Analysis and evaluation of short-term wind power interval forecast error based on K-means clustering

Songshan Li¹, Zhicheng Ma², Xiaoying Zhang¹, 4, Kun Wang³, Qiang Zhou², Jinping Zhang² and Wei Chen¹

¹ College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China
² State Grid Gansu Electric Power Research Institute, Lanzhou 730000, China
³ State Grid Gansu Electric Power Company Marketing Service Center, Department of Metering, Lanzhou 730000, China
⁴ E-mail: 245659219@qq.com

Abstract. The intermittency, volatility and randomness of wind power will inevitably affect the accuracy of wind power prediction. In order to improve the accuracy of wind power interval prediction and evaluation, a method of wind power interval prediction and evaluation based on K-means clustering is proposed. First, use K-means clustering to classify wind power prediction error data; second, combine the climate type to more accurately divide and process historical data, and apply multiple linear regression models to establish preliminary prediction models corresponding to each subset; Finally, according to the distribution characteristics of forecast errors, the results of wind power forecasting are analyzed and evaluated with interval coverage and interval average bandwidth as indicators. Comparing the K-means clustering algorithm with the BP neural network algorithm, the results show that the method proposed in this paper can better capture the characteristics of wind power error data, and can obtain more accurate wind power interval prediction and evaluation results.

1. Introduction
In recent years, the continuous use of fossil fuels such as petroleum and coal has led to problems such as energy shortages, environmental pollution and ecological degradation. In order to reduce air pollution and environmental damage caused by traditional energy sources and improve the stability of the power system, wind power as a renewable energy source Power generation has been widely used and developed all over the world, but due to the intermittent, random and volatility of wind power itself [1], it is difficult to accurately predict it. At present, my country is undergoing national grid interconnection, and the scale of the grid is increasing. For wind farms connected to a large power grid, the injection of wind power has little effect on the grid frequency. However, areas with abundant wind energy resources are sparsely populated, with small loads, and relatively weak grid structures. The injection of wind power has a great impact on local grids. In order to improve the controllability of large-scale wind power and the reliability when it is connected to the grid, interval assessment of the short-term wind power numerical fluctuation range of wind farms has very important practical significance [2].

The main evaluation methods of wind power forecast errors are parametric and non-parametric methods [3]. The parametric method assumes that the forecast error data obey a certain specific
distribution form, such as Gaussian distribution, t distribution [4], etc.; the non-parametric method does not need to presuppose the distribution form of the forecast error, but requires a large amount of data and more complicated calculations. Common methods include quantile regression, kernel density estimation [5-6] and so on.

At present, there are mainly two types of prediction models established for wind power: (1) physical models; (2) statistical models. For the physical model, the physical formula used in the modeling process itself has certain errors, and the prediction model has a large dependence on the system information of the wind power station, which makes the model's anti-interference ability poor. In contrast, prediction methods based on statistical models start from the data point of view, modeling is simple and easy to implement, and the model is more interpretable and transparent. Existing research considers the construction of a dispatch model based on factors such as load and wind power forecast error uncertainty to achieve the purpose of improving environmental economic dispatch; in addition, there are short-term power information of wind power for dynamic grouping of wind farms on a longer time scale, realize the reasonable distribution of active power commands among wind farms.

In response to the above problems, this paper proposes a method for analysing and evaluating wind power interval prediction errors based on K-means clustering [7]. First, use K-means clustering to classify the wind power forecast error data, combine the climate type to divide and process the historical data more accurately, and apply the multiple linear regression model to establish the preliminary forecast model corresponding to each subset, and according to the distribution characteristics of forecast errors, the forecast results are analysed and evaluated. The model deeply excavates the potential characteristics of the data, which can effectively improve the accuracy of wind power forecasting [8].

2. Wind power prediction model structure

The wind power prediction model proposed in this paper is mainly composed of two parts: a preliminary prediction model and an error evaluation analysis model. The overall structure of the model is shown in Figure 1.

