Multidimensional data analysis of load influencing factors in smart distribution network

Xu Pan1,a, Jinli Wang1, Yunshuo Li1 and Wei Xie2
1Power Distribution Department, China Electric Power Research Institute,100192, Beijing, China
2State Grid Shanghai Municipal Electric Power Company, 200122, Shanghai, China
a Corresponding author: yuqiaizi1986@163.com

Abstract. With the increasing scale of data in smart distribution network, the information contained in massive historical data is richer, and more valuable information can be mined. Based on the demand of electrical load analysis, the relationship between load and various potential factors in nature or society is analyzed from multiple dimensions such as time, space and meteorology, and then quickly matches the important factors. On the basis of matching the influencing factors, the intensity of association between factors and load is deeply excavated. And a more refined model, such as load-temperature model, is established to analyze the law of historical data change and predict the trend of future change, so as to support the advanced application of load forecasting.

1. Introduction
With the promotion of the construction of power network information, a large number of power data have been generated in the daily operation of the smart distribution network and these data have not been fully excavated and effectively utilized in a very long time. At present, experts and scholars at home and abroad have begun to study the application of big data in power grid, and have made some achievements. Studies [1-4] analyze some application scenarios in the background of the big data in smart distribution network, and it is considered that the effective use of the power data is an important means to improve the advanced applications such as electricity forecasting. In a wider range of applied research, studies [5-8] carry out large data analysis and mining through massive data of power generation, transmission, distribution and power consumption to realize the diagnosis, optimization and prediction of the power grid operation, and these measures provide guarantee for the safe, reliable, economical and efficient operation of the power grid.

This paper makes full use of the large data resources of electricity information in smart distribution network, and quickly matches the main factors from the potential factors in nature and society, and then constructs the model of load influence factors according to the factors that have been matched, so as to realize the specific quantification of the impact of various factors on the user's load.

2. Frame of multidimensional analysis on the influencing factors of load data
Taking weather and social economy as the main influencing factors, this paper combined with historical user load data to construct the load factor model, which makes the factors of load modeling more comprehensive and meticulous, and provides support for improving the precision of load building. The overall framework design is shown in Figure 1.
(a) Data Acquisition: Take electricity and external data into smart distribution network information platform, these data are derived from the Power User Electric Energy Data Acquire System, the Personal Information Management System, the Energy Management System, the SG186 Marketing System, the Meteorological Information System, and the Social and Economic Data System.

(b) Data Processing: Carry out some measures for load data, such as normalization, exception handling and data fitting, and enables load data to correspond to meteorological and socioeconomic data.

(c) Data Analysis: First of all, start to match various factors quickly according to the selected data. And the cumulative contribution rate of each factor is calculated by PCA algorithm to realize the extraction of main factors. Secondly, the data correlation analysis is carried out. The meteorological and load data are analyzed according to the main factors extracted, and on this basis, the electricity load model of meteorological, social and economic factors is constructed. Thirdly the three-dimensional dynamic display is published on the basis of the built electricity load model. Last but not the least, the model starts to self-learn and update according to the latest meteorological, social and economic and load data.

(d) Data Application: Use load models for load weather analysis and support the advanced application of load forecasting.

3. Process of multidimensional analysis on the influencing factors of load data

3.1 Data Acquisition
The data used in this paper originate from the internal and external systems of the power grid, these internal datas of the power grid include the user's electrical load data in the Power User Electric Energy Data Acquire System, the users of the Personal Information Management System, the power generation data of the Energy Management System, and the historical distribution data in the SG186 Marketing System. Besides, external datas include the meteorological data of the Meteorological Information System, and the socioeconomic data of the Social and Economic Data System, these external datas can also provide auxiliary decision-making support for the operation, management and service of the power grid.

