A spatial load forecasting method based on DBSCAN clustering and NAR neural network

Zhentao Han[1], Mengzeng Cheng[1], Fangxi Chen[2,4], Yanze Wang[3] and Zhuofu Deng[2]

[1]Economic Research Institute, State Grid Liaoning Electric Power Supply Co., Ltd., Shenyang, Liaoning, 110015, China

[2]College of Software, Northeastern University, Shenyang, Liaoning, 110169, China

[3]Kuandian Electric Power Supply Company, State Grid Liaoning Electric Power Supply Co., Ltd., Kuandian, Liaoning, 118200, China

[4]Corresponding author's e-mail: 924290841@qq.com

Abstract. In order to improve the accuracy of spatial load forecasting in power grid planning stage, a spatial load forecasting method based on density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm and nonlinear auto regressive (NAR) neural network is proposed. This method consists of three stages: cell division, clustering, and forecasting. At first, zones are divided into cellules that are taken as the basic unit of spatial load forecasting. Historical yearly load profiles, along with geographic information and land use types, are extracted from cells as features. Furthermore, similar cells are classified into several clusters according to these features. Finally, a NAR neural network is established to forecasting load one year ahead for each cluster, where the historical load profiles are taken as input. Experiments reveal that our proposed model decreases MAE by 45.95%, 42.04% and 47.49% respectively compared with linear regression, grey theory and exponential smoothing, showing great improvements in accuracy.

1. Introduction

Distribution planning is an important part of power system. The design of distribution network is related to the efficiency and quality of power supply. At present, distribution network planning not only needs to be able to predict the total electricity power load, but also to predict the spatial distribution of future load. Only by forecasting the growth and distribution of load, can we plan the location, capacity, feeder range, etc. of the substation. Spatial load forecasting (SLF) [1,2] is significant for guiding grid planning, and the accuracy of its results directly affects the rationality of the plan. The spatial load forecasting attracts ever-increasing attention from relevant departments.

Common SLF methods could be classified into four categories: land usage methods, load density index methods, multi-variate methods, and trend methods. Land usage methods are rough in obtaining load density, and it has no advantage in situation that the land usage is determinate [3]. Load density index methods encounter the problem that the growth of classified load is uneven in different cells [4-6]. The valid forecasting period of multi-variate methods is quite short. Trend methods ignore relationships between load and influence factors, making it difficult to catch the law of load change [7]. Nowadays, the development of artificial intelligence technologies, including neural networks, provides us with more excellent options [8].

In this paper, a spatial load forecasting method based on density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm and nonlinear autoregressive (NAR) [9,10] neural network is proposed. This method consist of three stage: cell division, clustering, and forecasting. Firstly, the area to be predicted is divided into cells with equal size, which is taken as the basic unit of forecasting. Historical yearly load profiles, along with geographic information and land use types, are extracted from cells as features. Next, According to these features, DBSCAN is adopted to cluster cells with similar features into a group, where outliers are excluded. Finally, the NAR prediction model is established for each kind of cell to fit characteristics of growth. Experiment in a city power grid shows that the proposed method is effective in improving forecasting accuracy.

2. Phases of the proposed method

Our proposed method is mainly concluded into three steps. At first, areas are divided into cellules that is taken as the basic unit of spatial load forecasting. Historical yearly load profiles, along with geographic information and land use types, are extracted from cells as features. Furthermore, similar cells are classified into several cluster according to these features. Finally, a NAR neural network is established to forecasting load one year ahead for each cluster, where the historical load profiles are taken as input. The flow chart is shown in the figure1.

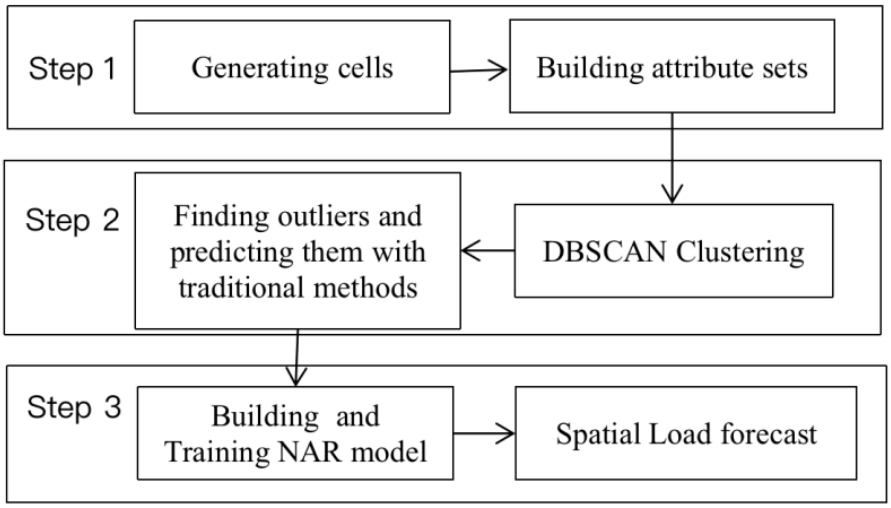

Figure 1. The flow chart of our proposed method

3. Cell features analysis

Area to be forecasted is firstly divided into cells with equal size, which is recorded as $C_{m,n}$, where m and n represent cell location in the m-th row and n-th column. Cell division is a necessary step in SLF, which aims to predict load growth related to spatial information for the planning of distribution network.