![Figure 1. Structure diagram of wind power prediction model.](image)

The first part is the preliminary prediction model of wind power, which provides the initial prediction of power, including the processing of climate and power data and the fitting of multiple
linear regression models. The climate history data is divided into multiple subsets according to the differences between different months, and standardized processing is performed with each subset as a unit. Through regression analysis, a multiple linear regression model corresponding to each subset was established as a preliminary prediction model. Using a linear model to fit the complex relationship between climate factors and output power will definitely produce a certain deviation. Therefore, the second part of the error analysis model is needed to deal with it.

2.1. Standardized processing of climate data
According to research, wind power is easily affected by factors such as geography and climate. Climate data includes climate data and weather data. The statistical results of climate data are called climate statistics. They are the basic data for analyzing and describing climate characteristics and their changing laws. Climatic statistics usually require longer records to make the statistical results more stable. Generally, records of more than 30 consecutive years are sufficient. The output power of wind farms varies greatly in different seasons, different regions, and different moments. In order to avoid weakening the factors with smaller values and ensure the accuracy of the analysis results, the Min-Max standardization method is selected to standardize the climate data:

\[ M_{ij} = \frac{N_{ij} - \min[N_{ij}]}{\max[N_{ij}] - \min[N_{ij}]} \]  

In the formula, \( N_{ij} \) represents the original value of the i-th climate factor in the j-th sample data, and \( M_{ij} \) represents the normalized value.

After standardizing the power data of each subset, the climate types are divided into simple and complex climate types, and the K-means clustering algorithm is used to cluster the power data of simple climate types.

2.2. Establishment of multiple linear regression model
Perform multiple linear regression analysis on the data set of each climate type in each subset, and establish the corresponding multiple linear regression model as a preliminary prediction model [9]. The established multiple linear regression model is shown in (2):

\[ Q_f = \beta_0 + \beta_1 V_1 + \cdots + \beta_m V_m \]  

In the formula: \( \beta_0, \beta_1, \cdots, \beta_m \) are regression coefficients; \( V_1, \cdots, V_m \) are the principal component variables of climate data; \( Q_f \) is the preliminary predicted value of power.

3. Error analysis based on probability distribution model
Due to the limited climatic factors considered when establishing the preliminary prediction model, and the complicated relationship between each factor and the output power, the multiple linear regression model is used to fit the model itself, and the influence of uncertain factors on the prediction results is also considered. It is necessary to study the statistical characteristics of preliminary forecast errors.

3.1. Complicated climate type error analysis model
Based on the statistical analysis of the preliminary prediction error data, a typical probability distribution model can be used to fit. Typical probability distribution models mainly include: normal distribution, t-Location-Scale distribution, etc. The fitting results are shown in Figure 2. This paper chooses the t-Location-Scale distribution with the highest degree of fit for fitting.

The t-Location-Scale distribution is a distribution containing location parameters and scale parameters. It contains three parameters: expected \( \mu \), variance \( \sigma \), and degrees of freedom \( \nu \). The probability density function of this distribution is:

\[ F(P_{err} = x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma \sqrt{\nu \pi} \Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\nu + (x - \mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}} \]  

Where: \( \Gamma(\cdot) \) is the gamma function.
3.2. Simple climate type error analysis model

The climatic factors of the simple climate type can remain relatively stable in each period, and the output power of wind power will not fluctuate greatly. In this case, the preliminary prediction error is mainly caused by the model deviation in the multiple linear regression model, and the influence of uncertainty factors on the error is relatively small. The size of the preliminary prediction error has a certain correlation with the preliminary prediction value $Q_f$, and the error distribution corresponding to different preliminary prediction values is shown in Figure 3.

According to the analysis in Figure 3, its probability distribution has the following characteristics:

1. The larger the preliminary prediction value, the larger the possible distribution interval of the corresponding prediction error, and the more concentrated the distribution within this interval;
2. Wind power forecast errors have the characteristics of "spikes and thick tails" [10-11]. Most error data are concentrated around a certain value, while some scattered error data have a wider range and larger value.

