Substation Load Characteristics and Forecasting Model for Large-scale Distributed Generation Integration

Zhengping Gao1,*, Jing Shi1, Hu Li1, Chen Chen1, Jian Tan1 and Lixin Liu2

1Economic and Technological Research Institute of State Grid Jiangsu Electric Power Company, Jiangsu 210008, China
2Beijing Tsingsoft Innovation Technology Co., Ltd., Beijing 100085, China

*Corresponding author e-mail: intel_ada1@163.com

Abstract. The large-scale access of distributed generation has a great impact on load forecasting in substation-area, which means the load curve in the substation-area cannot reflect the real load of users. Firstly, this paper considers the impact of distributed generation on the load curve, and proposes a load forecasting model based on data cleaning and deep learning in the substation-area. Secondly, considering the data missing during communication and transmission, the KNN algorithm is adopted to complete the missing data before the data input. And then, Pearson correlation coefficients are used for the correlation analysis of the input factors related to distributed generation, and the data is trained through Long-short Term Memory in deep learning. Finally, verified by load data of some substation-area in Jiangsu Province, the prediction model established in this paper has good prediction accuracy and stability.

1. Introduction
With the depletion of traditional fossil fuel energy, the environmental crisis becomes more and more prominent, which provides a good opportunity for the rapid development of distributed generation (DG). Therefore, DG has become an important way to use renewable energy. For example, with the rapid development of photovoltaic materials and related control technologies, the penetration rate of distributed photovoltaic in the distribution network is increasing year by year. By the end of December 2018, the cumulative installed capacity of distributed photovoltaic reached 50.61 gigawatt, and the newly increased installed capacity of distributed photovoltaic reached 20.96 gigawatt, which is the first time that the incremental scale of distributed photovoltaic exceeds that of centralized photovoltaic. The large-scale access of DG solves the consumption problem of renewable energy to a certain extent. However, the access of distributed power supply with high penetration ratio has a huge impact on the power flow distribution and load form in the substation-area, which makes the load in the substation-area is not the real situation of the user's power load. In this case, the load prediction in the substation-area is more difficult, because the impact of DG needs to be considered deeply.

In general, as the energy transfer place in power system, the substation is the link of receiving and distributing energy. The substation can be divided into step-up substation and step-down substation according to voltage level. The step-down substation is generally located at the user's side. After the power is reduced, it is sent to the corresponding distribution network for the user to use. When the power consumption in the substation-area exceeds the rated capacity, the main equipment enters the overload
operation state, which causes a huge impact on the substation equipment and related lines. Therefore, accurate load forecasting in the area of step-down substation will play a decisive role in the economic operation and safety and stability of the distribution network, and make load adjustment in advance to ensure the normal power supply during peak load [1].

This paper mainly studies the load forecasting problem in the area of step-down substation with large-scale access of distributed generation. At present, the accuracy of load forecasting in substation-area needs to be improved. The influence of DG access on load form and the randomness of load in step-down substation-area need to be considered. Compared with the power system level load, the user capacity of step-down substation is limited, and the load interaction among users is less smooth, which makes the overall load curve after aggregation fluctuate obviously [2]. All of these make the prediction more difficult.

For the traditional power system load forecasting problem, the existing research methods include neural network [4], random forest [5], deep learning [6], limit learning machine [7], etc. A short-term load forecasting method for users in the open electricity selling environment is proposed to calculate the typical daily load curve of users according to the historical load data of users. Kohonen neural network is used to mine the similarity between users' electricity consumption behavior. Considering the factors such as electricity price and temperature, online learning machine is used to predict the clustered users' load respectively [8]. The coupling characteristics of energy consumption information of electricity, heat and gas are extracted by using the deep confidence network, so as to effectively predict the power, heat, gas and other energy loads in the comprehensive energy system [9]. The load forecasting problem is abstracted as a non-linear mathematical problem, which considers the influence of production activity, supply and demand environment, macro society, economic indicators [10], etc. Based on the analysis of the correlation between bus load and weather characteristics, a bus load forecasting model based on the combination of numerical weather forecast and load classification is proposed [11].

