

# ***Long-term Wind Power Generation Prediction of Guangdong Province Based on Machine Learning Algorithm***

**Guolin Cao**

*International Education College, North China Electric Power University, Baoding, China  
Forest\_gzh@163.com*

**Abstract:** Wind power generation(WPG) plays a crucial role in China's "double carbon" strategic goal, boasting broad development prospects in the country. Guangdong Province, in particular, is endowed with rich WPG resources and great potential. Thus, forecasting long - term WPG in Guangdong is of great significance. This paper commences by introducing diverse machine - learning algorithms in the context of the "two - carbon" target and long - term wind power prediction. Subsequently, historical data on WPG in Guangdong Province from 2009 to 2023, along with its potential influencing factors, are collected. After screening these influencing factors, appropriate independent variables related to energy factors are identified as the feature input for machine - learning algorithms.. According to the fitting effect of different algorithms, those suitable for long-term wind power generation in Guangdong Province are selected, including RR, Enet, Sufit, Stepwise and SVM (linear kernel function). Finally, based on the historical data of energy factors, the paper models and predicts the WPG in the future of Guangdong Province under the five algorithms. The predicted values of WPG in Guangdong for 2030, 2035, 2050, and 2060 are obtained by taking the average of the results from these five algorithms.

**Keywords:** wind power generation, machine learning algorithm, Guangdong Province, energy factors

## **1. Introduction**

With the continuous growth of global energy demand, the development and utilization of renewable energy have drawn extensive attention globally. Among diverse renewable energy resources, wind energy become the substitution of the traditional ones gradually due to its abundant resources, wide distribution, eco-friendly and so on, and it is a vital force to transform the energy-utilize structure. China, being a major global energy producer and consumer, has seen its wind power generation (WPG) account for 56.74% of the total new energy generation [1]. In accordance with the "14th Five-Year Plan", China will vigorously develop wind power in the long term, and fulfill China's solemn commitment to the international community [2, 3]. Guangdong Province, located in southeastern part of China, leads the nation in economic aggregate. It boasts the longest coastline in China and has a subtropical monsoon climate, which is highly favorable for wind power generation [4]. Therefore, a reasonable prediction of the future WPG in Guangdong Province can better evaluate the potential of China's future wind power generation. At present, the mainstream methods for wind power generation prediction are divided into three categories: physical model, traditional statistical

model and machine learning model. In this paper, based on the wind power generation of Guangdong Province from 2009 to 2023 and its possible influencing factors, a variety of machine learning algorithms were used to model, and the model accuracy of each algorithm is obtained. Subsequently, the future predicted values of the selected energy-related independent variables are fed into the algorithms with good fitting performance to estimate the future wind power generation in Guangdong Province under different algorithms. Finally, the long-term wind power generation of Guangdong Province was predicted by taking an average to provide reference for the development of new energy.

## 2. Machine learning algorithm

### 2.1. Evaluation index

In this paper, the mean absolute error (MAE), root mean square error(RMSE) and fitting coefficient( $R^2$ ) were used to evaluate the fitting results of the regression model. The nearer the values of the mean absolute error and the sum of root - mean - square error are to zero, the more excellent the fitting effect of the regression model. Regarding the coefficient of determination, the closer its value gets to 1, the higher the prediction accuracy of the regression model. The expressions for MAE, RMSE and  $R^2$  are as follows:

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_i - f(x_i)| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^m (y_i - f(x_i))^2}{\sum_{i=1}^m (y_i - \bar{y})^2} \quad (3)$$

### 2.2. Machine learning algorithm

In this paper, the machine learning algorithms are used to predict the long-term wind power generation in Guangdong. The models adopted to predict WPG in this paper are: Sufit regression, elastic net regression (Enet), ridge regression (RR), stepwise regression, lasso regression, random forest (RF), support vector machine (SVM) regression, convolutional neural network-bidirectional long short term memory (CNN-BiLSTM). Among them, the first five belong to simple linear regression models, and the others belong to complex regression models. These models are widely applied and have been applied in many other fields and performed well [5].

