Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6682242, 14 pages
https://doi.org/10.1155/2021/6682242

Research Article

Prediction of Power Outage Quantity of Distribution Network Users under Typhoon Disaster Based on Random Forest and Important Variables

Min Li,[1] Hui Hou,[2] Jufang Yu,[2] Hao Geng,[2,3] Ling Zhu,[1] Yong Huang,[4,5] and Xianqiang Li[2]

[1]*Guangdong Power Grid Co., LTD., Guangzhou 510080, China*
[2]*School of Automation, Wuhan University of Technology, Wuhan 430070, China*
[3]*Electric Power Research Institute, Yunnan Power Grid Co., Ltd., Kunming 650200, China*
[4]*GuangDong Power GRID Co., Ltd., Electric Power Research Institute, Guangzhou 510080, China*
[5]*Power Remote Sensing Technology Joint Laboratory of China Southern Power Grid, Guangzhou 510080, China*

Correspondence should be addressed to Hui Hou; houhui@whut.edu.cn

Received 24 November 2020; Revised 22 December 2020; Accepted 28 December 2020; Published 6 January 2021

Academic Editor: Xiao-Shun Zhang

Typhoons can have disastrous effects on power systems. They may lead to a large number of power outages for distribution network users. Therefore, this paper establishes a model to predict the power outage quantity of distribution network users under a typhoon disaster. Firstly, twenty-six explanatory variables (called global variables) covering meteorological factors, geographical factors, and power grid factors are considered as the input variables. On this basis, the correlation between each explanatory variable and response variable is analyzed. Secondly, we established a global variable model to predict the power outage quantity of distribution network users based on Random Forest (RF) algorithm. Then the importance of each explanatory variable is mined to extract the most important variables. To reduce the complexity of the model and ease the burden of data collection, eight variables are eventually selected as important variables. Afterward, we predict the power outage quantity of distribution network users again using the eight important variables. Thirdly, we compare the prediction accuracy of a model called the No-model that has been used before, Linear Regression (LR), Support Vector Regression (SVR), Decision Tree Regression (DTR), RF-global variable model, and RF-important variable model. Simulation results show that the RF-important variable model proposed in this paper has a better effect. Since fewer variables can save prediction time and make the model simplified, it is recommended to use the RF-important variable model.

1. Introduction

Typhoon disasters may lead to a large area of power outage for distribution network users. The prediction of the power outage quantity of distribution network users under a typhoon disaster can effectively improve the accuracy of disaster prevention and reduction. It can also shorten the outage time of distribution network users, reduce power outage loss, and improve user satisfaction.

Under typhoon disaster, there are many factors affecting the power outage of distribution network users, including meteorological factors, geographical factors, power grid factors, and so on [1]. If the traditional model-driven method is used to predict the power outage quantity, the model will be complex and difficult to solve. In addition, with the increase and normalization of power outage data of distribution network users, it is possible to predict power outage quantity of distribution network users using a data-driven method [2, 3].

At present, some scholars have successfully used the data-driven method to assess the risk to power systems under typhoon disasters. Statistical learning models, such as linear model, are firstly applied to evaluate the power outage in hurricane weather in [4]. However, they mainly focused

on the fitting effect of the model, instead of prediction accuracy. The impact of soil and terrain on power outage of distribution network users based on classification and regression trees (CART) is studied in [5]. However, it did not pay attention to the improvement of prediction accuracy. To make the model more comprehensive, many scholars decided to take more influence factors into consideration. Considering the influence factors, such as maximum wind speed, wind speed duration, rainfall, etc., a cumulative time failure model was used to predict the power grid outage under hurricane in [6]. Considering meteorological, geographical, and social information, models of equipment failure rate under natural disasters were established in [7]. On this basis, data-driven methods are widely used to assess the risk to power systems under natural disasters. An ice cover risk assessment model for power systems based on fault tree was proposed in [8]. It involved the effective assessment of transmission line risk, line break, and tower collapse. Based on the relevant public data affecting the power system, prediction models of power outage rate under disasters through data mining were established in [9–11]. In addition, a method for predicting the risk level of power outage in distribution network was presented by [12], which takes into account the weather factors. However, the risk level was classified, while factors such as region were not taken into consideration. Based on support vector machine (SVM) and grey prediction technology, a reliability prediction model for transmission line operation was proposed in [13]. It considered factors such as the running time of components and the region where the components are located. Considering storm, rainstorm, high temperature, and other weather factors comprehensively, a prediction model of the original parameters of a power system based on fuzzy clustering and similarity degree was proposed in [14]. This model considered most climatic factors but did not further evaluate the damage to the power grid.

In order to improve the prediction accuracy, the prediction area was firstly meshed in [15]. To carry out distribution network planning in a scientific and reasonable way, a multistage grid division method for distribution network was proposed [16]. Then, based on geographical grid division, the negative binomial regression model was used to predict the power outages quantity of distribution network users under Hurricane [17]. Based on the data of weather and land cover type, the spatial distribution of power outage in the 2-kilometer grid was predicted by using the Boosted Trees [18]. In addition, the support vector machine was used to predict the number of distribution towers in a 3-kilometer grid [19]. However, due to the large grid division, the eigenvalues of variables in the grid vary to a great extent, resulting in the inaccuracy of the obtained sample data, which affect the final prediction accuracy.

In the light of the aforesaid scenario, this paper proposed a prediction method of power outage quantity of distribution network users based on Random Forest (RF) algorithm. The main innovative contributions of the paper can be summarized as follows:

(1) A data sample space with twenty-six explanatory variables covering meteorological factors, geographical factors, and power grid factors is constructed. In addition, to better understand the relationship between explanatory variables and response variables, correlation of each explanatory variable and response variable is analyzed.

(2) To take as many variables into account as possible, we established a RF-global variable model covering all the twenty-six explanatory variables to predict the power outages quantity of distribution network users.

(3) To accelerate the evaluation efficiency under disasters, the importance of each explanatory variable is mined in this paper. On this basis, we extract eight most important variables to establish a novel RF-important variable model to predict the power outages quantity of distribution network users.

(4) We compare the prediction accuracy of a model called the No-model that has been used before, Linear Regression (LR), Support Vector Regression (SVR), Decision Tree Regression (DTR), RF-global variable model, and RF-important variable model. The validity and accuracy of the method based on important variable model proposed in this paper is verified. Thus, the RF-important variable model can provide guidance for emergency repair work.

The remainder of this paper is organized as follows.

The framework of the prediction model proposed in this paper is described in Section 2. In Section 3, the data sample space is introduced, and the relationship between each explanatory variable and response variable is analyzed. The RF algorithm we mainly used and the evaluation indicators are described in Section 4. In Section 5, the prediction model based on all the 26 explanatory variables and RF is built. In Section 6, the prediction model based on 8 important explanatory variables and RF is built, and the errors of No-model, LR, SVR, DTR, and the proposed two models are analyzed. Finally, Section 7 is the conclusion.

2. Prediction and Evaluation Framework of Power Outage Quantity of Distribution Network Users

The prediction framework of power outage quantity of distribution network users established in this paper is shown in Figure 1.

