doi:10.6041/j.issn.1000-1298.2016.10.043

Landslide Spatial Prediction Based on Random Forest Model

Yu Kunyong[1,2] Yao Xiong[1,2] Qiu Qirong[3] Liu Jian[1,2]

(1. *University Key Laboratory for Geomatics Technology and Optimize Resources Utilization in Fujian Province*, *Fuzhou* 350002, *China*
2. *College of Forestry*, *Fujian Agriculture and Forestry University*, *Fuzhou* 350002, *China*
3. *School of Forestry and Resource Conservation*, *National Taiwan University*, *Taibei* 10617, *China*)

Abstract: Random forest (RF) is a non-parametric technology which was firstly proposed by Leo Breiman and Cutler Adele in 2001. It was used to deal with the classification and regression problems by gathering a large number of classification tree, which can improve the prediction accuracy. It was applied in the ecological field in recent years. Predicting the spatial distribution of landslide hazard was an important way to achieve disaster prevention and mitigation. The landslide dataset of Shunchang in Fujian province was taken as case to identify the relationship between mountain landslide occurrence and landslide factors by using RF model and logistic regression (LR) model respectively with landform, meteorological hydrology, soil and vegetation factors. The applicability of RF on landslide prediction in the southern mountain of China was tested by procedure of parameter selection and analysis of model accuracy. The result showed that the goodness of fit of RF was better than that of LR model. The prediction accuracy of RF on the landslide data was 90.8%, while the prediction accuracy of LR was 81.8%. The generalization of RF in the study area was better than that of LR model. The high risk areas and higher risk areas contained 66.05% of the total landslide, which was predicted by RF, while that of LR was 63.34%. The result of model comparison revealed that the RF model was superior to LR model on the mountain landslide prediction in the study area, thus it can be used in the landslide prediction and the division of landslide danger grade with the sample data. In addition, RF model could be applied to other relevant research.

Key words: landslide; random forest model; logistic regression model; spatial prediction

0 Introduction

Landslide is one of the most serious geological hazards in mountain area, and it has become a natural disaster problem which can not be ignored in the mountainous area[1-2]. Landslide is widely distributed in the mountainous rural areas of China, causing an average of nearly one thousand deaths each year. Landslide has seriously restricted the social and economic development of China[3]. Due to the global climate change, the frequent occurrence of extreme disasters and the increase of human economic activities, the occurrence of landslide would be more serious [4-5]. Based on this situation, researches on the relationship between the mountain landslide occurrence and landslide factors, thus realizing the accurate prediction of regional landslide, are of great significance for prediction of the damage degree caused by landslide[6].

Some researchers have studied the spatial prediction of landslides mainly with the traditional logistic regression model[7-9]. With the development of artificial intelligence, machine learning models are more widely used in the research of spatial prediction of landslides[1,10]. In 2001, BREIMAN and other researchers[11] proposed a new model of machine learning—random forest model with high learning ability and prediction accuracy. A few application cases on the spatial prediction of landslides with the random forest model have appeared in other countries in recent years[12-13], but few related applications is reported in China. So far, the random forest model is mainly used in the medicine, economics, ecology and other fields[14-17]. Due to the spatial heterogeneity of

Received date: 2016-03-29 Accepted date: 2016-05-12

Supported by National Natural Science Foundation of China (Grant No. 41401385), Fund Project of Fujian Province Education Department (Grant No. JA14126) and Youth Fund Project of Fujian Agriculture And Forestry University Forestry College

Corresponding author: Yu Kunyong, Associate Professor. E-mail: yuyky@126.com. Tel: +86-13313979361.

the study area, the conclusion that the superiority of random forest model in prediction of spatial prediction of random forest in the mountainous region is still to be discussed. Shunchang county in Fujian province was selected as the subject. The research on spatial prediction of landslides in the study area was carried out by random forest model and logistic regression model, and the overall performance of the two models in the study area was discussed. Based on the analysis of model fitting results, the adaptability of random forest model to the prediction of landslide in Shunchang county was analyzed, so as to provide an important evidence for further research and decision-making.

1 Materials and methods

1.1 General situation of study area

Shunchang is located in the northwest of Fujian province (117°29′ ~118°14′E, 26°38′ ~27°12′N), which is an important forest district in the south of China and a bamboo production base in Fujian province. The climate type of the study area is mid-subtropical maritime monsoon climate with moderate temperature and abundant precipitation. Average annual rainfall is approximately 2 051 mm and annual average temperature is nearly 19.1℃. Shunchang is hilly landform, the terrain is from the northeast, northwest and southwest to the middle of the tilt, the main type of soil is red soil. Total area of the county is approximately 2 000 km^2 and there are 11 villages and 3 towns, with a total of 4 neighborhood committees and 129 village committees.

1.2 Data sources

Data for research include: ①The data of spatial distribution of 1 478 landslide occurrences in the study area provided by Fujian Science and Technology Key Project (Item number 2012N0003). ② The ALOS multi-spectral remote sensing image data with spatial resolution of 10-meter level on November 8, 2010 and November 25, 2010 of Shunchang (Fig. 1). ③The digital elevation model (DEM) of Shunchang with resolution of 30-meter level provided by Geographic Data Cloud (http://www.gscloud.cn/) (Fig. 2). ④Rainfall data of the study area provided by the Shunchang Meteorological Bureau in June, 2010.

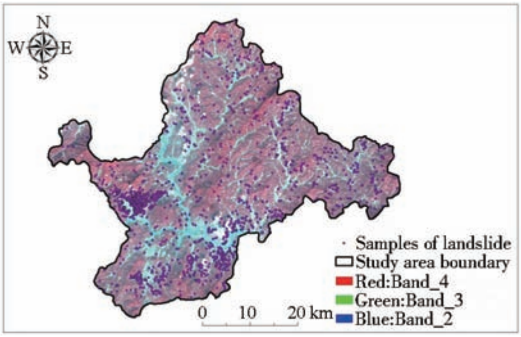


Fig. 1 Source imagine and landslides distribution map of study area

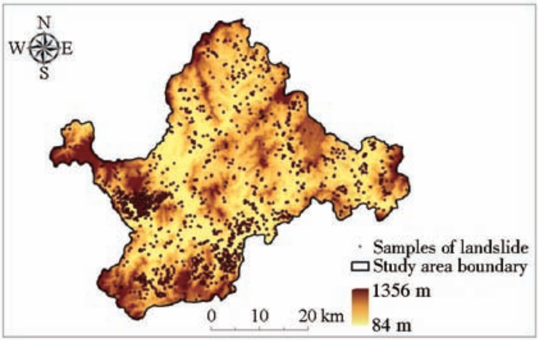


Fig. 2 DEM and landslides distribution map of study area

1.3 Extraction of landslide factor

Landslide is the product of many kinds of factors in nature, and each factor often has a causal relationship with each other. The landslide is influenced by the geographical environment of the area, including topography, soil environment and plant information. Generally, landslide disasters are not easy to occur in the smooth terrain. In the area of undulating terrain, the earthquakes, rainfall and other events would lead to landslides easily. In addition, the probability of landslide is increased with the increase of slope and elevation. The slope direction affects plant cover and rainfall path firstly, and then affects the stability of slope. The vegetation root system can strengthen the integrity of slope, and rainfall infiltration and slope sliding can be prevented effectively by the vegetation on surface of slope. Therefore, the plant coverage is a significant factor of landslide occurrence. Due to different levels of porosity of different types of soil, soil environment could directly affect the stability of slope. Precipitation is an important external factor that leads to the sliding of slope. The water pressure caused decrease of slope sliding resistance[18]. In addition, the closer the soil near water system, the greater content of water is in the soil. Higher water content always leads to landslides, therefore, the distance to the river system is a potential factor of landslides.