#### 2.2.1. CNN-BiLSTM

In a CNN, data usually needs to pass through the input layer, the hidden layer, and the output layer. The input layer, being the starting point of the CNN structure, preprocesses the input values, which is fundamental for their subsequent use in the network. The hidden layer is located at the second level, where these filters are applied to the output of the input layer by convolving the input with these filters and generating a feature map as the output. The final step of the CNN is the output layer, which learns the feature map and produces the output. To simplify data processing, it is common to add a pooled layer after each convolutional layer that contains a compressed form of the actual feature map

of the most important and relevant data to reduce the size of the feature map. Subsequently, based on the information from the convolution and pooling layers, dense layer learning is carried out [6].

The traditional LSTM algorithm is simpler, but the convergence rate is slow, and BiLSTM overcomes this shortcoming. The BiLSTM consists of two LSTM models that work in opposite directions. This technique is based on sequential learning in both forward and backward directions. One LSTM model processes the input sequence information in the forward direction, while the other processes data from the backward direction of the entire sequence. This approach not only speeds up the learning process but also provides additional context information for sequence learning [7].

### 2.2.2. RR, lasso regression and enet regression

In the long-term forecast of wind power generation, due to the generally high correlation between the predictor variables, the accuracy of the results of the general regression model may be affected in terms of multicollinearity data sets. Ridge regression algorithm can reduce this influence to a certain extent, and its objective function is as follows:

$$\text{obj}_{\text{RR}} = \min \|y - X\beta\|^2 + \lambda \|\beta\|^2 \quad (4)$$

Lasso regression is typically used for a large number of varieties. This regression model replaces the penalty term of the dependent variable  $y$  with the penalty term of the independent variable  $X = [x_{ij}]$ , so that the data set tends to select a specific subset and shrinks some model coefficients. The objective function of the model is:

$$\text{obj}_{\text{Lasso}} = \min \sum_i \|y - \beta X\|_y + \lambda \|\beta\|_x \quad (5)$$

Enet can deal with both multicollinearity problems and feature selection. It is a combination of RR and Lasso regression. Its objective function is:

$$\text{obj}_{\text{Enet}} = \min \sum_i \|y - \beta X\|_y + \lambda_1 \|\beta\|_x + \lambda_2 \|\beta\|_y \quad (6)$$

Parameters  $\lambda$ ,  $\lambda_1$  and  $\lambda_2$  can be obtained by cross-validation respectively.

### 2.2.3. Sufit and stepwise regression

In order to simplify the construction of the prediction model, it is usually necessary to select the most appropriate subset modeling in many variable quantum sets that affect wind power generation. Both the full subset regression model and the stepwise regression model are well - suited for this task. The full subset regression method divides the explanatory variables into different subsets and uses the least square method to regression the variables in each subset to get the most appropriate subset. The stepwise screening method, on the other hand, introduces variables one by one. Subsequently, a significance test is used to determine whether the newly introduced variable brings about significant changes to the model. If there is a significant change, when the original introduced variable no longer changes significantly due to the introduction of subsequent variables, this variable will be eliminated to ensure that only significant variables are included in the regression equation before each new variable is introduced, until no significant explanatory variable is selected into the regression equation and no insignificant explanatory variable is removed from the regression equation. Finally, an optimal set of variables is obtained.

#### 2.2.4. RF model

Random forest algorithm performs well in WPG [8]. It produces the final wind power by generating multiple wind power trees and orderly combining the predicted wind power generation of each tree. In this algorithm, the model training of each tree is carried out independently and simultaneously, and the sample features are randomly sampled, which has the characteristics of strong generalization ability.