This article mainly adopts the following two evaluation indicators [12].

3.3. Interval evaluation index

(1) Evaluation interval coverage

\[ I_{PICP} = \frac{1}{M} \sum_{\ell=1}^{M} K_{\ell} \]  

(4)
In the formula: \( M \) is the number of evaluation samples; if the evaluation target value falls within the evaluation interval, \( K_t \) is 1, otherwise it is 0. If the effect of a certain probability prediction is better or the reliability is higher, the interval coverage should be larger.

(2) Interval average bandwidth
The interval width of the evaluation interval is the difference between the upper bound \( X \) of the interval and the lower bound \( Y \) of the interval, namely:

\[
I_{PINAW} = \frac{1}{M} \sum_{i=1}^{M} (X - Y)
\] (5)

The average bandwidth of the interval can reflect the clarity of the evaluation inter.

4. Case analysis
The BP neural network algorithm is a neural network learning algorithm. Its principle is to use gradient search technology in the gradient descent method to minimize the error mean square error between the actual output value of the network and the expected output value. Its advantages are generalization ability, strong self-learning and adaptive ability, and it is particularly suitable for solving problems with complex internal mechanisms.

This section analyzes the performance of the K-means clustering algorithm [13] in predictive power error analysis by comparing with the BP neural network algorithm. The actual wind power data used June 2020 data from a wind farm in Gansu. According to the main sunshine time, the 8 periods of 8:00~16:00 are selected as the research objects. The climatic factors mainly consider irradiance, temperature and wind speed, and the data resolution is 1 hour.

![Figure 4. Comparison of predicted power error between two algorithms.](image)

It can be obtained from Figure 4 that the prediction power error of the K-means clustering algorithm is smaller than the BP neural network algorithm, so the prediction accuracy is higher.

This section compares the error evaluation of K-means clustering with BP neural network, and analyzes the performance of the method used in this paper in interval evaluation based on error classification. Using the interval evaluation model of t-distribution and normal distribution fitting error, the average bandwidth of the interval with a confidence [14] of 90% and the coverage rate of the evaluation interval are listed in Table 1.

| Algorithm | Normal Distribution | t-Location-Scale |
|-----------|---------------------|-----------------|
|           | \( I_{PINAW} \)     | \( I_{PICP} \)  | \( I_{PINAW} \) | \( I_{PICP} \) |
| BP        | 53.35               | 0.8765          | 64.43           | 0.8857          |
| K-means   | 49.87               | 0.8836          | 52.66           | 0.9064          |
According to the table analysis: Compared with the normal distribution, the interval coverage of the t distribution is higher, but the interval average bandwidth is larger. This is because the t distribution can more accurately fit the "spike and thick tail" characteristics of the prediction error than the normal distribution, so the evaluation interval of the t distribution is larger and the evaluation accuracy is higher. At the same time, it can be seen that the interval coverage under the K-means clustering algorithm is higher than that of the BP neural network algorithm, and the values are all over 88%. Under the t distribution, the interval coverage using the K-means clustering algorithm exceeds 90%. Therefore, the interval evaluation accuracy obtained by the interval evaluation method in this paper is higher. Improving the accuracy of the assessment will reduce the impact of wind power forecast errors on the balance of power generation, supply and consumption.

5. Conclusions
Aiming at the distribution characteristics of wind power forecast error data [15], this paper proposes an evaluation method of wind power interval forecast error. First, use K-means clustering to classify wind power prediction error data; second, combine the climate type to more accurately divide and process historical data, and apply multiple linear regression models to establish preliminary prediction models corresponding to each subset; By comparing with the interval evaluation based on t distribution and normal distribution, and comparing the evaluation results of K-means clustering and BP neural network, it can be concluded that the K-means error interval evaluation method is simpler and can get more accurate, the evaluation effect is not affected by the fitting distribution model. The core premise of prediction error evaluation is to master multiple data such as wind power operation, weather observation, and weather prediction. The accuracy of evaluation highly depends on the actual data quality, and how to ensure data quality through technology and management methods. The method in this paper breaks through the ideas of the original interval evaluation method, improves the accuracy of wind power interval evaluation, and provides a basis for wind power grid planning, operation and safety and stability analysis. However, the interval evaluation in this article is aimed at historical wind power and forecast errors. If input data, such as wind speed and direction and other influencing factors, are added, the accuracy of the interval evaluation may be further improved [16]. In the next step, we can study the problem of interval evaluation after integrating multiple influencing factors.