Firstly, create a data sample space. To consider as much as possible the collectible variables that may have an impact on the results, twenty-six explanatory variables are collected. The explanatory variables include meteorological factors (such as maximum wind speed, wind direction, rainfall, etc.), geographical factors (such as altitude, slope, underlay type, etc.), and power grid factors (such as number of distribution network users, number of box transformers, line

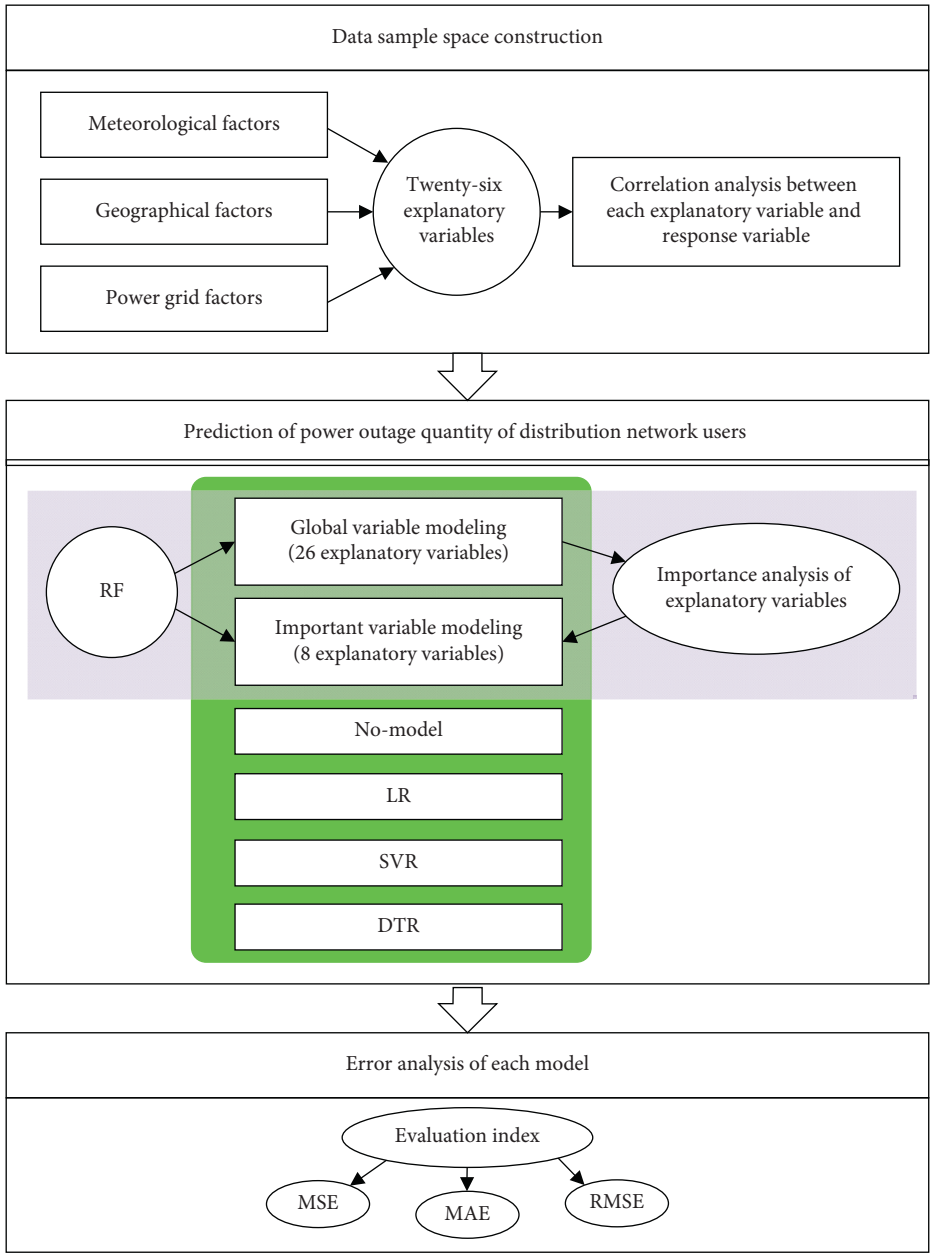

Figure 1: The prediction framework of power outage quantity of distribution network users.

length, etc.). Then, the correlation of explanatory variables and response variables are analyzed to mine the relationship of each variable.

Secondly, based on RF, the twenty-six explanatory variables (called global variables in this paper) are used to predict the power outage quantity of distribution network users. To reduce the complexity of the model, this paper analyzes the importance of all explanatory variables. The variables that have the greatest impact on the results are selected as important variables. This paper chose eight explanatory variables as important variables.

Finally, the important variables are used to conduct secondary modeling of power outage quantity prediction of distribution network users. In order to compare the pros and cons of the prediction results of each model, the results of RF-important variable model are compared with those of traditional No-model, LR, SVR, DTR, and RF-global variable model. Indicators for analyzing model errors include Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

3. Data Sample Space Construction

The power outage quantity of distribution network users under typhoon disaster is affected by many factors.

Therefore, the data of the prediction model is firstly described and the sample data space is constructed.

3.1. Analysis of Explanatory Variables. Similar to distribution network users' power outage spatial distribution [1], the factors affecting the power outage quantity of distribution network users under typhoon disaster include meteorological factors, geographical factors, and power grid factors. Among the power grid factors, the failure of the distribution network line mainly refers to the failure of the 10 kV overhead line. The cable is generally laid underground with insulation and protective layers, and its failure has little to do with the impact of typhoons and rainstorms. Therefore, this article only considers the power outage of the distribution network caused by the failure of the 10 kV overhead line exposed to the outdoor environment. In this paper, explanatory variables are added as much as possible to explore the relevant factors affecting the power outages quantity of distribution network users and to improve the accuracy of the prediction model. The selected explanatory variables of the prediction model are shown in Table 1 [20].

This paper establishes the prediction model of the power outage quantity of distribution network users on the basis of the sample data of three historical typhoons (Rammasun in 2014, Kalmaegi in 2014, and Mujigae in 2015) affecting Xuwen county, Guangdong province, China [21–23]. The data are provided by meteorological bureau and Electric Power Research Institute of Guangdong Power Grid Co., Ltd, China. In this paper, the study area is divided into 1641 samples; each sample represents a grid of 1 km × km. The variable X_1 is the maximum wind speed of each grid under the whole typhoon. Based on the regional grid division of 1 km × km, each typhoon produced 1641 samples with a total of 28 characteristic variables. Hence, the size of the entire sample space is $\Phi = (X, y)_{4923\times28}$. The variables in the meteorological factors and geographical factors are provided in the form of 1 km × 1 km data points. The Inverse Distance Weight Interpolation method is used to transform the data into continuous area data, and then the meteorological information and geographic information is extracted on this basis.

3.2. Analysis of Response Variable. In this paper, the power outage quantity of distribution network users under the typhoon disaster is predicted. Therefore, the power outage quantity of distribution network users Y_1 is taken as the response variable. The sample of descriptive statistics on the power outage quantity of distribution network users is shown in Table 2.

As shown in Table 2, the distribution range of the power outage quantity of distribution network users Y1 is 0~6121. The average predicted outage quantity is 70.51, and the standard deviation is 297.12. Three quartiles of 25%, 50%, and 75% are used to explore the distribution of results. It can be seen that the samples are mainly concentrated in the range of small data values. The probability distribution diagram of the response variable Y1 is shown in Figure 2. The samples are more concentrated in the range of small data values. The probability distribution diagram of tc response variable Y_1 is shown in Figure 2.

In order to eliminate the influence of the large coverage of power outage quantity of distribution network users, this paper normalizes this value and converts the response variable into the proportion of power outage. The proportion of power outage Y2 is equal to the number of power outage users Y1 divided by the number of distribution network users X20, Y2 = Y1/X20. Unless otherwise specified, the following response variables refer to the proportion of power outage Y_2.

3.3. Correlation Analysis between Each Explanatory Variable and Response Variable. In order to intuitively show the relationship between each explanatory variable and response variable Y_2, the scatter diagram between each explanatory variable and response variable is visualized, as shown in Figure 3.

As can be seen from Figure 3, there is no significant linear relationship between each explanatory variable and response variable, indicating that the effect of linear model will be poor. In order to further explore the relationship between each explanatory variable and response variable, Pearson correlation coefficient is used for quantitative correlation analysis. Assuming the existence of two variables, X and Y, the corresponding Pearson correlation coefficient [24] is calculated as follows.

$$r_{xy} = \frac{\mathrm{COV}(X, Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}, \tag{1}$$

where COV represents covariance and Var represents variance. If $|r_{xy}| < 0.4$, then X and Y are weakly correlated; if $0.4 \le |r_{xy}| < 0.7$, then X and Y are significantly correlated; if $0.7 \le |r_{xy}| < 1$, then X and Y are strongly correlated. The correlation analysis charts of variables are shown in Figures 4 and 5.