#### 2.2.5. SVM model

The main goal of SVM is to find a hyperplane that can divide space into different classes, whose equation is:

$$F(x) = \omega^T X + b = \left( \sum_{i=1}^n a_i y_i X_i \right)^T X + b = \sum_{i=1}^n a_i y_i (X_i^T X) + b \quad (7)$$

In the formula, sample  $X$  is an  $n$ -dimensional vector.  $X_i$  is the feature point of the support vector training sample.  $y_i$  is the category label of the support vector point  $X_i$ .  $a_i$  is the Lagrange multiplier.  $b$  is the displacement term that determines the distance from the origin to the hyperplane, and  $\omega = [\omega_1, \omega_2, \dots, \omega_d]$  is the normal vector of the hyperplane [9]. This algorithm needs to map the data dimension to a higher dimensional space, so the equation for the hyperplane can also be written as:

$$F(x) = \sum_{i=1}^n a_i y_i (\Phi(X_i) \cdot (\Phi(X))) + b \quad (8)$$

$\phi(X)$  represents the conversion function of vector  $X$  from the lower dimension of the original data to the higher dimension. It can be seen that the model accuracy of SVM depends on the number of support vectors and kernel parameters, and the model generalization ability is strong.

### 3. Examples

#### 3.1. Research idea

In order to simplify the establishment of wind power generation prediction modeling in Guangdong Province and achieve accurate prediction of future wind power generation in Guangdong Province, this study achieved long-term wind power generation prediction through the following steps:

- 2009~2023 is selected as the time interval for model establishment and evaluation, with natural year as the unit. Consult the independent variables of 6 energy factors: Guangdong's annual WPG( $Y$ ), Guangdong's economic aggregate( $X_1$ ), Guangdong's electricity generation( $X_2$ ), Guangdong's annual average wind speed( $X_3$ ), Guangdong's cumulative installed wind power capacity( $X_4$ ), Guangdong's total resident population( $X_5$ ), and Guangdong's annual total electricity consumption( $X_6$ ).
- Pearson correlation analysis was used to select factors with a correlation of more than 0.70 with the annual wind power generation in Guangdong Province as impact factors.
- Based on the impact factor of wind power generation, the data sets meeting the requirements are divided into 80% training set and 20% test set.
- In the training set, 10 algorithms, including convolution - bidirectional long and short term memory algorithm, Ridge regression algorithm Ridge regression model, Lasso regression algorithm, elastic network algorithm, full subset regression algorithm, step by

step regression algorithm, random forest algorithm and support vector machine (linear kernel, polynomial kernel, radial basis function kernel and Sigmoid kernel), are respectively adopted. The prediction model of wind power generation is constructed.

- Comprehensively evaluate the prediction accuracy of each model according to the MAE and RMSE in the test set , and the  $R^2$  in the test set, and select the prediction model with good performance.
- Based on China's future GDP growth rate and Pearson correlation coefficient mentioned in "Research on Low-carbon Transition Planning and Decision-making for Carbon Neutrality of Electricity", the changes of other influencing factors from 2025 to 2060 are reasonably calculated, and the forecast results of various models for future wind power generation are finally obtained [10]. Finally, the predicted values under various algorithms are averaged to obtain the estimated future wind power generation in Guangdong Province.

### 3.2. Data sources

The main data sources of wind power generation and various energy factors in Guangdong Province over the years are shown in the following table 1:

Table 1: The main data sources of wind power generation and various energy factors in Guangdong province over the years

| Factors              | Data Sources                                                                           |
|----------------------|----------------------------------------------------------------------------------------|
| $Y$                  | National Energy Administration (China), China Energy Statistical Yearbook              |
| $X_1, X_2, X_5, X_6$ | Guangdong Statistical Yearbook                                                         |
| $X_3$                | NOOA, NCEI                                                                             |
| $X_4$                | China Wind Power Installed Capacity Statistics, National Energy Administration (China) |

### 3.3. Model establishment and verification

#### 3.3.1. Influence factors of WPG

As shown in the figure, according to Pearson correlation coefficient, the influencing factors are ranked from the largest to the smallest: cumulative wind power installed capacity of Guangdong Province( $X_4$ ), total annual electricity consumption of Guangdong Province( $X_6$ ), economic aggregate of Guangdong Province( $X_1$ ), power generation of Guangdong Province( $X_2$ ), and total resident population of Guangdong Province( $X_5$ ) .