Acknowledgements
This paper is supported by the Gansu Province Science and Technology Major Special Project "Key Technologies and Demonstration Application of Market Consumption and Dispatching Control of Solar Thermal-Photovoltaic-Wind Power New Energy Base (19ZD2GA003)".

References
[1] Cui Xiaolin, Xing Haiyan, Sun Yan, et al. 2004 Research and Application of Frequency Conversion Technology of Electric Submersible Pump[J]. Petroleum Field Machinery 33(s1) 57-59
[2] Wu Wenzu, Qiao Ying, Lu Zongxiang, et al. 2017 Methods and prospects for probabilistic forecasting of wind power[J]. Automation of Electric Power Systems 41(18) 167-175
[3] Yan Jie, Li Ning, Li Yongqian, et al. 2019 Short-term uncertainty forecasting method for wind power based on real-time switching cloud model[J]. Automation of Electric Power Systems 43(3) 17-23
[4] Yang Mao, Yang Chunlin and Dong Juncheng 2019 Research on super short-term probability interval prediction of wind power based on prediction error distribution optimization model[J]. Acta Energiae Solaris Sinica 40(10) 2967-2978
[5] Zhou Bowen, Ma Xiangjin, Luo Yanhong, et al. 2019 Wind power prediction based on LSTM networks and nonparametric Kernel density estimation[J]. IEEE Access 7 165279-165292
[6] Zhang Xiaoying, Zhang Xiaomin, Liao Shun, et al. 2019 Prediction error analysis of wind
power based on clustering and non-parametric Kernel density estimation[J]. *Acta Energiae Solaris Sinica* **40**(12) 3594-3604

[7] Yang Junchuang and Zhao Chao 2019 Survey on K-means clustering algorithm[J]. *Computer Engineering and Applications* **55**(23) 7-14

[8] Qiao Ying, Lu Zongxiang and Min Yong 2017 Methods to improve the accuracy of wind power forecasting[J]. *Power System Technology* **41**(10) 3261-3269

[9] Liu Jie, Chen Xuemei, Lu Chao and Mao Hangyin 2020 Two-stage photovoltaic power prediction based on statistical characteristics of data considering error correction[J]. *Power System Technology* **44**(08) 2891-2897

[10] Dui Xiaowei, Zhu Guiping and Liu Yanzhang 2017 Research on battery storage sizing for wind farm considering forecast error[J]. *Power System Technology* **41**(2) 434-439

[11] Wang Chengfu, Wang Zhaqing, Sun Hongbin, et al. 2016 Model of unit commitment with wind farm considering time series characteristic of wind power forecast error[J]. *Proceedings of the CSEE* **36**(15) 4081-4089

[12] Yang Mao, Yang Chunlin and Dong Juncheng 2019 Research on super short-term probability interval prediction of wind power based on prediction error distribution optimization model[J]. *Acta Energiae Solaris Sinica* **40**(10) 2967-2978

[13] Wang Zhaoran 2006 Harmonic suppression and reactive power compensation[M]. Beijing: Machinery Industry Press

[14] Zang Yulong 2020 Research on optimization peak shaving method of thermal power unit considering the confidence of wind power prediction[D]. Changchun University of Technology

[15] Meng Yanfeng, Hu Shuju, Deng Wei, et al. 2013 Analysis and evaluation method of wind power predicted-error[J]. *Electric Power Construction* **34**(7) 6-9

[16] Han Li, Qiao Yan and Jing Huitian 2020 Wind power interval assessment based on error classification[J/OL]. *Automation of Power Systems* 1-12