A Short Term Forecasting Method for Regional Power Consumption Considering Related Factors

Chang Liu*, Yuanliang Zhang, Weisong Chen, Haitong Gu, Hui Li, Shaoliang Chen

Metrology Centre of Guangzhou Power Supply Bureau, Guangdong Power Grid Co., Ltd. Guangzhou, China

* Corresponding author: 18202710433@139.com

Abstract—Analysis and prediction of power consumption law is the basis of power grid planning and construction, and is also an effective guide for energy demand side management. With the rapid development of economy and the complex change of industrial structure in recent years, the internal structure of power demand is changing to some extent. Therefore, a short-term forecasting method of regional electricity consumption considering the related factors is proposed. Based on the analysis results, a short-term prediction model of regional electricity consumption considering the related factors is established, and the short-term prediction is realized by the calculation of the model. Through the example analysis, it is verified that the forecasting deviation of the short-term forecasting method is low and meets the basic requirements of electric quantity forecasting.

1. INTRODUCTION

The medium and long term load forecasting is mainly aimed at the load demand of monthly time scale and above. In recent years, China's reform and opening up into the deep-water zone, the domestic economic structure is undergoing profound adjustment, load demand changes and economic and social development. At the same time, with the advent of the information age, the statistical accuracy of data, the frequency of publication, and the degree of informatization continue to increase [1-3]. Detailed industry and business climate indexes provide sufficient data sources and data foundations for medium and long-term load forecasting.

In fact, the development of load demand can be divided into trend part, cycle part and other parts. Based on the monthly electricity consumption of the whole society, the different components are extracted by the method of seasonal decomposition, combined with a variety of economic data, and targeted models are used to forecast the electricity consumption [4-6]. However, the development trend of electricity demand of the whole society is relatively clear, and the macro external environment is relatively stable, and the change of influencing factors is relatively small. But the factors affecting the maximum load are more various, among which the regional load and regional meteorological factors is close [7]. At the same time, there is a complex nonlinear relationship between the related factors and load in the medium and long term prediction. Most of the above literatures use local nonlinear models to fit. Firstly, aiming at the problem that it is difficult to distinguish the dominant factors of the maximum load [8]. This paper adopts the method of constructing the correlation coefficient matrix to select the influential factors with strong correlation with the maximum load as the input variables; secondly, aiming at the problem that the regression model may overfit, this paper adopts principal component analysis (PCA) to reduce the dimension of the sample data, so as to avoid
the overfitting effect caused by the contradiction between the sample size and the type of influencing factors. Finally, aiming at the limitation of the local nonlinear model, this paper uses several nonlinear models to forecast, and analyzes the applicability of the model and the ability of data feature extraction.

2. **INFLUENCING FACTORS AND ANALYSIS OF RELATED FACTORS**

The definition of the Granger causal test in this case can be shown in Figure 1. Granger causality test was initially used to analyze the causality between economic variables. Later, with the convergence of various research fields, scholars gradually applied this method to the study of causality between economic variables and variables in other fields, such as the Granger causality test for electricity consumption and macroeconomics, as shown in Figure 1:

![Granger causality test of electricity and macroeconomics](image)

**Figure 1.** Granger causality test of electricity and macroeconomics

One of the prerequisites of Granger causality test for time series is that the series must be stationary, or false regression may occur. Therefore, the unit root test for the stationarity of each time series should be performed before performing the Granger causality test. The augmented Dicky-Fuller test (ADF unit root test) is a commonly used unit root test. The general steps of the ADF unit root test are as follows:

1. Verifying the original time series;
2. If the original time series fails to pass the test, that is, the original series is not stable, the smoothness shall be tested after the first order difference of the series;
3. If the first-order difference sequence has not passed the test, the second-order difference transformation shall be carried out continuously, and the time sequence after the second-order difference is generally stationary.

3. **SHORT-TERM FORECASTING MODEL OF REGIONAL ELECTRICITY CONSUMPTION CONSIDERING RELATED FACTORS**

In this paper, the Lagrangian interpolation method is used to process the data. The shortest number of edges on the path connecting the two nodes \( v_i \) and \( v_j \) in the network represents the distance between the two nodes \( d_{ij} \), and \( d_{ij} \) represents the efficiency between the nodes \( v_i \) and \( v_j \), and represents the speed of information transmission, and has \( \varepsilon_{ij} = 1/d_{ij} \). \( d_{ij} = 0 \), \( \varepsilon_{ij} = 0 \) denotes a pathless connection between two nodes. The diameter of the network \( D \) is expressed in terms of the maximum distance between any two nodes of the network, namely:

\[
D = \max d_{ij} \quad (1)
\]