Figure 1: Pearson correlation coefficient heat map

According to China's forecast GDP growth rate, the GDP growth rate of Guangdong Province is reasonably estimated, and the predicted value of  $X_1$  is obtained. Then, based on the historical data and the Pearson correlation coefficients of  $X_1$  and  $X_2$ ,  $X_4$  and  $X_6$ , a simple polynomial linear fitting is performed to obtain the future estimation. For  $X_5$ , the power function has the best fitting effect. The predicted value of the independent variable after fitting is:

Table 2: The predicted value of the argument

| Year | $X_1$<br>(Trillion yuan) | $X_2$<br>( $10^4$ tons of coal equivalent) | $X_4$<br>( $10^4$ kW) | $X_5$<br>( $10^8$ people) | $X_6$<br>( $10^8$ kWh) |
|------|--------------------------|--------------------------------------------|-----------------------|---------------------------|------------------------|
| 2030 | 18.8375                  | 11763.53                                   | 4198.38               | 1.3193                    | 10681.14               |
| 2035 | 22.4813                  | 14290.41                                   | 6832.83               | 1.3415                    | 12391.64               |
| 2050 | 31.7706                  | 20732.28                                   | 16491.85              | 1.3792                    | 13954.18               |
| 2060 | 35.9725                  | 23646.19                                   | 22249.67              | 1.3911                    | 18724.74               |

### 3.3.2. Simulation effect of wind power generation prediction model

The data of the dependent variable and independent variables were respectively input into the CNN-BiLSTM, RR, Lasso and Enet algorithms. It was discovered that the CNN - BiLSTM and Lasso algorithms performed poorly in wind power generation prediction. For the CNN - BiLSTM, this could be attributed to its typical application in processing image information. The latter may be because it is good at handling regression models with multiple independent variables, while this wind power generation only has 5 independent variables used as reference factors for model prediction. Conversely, the RR and Enet models, which exhibited better fitting results tend to have the same accuracy in wind power prediction, possibly because they both use similar cross-validation methods. Based on the data of wind power generation and its influencing factors in previous years, it is found that subset  $\{X_4\}$  has the best comprehensive performance after full subset regression for the non-null true subsets of the independent variables of energy factors. In contrast, the optimal subset selected by the stepwise regression model is  $\{X_2, X_4\}$ , both of which have high accuracy and are suitable for wind power generation prediction. Regarding the random forest algorithm, the model achieved the highest accuracy when 59 wind power trees were generated. However, its fitting level remained rather low. In the prediction of wind power generation in Guangdong Province, the most appropriate kernel function for the SVM algorithm was the linear kernel.

In summary, the algorithms suitable for future wind power generation prediction in Guangdong Province include RR, Enet, Sufit, Stepwise and SVM (linear kernel function). The accuracy of their model is shown below:

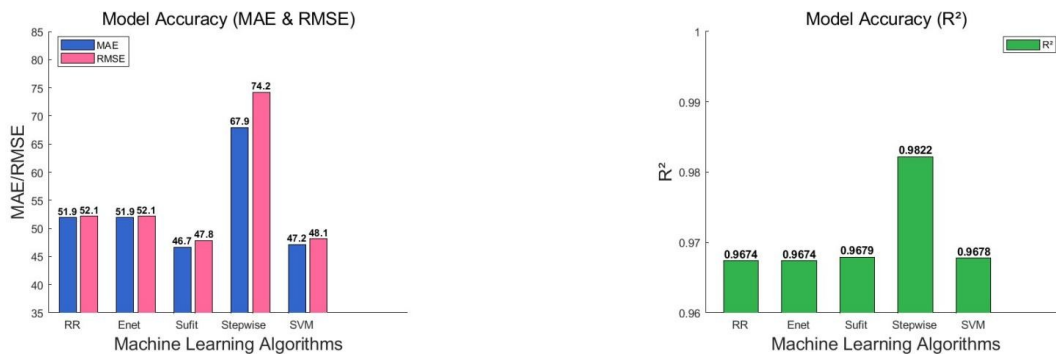


Figure 2:  $MAE$  and  $RMSE$  of machine learning algorithms & Figure 3  $R^2$  of machine learning algorithms

