Comparative U Of Machine Learning Model to Explore the Impact of Geographical Factors on Wind Speed

Zihe Zhao *

Brunel Lodon School, North China University of Technology, Beijing 100144, China

* Corresponding Author Email: 2281410@brunel.ac.uk

Abstract. At present, society's awareness of environmental protection has been enhanced, and the demand for simultaneous development and protection of environmental purification and sustainability is increasing. As a clean energy source, wind power generation can well replace a part of fossil energy and prevent the inevitable pollutants caused by the use of fossil energy. However, the generators that wind power depends on do not have the ability to migrate, so the location of wind turbines is particularly important. It can not only occupy the potential residence of human beings, but also provide a large and stable amount of power generation. In this paper, three different machine learning models are used to explore the impact of a number of geographical factors on wind speed and to predict the future trend of wind speed. By comparing the three different models, the most suitable model with the highest fitting rate is XGBoost model, and its corresponding $\boldsymbol{R^2}$ is 0.92. The scatter diagram of wind speed prediction and longitude and latitude produced in this paper can further assist in the location of wind turbines to maximize energy efficiency.

Keywords: Location of wind turbines; XGBoost model; Maximize energy efficiency.

1. Introduction

Comparative use of machine learning model to explore the impact of geographical factors on wind speed, with the growth of power demand, the strengthening of fossil fuel control and the intensification of environmental pollution, the public's demand for clean energy such as wind energy, is increasing. Not only that, wind energy can bring huge economic benefits, and wind energy will theoretically play a key role in energy supply in the future [1][2]. Wind speed, while the operation efficiency and economic benefits of wind power generation highly depend on the accuracy of wind speed prediction, and compared with traditional energy, wind energy is still more expensive, so it is particularly important to reduce costs and improve the efficiency of wind capacity [3][4][5]. Affected by many factors, its randomness, intermittency and uncontrollability also bring great challenges to practical applications [3][4][5]. Weyi Jiang features a sparse attention mechanism, and at the same time captures the macro and micro wind speed fluctuations to predict the interval value of spatio-temporal wind speed [6]. Adnan Saeed and others use the branch integration model architecture based on Long Short-Term Memory to predict the wind speed [7]. Mengning Wu proposed to predict wind speed by inputting atmospheric variables such as temperature and humidity divergence [8]. However, although these studies have improved in technology, there are still obvious deficiencies in the adaptability of multi-terrain prediction and long-term regression projects, and the performance is poor under special terrain conditions [9][10]. To solve these problems, three machine learning models, SVR, RF and XGBoost are compared and analyzed to improve the prediction accuracy and model robustness.

2. Data Set

2.1. Data Description

This paper uses the storms data set provided by the dplyr package, which is derived from NOAA's best track records of tropical cyclones (hurdat Series), covering the time interval observation of tropical storms and hurricanes in the Atlantic basin. The original data scale is 4629 observations and 14 variables. To be consistent with the wind speed prediction task, only six variables directly related

to the target are selected for modeling. The six variables used are shown in Table 1: wind, pressure, lat,long,tropicalstorm_force_diameter,hurricane_force_diameter. Other variables (such as name, year, month, day, hour, status, category, etc.) are used for background information and do not participate in model training.

In order to keep consistent with the caliber of the data source, this paper uses the original unit of data: the unit of wind speed is expressed in knots, the unit of air pressure is expressed in mbar, the unit of longitude and latitude is expressed in degree (decimal system), and the diameter of wind circle is expressed in nautical miles (NMI).

Table 1. Variables and units used in the model

Variable	Unit	Implication	type
wind	knots	Near surface wind speed (target variable)	value
pressure	mbar	Central/near central pressure of cyclone	value
lat	degree	Geographic latitude (North plus South minus)	value
long	degree	Geographic longitude (positive in the East and negative in the West)	value
tropicalstorm_force_diameter	nmi	≥34 kt Diameter of wind ring	value
hurricane_force_diameter	nmi	≥64 kt Diameter of wind ring	value

2.2. Data Reprocessing

2.2.1 Vvariable screening and missing value processing

According to the research objectives, only six variables directly related to wind speed prediction are retained: wind, pressure, lat, long, tropicalstorm_force_diameter, hurricane_force_diameter. Then, the deletion detection and processing were carried out only for these six variables, and the deletion method (na.omit) was used to eliminate the missing observations, so as to avoid the propagation of interpolation error to the subsequent modeling. In this step, the sample size was reduced from 4629 to 2086, and a total of 2543 lines (54.94%) were deleted. In order to keep the caliber consistent, the original units (knots, mbar, degree, nautical miles) are not changed in this paper.