Many factors could affect the load of cells, such as historical load, land type (industrial land, residential land, commercial land, public land, and others), geographic information (distances from cells to streets, subway stations or schools), density of population, temperature, regional development level, etc. For a certain administrative region of a city, the economic development level and temperature in each cell are basically the same, so we will not discuss it here. Therefore, the historical load, land use type and geographic information are recorded as the features $X=\{X_1, X_2, X_3\}$ of cells.

Among them, the historical loads of cells are provided by relevant departments, which are expressed as $X_1= \{x_{1.1}, x_{1.2}, ..., x_{1.l}\}$. Because a cell may contain multiple land types, the proportion of each land type in the cell is used as its features, which are recorded as $X_2= \{x_{2.1}, x_{2.2}, ..., x_{2.u}\}$. The distances between $C_{m,n}$ and the streets are recorded as $X_3= \{x_{3.1}, x_{3.2}, ..., x_{3.d}\}$.

4. DBSCAN

According to historical load, land use type and geographic information, similar cells are clustered into the same group, where they share the same prediction model, thus improving the forecasting accuracy. However, some cells hold attributions quite different with all of others, which are called outliers. Outliers lead to more noises in samples, hence increasing errors of forecasts. As a result, we employ DBSCAN to tackle this problem.

DBSCAN divides the area with enough density in the feature space into clusters, and finds clusters with arbitrary shape in the feature space with noises. It defines clusters as the largest set of points density-connected. The density based method can effectively deal with noise points by finding out outliers (because the outliers will not be included in any cluster, they are considered as noises). There are always some cells whose features are different from other groups with large density, so DBSCAN algorithm is the most suitable method to deal with this problem. Compared with K-means and balanced iterative reducing and clustering using hierarchies (BIRCH), DBSCAN is not only suitable for convex sample set, but also for non-convex sample set, and its computational speed is fast.

The essence of DBSCAN is to calculate how many objects (i.e. cells) are included in a given radius. The algorithm includes two parameters, *Eps* and *MinPts*. The core concepts of the algorithm are as follows:

(1) *Eps* domain: *Eps* is the maximum distance between two objects for one to be considered as in the neighbourhood of the other, and the *Eps* domain of an object indicate the space around the object with a radius *Eps*;

(2) Core object: Core objects refers to objects that the number of objects in its *Eps* domain is greater than or equal to *Minpts*;

(3) Direct density-reachable: For any set **C,** in which X^P is a core object and object X^D is in the *Eps* domain of object X^P, then object X^D is direct density reachable from X^P;

(4) Density-reachable: Given an object set $\{X^1, X^2, ..., X^N\}$, if any point X^i in the set is directly density reachable from X^{i-1}, then X^N is density-reachable from X^1;

(5) Cluster: A cluster of objects is defined as a set of objects that are density-reachable from an arbitrary core object;

(6) Noise object: An object is a noise object, if there are less than *MinPts* of objects in its *Eps* domain, and none of which is a core object.

The clustering principle of DBSCAN algorithm can be simply summarized as follows:

(1) According to the input parameters (*Eps*, *MinPts*) and the Euclidean distance between objects, all core objects are found out.

(2) An arbitrary core object is chosen as the initiative object, and all of the density-reachable objects are retrieved. The initiative object and all its density-reachable objects constitute a cluster, and they are marked as object clustered.

(3) Choosing next core object that is not marked yet, and repeating (2) until all core objects are clustered.

(4) Objects that do not belong to any cluster are noise objects.

Since clusters are generated by DBSCAN, we train a forecasting model for each cluster. Load of each noise cells are forecasted by the trend extrapolation method separately.

5. NAR neural network prediction model

It is well known that the load growth curve of a developing region is similar to a curve of logistics function. But more often, the increase of cells' load are phased, which is caused by staged construction or a combination of multiple land types. The load curve in this situation is shown in the figure2.

If the magnitude and period of growth in each phase are widely divergent, linear model such as auto regression cannot fit it precisely. Thereby, we introduce NAR neural network to tackle this problem. NAR neural network is a kind of nonlinear autoregressive model which uses itself as the regression variable, and predicts future outputs based on linear combination of historical load. Only historical load of cells (X_1) is taken as the model input. The structure of NAR neural network is shown in figure 3.