In formula (1) \( 1 < i < N \), \( N \) is the total number of network nodes, where \( N = 340 \). The diameter of the dynamic Bayesian network is calculated to be 14. The average distance between any two nodes represents the average path length \( L \) of the network:

\[
L = \frac{1}{C^2} \sum_{\gamma=1}^{n} d_{\gamma} \quad (2)
\]

The diameter of the network can be understood as the maximum relational chain in the network. The average path length of the dynamic Bayesian network is 4.8439. The neighbor of a node in the network is defined as the node to which it is directly connected by an edge. The cluster coefficient \( C_i \) of node
\( v_j \) is defined as the number of \( E_j \) edges that actually exist between neighbouring nodes of \( k_i \) in \( v_j \) divided by the total number of possible edges. In order to reduce the influence of each component forecast error on the overall forecast results, the relationship between each component is expressed in addition model:

\[
C_j = \frac{L}{C^2_{ki}} E_j \quad (3)
\]

Among them, the triple connected with node \( i \) refers to the three nodes including node \( i \), and there are at least two edges from node \( i \) to the other two nodes. As shown in Figure 2:

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure.png}
\caption{Two possible forms of triples with node \( i \) as one of the vertices}
\end{figure}

Spearman correlation coefficients indicate the correlation direction of data sets \( X \) and \( Y \), where \( X \) is called an independent variable and \( Y \) is a dependent variable. If the value of \( Y \) tends to increase with the increase of \( X \) value, the correlation coefficient of Spearman is positive. If the value of \( Y \) tends to decrease with the increase of \( X \) value, the correlation coefficient of Spearman is negative. If the Spearman correlation coefficient is zero, it indicates that there is no tendency for the value of \( Y \) to change as the value of \( X \) increases. If a node of the network \( v \) is passed by the shortest path of many other non-adjacent nodes, then the node's role in the network is irreplaceable. The influence and importance of a node is expressed by its intersection number \( B \):

\[
B_j = \sum_{j=1}^{n} \left[ \frac{n_{j,i}(i)}{n_{j,i}} \right] \quad (4)
\]

In formula (4), the shortest number of paths between nodes \( v_i \) and \( v_{j} \) shall be expressed by \( n \); the shortest number of paths between \( v_i \) and \( v_j \) passing through nodes \( v_j \) shall be expressed by \( n \) and \( (i) \), and the total number of nodes shall be expressed by \( N \). In order to avoid the influence of overfitting caused by various data types and complex regression models on the prediction accuracy, the dimensionality of each component data should be reduced. The main cause analysis (PCA) method is used to deal with the factors that have passed the cointegration test. Firstly, Z-score is used to standardize the various factors, and then the covariance matrix is constructed. Secondly, the covariance matrix is solved by using the eigenvalue, and the degradation dimension is determined by using Cumulative Partial Variance (CPV). Finally, the new dimension reduction matrix is reconstructed by using the PCA dimensionality reduction algorithm.

4. EXAMPLE TEST
In order to verify the reliability and validity of the annual electricity consumption forecasting model based on weighted partial least squares regression, this paper uses the annual social and economic development index of a certain area and the data of the whole society to model and forecast. Verify the practicability and reliability of the short-term prediction method of regional electricity consumption which is designed in this paper.

4.1. Case analysis data sources
The data source still uses the monthly data on electricity consumption, weather, economy and transportation in XX city. After processing the missing values of the input data using the hybrid information patching algorithm, the model is constructed using the improved DBN method introduced in 3.3. First, in the first stage of the reasoning process, a first-order dependent scoring matrix, \( Si \) (340 ×
340), is obtained. The element in S1 is the score of each edge of a directed acyclic graph, which is the inference value. The lower the score of DBN, the more significant the first order dependency is. The first five rows and the first five columns of the first step scoring matrix, S, are listed here. V plus subscripts represent variable names, as shown in Table 1:

| TABLE 1. ESTABLISHMENT OF SCORING MATRIX |
|------------------------------------------|
|   | V1   | V2   | V3   | V4   | V5   |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| V1               | 0.730247        | 0.999279        | 0.011232        | 0.007797        | 0.197536        |
| V2               | 0.089954        | 0.000061        | 0.000134        | 0.254780        | 0.245780        |
| V3               | 0.745219        | 0.384123        | 0.854216        | 0.001245        | 0.236980        |
| V4               | 0.145863        | 0.256874        | 0.315522        | 0.245879        | 0.236587        |
| V5               | 0.124580        | 0.264781        | 0.214587        | 0.236985        | 0.214586        |