2.2.2 Abnormal value identification and handling

In order to reduce the interference of extreme observations on model training, the interquartile range (IQR) method is used to identify outliers for the target variable wind. Remember that Q_1, Q_3 are the 1st and 3rd quartiles respectively, and IQR is defined as

$$IQR = Q_3 - Q_1 \tag{1}$$

According to Tukey's "fence" rule, set the judgment interval as

$$Lowerbound = Q_1 - 1.5IQR \tag{2}$$

$$upperbound = Q_3 + 1.5IQR \tag{3}$$

Observations falling outside$[Q_1 - 1.5\,IQR\,,\ Q_3 + 1.5\,IQR]$are recorded as outliers and eliminated. After this step, the sample size is adjusted from 2086 to 2086, and a total of 0 rows (0%) are deleted.

When detecting outliers, this paper uses the interquartile distance (IQR) method to test the wind speed variable (wind). After calculation, the distribution range of wind speed in the cleaned data set is 65 – 135 knots, and the abnormal value judgment interval given by the quartile distance method is$[Q_1 - 1.5\,IQR\,,\ Q_3 + 1.5\,IQR] \approx$ [lower bound, upper bound]. Since all wind speed observations fall within this range, no outliers are detected.

In terms of physical and geographical factors, the manifestation of wind speed is relatively stable and has no significant outliers. In addition, a large number of deleted data lead to the reduction of the amount of remaining data, and the probability of presenting anomalies is smaller. The data cleaning effect is shown in Figure 2.

2.2.3 Dataset construction and reproducibility

After cleaning, the analysis data set "data_clean" (2086×6) is obtained and exported as cleaned_stores.csv to support replication. Data processing is completed in the environment of R version 4.5.0 (2025-04-11 ucrt) and dplyr (1.1.4). The subsequent standardization and model parameter adjustment are implemented based on the statistics of the training set, and the test set is only used for the final evaluation to avoid information leakage. Figure 1 shows the results of data Visualization.

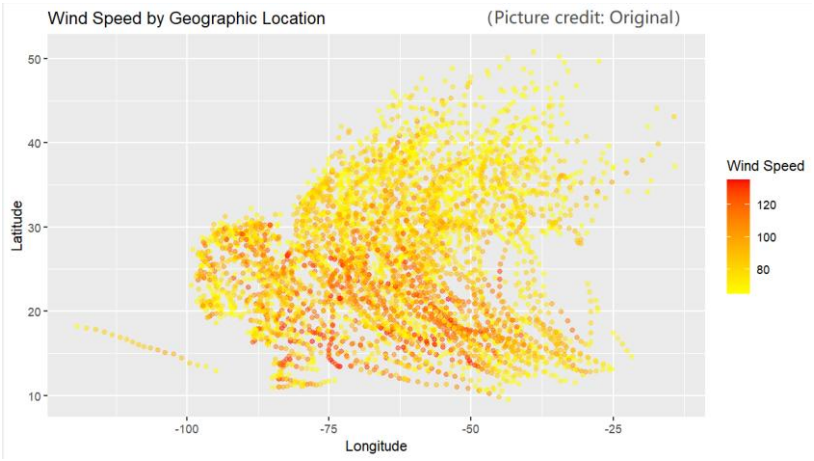

Fig. 1 Data visualization (Data from: storms dataset)