As can be seen from Figure 4, among the explanatory variables, there is a strong positive correlation between the distribution network users (X_{12}), maximum wind speed (X_1), wind speed duration (X_6, X_7), rainfall (X_3), and the power outage proportion (Y_2), while correlation between the other explanatory variables and the power outage proportion is weak.

In order to find out whether there is a correlation among the explanatory variables, the correlation heat map is shown in Figure 5.

As can be seen from Figure 5, there is a strong positive correlation between maximum wind speed (X_1) and rainfall (X_3), wind speed duration (X_6, X_7) and landing area (X_{11}). That is, when a typhoon lands in the study area, it will be accompanied by high wind speed and precipitation. And high wind speed makes the wind speed last longer.

4. The Prediction Principle of Power Outage Quantity of Distribution Network Users

4.1. Principle of Random Forest Algorithm. The main objective of supervised learning is to estimate the unknown

Table 1: The explanatory variables of the prediction model.

Factors	Variable name	Symbol	Remarks
Meteorological factors	Maximum wind speed	X_1	Maximum wind speed of each grid during a typhoon
	Wind direction	X_2	Corresponding wind direction at maximum wind speed
	Rainfall	X_3	Cumulative rainfall during a typhoon
	Temperature	X_4	Maximum temperature during a typhoon
	Humidity	X_5	Average humidity during a typhoon
	Wind speed duration of 20 m/s	X_6	Accumulated time when wind speed exceeds 20 m/s during a typhoon
	Wind speed duration of 30 m/s	X_7	Accumulated time when wind speed exceeds 20 m/s during a typhoon
	Wind class	X_8	Typhoon landing force
	Ten-level wind circle radius	X_9	Typhoon class 10 wind circle radius
	Landing time	X_{10}	Time interval from the last typhoon landing, in months
	Landing area	X_{11}	Indicator variable; if it is logged in the study area, it will be recorded as 1, otherwise it will be recorded as 0
Geographical factors	Whether there are distribution users	X_{12}	Indicating variable; the existence of distribution network users is recorded as 1, otherwise it is recorded as 0
	Altitude	X_{13}	/
	Slope	X_{14}	/
	Slope direction	X_{15}	/
	Underlay type	X_{16}	/
	Surface type	X_{17}	/
	Longitude	X_{18}	Longitude of grid center (LON)
	Latitude	X_{19}	Latitude of grid center (LAT)
Power grid factors	Number of distribution network users	X_{20}	Number of distribution network users in the grid
	Number of box transformers	X_{21}	Number of box transformers in the grid
	Number of desktop transformers	X_{22}	Number of desktop transformers in the grid
	Number of power towers	X_{23}	Number of 10 kV pole towers in the grid
	Number of pulling-line	X_{24}	/
	Number without pulling-line	X_{25}	/
	Line length	X_{26}	10 kV line length in the grid
	The power outage quantity of distribution network users	Y_1	Response variable
	Power outage proportion	Y_2	The ratio between the number of distribution network users' power outages and the number of distribution network users (response variable)

Table 2: Descriptive statistics of response variable.

	Mean	Std	Min	25%	50%	75%	Max
Y_1	70.51	297.12	0	0	0	18	6121

function f of the prediction variable Y (such as the power outage quantity) by using the d-dimensional vector of relevant input X (such as meteorological features, geographical features, and power grid features). For example, $Y = f(X) + e$, and e is the irremediable error. By minimizing the loss function L that represents the deviation between the observed value and the predicted value, the best unknown function f can be selected to make the prediction work best. This is the idea of supervised regression learning algorithm.

Random Forest (RF) is a nonparametric integrated data mining algorithm based on tree. Unlike a single regression tree with high variance and low bias, RF overcomes the problem of high variance by using model average. In addition, when the number of input variables is large, RF has better precision than other classical machine learning algorithms [7]. Hence, this paper establishes a prediction model for the power outage quantity of distribution network users based on the RF algorithm. The final RF output estimate is the predicted average of all the trees, expressed as follows:

$$f_{rf}(x) = \frac{1}{M} \sum_{m=1}^{M} T_m(x), \tag{2}$$

where M is the number of regression trees in RF, and $T_m(x)$ represents the model constructed by the m-th regression tree. The advantage of this method is that it can capture the nonlinear structure of data well, and it is robust to outliers and noise with a strong prediction accuracy.

4.2. The Evaluation Indicators. After the construction of the prediction model for the power outage quantity of distribution network users under typhoon disaster, it is necessary to evaluate the advantages and disadvantages of the model. In this paper, the evaluation indexes of the regression model are Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). Suppose the data set is $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, and the prediction regression function is $f(x)$, then the various error expressions are as follows:

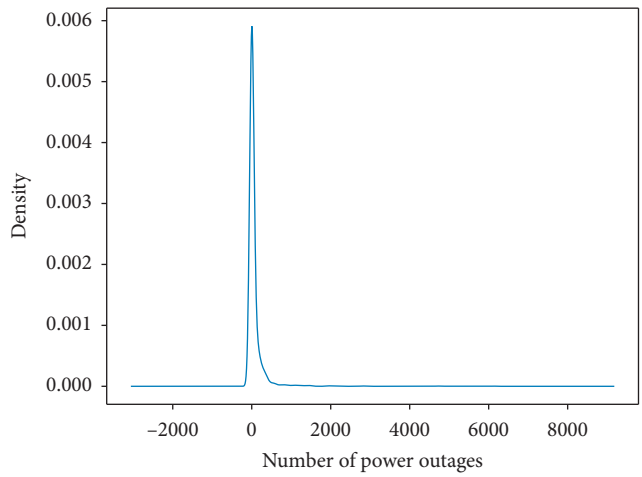

Figure 2: The probability distribution diagram of response variable.