3.2 Data Processing

3.2.1 Data Cleaning
In order to ensure the integrity and accuracy of load data, data loading data need to be cleaned. Taking the load data in the Power User Electric Energy Data Acquire System as an example, the present situation of the load data is as follows:

- The number of data points collected by different types of transformers in one day is different, which is divided into 24 points (one data per 1H), 48 points (one data per 0.5h), and 96 points (one data per 15min).
- Part of the data is not stored in time, resulting in missing data points.
- Some individual data points that have been stored are obviously larger or smaller than the normal load range and these belong to abnormal data.

The load data cleaning is shown in Figure 2.

![Diagram of load data cleaning process](image)

**Figure 2. Frame of load influencing factor model**

(a) Data Judgment: In the processing of super-large or super-small data, the threshold $N$ is determined first, and then judge whether it is abnormal or not in a continuous load data sequence. The formula for determining whether the data is abnormal is as follows:

$$L_i = \begin{cases} 
\text{super-large data} & \text{if } L_i > L_{r+1} \times N & \text{or } L_{i-1} \times N \\
\text{super-small data} & \text{if } L_i < L_{r+1} / N \text{ & } L_i > L_{r+1} / N \\
\text{normal data} & \text{Access to the database}
\end{cases} \quad (1)$$

The revised formula for super-large data and super-small data is

$$L_i = \left( L_{r-1} + L_{r+1} + L_{i-1} + L_{i+1} \right) / 4 \quad (2)$$

When the load data $L_i$ and $L_j$ have been successfully collected, but there are $j-i-1$ points between $L_i$ and $L_j$ which have not been collected successfully. In this case, there exist some missing data and need to be completed. The method of completion is to construct two point definite linear equations, and fill in the missing load values.

$$L_{g+1} = L_i + k \times \frac{L_j - L_i}{j-i} \quad (k = 1,...,j-i-1) \quad (3)$$

(b) Data Fitting: 24 or 48 original load data are normalized and converted to 96 load data. The data processing needs to use the data at 0 o'clock in the next day, but the acquisition system has not collected the data of the next day, so the data at 0 o'clock is copied to the end of the 24 load data sequence, and the 25 data sequence is finally formed. If the 25 data sequence is \{L_1, L_2, \ldots, L_{96}, L_{97}\}, the data sequence of the middle vacancy is the data that needs to be complemented, and the formula is

$$\begin{align*}
L_{24-i} &= \left( 3L_{4i-3} + L_{4i+1} \right) / 4 \\
L_{4i-1} &= \left( L_{4i-3} + L_{4i+1} \right) / 2 \quad (i = 1,\ldots, 24) \\
L_{4i} &= \left( L_{4i-1} + 3L_{4i+1} \right) / 4
\end{align*} \quad (4)$$

Similarly, 48 load data sequences need to add the load data at 0 o'clock on that day, and finally form 49 data sequence which is \{L_1, L_2, \ldots, L_{48}, L_{49}\}. The data sequence of the middle vacancy is the data that needs to be complemented, and the formula is

$$L_{24-i} = \left( L_{24-i} + L_{24-i+1} + L_{24-i+2} \right) / 4 \quad (i = 1,\ldots, 24) \quad (5)$$

(c) Data Verification: Check user load data and exclude outliers which do not meet the capacity according to user capacity parameters. At the same time, the condition of super-large or super-small is checked, and the abnormal data are converted. Finally, 96 load data of users in one day after cleaning are obtained.
3.2.2 Data Adjustment

In order to ensure the time correspondence and relationship correspondence between meteorological, socioeconomic data and load data, data adjustment is carried out. Taking temperature meteorological data as an example, the temperature data is collected from the real time meteorological data of the regions of the province, and it is a data per 10 min. Therefore, the corresponding relationship between the load data and the temperature data should be determined, for example, the load data at 00:15 should correspond to the temperature data at 00:10 or temperature data at 00:20. The specific corresponding relationship is shown in Figure 3.

![Figure 3. Corresponding relationship between load and temperature data](image)

3.3 Data Analysis

The load is influenced by many factors such as meteorological factors, socioeconomic factors and so on. This paper selects the appropriate factors to analyze and quantifies the impact of various factors on the load, and constructs a multi-dimensional load influencing factor model with self-learning ability.