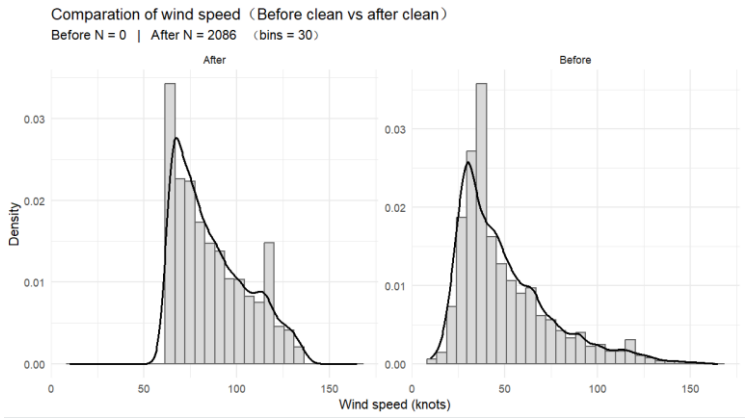

Fig. 2 Effect diagram of data cleaning (Data from: storms dataset)

3. Experimental Method

In this paper, three kinds of models are used for experiments, and the experimental results and model evaluation indices such as $\boldsymbol{R^2}$ and MSE are compared and analyzed. The advantages and disadvantages of each model and the most suitable model for this experiment are discussed.

3.1. Support Vector Regression

Compared with the linear separable model (SVM), the SVR model area does not need to completely and correctly classify all data, that is, the "hard interval method", but chooses to introduce a tolerance interval ε (ε -tube), and only the data exceeding the tolerance interval will be punished. At the same time, use $\frac{1}{2}||\boldsymbol{w}||^2$ to control the complexity of the model. Its original form is:

$$\min_{\boldsymbol{w},b,\xi,\xi^*} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{4}$$

$$s.t.\, y_i - (\boldsymbol{w}^{\top}\phi(\boldsymbol{x_i}) + b) \leq \varepsilon + \xi_i \tag{5}$$

$$(\boldsymbol{w}^{\top}\boldsymbol{\phi}(\boldsymbol{x_i}) + b) - y_i \leq \varepsilon + \xi_i^* \tag{6}$$

Where $\xi_i, \xi_i^* \geq 0, i = 1, \ldots, n$ and $C > 0$; $\xi \geq 0$, $L_\xi(y, y\hat{}) = max\{0, |y - y\hat{}| - \xi\}$

Dual the above formula and obtain two sets of Lagrange multipliers: α_i , $\alpha_i^* \in [0, C]$. "*" is used to distinguish two groups of different variables. The variables using * are the high part, while those not used are the low part.

The final regression function is written as the weighted sum of support vectors:

$$\hat{y}(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{7}$$

Where $\boldsymbol{K(x, z) = \langle \phi(x), \phi(z) \rangle}$ is the kernel function. The radial basis function (RBF) kernel is used in this paper

$$K(x, z) = exp\left(-\Upsilon \left| |x - z| \right|^2\right), \Upsilon = \frac{1}{2\sigma^2} \tag{8}$$

R Only the data exceeding the "tolerance interval", that is, the samples meeting$|\ \boldsymbol{y_i - \hat{y}(x_i)}\ | > \boldsymbol{\varepsilon}$, will be called support vectors and become the main contribution data of the training model. Therefore, the model is naturally sparse, and similar historical support vectors contribute more to wind speed prediction.

3.2. Random Forest

Random forest is an ensemble learning method, which can make a regression by averaging the prediction results of multiple decision trees. By bootstrap the training set, each tree can train the model on the training subset after scrambling and sampling. The prediction of the new sample x is the average output of each subtree:

$$\hat{y}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{9}$$

Where $\boldsymbol{T_b(x)}$ is the prediction of the b tree to x.

The regression tree selects the $\boldsymbol{mtry}$ feature at each node, and selects the partition that makes the MSE drop the most. Set the current node sample set as S purity can be defined as:

$$\mathcal{J}(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \overline{y_S})^2 \tag{10}$$

After S is divided into left and right subsets S_L and S_R according to a certain characteristic threshold, the impurity is:

$$\mathcal{J}_{\text{split}} = \frac{|S_L|}{|S|} \mathcal{J}(S_L) + \frac{|S_R|}{|S|} \mathcal{J}(S_R) \tag{11}$$

Maximize impurity reduction:

$$\Delta\mathcal{J} = \mathcal{J}(S) - \mathcal{J}split \tag{12}$$

Usually, no pruning or weak pruning is used to reduce the deviation, and the variance is suppressed by the ensemble average.