This paper presents a load forecasting method of substation-area considering the large-scale access of DG, based on the load data cleaning in the substation-area. The remainder of this paper is organized as follows. The following section introduces the data cleaning method based on KNN and Long-short Term Memory. Section 3 proposed the load forecasting model of substation-area considering the large-scale access of distributed generation. Case studies and analysis are presented in Section 4, the results show the completion of missing data can maximize the accuracy of the prediction. The Pearson coefficient effectively analyzes the influencing factors of the load. The algorithm in this paper has a good application effect on the load forecasting of the substation-area with large-scale access of distributed generation. Conclusions are outlined in the last section.

2. Prediction Algorithm Mechanism

2.1. Data Cleaning Method based on KNN

The step-down substation is usually located in the distribution network or the transmission network with lower voltage level. It belongs to the end of the power system, and the power grid cannot be managed as finely as the large power grid at this voltage level. The development of communication and information system in distribution network or transmission network with lower voltage level is relatively lagging behind, which makes the fault rate of measurement or transmission equipment higher and the reliability worse. In addition, the professional level of manager and maintenance personnel in the distribution network still needs to be improved, and the data quality caused by human factors should also be highly valued. The load data in the substation-area is very easy to be missing. Therefore, if the missing data is not cleaned effectively, the missing data set is directly nested into the model for training, which will have a huge impact on the model and greatly reduce the accuracy of load forecasting in the substation-area.

The load data in the substation-area is very easy to be missing, so if the missing data is not cleaned effectively, the missing data set is directly nested into the model for training, which will have a huge
impact on the model and greatly reduce the accuracy of load forecasting in the substation-area. In view of the problem of data missing in step-down substation, this paper uses KNN algorithm to clean the data. KNN is more effective in missing data cleaning, and it is also an algorithm with low computational complexity. It has been widely used in data preprocessing of machine learning. In KNN model, the closer the Euclidean distance is, the higher the similarity between data is. NN selects k complete nearest neighbor data of missing data in the data set to complete the missing values. The distance is calculated as shown in formula (1).

\[ d(Y_i, Y_j) = \sqrt{\sum_{r=1}^{m} (y_{ir} - y_{jr})^2} \]  

Where \( Y_i = \{y_{i1}, y_{i2}, y_{i3}, \ldots, y_{im}\} \) represents the first m-dimensional characteristics of the j-th sample point. \( y_{ir} \) represents the r-th dimension attribute of the i-th sample point.

2.2. Long-short Term Memory

LSTM is a kind of deep learning algorithm widely used in natural language processing. LSTM is generally composed of input layer, output layer and hidden layer. LSTM has a great improvement in control and storage compared with the traditional cyclic neural network. The framework of LSTM unit is shown in Figure 1.

![Figure 1. Framework of LSTM unit](image-url)

Each LSTM unit has a cell, which has memory function, and its state at time t is recorded as \( c_t \). LSTM unit receives the current state \( x_t \) and the last tuple state \( h_{t-1} \) through in-gate \( i_t \), forgetting-gate \( f_t \) and output gate \( o_t \). At the same time, the state \( c_t \) of the memory unit is input to each gate as internal information. After receiving the input information, the in-gate, forgetting-gate and out-gate perform internal operation to determine whether to activate the cell tuple. After the nonlinear function transformation, the in-gate signal is superposed with the memory unit state processed by the forgetting-gate to form a new memory unit state \( c_t \). Finally, the output \( h_t \) of the LSTM is formed by the operation of the nonlinear function and the dynamic control of the output gate. Each variable is calculated as follows.

\[ i_t = \text{sigmoid}[W_{hi} \times h_{t-1} + W_{xi} \times x_t + W_{ci} \times c_{t-1} + b_i] \]
$$f_t = \text{sigmoid}[W_{hf} \times h_{t-1} + W_{xf} \times x_t + W_{cf} \times c_{t-1} + b_f]$$  \hspace{1cm} (3)$$

$$o_t = \text{sigmoid}[W_{ho} \times h_{t-1} + W_{xo} \times x_t + W_{co} \times c_t + b_o]$$  \hspace{1cm} (4)$$

$$c_t = f_t c_{t-1} + t_t \tanh[W_{hg} \times h_{t-1} + W_{xg} \times x_t + b_c]$$  \hspace{1cm} (5)$$

$$h_t = o_t \tanh[c_t]$$  \hspace{1cm} (6)$$

Where W is the weight coefficient, b is the vector of bias. In general, the LSTM algorithm uses the classic error back-propagation algorithm to train the expanded network, which will not be discussed here.