In sum, under the same individual conditions, based on existing research, with comprehensive consideration of the natural geographical conditions of the study area, from the topography (including the aspect of slope, elevation, topography, plane curvature, profile curvature and slope gradient), meteorological and hydrological factors (including rainfall on the day of landslide, rainfall before 1 d of landslide, rainfall before 2 d of landslide, distance to water), and soil vegetation (NDVI, soil type), totally 12 factors are selected in 3 aspects (Fig. 1). The thematic information of landform and geomorphology is extracted from DEM data. The rainfall of some days before the landslide is mainly obtained through the inverse distance weighted interpolation method in ArcGIS 9.3 software[18]. The distance data of water system is used to calculate the Euclidean distance of the study area to the river with ArcGIS 9.3 spatial module. NDVI information was obtained through ERDAS 9.2 modeling tool.

Tab. 1 Source of landslide factors

Types	Variables	Sources	Codes
Topographic features	Aspect	DEM	ASP
	Altitude	DEM	ELE
	Topographic relief	DEM	DXD
	Plan curvature	DEM	PLC
	Profile curvature	DEM	PRC
	Slope	DEM	SLO
Meteorology and hydrology	Rainfall on the day of landslide	The Meteorological Bureau of Shunchang	RA0
	Rainfall before 1 d of landslide	The Meteorological Bureau of Shunchang	RA1
	Rainfall before 2 d of landslide	The Meteorological Bureau of Shunchang	RA2
	Distance to water	Topographic map	DSX
Soil vegetation	NDVI	Image data of ALOS	NDVI
	Soil type	1:250 000 soil type data in Fujian province	TRLX

1.4 Model introduction

Logistic regression is a kind of multiple statistical analysis model, it has been widely used in the prediction of landslide space. The main idea is to use the maximum likelihood to construct the relationship between the predictor variables and results of two-category, to ensure every point is optimum fitting. Supposing the probability of landslide occurrence is P, then the probability of nonoccurrence landslide is $(1-P)$. The regression between probability of landslide occurrence and independent variables (landslide influence factors) is[6]

$$P = \frac{1}{1+\exp(-(\beta_0+\beta_1 X_1+\cdots+\beta_m X_m))} \quad (1)$$

where P is probability of landslide; β_i is logistic regression coefficients based on training samples, $i=0,1,2,\cdots,m$; X_i is independent variables, $i=0,1,2,\cdots,m$.

Random forest (RF) model is an algorithm based on classification and regression tree. The main idea of random forest model is to extract k sample from the original training samples by bootstrap sampling, the size of each sample is consistent with the original training. And then the samples of each decision tree are modeled respectively to get k modeling results. Finally, the final classification results are determined by voting with the modeling results of all decision trees[14].

Compared with traditional methods of landslide prediction (logistic regression method, support vector machine, etc.), random forest model is needless to check whether the interaction of variables is significant or not, as got two random sampling, it has a high degree of tolerance in outlier and noise. Moreover, it is uneasy to appear over fitting and has accurate prediction[19].

1.5 Model establishment

Selecting training samples data for model establishment before the landslide prediction by using the quantitative model. In previous studies, the training samples data usually considered the positive sample (landslide point data) only, and the influence of negative samples (non-landslide point data) on the prediction of landslide was ignored. The results from the model of training samples data can not reflect the mechanism of landslide poperly. To this end, the ArcGIS 9.3 software is used to randomly generate the same number of points as non-landslide points in the known landslide points over 100 m. Then, the quality value of each landslide factor is extracted as positive samples and negative samples, and the training samples are selected from total samples (composed of positive and negative samples).

In order to reduce the impact of a single sampling method on the model establishment, the total samples data (1 478 positive samples and 1 478 negative samples) were randomly divided into two parts: 60% of sample data as training sample and 40% of the sample data as test sample, repeating 5 times of random division can get 5 groups of different samples. Then the R statistical software and SPSS software (Tests 1 ~5) were used to calculate the random forest and logistic regression model for the 5 groups of different samples. Finally, the two models of the significant variables in the 5 trials and the number of times greater than or equal to 3 times were used as the fundamental to determine final variables for calculating the full sample model (Test 6) respectively.

1.6 Model evaluation

The results of logistic regression model are evaluated by receiver operating characteristic curve (ROC) of area under the curve (AUC). Value of AUC is between 0.5 and 1, as it is close to 1, the forecast effect is getting better, and the model is the best one when it is 1. In addition, the sensitivity and specificity were calculated on the basis of ROC curve analysis method that the Youden index can be obtained. The Youden index is equal to sensitivity plus specificity and minus 1, and then the best critical value can be determined. According to the critical value, it can be used to judge the occurrence of landslide, if the probability of landslide is greater than critical value, landslide occurs, but if it is less than critical value, landslide does not occur[20].

The effect of random forest model was evaluated by Kappa coefficient, and the specific calculation formulas are found in the literature[21].

In order to accurately evaluate the overall performance of the two models, the spatial distribution of landslide is obtained by spatial interpolation probability of landslide and non landslide points. It is pointed out in some researches[22] that favorable prediction model of landslide space should satisfy two criteria: ①Landslide hazards should be located in the high risk area of landslide as much as possible. ②The high risk area of landslide should be as small as possible in forecast map. According to these, the risk of landslide occurrence is divided into 4 grades: lower (0 ~0.25), low (0.25 ~0.5), high (0.5 ~0.75) and higher (0.75 ~1). According to the classification of the danger grade map, the landslide ratio in each grade area is calculated, and combined with the area of each risk level to evaluat the generalization ability of the model.