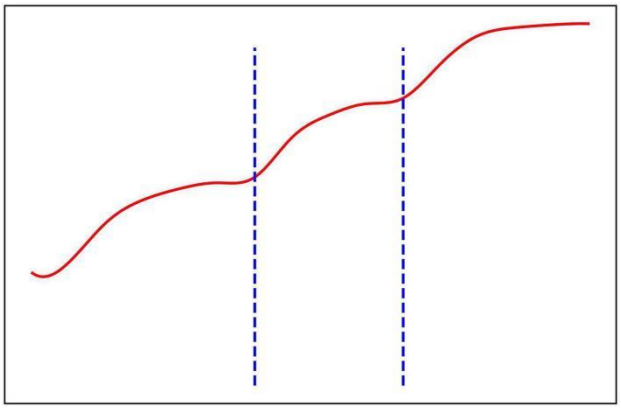

Figure 2. Phases of cells' load growth

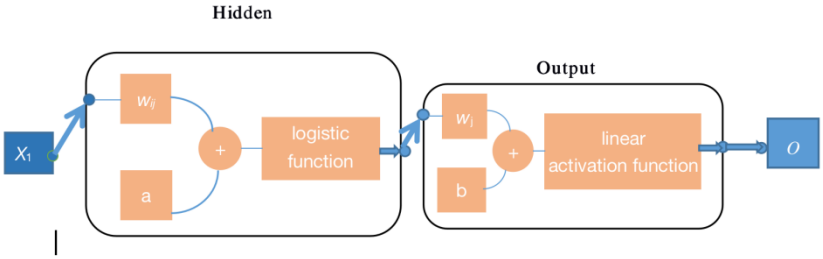

Figure 3. Structure of NAR neural network

For the load curve $X_1 = \{x_{1.1}, x_{1.2}, ..., x_{1.l}\}$, we firstly get the hidden neurons $H = \{h_1, h_2, \ldots, h_k\}$ by:

$$h_j = f\left(\sum_{i=1}^{l} w_{ij} x_{1.i} + a_j\right), \quad j = 1,2,...,k \tag{1}$$

where i represents the time, k represents the number of neurons in the hidden layer, $f(\cdot)$ denotes the activation function (logistic function) of the hidden layer, w_{ij} represents the connection weight between the i-th input and the j-th neuron in the hidden layer, and a_j represents the bias of the j-th neuron in the hidden layer.

Next, the output O is calculated according to the output h_j of the hidden layer:

$$O = g\left(\sum_{j=1}^{l} w_j h_j + b\right) \tag{2}$$

where w_j is the weight of the connection between the j-th neuron in the hidden layer and the output, b represents bias in the output layer, and $g(\cdot)$ is a linear activation function.

In this structure, k logistic functions $f(\cdot)$ are deployed to fit different phases of growth in curves, and the formula of $f(\cdot)$ is:

$$f(x) = \frac{1}{1 + e^{\beta + \gamma x}} \tag{3}$$

where β and γ are learnable parameters, by which the network could fit various situation.

6. Experiment and Analysis

In this experiment, we forecast an area of a city from china. The forecasted area is divided equally into cells with side length of 300m, as shown in the figure 3. The historical maximum cell load $X_1 = \{x_{1.1}, x_{1.2}, ..., x_{1.16}\}$ is acquired from 2000 to 2015. In this experiment, we obtain five types of load usage $X_2 = \{x_{2.1}, x_{2.2}, ..., x_{2.5}\}$ from urban planning department, which are industrial land, residential land, commercial land, public land, and other land. In this area, there are 3 main streets including Zhihui Street

and Shangyuan Street, Yinxing Street. The nearest distance between the cell and each street is taken as $X_3 = \{x_{3.1}, x_{3.2}, x_{3.3}\}$, which got from GIS. DBSCAN clustering analysis is carried out according to the cell feature set, and the cell clustering results are shown in Figure 4. As shown, cells in this area are clustered in 7 group, which are displayed in 7 different colours respectively. Outliers is shown in black.

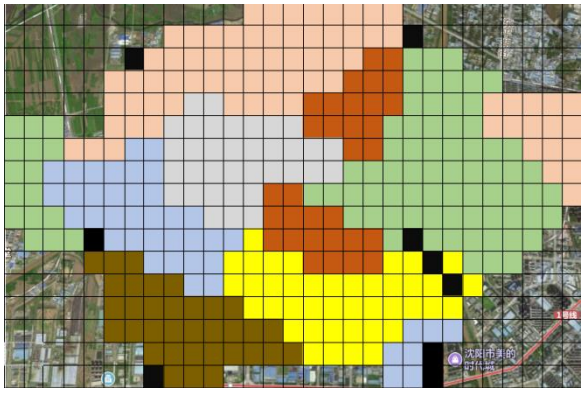

Figure 4. The clustering result of cells.