Use OOB error built-in verification to take the complement set other than the training set used by each tree as the test set of the model to roughly estimate the generalization ability of RF model.

OOB error is:

$$\text{MSE}_{\text{OOB}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{y^{(-i)}}(x_i)\right)^2 \tag{13}$$

The replacement importance is based on OOB test set, and the J feature is randomly replaced to obtain the error increment:

$$\mathrm{VI_j} = \mathrm{MSE_{perm(j)}} - \mathrm{MSE_{base}} \tag{14}$$

The larger the error increment, the greater the contribution of the j-feature to the generalization model.

3.3. Extreme Gradient Boosting

R XGBoost is an improved algorithm of gradient boosting trees, which improves the prediction accuracy by iterating to build a weak learner and gradually optimizing the residual.

In the relationship between wind speed and pressure, wind circle diameter, latitude and longitude and other variables studied in this paper, XGBoost step-by-step tree addition, second-order approximation and regularization can learn the local structure of different regions, which is in line with the strong nonlinear and threshold scenarios discussed in this paper.

Xgboost overlays the regression trees in a "tree by tree" manner to approximate the objective function. The $\boldsymbol{t}$ round model learns a new tree f_t (learning rate η) based on the previous prediction:

$$\widehat{y^{(t)}}(x) = \widehat{y^{(t-1)}}(x) + \eta\, f_t(x), f_t \in \mathcal{F} \tag{15}$$

Objective function to minimize "loss+model complexity" during training:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t) \tag{16}$$

$$\Omega(f) = \gamma T + \frac{\lambda}{2}\sum_{j=1}^{T} w_j^2 \tag{17}$$

Where T is the number of leaf nodes and W_j is the weight of the J leaf; $\boldsymbol{\gamma}$ is the penalty coefficient, $\boldsymbol{\lambda}$ is L_2 of leaf weight.

Regularization, both of which suppress overfitting.

For efficient solution, XGBoost performs second-order Taylor expansion on $\mathcal{L}^{(t)}$at $\widehat{y^{(t-1)}}$:

$$g_i = \left.\frac{\partial l(y_i, \hat{y})}{\partial \hat{y}}\right|_{\hat{y}=\widehat{y_i^{(t-1)}}} \tag{18}$$

$$h_i = \left.\frac{\partial^2 l(y_i, \hat{y})}{\partial \widehat{y^2}}\right|_{\hat{y}=\widehat{y_i^{(t-1)}}} \tag{19}$$

$$G_j = \sum_{i\in I_j} g_i\,, H_j = \sum_{i\in I_j} h_i \tag{20}$$

Then the approximate target is decomposed into each leaf, and the optimal leaf weight and the corresponding splitting gain are:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{21}$$

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda}\right) - \gamma \tag{22}$$

During the training, the partition with the maximum gain is selected in the candidate segmentation to generate the long tree.

4. Experiments

4.1. Experimental Configuration

Table 2. Experimental Configuration

Project	Configuration
Hardware environment	AMD Ryzen9 8945HX; RTX 5060 GPU 8 GB; Windows 11
Software environment	R 4.5.0, caret, RandomForest, XGBoost
Dataset	cleaned_storms.csv, 10000 records, 6 features
Preprocessing	Delete missing values, standardize continuous variables, and eliminate outliers using IQR method
Training/Testing Division	80% training, 20% testing
Model Configuration	SVR (RBF , Grid Search),RF (ntree=800),XGBoost (η=0.1, max_depth=6, subsample=0.8, nrounds=200)
Evaluation metric	R^2, MSE, RMSE
Hardware environment	AMD Ryzen9 8945HX; RTX 5060 GPU 8 GB; Windows 11

Table 2 shows the experimental configuration used in this paper, as well as the parameter settings of the model and the evaluation indexes of the model in this paper.