2 Results and analysis

2.1 Fitting result analysis of logistic regression model

2.1.1 Diagnosis of multi-collinearity diagnosis

Because of the high correlation between independent variables in the linear regression model, it may lead to the failure of model's prediction function or increase difficult to estimate the result accurately without elimination of its correlation. Therefore, multi-collinearity diagnosing of independent variables and eliminating the significant variables are necessary before the model operation. VIF and TOL are wildly used as common diagnostic indicators. These two indicators are the inverse of each other. Generally speaking, when VIF is higher than 10 (i. e. TOLis less than 0.1) , the multi-collinearity of selected variables are more serious. Tab. 2 gives the factors affecting landslide occurrence figured out by SPSS 21.0.

Tab. 2 Multi-collinearity diagnosis indexes for variables

Variables	TOL	VIF
RA0	0.600	1.667
RA1	0.207	4.839
RA2	0.076	13.170
ASP	0.978	1.022
ELE	0.581	1.720
DSX	0.047	21.375
DXD	0.041	24.390
NDVI	0.767	1.304
PLC	0.785	1.275
PRC	0.772	1.296
SLO	0.465	2.151
TRLX	0.089	11.207

Tab. 2 shows that VIF of "Rainfall before 2 d of landslide", "Distance to water", "Topographic relief" and "soil type" exceeded 10, so these four variables would be removed. Finally, eight variables of "Rainfall on the day of landslide", "Rainfall before 1 d of landslide", "Aspect", "Altitude", "NDVI", "Plan curvature", "Profile curvature" and "slope" would enter fitting stage of the model.

2.1.2 Analysis of model fitting results

The landslide occurrence in the study areawas researched and corresponding factors after collinearity were calculated by logistic regression model. Firstly, totally 5 training samples were fitted to the model and 5 different notable factors are obtained after 5 experiments. Then the factors which appeared three or more times in the experiments would be chosen as whole sample data to fit the model (Test 6). The notable factors in each test are shown in Tab. 3.

Tab. 3 Significant factors in logistic regression model for each experiment data

Factors	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
RA0	Y	Y	Y	Y	Y	Y
RA1	Y	N	N	Y	N	N
ASP	N	Y	Y	N	Y	Y
ELE	N	Y	Y	Y	Y	Y
NDVI	Y	Y	Y	Y	Y	Y
PLC	Y	Y	N	Y	Y	Y
PRC	N	N	Y	N	Y	N
SLO	N	N	Y	N	N	N

Note: Y indicates that the test is significant, and N indicates that the test is not significant; the same below.

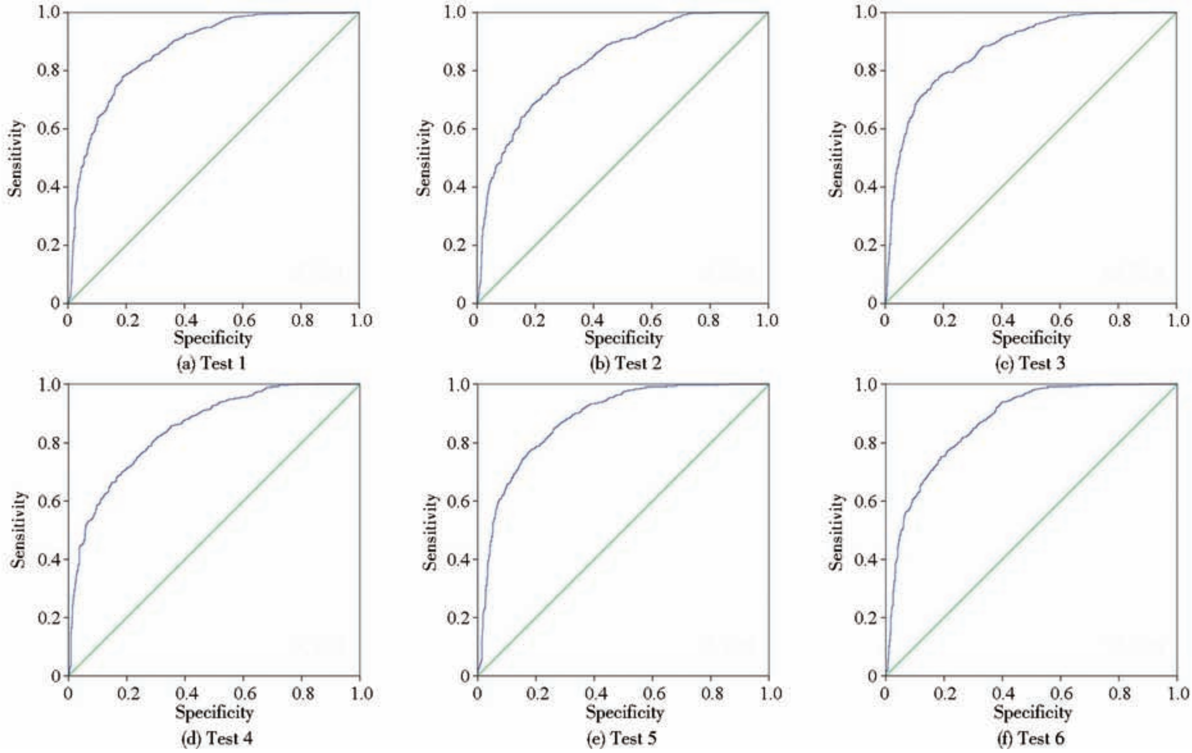
The full sample logic regression fitting results of Test 6 showed that Cox&Snell R^2 of model is 0.542, Nagelkerke R^2 is 0.723, which indicates the overall effectiveness of the model is well. It can be known from the model parameter fitting results (Tab. 4) that final model variables are significantly correlated with landslides at $P < 0.01$ level.

Tab. 4 Fitting results of logistic regression model

Independent variables	Regression coefficient	Standard error	Wals value	Degree of freedom	Significance level
RA0	8.001	0.022	17.021	1	<0.01
ASP	-0.007	0.001	29.606	1	<0.01
ELE	0.003	0.001	22.746	1	<0.01
NDVI	-12.030	0.086	19.361	1	<0.01
PLC	-1.286	0.085	18.180	1	<0.01
constants	4.843	0.061	13.732	1	<0.01

2.1.3 Model checking

ROC curves of each test are shown in Fig. 3 based on SPSS software, the AUC value, critical value and prediction rate of each test model are presented in Tab. 5. The AUC values of tests 1 ~ 6 are 0.872, 0.830, 0.876, 0.849, 0.881 and 0.873, respectively, the



(a) Test 1 (b) Test 2 (c) Test 3

(d) Test 4 (e) Test 5 (f) Test 6

Fig. 3 ROC curves of logistic regression model

results show that the logistic regression model established is effective and can be used in the prediction of landslides. In addition, the prediction rate of each test group is 81. 7% ~82. 8%. The results are calculated by establishment of model and combined the optimal critical value with the Youden index obtained from the prediction rate of each test group.

Tab. 5 Testing results of logistic regression model

Parameters	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
AUC value	0. 872	0. 830	0. 876	0. 849	0. 881	0. 873
Critical value	0. 818	0. 814	0. 816	0. 820	0. 832	0. 823
Prediction rate/%	81. 7	82. 1	82. 0	82. 8	82. 0	81. 8

Note: prediction rate is calculated from 40% of the test samples, the same below.