We train 7 NARs for 7 groups respectively, and then forecast load in 2016. In addition, trend extrapolation is used to predict outliers separately. At the same time, the prediction methods of linear regression, grey theory and exponential smoothing are used for comparative analysis with the proposed method. Table 1 shows the actual load of the randomly selected cells in 2016 and the load forecasts of each method.

Experiment results are shown in table 1, from which we can see that the Mean Absolute Error (MAE) of proposed method is obviously lower than compared methods. Specifically, compared with linear regression, grey theory and exponential smoothing, our model decreases the MAE by 45.95%, 42.04% and 47.49%, respectively. The compared methods do not consider the characteristics of the spatial load, thus getting forecasts with large MAE. On the contrary, our proposed method precisely fit the phased characteristic of spatial load profile, hence improving the accuracy of forecasts.

Table 1. Load forecasts and MAE of several methods.

Cell Number	Ground truth	Linear Regression		Grey Theory		Exponential Smoothing		Proposed	
	Load (KW)	Load (KW)	MAE (KW)	Load (KW)	MAE (KW)	Load (KW)	MAE (KW)	Load (KW)	MAE (KW)
$C_{8,5}$	317.35	221.37	95.98	245.32	72.03	198.38	118.97	277.73	39.62
$C_{9,12}$	398.27	320.23	78.04	343.73	54.54	299.18	99.09	361.56	36.71
$C_{10,23}$	408.32	300.71	107.61	333.92	74.4	299.18	109.14	357.26	51.06
$C_{11,2}$	397.73	250.53	147.2	297.62	100.11	279.63	118.1	301.52	96.21
$C_{11,17}$	501.48	433.36	68.12	622.71	121.23	420.82	80.66	450.64	50.84
$C_{12,3}$	424.37	310.26	114.11	276.36	148.01	320.17	104.2	350.48	73.89
$C_{13,20}$	527.37	403.17	124.2	642.72	115.35	400.73	126.64	478.28	49.09
Average			105.04		97.95		108.11		56.77

7. Conclusion

In this paper, we proposed a new three steps method based on DBSCAN and NAR neural network to tackle SLF. Experiments show that the proposed method enormously improves the accuracy of SLF. These improvements can be attributed to:

(1) According to features of cells, we cluster similar cells into several groups. Cells in one cluster share the same forecasting model, thus improving the forecasting accuracy.

(2) In clustering stage, outliers will increase errors of forecasts. This problem is well solved by introducing DBSCAN to find outliers, which are forecasted by the trend extrapolation method separately.

(3) Linear models such as auto regression cannot fit the load profile precisely when magnitude and period of growth in each phase are widely divergent. Therefore, we adopt NAR neural network to tackle this problem successfully.

References

[1] Willis H L 1997 Spatial electric load forecasting [book review]. *Computer Application in Power IEEE* .**10(2)** 58-59

[2] Willis H L and Tram H 1983 A cluster based V.A.I. method for distribution load forecasting. *Power Engineering Review IEEE.* PER-**3(8)** 47-47

[3] Liu Z F, Pang C C, Wang Z L and Li K 2013 Spatial load forecasting for distribution network based on cloud theory and cellular automata. *Proceedings of the CSEE.* **33(10)** 98-105

[4] Xiao B, Yang X Y, Mu G 2014 The load density index method based on the historical data of cellular. *Power System Technology.* **38(4)** 1014-19

[5] Zhu F J, Wang Z D, Lu J, Zhang Q X, Xiang T T 2012 Disequilibrium development areas based classification and subarea method for spatial load forecasting. *Automation of Electric Power Systems.* **36(12)** 41-47

[6] Liu Z F, Pang C C, Wei J and He J 2013 Index calculation of load density based on IAHP and TOPSIS method. *Automation of Electric Power Systems.* **36(13)** 56-60

[7] Xiao B, Mu Gang, Li P, Yang Y H, Yan H and Yang C J 2010 A time series mismatch corrective method for spatial load forecasting. *Automation of Electric Power Systems.* **34(16)** 50-54

[8] Deng Z F, Wang, B B, Xu Y L, Xu T T, Liu C X, Zhu Z L 2019 Multi-Scale convolutional neural network with Time-Cognition for Multi-Step Short-Term load forecasting. *IEEE Access.* **7** 88058 -71

[9] Harris G H and Lapidus L 1967 Identification of nonlinear systems. *Industrial & Engineering Chemistry.* **59(6)** 66-81

[10] Eskinat E, Johnson S H and Luyben W L 2010 Use of Hammerstein models in identification of nonlinear systems. *Aiche Journal.* **37(2)** 255-68