4.2. Experimental Results

Table 3. Experimental results

Model	R^2	MSE	RMSE
SVR	0.8922	38.25	6.18
Random Forest	0.9180	29.0236	5.3874
XGBoost	0.9229	27.2615	5.2213

Table 3 shows that the inspection index is: $\boldsymbol{R^2}$ =0.9229; MSE = 27.2615; XGBoost with RMSE=5.2213 has the maximum $\boldsymbol{R^2}$ value and the minimum MSE and RMSE values. Therefore, XGBoost is the model with the best prediction effect, while SVR model has the least satisfactory fitting effect in this experiment, and random forest has medium performance.

The main theoretical advantage of XGBoost is that it gradually reduces the deviation through fitting residuals, which makes it easier to approach the real function in nonlinear scenarios such as predicted wind speed, and XGBoost uses a more comprehensive parameter adjustment method. Its L1/L2 regularization + early stop method can well inhibit overfitting, while for the SVR model, it mainly depends on individual super parameters to adjust parameters, which has a small adjustment space and is easy to overfit. The optimization goal of XGBoost is the same as MSE/RMSE, and the second-order approximation is used to do a more refined gradient update, while the soft interval method used by SVR is not sensitive to small errors, and the overall deviation is high. The partition + rule of the tree model is more suitable for this kind of non-stationary and segmented relationship.

5. Conclusion

This paper has completed the wind speed prediction project based on geographical factors, which can be more intuitive and convenient to find the geographical location where wind speed can be efficiently used, and the XGBoost model used has strong stability, at the same time, the linear relationship between wind speed and geographical factors has a small change range, which can be better put into use. Prospects for the future: if it is to be put into use in a large area, it is also necessary

to add other considerations after the predicted wind speed results are obtained, such as the impact of high wind speed on the life of the generator, the corrosion of the marine geographical location on the generator and the maintenance cost. In the future, it is expected to put a variety of factors into the wind speed prediction model to complete the one-stop prediction task. To complete this task, people need to rely on a more comprehensive model and more advanced technology, such as the LSTM model and attention mechanism mentioned above.

References

[1] C Shorabeh S N, Firozjaei H K, Firozjaei M K, et al. The site selection of wind energy power plant using GIS-multi-criteria evaluation from economic perspectives. Renewable and Sustainable Energy Reviews, 2022, 168: 112778.

[2] Msigwa G, Ighalo J O, Yap P S. Considerations on environmental, economic, and energy impacts of wind energy generation: Projections towards sustainability initiatives. Science of The Total Environment, 2022, 849: 157755.

[3] Roga S, Bardhan S, Kumar Y, et al. Recent technology and challenges of wind energy generation: A review. Sustainable Energy Technologies and Assessments, 2022, 52: 102239.

[4] Liu J, Song D, Li Q, et al. Life cycle cost modelling and economic analysis of wind power: A state of art review. Energy Conversion and Management, 2023, 277: 116628.

[5] Meyers J, Bottasso C, Dykes K, et al. Wind farm flow control: prospects and challenges. Wind Energy Science Discussions, 2022, 2022: 1-56.

[6] Jiang W, Wang J, He X. Interval multi-feature sparse transformer-CNN: A synergistic approach to precise and efficient spatio-temporal wind speed interval-value prediction. Alexandria Engineering Journal, 2025, in press. DOI: 10.1016/j.asoc.2025.113910.

[7] Saeed A, Li C, Rubaiee S, Danish M, Anwar S. Enhanced wind speed forecasting for sustainable power systems: A deep learning framework unifying deterministic predictions and uncertainty quantification. Energy, 2025, in press. DOI: 10.1016/j.energy.2025.137979.

[8] Wu M. Wind speed forecasting by spatial-temporal data-driven models using atmospheric input variables. Ocean Engineering, 2024, 277: 118191. DOI: 10.1016/j.oceaneng.2024.118191.

[9] Al-Selwi S M, Hassan M F, Abdulkadir S J, et al. LSTM inefficiency in long-term dependencies regression problems. Journal of Advanced Research in Applied Sciences and Engineering Technology, 2023, 30(3): 16-31.

[10] Elkadeem M R, Younes A, Mazzeo D, et al. Geospatial-assisted multi-criterion analysis of solar and wind power geographical-technical-economic potential assessment. Applied Energy, 2022, 322: 119532.