2. 2 Fitting result analysis of random forest model

2. 2. 1 Selection of model characteristic variables

Random forest models can be used to select the characteristic variables, which is based on the minimum error of out bag data. The landslide occurrence random forest model was studied and the corresponding landslide factors are calculated. Firstly, the model characteristic variables of 5 training samples were calculated though the program package varSelRF in R statistical software, and the different significant factors were obtained from 5 tests. The factors which appeared three or more times in 5 tests would be chosen as whole sample data and fitted model similarly.

Tab. 6 Significant factors in random forest model

Factors	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
RA0	Y	Y	Y	Y	Y	Y
RA1	Y	Y	N	Y	N	Y
RA2	N	N	N	N	N	N
ASP	Y	N	N	N	Y	Y
ELE	N	Y	Y	Y	N	Y
DSX	N	N	N	N	Y	N
DXD	N	Y	N	N	N	N
NDVI	Y	Y	Y	Y	Y	Y
PLC	Y	Y	Y	Y	N	Y
PRC	N	N	N	N	N	N
SLO	N	N	Y	Y	N	N
TRLX	N	Y	N	N	N	N

2. 2. 2 Importance ranking of model characteristic variables

The random forest model can give the importance of landslide factors by mean decrease accuracy, which through disrupting the value of a landslide factor, and then compared the reduction degree of the random forest prediction accuracy rate with disruption before. The importance of factors was increased with the increase of reduction degree. After selection of characteristic variables by random forest model, the importance of the landslide factors can be got by fitting training of random forest model through 6 test group data in experiment (Fig. 4). From the whole sample model (Fig. 4f), the most important factor that affectes the landslide in the study area was NDVI, and secondly israinfall on the day of landslide (RA0). Rainfall before 1 d of landslide (RA1) has the lightest influence. The effects of NDVI and RA0 on the occurrence of landslides are higher than those of other variables from the 6 fitting test results.

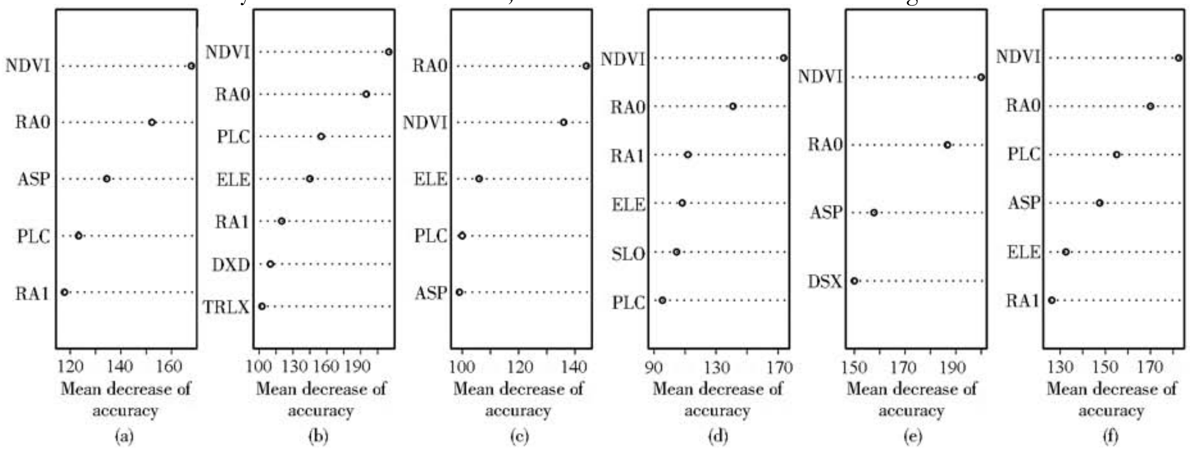


Fig. 4 Importance sorting of landslide factors in random forest model

2.2.3 Model checking

Tab. 7 presentes the Kappa coefficients and prediction rates for each test model in the random forest algorithm. The results indicates that the Kappa system coefficient value of tests 1 ~ 6 are 0.812 6, 0.836 1, 0.834 3, 0.807 2, 0.810 5 and 0.824 7, respectively. It shows that the fitting effect of random forest model has good results and it can be used to predict the landslide space. In addition, it can be known from the test samples of the prediction rate that the forecast rate range of each test group is 86.0% ~ 92.1%, which fully indicates that the random forest model has good generalization ability.

Tab. 7 Testing results of random forest model

Parameters	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Kappa coefficient	0.812 6	0.836 1	0.834 3	0.807 2	0.810 5	0.824 7
Prediction rate/%	88.7	92.1	91.5	86.0	87.9	90.8

2.3 Fitting results comparison of logistic regression model and random forest model

2.3.1 Comparative analysis of model prediction

According to the results of model variables, the prediction accuracy rate of logic regression model and the random forest model are calculated respectively (Tab. 8). Research shows that the correct discrimination rate of the random forest model is higher than that of logistic regression model in these 6 tests. In the training and test samples of the 5 sub samples (Tests 1 ~ 5), prediction accuracy rate of the random forest model is 3.2% ~ 10.9% higher than that of logistic regression model. But compared with whole sample, the random forest prediction rate is about 7.7% and 9% higher than that of the logistic regression model (Test 6). The results show that the fitting effect of random forest algorithm is more suitable for Shunchang landslide of Fujian than the traditional logistic regression model, which can be used to predict the landslides occurrence in this area.

2.3.2 Comparative analysis of model generalization ability

After testing and evaluating the effect of modeling, the logistic regression model and random forest model will be used to predict the spatial distribution of landslide risk in the whole study area based on the whole sample data and the obtained 2 models based on ratings system (Fig. 5, Fig. 6). For each model, the proportion of landslide and its contribution in the 4 hazardous areas are shown in Tab. 9.