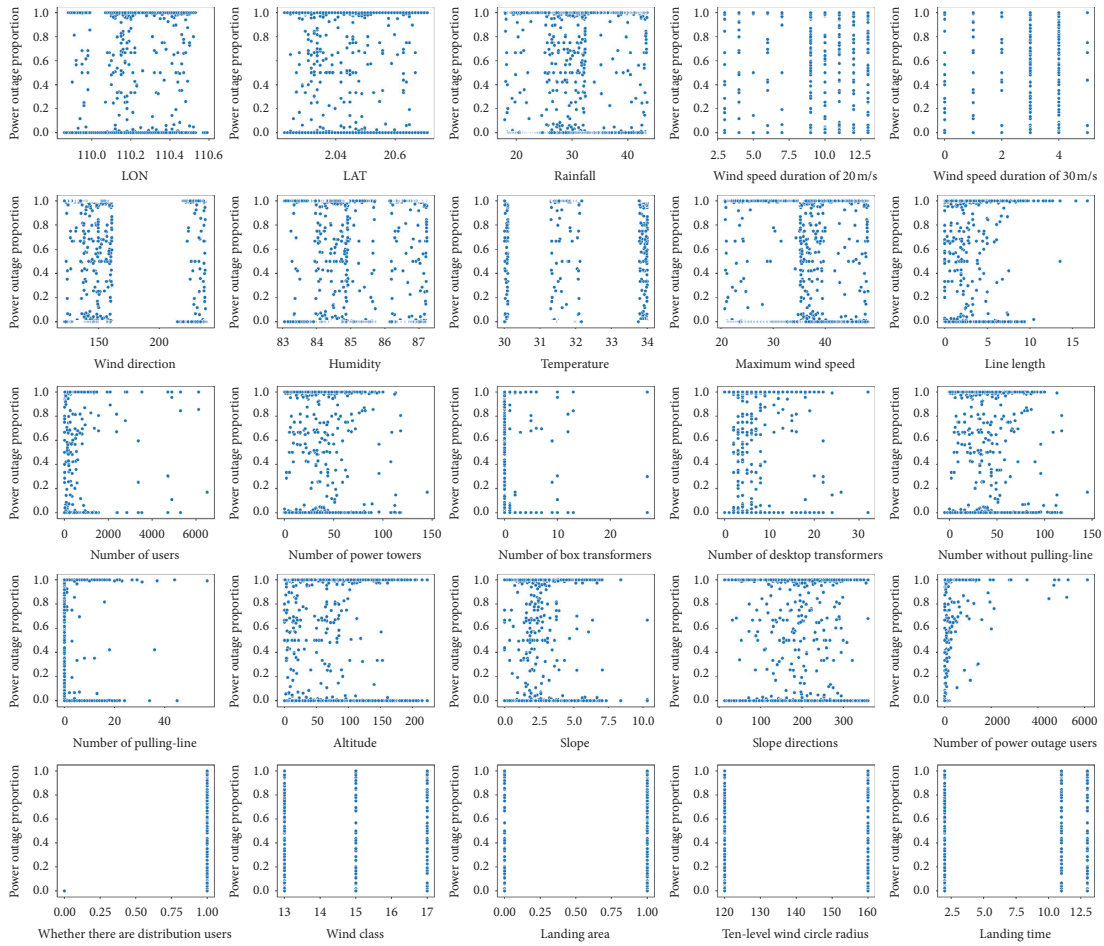

Figure 3: The scatter diagram between each explanatory variable and response variable.

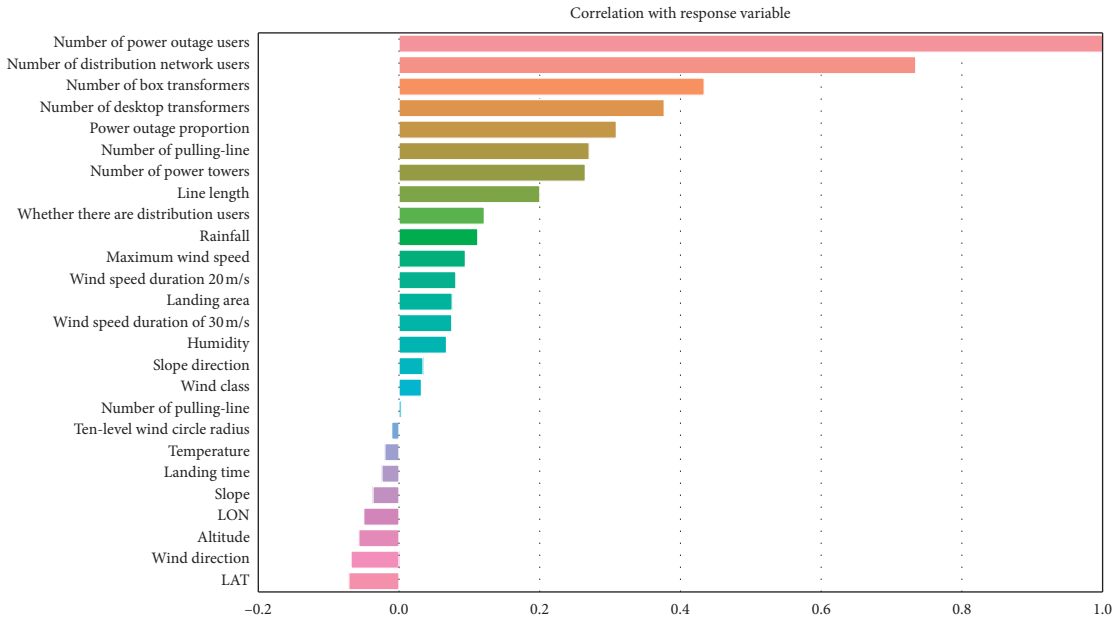

Figure 4: Correlation analysis of explanatory variables and response variables.

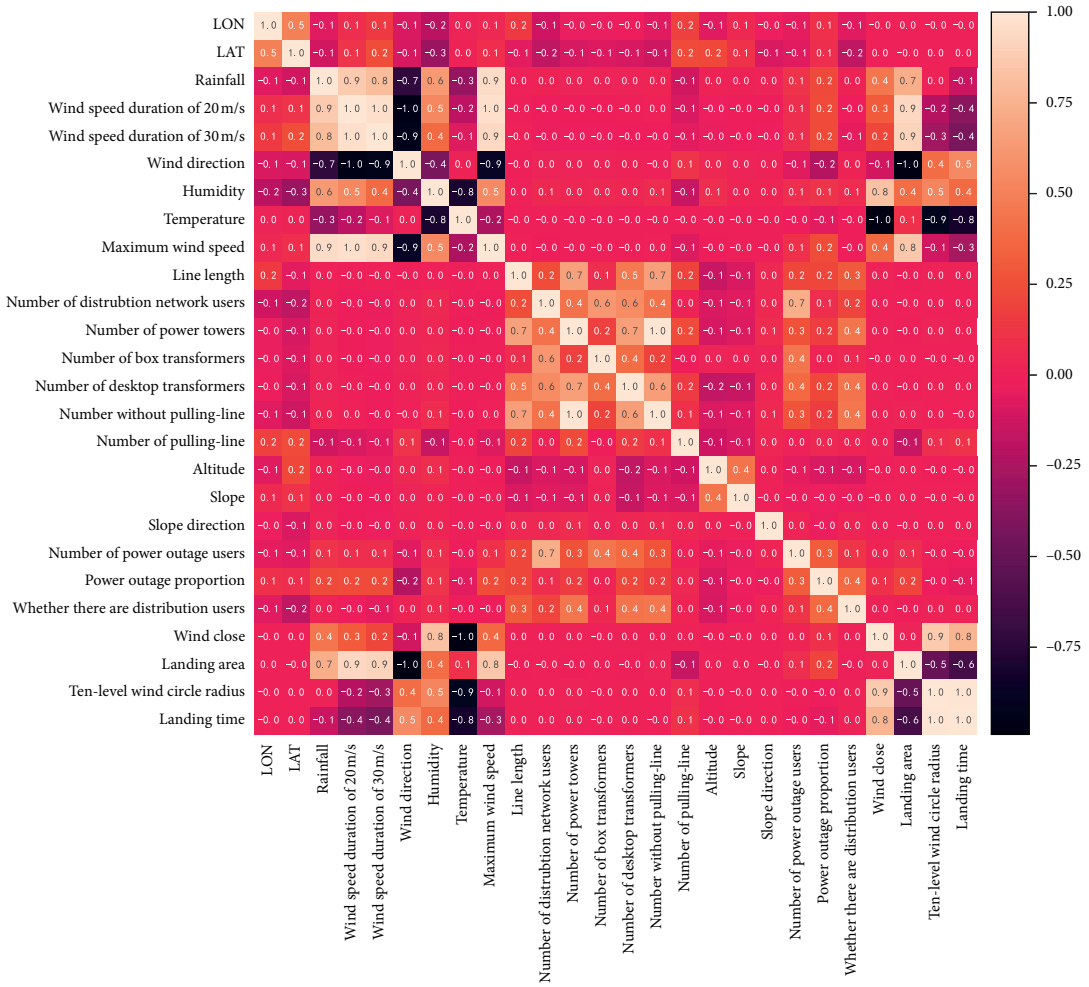

Figure 5: Correlation analysis among explanatory variables.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - f(x_i)\right|,$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2, \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2}.$$

In this paper, y_i represents the actual power outage quantity of distribution network users in i-th grid, and $f(x_i)$ represents the predicted power outage quantity of distribution network users in the i-th grid.