Enter the time series data of the first step scoring matrix S and 340 variables, the given time window width = 0.1, if the number of time slices in the time window is still less than the time window width, the process is repeated after the evidence information of the next time slice is obtained. When the number of time slices is equal to the width of the time window, and the reasoning results are updated, then the reasoning process of the second stage is entered. The first five rows and the first five columns of the Sz scoring matrix are listed here. Since a lower score indicates a more significant causal relationship, an inference value greater than 0.1 is set to NA. As shown in Table 2:

| TABLE 2. ESTABLISHMENT OF SCORING MATRIX2 |
|------------------------------------------|
|   | V1   | V2   | V3   | V4   | V5   |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| V1               | NA              | NA              | 0.011232        | NA              | NA              |
| V2               | 0.089954        | 0.000061        | 0.000134        | NA              | 0.245780        |
| V3               | NA              | NA              | NA              | NA              | 0.236980        |
| V4               | NA              | NA              | NA              | NA              | NA              |
| V5               | 0.124580        | NA              | NA              | NA              | NA              |

Visualize the DBN complex network using the sna package in R [50-51]. From the scoring matrix S, we can get the directed edges of the network, which are sorted according to their scores from small to large. The first column is the predictive variable, the second is the explanatory variable, and the third is the corresponding edge score. There are 650 dependencies. The adjacency matrix A (340 × 340) can be constructed from the directed edge case of a network, and the element a in the matrix is 0 or 1. A_{i,j} = 1 means that the j variable is linearly dependent on the i variable; a = 0 means that the j variable is not linearly dependent on the i variable. Only the first five rows and the first five columns of the network adjacency matrix are listed here, as shown in Table 3:

| TABLE 3. ADJACENCY MATRIX OF NETWORK |
|--------------------------------------|
|   | V1   | V2   | V3   | V4   | V5   |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| V1               | 0               | 0               | 1               | 0               | 0               |
| V2               | 1               | 1               | 1               | 0               | 1               |
| V3               | 0               | 0               | 0               | 0               | 1               |
| V4               | 0               | 0               | 0               | 0               | 0               |
| V5               | 1               | 0               | 0               | 0               | 0               |

The proportion of all the shortest paths that pass through the node in the network is the number of intervals. List the top 10 variables in number of mediations, as shown in Table 4:

| TABLE 4. TOP TEN INTERMEDIARIES |
|--------------------------------|
| Serial number | Variable name                              | Medium number |
| 1             | Domestic electricity consumption of urban and rural residents | 2635.1        |
| 2             | Electricity consumption of secondary industry | 2041.78       |
| 3             | Mean atmospheric pressure                  | 1925.41       |
| 4             | Electricity consumption of the whole society | 1360.24       |
| 5             | Electricity consumption in textile         | 1266.32       |
industry and light textile industry
6 Electricity consumption in construction industry. 1118.56
7 Electricity consumption in smelting nonferrous metals 1059.42
8 Total electricity consumption of physical stores 941.23
9 Total electricity consumption of gas industry 841.23
10 Electricity for animal husbandry. 814.52

For the data in Table 4, there is a complex non-linear relationship between load density and load density because of the close relationship between load density and load density. If the input sample data of the prediction model is not properly selected, the prediction accuracy of the model will be greatly reduced. Therefore, it is necessary to introduce the appropriate sample at the beginning of the prediction model. In this section, based on the spatial load forecasting of LSSVM load density method, grey relational analysis is introduced to improve the forecasting precision and speed of LSSVM.

4.2. Application results
The steps of selecting LSSVM model training samples with grey incidence degree are: 1) nondimensionality of the sample data of original influencing factors; 2) difference sequence; 3) maximum and minimum difference between two poles; 4) correlation coefficient; 5) calculation of incidence degree. Taking the load density prediction of a certain residential area as an example, the validity of the model proposed in this paper is analyzed and verified. A large sample of residential load and its influencing factors is collected as shown in Table 4.1. Among them: A, population density; A, per capita income; A, per capita electricity; A, per capita coal-electricity price ratio growth rate; D, load density. As shown in Table 5:

| Serial number | A1/(person/km²) | A2/(yuan) | A3/(Kw·h) | A4/(%) | D/(Kw·h) |
|---------------|-----------------|-----------|-----------|--------|-----------|
| 1             | 30651           | 1325      | 519       | 1.2547 | 11.2364   |
| 2             | 16150           | 4837.42   | 1033      | 1.0635 | 27.62     |
| 3             | 29410           | 1122      | 537       | 1.677  | 10.3      |
| 4             | 25413           | 1000.3    | 834       | -1.313 | 11.58     |
| 5             | 16600           | 1222.7    | 1019      | 2.365  | 2.54      |
| 6             | 15855           | 2608.4    | 628       | 1.6587 | 1.117     |