Tab. 8 Prediction accuracy of logistic model and random forest model

Test group	Logistic regression model		Random forest model	
	Training sample/%	Test sample/%	Training sample/%	Test sample/%
Test 1	80.2	81.7	91.1	88.7
Test 2	81.9	82.1	90.4	92.1
Test 3	80.8	82.0	89.7	91.5
Test 4	83.4	82.8	88.2	86.0
Test 5	81.3	82.0	84.9	87.9
Test 6	82.6	81.8	90.3	90.8

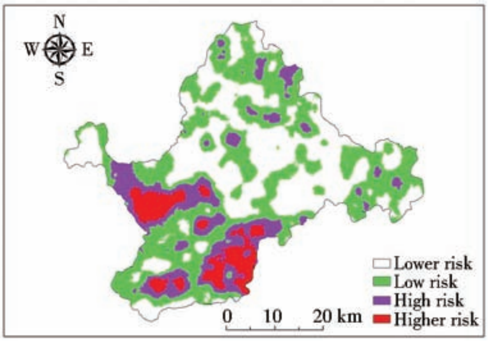


Fig. 5 Landslide susceptibility map based on logistic regression model

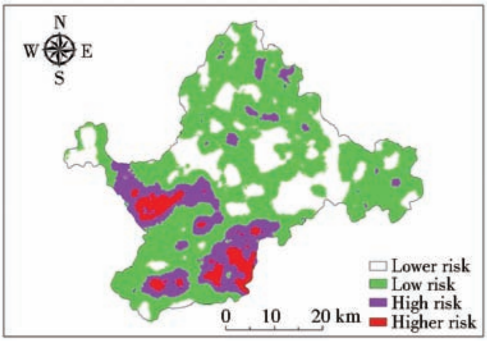


Fig. 6 Landslide susceptibility map based on random forest model

As seen from Tab. 9, contribution of risk zoning predicted by the two models is increased with the increase of risk level, it shows that the model is consistent with thehistorical landslide data, and the model results are consistent with the actually landslide distribution in the study area. The logistic regression model predicts that the high risk area and the higher risk area account for 21.11% of total study area, which is 21.02% higher than that of the random forest model. For the random forest model, the prediction of the high risk area and the higher risk area contains 66.05% of the total landslide, higher than the logistic regression model results about 63.34%. Both of the random forest model contribution values of the high risk area are higher risk area is higher than the that of logistic regression model. According to the above research conclusion, a good landslide space prediction model not only requires landslide disaster must be included in high risk area as much as possible but also requires the prediction of the high risk area of the landslide area as small as possible, which means ability of random forest model is better than that of logistic regression model. This result is consistent with the prediction rate of the 2 models, which further shows that the prediction performance of the random forest model is better and can be used to landslide prediction.

Tab. 9 Proportions of landslide and contribution values of different susceptibility classes for two models

Levels	Logistic regression model					Random forest model				
	Area/ km^2	Area ratio/%	Landslide quantity/each	Landslide ratio/%	Contribution value	Area/ km^2	Area ratio/%	Landslide quantity/each	Landslide ratio/%	Contribution value
Ⅰ(Lower risk)	775.71	38.94	129	8.74	0.22	447.01	22.44	45	3.05	0.14
Ⅱ(Low risk)	795.77	39.95	412	27.91	0.70	1 126.28	56.54	456	30.89	0.55
Ⅲ(High risk)	284.30	14.27	323	21.88	1.53	234.26	16.78	439	29.74	1.77
Ⅳ(Higher risk)	136.22	6.84	612	41.46	6.06	84.45	4.24	536	36.31	8.57

Note: in order to eliminate area difference, contribution value is introduced, contribution value is the ratio of landslide ratio to area ratio.

3 Conclusion

(1) It can be seen from the model variable selection results that those five landslide factors, i. e., "NDVI", "rainfall on the day of landslide", "Plan curvature", "Aspect" and "Altitude", progress to full sample of logistic regression model and random forest model (Test 6). These five landslide factors have important influence on landslide occurrence.

(2) The prediction results of the model show that AUC value and the Kappa coefficient have statisticalsignificance. Among the five samples of the two models, prediction rate of the random forest model is 3.2% ~ 10.9% higher than that of logistic regression model. In the training and test of the whole sample, the random forest prediction rate is about 7.7% and 9% higher than that of logistic regression model. The generalization ability of the model further proved that the random forest model has superior prediction effect than the logistic regression model.

(3) The random forest model predicts that the higher risk area and the high risk area account for 66.05% of the totallandslide area and the accuracy is relatively low. The main reason includes: ①A high threshold of risk classification system for the research. ② Just considering topography, meteorology and hydrology, soil and plant as the main factors, but ignoring influence of geological conditions, human activities, social and economic conditions and other potential factors on landslide that may cause some errors in the prediction of the model.

References

[1] ZHANG Fanyu, LIU Gao, CHEN Wenwu, et al. Large landslide susceptibility assessment by multivariate statistical analysis in the Longnan area affected by the Wenchuan earthquake [J]. Journal of Central South University: Science and Technology, 2012, 43(9): 3595 – 3600. (in Chinese)

[2] CHEN Zhanpeng, LEI Tingwu, YAN Qinghong, et al. Estimation of erosion from earthquake landslides in Wenchuan area [J]. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(4): 195 – 200. (in Chinese)

[3] LIU Guangxu, XI Jianchao, DAI Erfu, et al. Loss risk assessment of the hazard-affectted body of landslides in China [J]. Journal of Natural Disasters, 2014, 23(2): 39 – 46. (in Chinese)

[4] TAN Long, CHEN Guan, ZENG Runqiang, et al. Application of artificial neural network in landslide susceptibility assessment [J]. Journal of Lanzhou University: Natural Sciences, 2014, 50(1): 15 – 20. (in Chinese)

[5] SONG Shengyuan, WANG Qing, PAN Yuzhen, et al. Evaluation of landslide susceptibility degree based on catastrophe theory [J]. Rock and Soil Mechanics, 2014, 35(Supp. 2): 422 – 428. (in Chinese)

[6] XU Chong, DAI Fuchu, XU Suning, et al. Application of logistic regression model on the Wenchuan earthquake triggered landslide hazard mapping and its validation [J]. Hydrogeology & Engineering Geology, 2013, 40(3): 98 – 104. (in Chinese)

[7] HONG Haoyuan, PRADHAN B, XU Chong, et al. Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression,

alternating decision tree and support vector machines [J]. CATENA, 2015, 133: 266 - 281.

[8] OZDEMIR A, ALTURAL T. A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan Mountains, SW Turkey [J]. Journal of Asian Earth Sciences, 2013,64:180 - 197.

[9] WANG L J, SAWADA K, MORIGUCHI S. Landslide susceptibility analysis with logistic regression model based on FCM sampling strategy [J]. Computers & Geosciences, 2013, 57: 81 - 92.

[10] XU Chong, XU Xiwei. The 2010 Yushu earthquake triggered landslides spatial prediction models based on several kernel function types [J]. Chinese Journal of Geophysics,2012, 55(9): 2994 - 3005. (in Chinese)

[11] BREIMAN L. Random forest[J]. Machine Learning, 2001, 45(1): 5 - 32.

[12] CHEN Weitao, LI Xianju, WANG Yanxin, et al. Forested landslide detection using LiDAR data and the random forest algorithm: a case study of the Three Gorges, China[J]. Remote Sensing of Environment, 2014, 152: 291 - 301.

[13] STUMPF A, KERLE N. Obeject-oriented mapping of landslides using random forests[J]. Remote Sensing of Environment, 2011(115): 2564 - 2577.