5. RF-Global Variable Modeling and Analysis

In order to as far as possible explore the potential relationship between each explanatory variable and response variable, the global variables (all explanatory variables) are used in this section, and the importance of variables is analyzed to identify the contribution of variables in the prediction model.

5.1. Analysis of Prediction Results. Firstly, 80% samples are randomly selected from the sample data for model training, and the remaining 20% samples are conducted for model test. Then, it is recycled 100 times. At last, the average values of MAE, MSE, and RMSE are obtained, as shown in Table 3.

As shown in Table 3, the prediction model of the power outage quantity is constructed with the proportion of power outage as the response variable. The MAE, MSE, and RMSE in the test errors are up to 0.1497, 0.0613, and 0.2474, respectively. To intuitively reflect the prediction effect, new model evaluation indexes ±100/±200/±300 (if the deviation between the predicted quantity and the actual quantity is within 100/200/300, the prediction is considered accurate) and ±10%/±20%/±30% (if the proportion of the deviation between the predicted quantity and the actual quantity is within ±10%/±20%/±30%, the prediction is considered accurate) are added. The accuracy analysis of the power outage quantity prediction model is shown in Table 4.

As shown in Table 4, the accuracy rate of prediction error within ±100/±200/±300 is higher than 90%. However, considering the small number of users of distribution network in most actual grids, evaluating the model with a fixed error may overestimate the predictive effect of the model. Therefore, the evaluation index ±10%/±20%/±30% based on floating error is constructed, in which the accuracy of the error within ±10% is 0.7546, within ±20% is 0.8320, and within ±30% is 0.8660. As can be seen from Tables 3 and 4, the prediction method of power outage quantity of distribution network users based on RF proposed in this paper has better performance.

5.2. Assessment of Variable Importance. As many explanatory variables as possible were selected in the early stage of modeling. However, this may lead to a large workload of data collection and processing in the actual application of the model. In order to evaluate the contribution of each explanatory variable in the prediction model and reduce the pressure of data collection, the importance of explanatory variables is evaluated.

Table 3: Training and test error analysis.

Evaluation index	MAE	MSE	RMSE
Training set	0.0550	0.0083	0.0913
Test set	0.1497	0.0613	0.2474

Table 4: Model accuracy analysis.

Evaluation index	±100	±200	±300	±10%	±20%	±30%
Accuracy	0.9279	0.9706	0.9831	0.7546	0.8320	0.8660

In the RF model, the importance ranking is calculated based on the degree of chaos (Impurity/Gini coefficient). That is to say, the criterion to measure the importance of a feature is to see how much chaos the feature reduces in the process of building a random forest through the decision tree [25]. After synthesizing all the trees, the greater the average decrease is determined as the more important feature. But the problem is that when features are continuous or there are many categories of classification factors (High-cardinality category variables), the method of feature importance analysis mentioned above will increase the importance of these features. Thus, the Permutation Importance Measure is used in this paper to solve this problem. The specific method of variable importance evaluation based on RF is as follows:

(1) The original accuracy of test data or OOB (out of bag) data in random forest (such as the OOB data error, denoted as err_{OOB1}) is taken as an accuracy baseline.

(2) One of the features that need to be measured is permuted; that is, scrambling the data and rearranging them. Then run the model again with the test data (the same data set) to calculate the new accuracy rate, denoted as err_{OOB2}.

(3) Calculate the difference between the new accuracy and the baseline accuracy. The larger the difference, the more important the feature is. Assuming that there are n trees in RF, the importance of the characteristic is $1/n\sum(err_{OOB2} - err_{OOB1})$.

In this process, the data do not need to be standardized, and the final importance ranking is not 1 but a relative ranking.

The importance analysis diagram of global variables is shown in Figure 6.

As can be seen from Figure 6, the explanatory variables such as longitude, latitude, maximum wind speed, wind direction, rainfall, number of users of distribution network, line length, and altitude contribute greatly to the accuracy of the prediction model. However, the explanatory variables such as landing time, landing area (whether landing in the

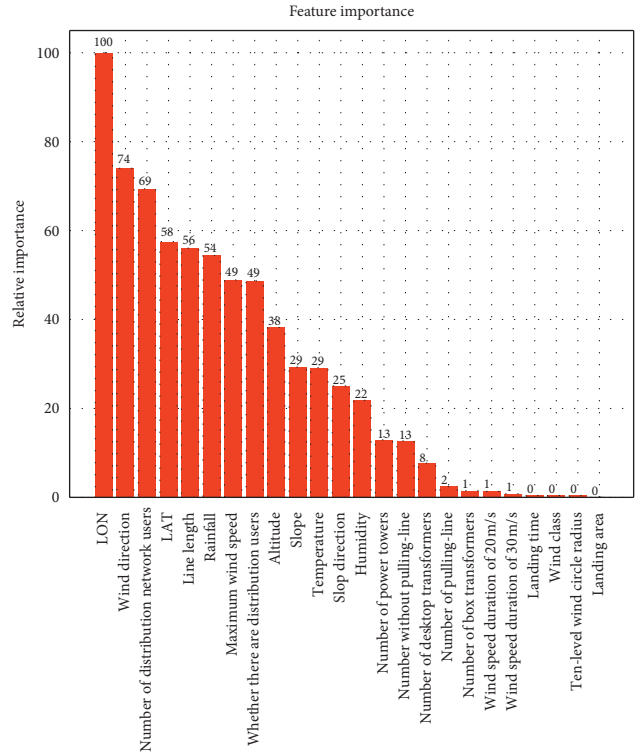

Figure 6: The importance analysis diagram of global variables.

research area), and wind level contributed little to the accuracy of the prediction model. Therefore, this paper focuses on the analysis of the variables that contribute a lot to the prediction model, and analyzes their impact on the power outage quantity of distribution network users. On this basis, the variation in the accuracy of the RF-important variable model and the RF-global variable model is analyzed.

5.3. Variable Dependency Analysis. The classical Partial Dependence Plots (PDP) [26] help visualize the average relationship between the response variable and one or more of the characteristics. When a specified characteristic changes in its marginal distribution, the PDP plots change in the average predicted value. With the help of the PDP, the trained supervised learning model can be better understood.

In order to formally define the PDP, let S ⊂ {1,..., p}, C be the complement of S, and $S \cup C = m$. And m is the set of all characteristics. Then, the partially dependent function f of the partial characteristics set x_S is as follows:

$$f_S = \mathrm{E}_{xc}\left[f\left(x_S, x_C\right)\right] = \int f\left(x_S, x_C\right) \mathrm{dP}\left(x_C\right). \tag{4}$$

Since f and $\mathrm{dP}(x_C)$ are unknown, equation (4) can be estimated by the following equation:

$$\widehat{f}_S = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}\left(x_S, x_{Ci}\right), \tag{5}$$

where, n is the number of samples of the training set, $\{x_{C1}, \ldots, x_{Cn}\}$ represents different values of the characteristic set x_C of the training set. When the characteristic set x_S contains only one characteristic variable x_j, $j = 1, 2, \ldots, m$, the partial dependency function of x_j is:

$$\widehat{f}_j\left(x_j\right) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_j\left(x_j, x_{-j,i}\right), \tag{6}$$

where, the PDP value $\widehat{f}_j(x_j)$ of the characteristic variable x_j represents the average value of the output value of the regression prediction function when x_j is fixed and changes along its marginal distribution.

To analyze the impact of the characteristics of variables on the response variable, this paper analyses the nine most important explanatory variables for modeling (longitude X_{18}, latitude X_{19}, number of distribution network users X_{20}, maximum wind speed X_1, rainfall X_3, line length X_{26}, whether there are distribution users X_{12}, wind direction X_2, and altitude X_{13}.) based on variable importance analysis. The partial dependency is shown in Figure 7.