3.3.1 Fast Matching

The influencing factors of load include temperature, humidity, cloud amount, wind speed, wind direction, GDP, electricity price and so on. Assuming that there are \( m \) load data samples, and each sample has \( n \) influencing factors. Finally, a \( m \times n \) load influencing factor matrix is constructed.

\[
A = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}
\] (6)

In order to grasp the inherent regularity of the load from many factors, it is necessary to reduce the dimension of the factor matrix of the load. In other words, it is to replace the original variable with a few variables, and these few variables can reflect most of the information of the original variable. In this paper, the Principal Component Analysis (PCA) is used to calculate the influence matrix, and the complex factor is reduced to several main components, which makes the problem simple and the result is more scientific and effective. The specific steps are as follows.

(a) The meteorological, and socioeconomic data need to be converted, standardized, and eventually unified. The conversion formula is

\[
X = \frac{A_j - \bar{A}_j}{S_j}
\] (7)

\[
\bar{A}_j = \frac{1}{m} \sum_{j=1}^{m} A_{ij}
\] (8)

the \( A_{ij} \) indicates the \( i \) component of the \( j \) index, and the sample standard deviation is

\[
S = \sqrt{\frac{1}{m-1} \sum_{j=1}^{m} (A_{ij} - \bar{A}_j)^2}
\] (9)

\( X \) is obtained after the above calculation, and this standardization method effectively reduces the influence of data dimension on data extraction.

(b) The covariance matrix \( R \) is calculated based on the normalized matrix \( X \).
\[
R = \frac{1}{m-1} X^T \cdot X = (r_{ij})_{m \times m}
\]  
(10)

\[
r_{ij} = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)^2} \cdot \sqrt{\sum_{k=1}^{n} (x_{jk} - \bar{x}_j)^2}}
\]  
(11)

(c) Calculating eigenvalues of covariance matrix \( R \) by Jacobi iteration method.

(d) The eigenvalues are arranged in descending order and get \( \lambda'_1, \lambda'_2, \ldots, \lambda'_n \), then calculate the cumulative contribution rate of eigenvalue \( B_1, B_2, \ldots, B_n \).

\[
B_i = \frac{\lambda'_i}{\sum_{k=1}^{n} \lambda'_k}
\]  
(12)

(e) According to the given extraction efficiency \( p \), if \( B_i \geq p \), then \( t \) main factors are extracted. And \( t \) is usually 85\%, and can be determined according to actual demand.

Taking the load factors of a city as an example, the main factors are temperature, humidity and GDP. Their cumulative contribution rate has reached over 85\%. The specific data can be found in Table 1.

Table 1. The proportion of main factors affecting load in a certain area

| City | Temperature | Humidity | GDP     | Contribution rate |
|------|-------------|----------|---------|-------------------|
| A    | 41.283%     | 28.239%  | 19.145% | 88.667%          |
| B    | 40.287%     | 30.122%  | 22.342% | 92.751%          |
| C    | 44.344%     | 25.839%  | 16.233% | 86.416%          |
| D    | 47.238%     | 23.239%  | 18.347% | 88.824%          |
| E    | 42.239%     | 27.298%  | 17.323% | 86.860%          |
| F    | 43.287%     | 25.344%  | 14.393% | 83.024%          |
| G    | 42.219%     | 30.344%  | 18.540% | 91.103%          |

As can be seen from the table, temperature is the primary factor affecting the load which accounts for more than 40\%, humidity is the second, and GDP is the third.

3.3.2 Correlation Analysis

According to the main influencing factors by fast matched, this paper carries out correlation analysis to quantify the degree of influence of each major influencing factor on the load. Taking the temperature-load data as an example, this paper quantifies the influence of temperature on the load, so as to improve the accuracy of the user load meteorological model. In order to eliminate the influence factors of annual natural growth rate and holidays, the load temperature model is constructed based on the load data of three years or more historical working days. The specific steps are as follows.