[14] FANG Kuangnan, WU Jianbin, ZHU Jianping, et al. A review of technologies on random forests [J]. Statistics & Information Forum, 2011, 26(3): 32 - 38. (in Chinese)

[15] LI Fenling, WANG Li, LIU Jing, et al. Remote sensing estimation of SPAD value for wheat leaf based on GF-1 data [J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(9): 273 - 281. (in Chinese)

[16] ZHANG Lei, WANG Linlin, ZHANG Xudong, et al. The basic principle of random forest and its applications in ecology: a case study of *Pinus yunnanensis* [J]. Acta Ecologica Sinica, 2014, 34(3): 650 - 659. (in Chinese)

[17] MA Yue, JIANG Qigang, MENG Zhiguo, et al. Classification of land use in farming area based on random forest algorithm [J]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(1): 297 - 303. (in Chinese)

[18] YAO Xiong, YU Kunyong, LIU Jian, et al. Application of random forest model on the landslide spatial prediction caused by precipitation [J]. Journal of Fujian Agriculture and Forestry University: Natural Sciences Edition,2016, 45(2): 219 - 227. (in Chinese)

[19] LI Xinhai. Using "random forest" for classification and regression[J]. Chinese Journal of Applied Entomology, 2013, 50(4): 1190 - 1197. (in Chinese)

[20] DONG J J, TUNG Y H, CHEN C C, et al. Logistic regression model for predicting the failure probability of a landslide dam[J]. Engineering Geology, 2011,117(1 - 2):52 - 61.

[21] SI Kangping, TIAN Yuan, WANG Daming, et al. The comparison of three statistical methods on landslide susceptibility analysis: a case study of Shenzhen City [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2009,45(4): 19 - 26. (in Chinese)

[22] CAN T, NEFESLIOGLU H A, GOKCEOGLU C, et al. Susceptibility assessments of shallow earthflows triggered by heavy rainfall at three catchments by logistic regression analyses [J]. Geomorphology, 2005, 72(1 - 4): 250 - 271.

doi:10.6041/j.issn.1000-1298.2016.10.043

基于随机森林模型的山体滑坡空间预测研究

余坤勇[1,2] 姚 雄[1,2] 邱祈荣[3] 刘 健[1,2]

（1.3S 技术与资源优化利用福建省高校重点实验室，福州 350002；2.福建农林大学林学院，福州 350002；
3.台湾大学森林环境暨资源学系，台北 10617）

摘要： 滑坡灾害空间分布的准确预测是实现防灾减灾的重要途径。以2010 年福建省顺昌地区滑坡资料为基础数据，分别应用随机森林模型和逻辑回归模型对福建顺昌地区山体滑坡发生与滑坡因子之间的关系进行实证分析，通过模型变量筛选、模型精度分析，探讨了随机森林模型在我国南方山体滑坡空间预测中的适应性。结果表明：随机森林模型对滑坡发生数据的拟合效果比逻辑回归模型好，其对顺昌地区滑坡发生数据的预测精度为90.8%，而逻辑回归模型的预测精度为81.8%；随机森林模型对研究区滑坡发生的泛化能力比逻辑回归模型好，其预测出高危险区和较高危险区所包含的滑坡比总和为66.05%，而逻辑回归模型为63.34%。研究结果表明随机森林模型的性能优于逻辑回归模型，可用于顺昌地区基于滑坡因子的未来滑坡发生的预测预报。

关键词： 山体滑坡；随机森林模型；逻辑回归模型；空间预测

中图分类号： P642.22；X43 **文献标识码：** A **文章编号：** 1000-1298(2016)10-0338-08

Landslide Spatial Prediction Based on Random Forest Model

Yu Kunyong[1,2] Yao Xiong[1,2] Qiu Qirong[3] Liu Jian[1,2]

(1. *University Key Laboratory for Geomatics Technology and Optimize Resources Utilization in Fujian Province*, *Fuzhou* 350002, *China*
2. *College of Forestry*, *Fujian Agriculture and Forestry University*, *Fuzhou* 350002, *China*
3. *School of Forestry and Resource Conservation*, *National Taiwan University*, *Taibei* 10617, *China*)

Abstract: Random forest (RF) is a non-parametric technology which was firstly proposed by Leo Breiman and Cutler Adele in 2001. It was used to deal with the classification and regression problems by gathering a large number of classification tree, which can improve the prediction accuracy. It was applied in the ecological field in recent years. Predicting the spatial distribution of landslide hazard was an important way to achieve disaster prevention and mitigation. The landslide dataset of Shunchang in Fujian Province was taken as case to identify the relationship between mountain landslide occurrence and landslide factors by using RF model and logistic regression (LR) model respectively with landform, meteorological hydrology, soil and vegetation factors. The applicability of RF on landslide prediction in the southern mountain of China was tested by procedure of parameter selection and analysis of model accuracy. The result showed that the goodness of fit of RF was better than that of LR model. The prediction accuracy of RF on the landslide data was 90.8%, while the prediction accuracy of LR was 81.8%. The generalization of RF in the study area was better than that of LR model. The high risk areas and higher risk areas contained 66.05% of the total landslide, which was predicted by RF, while that of LR was 63.34%. The result of model comparison revealed that the RF model was superior to LR model on the mountain landslide prediction in the study area, thus it can be used in the landslide prediction and the division of landslide danger grade with the sample data. In addition, RF model could be applied to other relevant research.

Key words: landslide; random forest model; logistic regression model; spatial prediction

收稿日期：2016－03－29 修回日期：2016－05－12

基金项目： 国家自然科学基金项目(41401385)、福建省教育厅基金项目(JA14126)和福建农林大学林学院青年基金项目

作者简介： 余坤勇(1980—)，男，副教授，主要从事资源监测及3S 技术研究，E-mail：yuyky@126.com

通信作者： 刘健(1963—)，男，教授，博士生导师，主要从事森林经营管理与3S 技术应用研究，E-mail：fjliujian@126.com

引言

滑坡是山区环境中破坏最严重的地质灾害之一，已成为影响和危害山区民众不可忽视的自然灾害问题[1-2]。我国是一个多山的国家，滑坡在我国广泛分布，平均每年造成近千人死亡，严重制约着我国社会经济发展[3]。由于全球气候变迁和极端灾害频发以及人类经济活动的不断加剧，滑坡灾害发生情况可能会更加严重[4-5]。因此，研究区域滑坡灾害与滑坡因子之间的关系，实现区域山体滑坡的准确预测，对于滑坡发生的可能及危害程度的预警预报、防灾减灾、灾害管理等具有重要意义[6]。