It can be seen from Figure 7 that the longitude and latitude have a positive influence on power outage of distribution network users; that is, the increase of longitude and latitude leads to an increase in its influence on distribution network users. The main reason may be that the region mentioned in this paper is a coastal region. The closer a region is to the sea, the stronger the typhoon attacks on its distribution network users, and the more serious the impact. However, the dependence of the model on the number of distribution network users is not obvious and the influence is relatively stable. Moreover, the greater the maximum wind speed and rainfall of a typhoon, the greater the impact of the typhoon on distribution network users. In the geographic

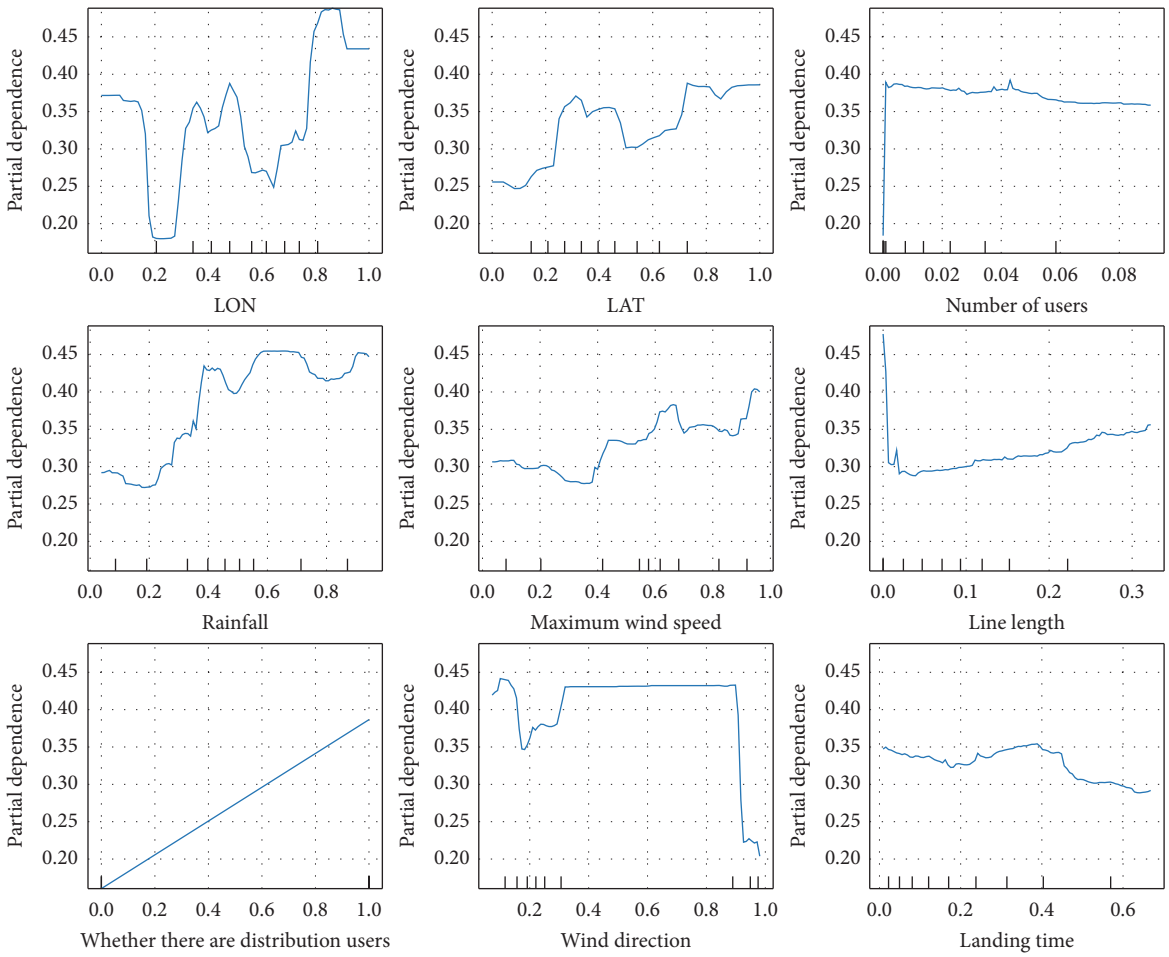

Figure 7: The PDP of explanatory variables.

information, the influence of altitude on power outage of distribution network users is negatively correlated; that is, the higher the altitude in the region, the smaller the influence on power outage of distribution network users, which is consistent with the influence trend of longitude and latitude. As for the line length, its influence is positively correlated with the increase of the line length. The longer the line length, the higher the probability of power outage of distribution network users will be. For classification variables with or without distribution network users, there is a relatively obvious positive correlation, because only if there are distribution network users in the grid, the distribution network users may have a power outage accident under the typhoon disaster. For the wind direction, there is no obvious correlation shown in the PDP chart. The main reason may be that the wind direction data changes rapidly and the model is not able to capture its performance characteristics. Besides, the wind direction is not a constant value under a typhoon disaster; it is difficult to select an appropriate quantitative description. Thus, we decided not to take it as one of the values in RF-important variable model.

Since longitude and latitude, wind speed and direction, wind speed and rainfall often occur simultaneously, the characteristic dependence of the two variables of these combinations is analyzed, as shown in Figures 8–10.

As shown in Figure 8, the combination of longitude and latitude can locate an area. When the longitude is large and the latitude is small, it has a greater impact on power outage of distribution network users. The region is located in the southeast corner of the study area, closer to the landfall area of the typhoon.

In general, high wind speed tends to bring rain and aggravate the impact on power distribution network users. As shown in Figure 9, the greater the wind speed and greater the rainfall, the greater the impact on power distribution network users.

As shown in Figure 10, there is no obvious correlation between wind direction and power outage of distribution

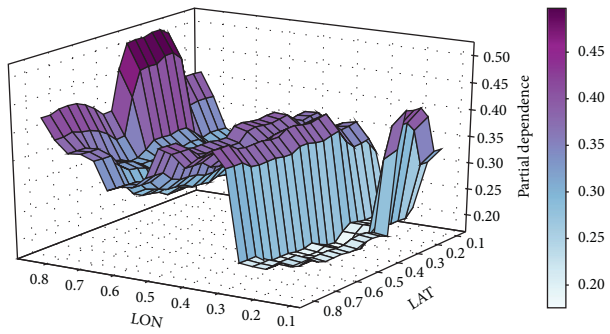

Figure 8: The combination of longitude and latitude.

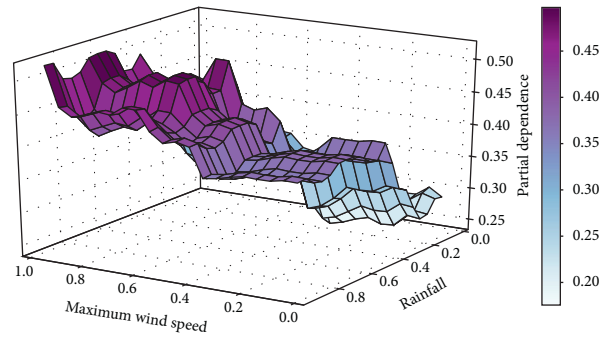

Figure 9: The PDP of max wind and rainfall.

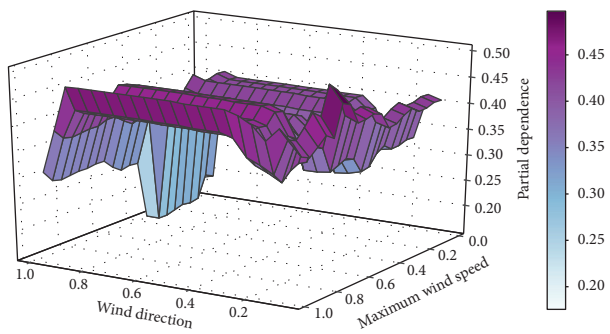

Figure 10: The PDP of max wind and wind direction.

network users. It shows that the wind direction has little influence on the power outage of users, so it can be removed. In addition, the higher the wind speed, the greater the probability of power outage for distribution network users.