By substituting the selected models into the predicted values of the independent variables of energy factors, the predicted values of Y under the five algorithms are finally obtained as follows table 3: (unit: billion kWh)

Table 3: Prediction of WPG

| Year | RR     | Enet   | Sufit  | Stepwise | SVM    | average |
|------|--------|--------|--------|----------|--------|---------|
| 2030 | 653.1  | 653.0  | 696.2  | 1957.9   | 609.6  | 913.96  |
| 2035 | 1060.4 | 1060.3 | 1135.6 | 2379.4   | 735.8  | 1274.30 |
| 2050 | 2550.7 | 2550.5 | 2746.5 | 3453.7   | 1064.3 | 2473.14 |
| 2060 | 3438.1 | 3437.8 | 3706.8 | 3939.7   | 1216.0 | 3147.68 |

#### 4. Conclusion and prospect

The correlation between the annual average wind speed of Guangdong Province and the predicted value of long - term wind power generation, as well as other energy factors, is low. This is due to the fact that the long - term wind power generation forecast in this paper relies on regular historical data accumulated over the years, which effectively minimizes the interference from wind speed and other chaotic system elements. In the context of wind power generation prediction in Guangdong, algorithms like RR, Enet, Sufit, Stepwise, and SVM (linear kernel function) exhibit a good fitting performance, rendering them appropriate for predicting future wind power generation in the province. Among these, the SVM (linear kernel function) algorithm demonstrates the most favorable comprehensive fitting results, so the relevant power generation industry can apply SVM for medium - term forecasting. On the contrary, in the long - term wind power generation prediction model for Guangdong Province, CNN - BLSTM, Lasso, RF, SVM (polynomial kernel function), and SVM (radial kernel function) show poor fitting effects and thus are not suitable for long-term wind power generation prediction. Additionally, it is worth attempting to combine different algorithms to leverage their respective strengths and offset weaknesses. For example, the Sufit and SVM algorithms can be organically integrated; potentially, the new algorithm SUFIT-SVM could achieve higher accuracy and a faster convergence rate.

#### References

- [1] Yang K. *Research on low-carbon Transformation of Power Structure under the goal of "dual carbon"* [D]. Shanxi University of Finance and Economics, 2022.
- [2] Xu Yang.(2023). *Research on Deep Learning method for ultra-short term prediction of wind power* (Master Dissertation, Beijing University of Posts and Telecommunications). Master of <https://link.cnki.net/doi/10.26969/d.cnki.gbydu.2023.002826doi:10.26969/d.cnki.gbydu.2023.002826>.
- [3] Gao A. *Research on the optimization path of China's clean energy development driven by the "dual carbon" goal* [D]. Jiangsu University, 2021.
- [4] The general office of the guangdong people's government about print and distribute to promote sustainable development of the offshore wind and orderly development and related industry plan to inform [http://www.gd.gov.cn/zwgk/gongbao/2021/17/content/post\\_3367237.html](http://www.gd.gov.cn/zwgk/gongbao/2021/17/content/post_3367237.html)
- [5] You S B, Yan Y. *Stepwise Regression Analysis and Its Application*[J]. *Statistics & Decision*, 2017 (14): 31-35.
- [6] Noman Khan, Samee Ullah Khan, Sung Wook Baik. *Deep dive into hybrid networks: A comparative study and novel architecture for efficient power prediction*. *Renewable and Sustainable Energy Reviews* 182 (2023) 113364.
- [7] Lei J, Liu C, Jiang D. *Fault diagnosis of wind turbine based on Long Short-term memory networks*. *Renew Energy* 2019;133:422–32.
- [8] Wangjie Liu, Li Jia. *Wind Power Prediction Based on the Stacking Model of XGBoost and Random Forest*. 2022 IEEE 11th Data Driven Control and Learning Systems Conference. August 5-7, 2022, Emeishan, China
- [9] Chen T, Guestrin C, *Xgboost: A scalable tree boosting system*, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016:785-794.
- [10] Zhang Haonan. *Research on Low-carbon Power Transition Planning and Decision-making for Carbon Neutrality* [D]. North China Electric Power University (Beijing), 2022.