已有学者对山体滑坡进行了空间预测研究，主要采用的是传统逻辑回归模型[7-9]。随着人工智能的发展，机器学习模型已越来越多地应用于山体滑坡空间预测研究中[1,10]。2001 年 BREIMAN 等[11]提出了一种新的机器学习模型——随机森林模型，具有很高的学习能力和预测精度。近年来国外有少数研究应用随机森林模型进行山体滑坡空间预测[12-13]，而国内关于随机森林模型在滑坡灾害空间预测领域的应用还鲜有报道，目前，随机森林模型主要应用于医学、经济学、生态学等领域[14-17]。由于研究区域的空间异质性，国外关于随机森林模型在山地滑坡空间预测上的优越性结论能否适用于我国的山地还有待探讨。为此，本研究选择福建省顺昌县为研究对象，分别选择随机森林模型和逻辑回归模型对研究区进行山体滑坡空间预测的研究，探讨 2 种模型在研究区中的整体性能。通过模型拟合结果的分析，判断随机森林模型在顺昌县滑坡预测中的适应性，以期为深入开展区域滑坡预警预报研究和决策提供重要依据。

1　材料与方法

1.1　研究区概况

顺昌县地处福建省西北部(117°29′~118°14′E，26°38′~27°12′N)，是我国南方重要林区，也是福建省毛竹生产基地县。研究区属中亚热带海洋性季风气候，气候温和，降水丰富，年均降水 2 051 mm，年均气温 19.1℃，地貌以低山丘陵为主，地势从东北、西北和西南向中部倾斜，土壤类型以红壤为主。该县总面积约 2 000 km^2，下辖 11 个乡和 3 个镇，共有 4 个居委会和 129 个村委会。

1.2　数据来源

研究采用的数据包括：①福建省科技计划重点项目 2012N0003 提供的研究区 1 478 处滑坡空间分布数据。②顺昌县 2010 年 11 月 8 日和 2010 年 11 月 25 日空间分辨率为 10 m 的 ALOS 多光谱遥感影像数据(图 1)。③地球空间数据云平台(http://www.gscloud.cn/)提供的顺昌县 30 m 分辨率的数字高程模型(Digital elevation model，DEM)数据(图 2)。④顺昌县气象局提供的研究区 2010 年 6 月份分时降雨量数据。

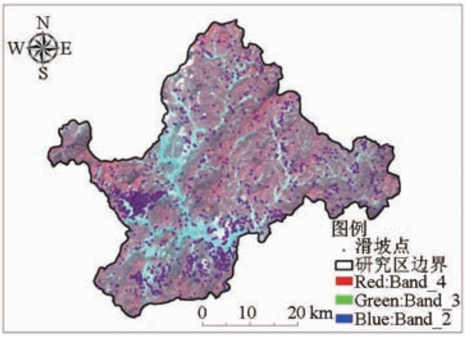


图 1　研究区原始影像与滑坡点分布

Fig. 1　Source imagine and landslides distribution map of study area

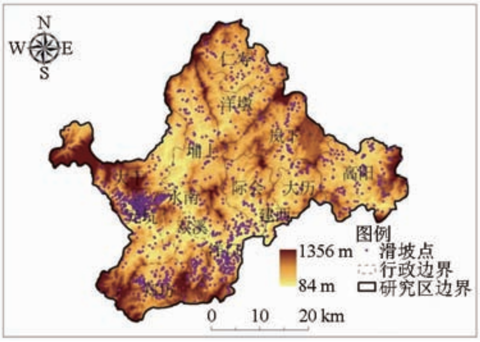


图 2　研究区 DEM 与滑坡点分布

Fig. 2　DEM and landslides distribution map of study area

1.3　滑坡因子提取

滑坡是自然界多种因子综合作用的结果，各个因子之间往往有着相互关联的因果关系。滑坡发生与否，首先受到自身地理环境的影响，包括地形地貌、土壤环境和植被信息等。一般情况下，在地势平坦区不易发生滑坡灾害，在地形起伏大的区域会随着地震、降雨等事件而发生滑坡，且滑坡发生的概率会随着坡度和高程的增加而增加；坡向可以通过影响植被覆盖和降雨路径，进而影响坡体的稳定性。植被根系可以加强坡体整体性，同时坡面植被可以有效阻止降雨入渗，防止坡体发生滑动，因此植被覆盖度是滑坡发生的一个影响因子。由于不同类型的土壤疏松程度不一，因此土壤环境可以直接影响坡体的稳定。降水是导致坡体滑动的重要外在因子，主要表现在降水形成的水压力引起坡体抗滑力减小[18]。另外，离水系越近的土壤，其含水性越大，而含水性越大往往有利于滑坡的发生，因而，到水系距离也是滑坡发生的一个潜在因子。

综上分析,在个别条件相同及参阅已有研究的基础上,综合考虑研究区的自然地理条件,分别从地形地貌(包括坡向、高程、地形起伏度、平面曲率、剖面曲率、坡度)、气象水文(包括滑坡发生当天降雨量、滑坡发生前1 d降雨量、滑坡发生前2 d降雨量、到水系的距离)、土壤植被(包括归一化植被指数、土壤类型)3个方面选取了12个因子作为滑坡影响因子(表1)。其中地形地貌的专题信息主要利用DEM数据提取;滑坡发生前n d的降雨量主要通过ArcGIS 9.3软件中的反距离加权插值法(Inverse distance weighted,IDW)获取[18];到水系距离数据利用ArcGIS 9.3空间模块计算研究区到河流的欧氏距离;归一化植被指数(Normalized difference vegetation index,NDVI)专题信息通过ERDAS 9.2建模工具获取。

表1　滑坡因子来源

Tab.1　Source of landslide factors

类型	变量	来源	代码
地形地貌	坡向	DEM	ASP
	高程	DEM	ELE
	地形起伏度	DEM	DXD
	平面曲率	DEM	PLC
	剖面曲率	DEM	PRC
	坡度	DEM	SLO
气象水文	滑坡发生当天降雨量	顺昌县气象局	RA0
	滑坡发生前1 d降雨量	顺昌县气象局	RA1
	滑坡发生前2 d降雨量	顺昌县气象局	RA2
	到水系的距离	地形图	DSX
土壤植被	归一化植被指数	ALOS影像数据	NDVI
	土壤类型	福建省1:250 000土壤类型数据	TRLX

1.4　模型介绍

逻辑回归(Logistic regression,LR)模型是一种多元统计分析模型,已经被广泛地应用于滑坡空间预测研究。主要思想是利用最大似然法来构建预测变量与二分类结果之间的关系,保证每一点均为最优拟合。设滑坡发生的概率为P,则没有滑坡发生的概率为$1-P$。滑坡发生的概率P与自变量(滑坡影响因子)之间的回归关系[6]为

$$P=\frac{1}{1+\exp(-(\beta_0+\beta_1X_1+\cdots+\beta_mX_m))}\qquad(1)$$

式中　β_i——基于训练样本得到的逻辑回归系数,$i=0,1,2,\cdots,m$

X_i——自变量,$i=1,2,\cdots,m$

随机森林(Random forest,RF)模型是一种基于分类回归树的算法。随机森林模型的主要思想是通过自助法(bootstrap)抽样从原始训练集中抽取k个样本,且每个样本的样本容量均与原始训练集的大小一致;然后对每个样本分别进行决策树建模,得到k个建模结果;最后,利用所有决策树的建模结果,通过投票表决决定其最终的分类结果[14]。