3. Load Forecasting Model of Substation-Area Considering Distributed Generation

Traditional load forecasting methods do not consider the impact of DG access, but only consider the impact of load change. In fact, the energy generated by DG is absorbed locally in the distribution network, which changes the trend of load change. In order to further analyze the load curve in the substation-area, this paper further excavates the influence of DG factors on the load side. In this paper, Pearson correlation coefficient is adopted to calculate the correlation between input features and output information, so as to analyze the influence of DG on the accuracy of load prediction. Pearson correlation coefficient is calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{m} (y_i - \bar{y})^2}}$$  \hspace{1cm} (7)$$

Where $\bar{x}$, $\bar{y}$ are the average value of the elements in each vector respectively.

The overall training process of the model is as follows.

(1) Use Pearson correlation coefficient to analyze the correlation between photovoltaic, wind power output and load in substation-area.

(2) Fill in the missing data with KNN algorithm.

(3) Input the complete data into the LSTM model to obtain the final load prediction result.

The block diagram is shown in Figure 2.
4. Case Study

Load data from a certain area of Jiangsu is selected for verification. In order to test the effectiveness of the algorithm in this paper in a variety of scenarios, a variety of application scenarios are selected, in which the access capacity of distributed photovoltaic reaches a high proportion. Scenario 1 refers to the residential load in the substation-area, and scenario 2 refers to the industrial load in the substation-area. In each scenario, the validity of the prediction model is verified at two voltage levels of 10kV and 110kV. The forecast targets are the total load of the substation on the second day. The error index used in the case includes Mean Relative Percent Error (MAPE) and Root Mean Square Error (RMSE), as shown in formula (8) and (9).

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{a_i - b_i}{a_i} \right| \times 100\%$$  \hspace{1cm} (8)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (a_i - b_i)^2} \times 100\%$$  \hspace{1cm} (9)$$

Where $n$ is the number of samples, $a_i$ and $b_i$ are the actual load and the predicted load at time $i$, respectively.

4.1. Comparison of Data Preprocessing Schemes and Analysis of Feature Correlation

In order to analyze the pros and cons of different data filling schemes, this section compares KNN completion algorithms, linear interpolation completion methods, and non-completion algorithms. The
model is trained in three ways, and the impact of different data completion algorithms on the prediction results is shown in Table 1.

| Scenario | Voltage level of substation | Data Missing Percentage | MAPE Non-completion | Linear interpolation completion | KNN completion |
|----------|-----------------------------|-------------------------|---------------------|---------------------------------|----------------|
| Scenario 1 | 10kV                        | 5.67%                   | 24.76%              | 8.66%                           | 5.11%          |
|          | 110kV                       | 3.27%                   | 16.54%              | 5.26%                           | 3.21%          |
| Scenario 2 | 10kV                        | 4.84%                   | 19.76%              | 6.26%                           | 3.31%          |
|          | 110kV                       | 3.12%                   | 10.54%              | 5.31%                           | 2.55%          |

It can be seen from Table 1 that the lower the voltage level in a substation jurisdiction, the more serious the lack of load data is, indicating that the management level of the distribution network is related to the voltage level, and the lower the management level of a substation with a lower voltage level. Different types of loads present different error distributions, and the prediction errors of industrial loads are generally smaller than those of residential loads. In addition, the non-completion algorithm has a great impact on the prediction accuracy. This is because the information of the missing data is deleted, the original internal characteristics of the data are destroyed, and the prediction accuracy is greatly reduced. When the linear interpolation method is used, only the stock data structure of a single feature variable in the input data can be considered. For missing data, the linear method is simply used to complete, and the prediction accuracy is limited. The KNN completion algorithm fills in missing data by clustering different data attributes, analyzes the coupling relationship between multiple features, and can better restore the original shape of the data. The accuracy of substation-area load prediction is highest with the data of KNN completion.