Select these four variables as explanatory variables to forecast the 12 months power consumption of Shanghai in 2020. The gap between the actual and predicted values of the method designed in this article is shown in Figure 3:

![Figure 3. Prediction results](image)

From the Figure 3, we can draw the conclusion that the improved DBN model proposed in this paper is more effective than the other models. The classical model VAR is easy to calculate, but the
error is also the largest among all the models. It shows that the prediction of 12 time points has exceeded the range of its accuracy. BP network does not have memory, at the same time only consider the current time point data, and do not consider the characteristics of the time series before and after the connection, so the error is second only to the VAR model, is also very large. The traditional DBN combines the causal relationship and the Markov model to describe the probability transfer of time series, so the error is obviously reduced. But it is a Bayesian network based on steady-state assumption. Elastic network and Granger causality test model analyze time sequence before causality, but DBN model can not analyze time sequence and causality at the same time. First, the output data is processed by using mixed information patching algorithm, then the improved DBN model is built by using data analysis software R. By relaxing the traditional DBN's assumption of time homogeneity, the disadvantage that the traditional DBN can not model the unsteady process is solved. Finally, the monthly data of electricity consumption, weather, economy and transportation in Shanghai from 2011 to 2017 are used to forecast the electricity consumption of the whole society for 12 months in 2020. Compared with VAR (15.1%), BP (10.8%), traditional DBN (7.37%) and elastic network model (3.07%), the improved DBN model proposed in this paper is superior to other algorithms in the prediction accuracy of monthly electricity consumption (1.63%).

5. Conclusion
In this paper, a short-term forecasting method of regional electricity consumption considering related factors is proposed. The main conclusions are as follows: (1) In this paper, a monthly data set of 340 variables including electricity consumption, meteorology, economy and transportation, covering 96 time points from 2011 to 2020 is established. (2) A prediction method for monthly electricity consumption of elastic network model is proposed. After processing the missing value of the input data by multiple interpolation method, the paper uses elastic network and Granger to carry out two-step factor screening so as to achieve the goal of dimension reduction. (3) An improved dynamic Bayesian network forecasting method for monthly electricity consumption is proposed to further improve the forecasting accuracy. By relaxing the traditional assumption of time homogeneity in dynamic Bayesian networks, the unsteady process can be modeled, and the algorithm flow is designed. At the same time, the complex network method is used to analyze the dynamic Bayesian network. Compared with VAR (15.1%), BP (10.8%), traditional DBN (7.37%) and elastic network model (3.07%), the improved DBN model is superior to other algorithms in prediction accuracy (1.63%), and the advanced property of the improved DBN is verified by actual data.

Acknowledgment
The study was supported by “Science and Technology Project of Guangzhou Power Supply Bureau, China (Grant No. GZHKJXM20200036)”.

References
[1] Aladin S , Tran A , Allogba S , et al. (2020) Quality of Transmission Estimation and Short-Term Performance Forecast of Lightpaths. Journal of Lightwave Technology, , PP(99):1-1.
[2] Hong D Y , Ji T Y , Li M S , et al. (2019) Ultra-short-term forecast of wind speed and wind power based on morphological high frequency filter and double similarity search algorithm. International Journal of Electrical Power & Energy Systems, 2019, 104(JAN.):868-879.
[3] Kang H , Lee K , Shin D H . (2019) Short-term Forecast Model of Apartment Jeonse Prices using Search Frequencies of News Article Keywords. KSCE Journal of Civil Engineering, 23(12):4984-4991.
[4] Roushangar K , Alizadeh F . (2019) Scenario-based prediction of short-term river stage-discharge process using wavelet-EEMD-based relevance vector machine. Journal of Hydroinformatics, 21(1-2):56-76.
[5] Allogba S , Aladin S , Tremblay C . (2021) Multivariate Machine Learning Models for Short-Term Forecast of Lightpath Performance. Journal of Lightwave Technology, PP(99):1-1.
[6] Krylov S S, Perepelkin V V, Soe W Y. (2020) Short-Term Forecast of the Earth's Pole Motion, Taking Into Account Lunar Disturbances. Mechanics of Solids, 55(6):892-897.

[7] Tang Q, Yang M, Yang Y. (2019) ST-LSTM: A Deep Learning Approach Combined Spatio-Temporal Features for Short-Term Forecast in Rail Transit. Journal of Advanced Transportation, 2019:1-8.

[8] Kka C, Aba C, Mgb C. (2021) Analysis of short-term operational forecast deviations and controllability of utility-scale photovoltaic plants. Renewable Energy, 167:343-358.