6. Analysis of Modeling Important Variables

In Section 5, global variables are used for modeling, and the prediction results of power outage quantity of distribution network users are evaluated and analyzed. Based on historical data, more explanatory variables could be mined to support the accuracy of power outage quantity prediction. However, in reality, some explanatory variables are difficult to obtain, such as wind speed duration of 20 m/s and 30 m/s. In addition, many variables contribute little to prediction accuracy. Therefore, this section analyzes and compares the prediction accuracy of models considering global variables and important variables, so as to increase the efficiency and availability of the model.

6.1. Model Training Test Analysis. According to the analysis results of the above section, in this section, the eight explanatory variables that are most important to the predicted results are selected as explanatory variables to carry out the training of power outage quantity prediction model: longitude X_{18}, latitude X_{19}, maximum wind speed X_1, rainfall X_3, distribution network user X_{20}, line length X_{26}, whether there are distribution users X_{12}, and altitude X_{13}. For all the samples, 80% are randomly selected as the training set and the remaining 20% as the test set, with random recycling for 100 times. The error results of the training test are shown in Table 5. The accuracy of the change of the evaluation index is shown in Table 6.

It can be seen from Table 5 that the test set MAE is 0.1366, MSE is 0.0580, and RMSE is 0.2406 for modeling analysis with important characteristic variables, and the overall prediction effect is good. The prediction accuracy of the model calculated when changing the evaluation index is shown in Table 6.

Table 6 shows that eight important variables are used for prediction model training; the accuracy of 100/±200/±300 reaches 0.9346, 0.9706, 0.9852, and the accuracy of ±10%/±20%/±30% is 0.7582, 0.8345, and 0.8822, respectively. The prediction accuracy of the model is close to that of the RF-global variable model, indicating that building a prediction model with less important variables does not significantly reduce the accuracy of the model but makes the process of predicting and evaluating the power outage quantity simpler and faster (saving time for collecting and sorting out the remaining variables). Furthermore, it accelerates the assessment of the power outage quantity of distribution network users under the typhoon disaster and prepares the conditions for further emergency decision-making. It can be seen from Figure 11 that, except for a few points, the difference between the actual value and the predicted value of most points is around 0. It indicates that the fitting data of the user outage number prediction model of the distribution network is good.

6.2. Comparative Analysis of Models. In order to further analyze the model built based on important variables in this paper, a No-model [27] and three other machine learning algorithms are used to compare with the trained RF model based on global variables and important variables, as shown in Table 7. The average values of the samples are used as the prediction value in No-model, LR, SVR, and DTR. At the same time, in order to visually demonstrate the prediction effect of each model, a histogram of the error analysis of each model is shown in Figure 12.

Table 7 and Figure 12 show that the prediction model of power outage quantity of distribution network users based on RF in this paper has a better prediction effect. Whether based on global variables or import variables, its MAE, MSE, and RMSE are all smaller than that of the other three

Table 5: Partial variable training and testing errors.

Evaluation index	MAE	MSE	RMSE
Training set	0.0503	0.0080	0.0892
Test set	0.1366	0.0580	0.2406

Table 6: Model accuracy analysis.

Evaluation index	±100	±200	±300	±10%	±20%	±30%
Accuracy	0.9346	0.9706	0.9852	0.7582	0.8345	0.8822

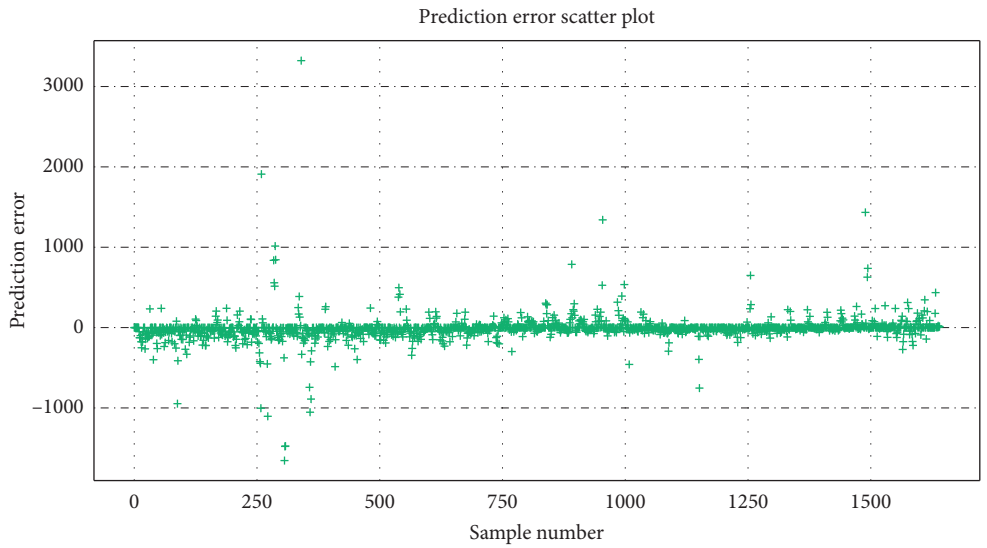

Figure 11: Scatter distribution of prediction error.

Table 7: Comparative analysis of six models.

Evaluation index	MAE	MSE	RMSE
No-model	0.436	0.215	0.464
LR	0.352	0.160	0.400
SVR	0.235	0.117	0.342
DTR	0.174	0.112	0.334
RF-global variables	0.150	0.061	0.247
RF-important variables	0.137	0.058	0.241

machine learning models and the No-model. Among them, the RF algorithm has the best effect, followed by DTR, SVR, LR, and the No-model. Compared with the No-model, MAE of the RF-global variable model decreased by 66%, MSE by 72%, and RMSE by 47% on average, which shows the effectiveness of the RF-global variable model trained in this paper. And, compared with the RF-global variable model, MAE of RF-important variable model (only eight variables are considered) was reduced by 8.8%, MSE by 18.4%, and RMSE by 2.7% on average, showing that the error based on the important variables is smaller than that based on the global variables. The main reason is that during the RF-global variable modeling, more explanatory variables with strong correlation are introduced, leading to certain deviation of the trained prediction model.

To sum up, the prediction model of power outage quantity based on RF has a good effect. The errors of both RF-global variable model and RF-important variable model are lower than that of No-model, LR, SVR, and DTR. Meanwhile, the effect of RF-global variable model is close to that of RF-important variable model, and the prediction effect of RF-important variable model is better. Moreover, it takes less time to collect and sort out the original data of important variables. This improves the efficiency of

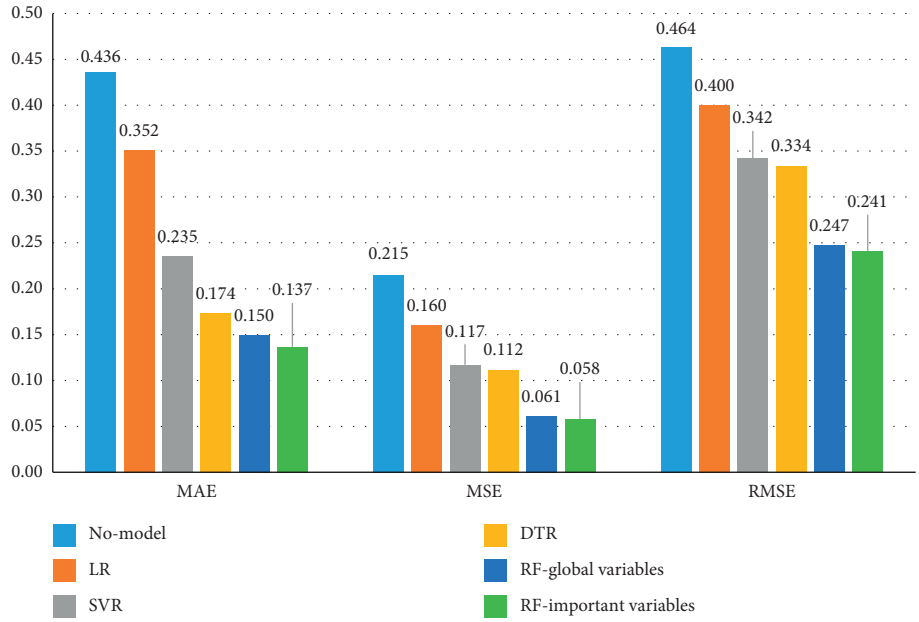

Figure 12: Error analysis of each model.

prediction and evaluation, and provides an effective basis for the early allocation of emergency repair resources, the reduction of power outage loss, and the improvement of distribution network user satisfaction.