Due to the high DG permeability in the selected area, the DG output has a certain effect on the load pattern in the substation-area. This paper uses the photovoltaic and wind power data at the same time in the jurisdiction as the characteristic information, analyzes the correlation between load forecasting, photovoltaic output and wind power output, and analyzes the correlation between the DG output and the true value of the load on the basis of data cleaning. Table 2 shows the Pearson coefficient relationship between the load of different levels of substations and wind power and photovoltaic output in two scenarios. When the access capacity of DG reaches a high proportion, the relevant factors affecting the output of DG have a greater correlation with the load change in the substation-area. The higher the DG permeability, the greater the impact of DG-related information on load prediction.
4.2. Prediction of Industrial Load and Residential Load

In order to further compare the impact of DG output on the load forecasting of the substation, this section analyzes the load forecasting results in two scenarios, and selects a period of one week (168 hours) for analysis, and uses DG data and unused DG data. The comparison of the load forecast results is shown in Figure 4. Figure 4 shows the power consumption of residential loads in the jurisdictions of 10kV and 110kV substations, and Figure 5 shows the power consumption of industrial loads in the jurisdictions of 10kV and 110kV substations.
Figure 5. Substation-area load forecasting results based on industrial type

From the prediction results, it can be known that for a substation-area with a high DG penetration rate, the load prediction method considering DG output data can meet the needs of various scenarios, and can obtain higher prediction accuracy compared with models that do not consider DG data. This also shows that the high proportion of renewable energy access has a certain impact on the load pattern.

4.3. Comparative Analysis of Prediction Results of Multiple Algorithms

In order to further verify the superiority of the algorithms in this paper, the prediction results of the deep learning algorithm LSTM in this paper are compared with the prediction results of neural networks and ARIMA (Autoregressive Integrated Moving Average model). These algorithms use the relevant information of DG. The prediction results are shown in Table 2.

| Load type         | Prediction algorithm | MAPE   | RMSE  |
|-------------------|----------------------|--------|-------|
| 10kV Residential Load | LSTM                 | 5.67%  | 4.31% |
|                   | NN                   | 6.95%  | 6.43% |
|                   | ARIMA                | 8.50%  | 8.22% |
| 10kV Industrial Load | LSTM                 | 3.52%  | 2.99% |
|                   | NN                   | 4.57%  | 4.83% |
|                   | ARIMA                | 7.57%  | 7.31% |
| 110kV Residential Load | LSTM                | 3.25%  | 3.79% |
|                   | NN                   | 5.30%  | 5.41% |
|                   | ARIMA                | 7.76%  | 7.64% |
| 110kV Industrial Load | LSTM                | 2.62%  | 2.54% |
|                   | NN                   | 3.57%  | 3.83% |
|                   | ARIMA                | 7.57%  | 7.31% |

It can be seen from Table 2 that the randomness of residential power consumption is also strong, which also increases the difficulty of forecasting. The accuracy of residential load forecasting is generally lower than the industrial load forecasting accuracy. The higher the voltage level in a substation-area, the higher the accuracy of load prediction.

LSTM-based deep learning algorithms can track changes in residential or industrial loads. The algorithm in this paper has obtained the highest accuracy in various scenarios, because the deep learning algorithm improves the generalization and robustness of the model. NN belongs to the traditional shallow model, and the generalization has reached the bottleneck. Compared with other algorithms, NN
has also achieved good prediction accuracy, but it is still lower than the method used in this paper. This is because NN belongs to the traditional shallow model, and the generalization has reached the bottleneck. In addition, the optimization process of the NN model has the risk of falling into a local minimum, which makes the prediction result more general.

5. Conclusion
With the rapid development of smart grid and renewable energy, the access of high proportion of renewable power makes the substation level load more difficult to predict. The volatility and randomness of DG make the substation level load is not the real power consumption form of users. It is necessary to accurately predict the trend of the load in the substation when the DG output is considered.

