wesley Publication, New Jersey

[40] Tong X, Kang CQ, Chen QX et al (2014) Virtual bus technique and its application Part II: virtual bus load forecasting. Proc CSEE 34(7):1132–1139

[41] Shanghai Mulniciple Statistics Bureau (2015) Shanghai statistical yearbook. China Statistics Press, Beijing

[42] Kang CQ, Xia Q, Liu M (2007) Power system load forecasting. China Electric Power Press, Beijing

Yiyan LI received the B.E. degree in Electrical Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014, where he is currently pursuing the Ph.D. degree. His current research focuses on the mid-long term load forecasting.

Dong HAN received the B.E. degree and the M.E. degree in Harbin Institute of Technology University, Harbin, China, in 2007 and 2009, respectively, and the Ph.D. degree in Shanghai Jiao Tong University, Shanghai, China in 2016. His current research interests include optimal operation of power system and electricity market.

Zheng YAN received the B.E. degree in Shanghai Jiao Tong University, Shanghai, China, in 1984, the M.E. degree and the Ph.D. degree in Tsinghua University, Beijing, China, in 1987 and 1991, respectively. His main research interests include optimal operation of power system, electricity market and stability analysis of power system.