7. Conclusion

In this paper, the prediction and evaluation method of power outage quantity of distribution network users under typhoon disaster is studied, and the prediction model of power outage quantity of distribution network users based on RF is proposed.

(1) In order to make the evaluation process more convenient, this paper selects the eight most important explanatory variables for model training. The results show that the model errors do not increase seriously but decrease slightly, providing auxiliary guidance for rapid prediction.

(2) The prediction and evaluation with the important variable model based on RF reduces the time spent collecting and processing other variables and improves the prediction efficiency of the power outage quantity of distribution network users.

(3) Compared with the No-model, LR, SVR, and DTR, it is found that the RF-global variable model and RF-important variable model trained in this paper are better, and their MAE, MSE, and RMSE are significantly reduced. And the prediction effect of the RF-important variable model is slightly better than that of the RF-global variable model, which can provide an effective basis for disaster prevention and reduction of power grid.

(4) In the actual application process, the predicted maximum gust wind speed of 72 hours, 48 hours, and 24 hours before typhoon landing can be used as model inputs, respectively. The prediction results can provide some guidance for the formulation of pre disaster emergency dispatching strategy.

Data Availability

The datasets used or analyzed during the current study are available from the authors upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the University-Industry Collaborative Education Program of the Ministry of Education (grant no. 201902056044) and Guangdong Power GRID Co., Ltd., Electric Power Research Institute (grant no. GDKJXM20198382).

References

[1] H. Hui, G. Hao, X. Xiang et al., "Prediction and evaluation of outage area of distribution network users under Typhoon Disaster," *Power System Technology*, vol. 43, no. 06, pp. 1948–1954, 2019.

[2] L. Xi, J. Wu, Y. Xu, and H. Sun, "Automatic generation control based on multiple neural networks with actor-critic strategy," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 99, pp. 1–11, 2020.

[3] L. Xi, L. Yu, Y. Xu, S. Wang, and X. Chen, "A novel multi-agent DDQN-AD method-based distributed strategy for automatic generation control of integrated energy systems," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, pp. 2417–2426, 2020.

[4] H. Liu, R. A. Davidson, and T. V. Apanasovich, "Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms," *Reliability Engineering & System Safety*, vol. 93, no. 6, pp. 897–912, 2008.

[5] S. M. Quiring, L. Zhu, and S. D. Guikema, "Importance of soil and elevation characteristics for modeling hurricane-induced power outages," *Natural Hazards*, vol. 58, no. 1, pp. 365–390, 2011.

[6] H. Liu, R. A. Davidson, and T. V. Apanasovich, "Statistical forecasting of electric power restoration times in hurricanes and ice storms," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 2270–2279, 2007.

[7] H. Hui, Y. Shiwen, W. Hongbin et al., "Risk assessment and its visualization of power tower under typhoon disaster based on machine learning algorithms," *Energies*, vol. 12, no. 2, pp. 1–23, 2019.

[8] T. B. Gutwin, "Application of BCTC standardized risk estimation model to assess risk due to ice storms," in *Proceedings of the 8th International Conference On Probabilistic Methods Applied To Power Systems*, pp. 970–974, Ames, lowa, USA, September 2004.

[9] S. D. Guikema, R. Nateghi, S. M. Quiring, A. Staid, A. C. Reilly, and M. Gao, "Predicting hurricane power outages to support storm response planning," *IEEE Access*, vol. 2, pp. 1364–1373, 2014.

[10] S. D. Guikema and S. M. Quiring, "Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data," *Reliability Engineering & System Safety*, vol. 99, pp. 178–182, 2012.

[11] D. T. Radmer, P. A. Kuntz, R. D. Christie, S. S. Venkata, and R. H. Fletcher, "Predicting vegetation-related failure rates for overhead distribution feeders," *IEEE Transactions on Power Delivery*, vol. 17, no. 4, pp. 1170–1175, 2002.

[12] Z. Wen, S. Wanxing, L. Kexue et al., "Prediction method of distribution network fault risk level considering weather factors," *Power System Technology*, vol. 42, no. 08, pp. 2391–2398, 2018.

[13] Z. Qingqing, Y. Zheng, J. Yanbing et al., "Reliability prediction of transmission line operation," *Automation of Electric Power Systems*, vol. 34, no. 24, pp. 18–22, 2010.

[14] Z. Yongjun, C. Chao, and X. Liang, "Estimation of original parameters of power system reliability based on fuzzy clustering and similarity," *Power System Protection and Control*, vol. 39, no. 8, pp. 1–5, 2011.

[15] C. Bin, S. Shengwen, H. Haikun et al., "Research progress on early warning technology of transmission and distribution lines against strong typhoon in coastal areas," *High Voltage Appliances*, vol. 54, no. 07, pp. 64–72, 2018.

[16] X. Juqin, Z. Linyao, W. Guilian et al., "Multilevel grid division of urban distribution network based on hierarchical spatial reasoning," *Science and Technology Bulletin*, vol. 36, no. 02, pp. 55–58, 2020.

[17] L. Haibin, R. A. Davidsona, D. V. Rosowsky et al., "Negative binomial regression of electric power outages in hurricanes," *Journal of Infrastructure Systems*, vol. 11, no. 4, pp. 258–267, 2005.

[18] D. W. Wanik, E. N. Anagnostou, B. M. Hartman, M. E. B. Frediani, and M. Astitha, "Storm outage modeling for an electric distribution network in Northeastern USA," *Natural Hazards*, vol. 79, no. 2, pp. 1359–1384, 2015.

[19] Z. Linming, S. Shengwen, C. Bin et al., "Prediction method of 10 kV tower damage based on lattice and support vector machine in strong typhoon environment," *High Voltage Technology*, vol. 46, no. 01, pp. 42–51, 2020.

[20] H. Hui, Y. Jufang, G. Hao et al., "Data-driven prediction for the number of distribution network users experiencing typhoon power outages," *IET Generation, Transmission & Distribution*, vol. 14, no. 24, pp. 5844–5850, 2020.

[21] Electric Power Research Institute, *Analysis Report of 'Rammasun*, Guangdong Power Grid Co., Ltd., Guangdong, China, 2014.

[22] Electric Power Research Institute, *Analysis Report of 'Kalmaegi*, Guangdong Power Grid Co., Ltd., Guangdong, China, 2014.

[23] Electric Power Research Institute, *Analysis Report of 'Mujigae*, Guangdong Power Grid Co., Ltd., Guangdong, China, 2015.

[24] Z. Hailong, Z. Dandan, H. song et al., "Analysis of the relationship between lightning flash density and lightning stroke fault in Hainan Province Based on Pearson correlation coefficient," *High Voltage Appliances*, vol. 55, no. 08, pp. 186–192, 2019.

[25] C. Bihan, *Research on Variable Selection in Data Mining*, Huazhong University of science and technology, Wuhan, China, 2018.

[26] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[27] R. Nateghi, S. D. Guikema, and S. M. Quiring, "Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes," *Risk Analysis*, vol. 31, no. 12, pp. 1897–1906, 2011.