Considering the characteristics of PV and wind power, this paper proposes a method of power load forecasting based on data cleaning and deep learning, and analyzes the difference between substation level load forecasting and traditional load forecasting. Before the prediction, the data is fully cleaned. Based on the complete data set, the prediction model based on Long-short Term Memory is proposed. Finally, verified by the case, the results show the superiority of data cleaning and deep learning algorithm in prediction, which have a good effect on industrial load and residential load at different voltage levels.

Acknowledgments
This work was financially supported by Science and Technology Project (SGJSJY00GHJS1900035) of State Grid Jiangsu Electric Power Company.

References
[1] Z. B. Jiang, H. Wu, X. Cheng, W.Z. Sun, J.Y. Shang, Analysis of Substation Characteristics Based on Multivariate Clustering Model and Two-stage Clustering-correction Algorithm, J. Automation of Electric Power Systems. 2018, 42 (15): 157 - 163+243-244.
[2] Q. X. Yang, DEVELOPMENT TREND OF INTEGRATED PROTECTION AND CONTROL IN POWER SUBSTATIONS, J. Automation of Electric Power Systems. 1995 (10): 7 - 9.
[3] M. C. Chao, Y. P. Wang, X.W. Li, F. Wang, W. F. Cai, Smart substation and technical characteristics analysis, J. Power System Protection and Control. 2010, 38 (18): 59 - 62+79.
[4] X. N. Su, T.Q. Liu, H. Q. Cao, H.M. Jiao, Y.G. Yu, C. He, J. Shen, A Multiple Distributed BP Networks Approach for Short-term Load Forecasting Based on Hadoop Framework, J. Proceedings of CSEE. 2017, 37 (17): 4966 - 4973+5216.
[5] J. Li, J. P. Liu, J. J. Wang, Mid-long Term Load Forecasting Based on Simulated Annealing and SVM Algorithm, J. Proceedings of the CSEE. 2011, 31 (16): 63 - 66.
[6] R. Z. Wu, Z.R Bao, X.Y. Song, W. Deng, Research on Short-term Load Forecasting Method of Power Grid Based on Deep Learning, J. Modern Electric Power. 2018, 35 (02): 43 - 48.
[7] G. Z. Li, W. Y. Liu, H. Z. Yun, Y. H. Gao, A New Data Preprocessing Method for Bus Load Forecasting, J. Power System Technology. 2010, 34 (2): 149 - 154.
[8] B. C. Yang, J. Zhang, K. P. Yu, User Short-Term Load Forecasting Method under the open sales environment, J. Advanced Technology of Electrical Engineering and Energy. 2018: 1.
[9] J. Q. Shi, T. Tao, J. Guo, Y. Liu, J. H. Zhang, Multi-Task Learning Based on Deep Architecture for Various Types of Load Forecasting in Regional Energy System Integration, J. Power System Technology. 2018, 42 (03): 698 - 707.
[10] P. Wang, Q. X. Chen, Q. Xia, X.Y. Wang, Correlation Analysis and Forecasting Method on Industrial Electricity Demand Based on Vector Error Correction Model, J. Proceedings of the CSEE, 2012, 32 (4): 100 - 107.
[11] Bo. Li, D. Y. Men, Y. Q. Yan, J. F. Yang, J.Y. Zhou, Z.Q Luo, L.L. Zeng, Bus Load Forecasting Based on Numerical Weather Prediction, J. Automation of Electric Power Systems. 2015, 39 (1): 137 - 140.
[12] W. C. Shen, H. I. Jen, A.Y. Wu, New Ping-Pong Scheduling for Low-Latency EMD Engine
Design in Hilbert-Huang Transform, J. IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-II: EXPRESS BRIEFS. 2013, 60 (8): 532 - 536.

[13] A. H. Nizar, Z. Y. Dong, Y. Wang, Power Utility Nontechnical Loss Analysis With Extreme Learning Machine Method, J. IEEE TRANSACTION ON POWER SYSTEMS, 2008, 23 (3): 946 - 955.

[14] K. A. Toh, Deterministic Neural Classification, J. Neural computation, 2008, 20 (6): 1565 - 1595.