(a) Extract workload data in 5 years, and calculate the average of 96 loads in 5 years.

\[
\bar{Q}_{(y,d,i)} = \frac{\sum_{i=1}^{96} Q_{(y,d,i)}}{D_y}
\]  
(13)

the variable \( Q_{(y,d,i)} \) indicates the \( i \) load on the \( d \) working day of the \( y \) year.

(b) Calculate the influence rate of each load per day.
\[ V_{(y,d,i)} = \frac{Q_{(y,d,i)} - \overline{Q}_{(y,d)}}{\overline{Q}_{(y,d)}} \]  

(14)

the variable \( V_{(y,d,i)} \) indicates the influence rate of the \( i \) load on the \( d \) working day of the \( y \) year.

(c) According to the temperature range, 96 load influence rates on all working days of the \( y \) years are classified into corresponding temperature range. In view of the fact that the temperature is divided into 45 gears of > 40, < - 4 and - 4 ~ 40, 45 \( \times \) 96 temperature - load influence rate sequences \( V_{(y,d,i,t)} \) are finally formed. As the temperature corresponds to the influence rate of each load, \( V_{(y,d,i)} \), becomes \( V_{(y,d,i,t)} \) and the load influence data of time \( i \) and temperature \( t \) are all included in the set \( W_{(i,t)} \).

\[ V_{(y,d,i,t)} \in W_{(i,t)} \]  

(15)

(d) In order to reflect the effect of temperature on load more effectively, the average value of load influence rate at each temperature range in the set \( W_{(i,t)} \) is calculated to obtain the comprehensive influence rate \( C_{(i,t)} \).

\[ C_{(i,t)} = \frac{\sum_{m=1}^{m} S_{(i,t)}}{m} \]  

(16)

(e) After the above calculation, the temperature load model has been basically completed, and the three-dimensional display of the model is shown in Figure 4.

![Figure 4. 3D view of temperature-load model](image)

3.4 Data Application

Just like the construction process of temperature-load model, we can continue to build humidity-load model and GDP-load model, and then use the three models to support load forecasting. In view of the relevant experts and scholars have done a lot of research on load forecasting, this paper does not elaborate. The innovation and advantage of this paper is that the forecasting results can be modified by combining the influence ratio of each factor, so as to get more accurate load forecasting results.

4. Conclusions

This paper makes full use of the large data resources of electricity information, and based on the demand of electrical load analysis, this paper quickly matches the main factors from the potential factors in nature and society, and then constructs the model of load influence factors according to the factors that have been matched, so as to realize the specific quantification of the impact of various factors on the load. And taking the temperature-load data as an example, this paper establishes temperature-load model which proves the effectiveness of the proposed method.

Acknowledgements

Project Supported by National Key R&D Program of China (2017YFB0902800).
References

[1] Ding X, Sun H, Zheng H Y, et al. Distribution and Consumption Big Data Based Short-term Load Forecasting[J]. Electric Power Engineering Technology, 2018(3).

[2] Liu K Y, Sheng W X, Zhang D X, et al. Big data application requirements and scenario analysis in smart distribution network[J]. Proceedings of the CSEE, 2015,35(2):287-293.

[3] Wang D J. Short-term load analysis and forecast based on big data technology[J]. East China Electric Power, 2014,42(10):2007-2010.

[4] Zhang S X, Zhao B Z, Wang F Y, et al. Short term power load forecasting based on big data[J]. Proceedings of the CSEE, 2015,35(1):37-42.

[5] Xu Q, Zeng L. Analysis on load and temperature sensitivity of Nanjing Grid[J]. Jiangsu Electrical Engineering, 2012,31(5):55-57, 61.

[6] He C G. Correlation analysis on system peak load and temperature[J]. Hebei Electric Power, 2011, 30(5):15-17.

[7] Diao Y L, Sheng W X, Liu K Y, et al. Research on online cleaning and repair methods of large-scale distribution network load data[J]. Power System Technology, 2015,39(11):3134-3140.

[8] Lan L. Analysis of fault evolution in distribution lines[J]. Zhejiang Electric Power, 2016,35(4):28-31.