与传统的滑坡预测方法相比(逻辑回归方法、支持向量机等),随机森林模型不需要检查变量的交互作用是否显著,由于它进行了两次随机抽样,故其在异常值和噪声方面具有较高的容忍度,而且不容易出现过拟合现象,具有很高的预测精度[19]。

1.5　模型建立

应用定量模型进行滑坡预测之前,需要选择训练样本数据用于模型的建立。以往的研究中训练样本数据往往只考虑正样本(即滑坡点数据),忽略了负样本(即非滑坡点数据)在滑坡预测中的作用,以这种训练样本数据建立的滑坡预测模型所得到的结果不能很好地体现滑坡的影响机制。为此,本研究利用ArcGIS 9.3软件在已知滑坡点100 m外随机生成同样数目的点作为非滑坡点,进而提取滑坡点和非滑坡点各滑坡因子的属性值数据作为正样本和负样本,并从总体样本中(由正样本和负样本组成)选择训练样本。

为了减少单一的取样方式对模型建立产生的影响,本研究将总体样本数据(1 478个正样本和1 478个负样本)随机划分为60%的样本数据作为训练样本集和40%的样本数据作为检验样本集,并重复进行5次随机划分,以得到5组不同的样本组。然后,选择R统计软件和SPSS软件分别对5组不同的样本组进行随机森林和逻辑回归模型计算(试验1～5)。最后,以2种模型的显著变量在5次试验中出现次数大于或等于3次为原则,确定最终变量,分别进行全样本的模型计算(试验6)。

1.6　模型评价

采用受试者工作特征曲线(Receiver operating characteristic curve,ROC曲线)的曲线下面积(Area under the curve,AUC)对逻辑回归模型结果进行评价。AUC值介于0.5～1之间,其值越接近于1,表示模型的预测效果越好,当AUC值为1时,表示模型是一个效果最佳的预测。另外,根据ROC曲线分析法计算的敏感度和特异度,可求得约登指数,约登指数即为敏感度与特异度之和减去1,进而判断最佳临界值。在山体滑坡预测中便可根据临界值来判断滑坡发生与否,如果滑坡发生的预测概率大于该临界值则认为有滑坡发生,小于该临界值则认为无滑坡发生[20]。

采用Kappa系数对随机森林模型效果进行评价,其具体计算公式参照文献[21]。

为了更精确评价 2 种模型的总体性能，本文对滑坡及非滑坡点的滑坡发生概率进行空间插值得到滑坡灾害的空间分布预测图。有研究指出[22]，良好的滑坡空间预测模型应满足 2 个准则：①滑坡灾害应尽可能多的位于滑坡高危险区。②预测图上滑坡高危险区应尽可能小。依据这 2 个准则，本文将预测图上滑坡发生的危险性划分为低（0 ~ 0.25）、较低（0.25 ~ 0.50）、较高（0.50 ~ 0.75）、高（0.75 ~ 1.00）4 个等级。根据划分后的危险性等级图，计算各等级区域内的滑坡比，结合各危险性等级区域面积对模型的泛化能力进行评价。

2　结果与分析

2.1　逻辑回归模型的拟合结果分析

2.1.1　多重共线性诊断

由于线性回归模型中自变量之间可能存在精确相关关系或高度相关关系，如果不加以剔除，会导致模型的预测功能失效或结果难以估测准确。因此，在模型运算之前，要进行自变量的多重共线性诊断，剔除有显著共线性的变量。常用的共线性诊断指标主要有方差膨胀因子（Variance inflation factor，VIF）和容忍度（Tolerance，TOL），这 2 个指标互为倒数，一般说来，VIF 大于 10（即 TOL 小于 0.1）时，表明所选择的变量多重共线性较严重。利用 SPSS 21.0 软件对影响滑坡发生的因子数据进行共线性检验，结果见表 2。

表 2　多重共线性诊断结果

Tab. 2　Multicollinearity diagnosis indexes for variables

变量	TOL	VIF
RA0	0.600	1.667
RA1	0.207	4.839
RA2	0.076	13.170
ASP	0.978	1.022
ELE	0.581	1.720
DSX	0.047	21.375
DXD	0.041	24.390
NDVI	0.767	1.304
PLC	0.785	1.275
PRC	0.772	1.296
SLO	0.465	2.151
TRLX	0.089	11.207

表 2 显示，滑坡发生前 2 d 降雨量、到水系距离、地形起伏度、土壤类型 4 个变量的 VIF 大于 10，因而将这 4 个变量剔除，最终滑坡发生当天降雨量、滑坡发生前 1 d 降雨量、坡向、高程、归一化植被指数、平面曲率、剖面曲率、坡度共 8 个变量进入模型的拟合阶段。

2.1.2　模型拟合结果分析

本文对研究区滑坡发生和对应的经共线性检验后的滑坡因子进行逻辑回归模型计算，先对 5 个训练样本进行模型拟合，得到 5 次试验中不同的显著因子，在此基础上选择 5 次试验中出现 3 次及以上的因子进入全样本数据进行拟合（试验 6），各试验中的显著因子见表 3。

表 3　逻辑回归模型中各试验数据拟合中的显著因子

Tab. 3　Significant factors in logistic regression model for each experiment data

因子	试验 1	试验 2	试验 3	试验 4	试验 5	试验 6
RA0	Y	Y	Y	Y	Y	Y
RA1	Y	N	N	Y	N	N
ASP	N	Y	Y	N	Y	Y
ELE	N	Y	Y	Y	Y	Y
NDVI	Y	Y	Y	Y	Y	Y
PLC	Y	Y	N	Y	Y	Y
PRC	N	N	Y	N	Y	N
SLO	N	N	Y	N	N	N

注：Y 表示试验显著因子，N 表示试验非显著因子；下同。

基于试验 6 全样本数据的逻辑回归拟合结果显示，模型的 Cox&Snell R^2 为 0.542，Nagelkerke R^2 为 0.723，说明模型的整体有效性良好，从模型的参数拟合结果可知（表 4），最终的模型变量均与滑坡发生有显著相关性且均在 0.01 水平上显著相关。

表 4　逻辑回归模型拟合结果

Tab. 4　Fitting results of logistic regression model

自变量	回归系数	标准误差	Wals 卡方值	自由度	显著水平
RA0	8.001	0.022	17.021	1	<0.01
ASP	−0.007	0.001	29.606	1	<0.01
ELE	0.003	0.001	22.746	1	<0.01
NDVI	−12.030	0.086	19.361	1	<0.01
PLC	−1.286	0.085	18.180	1	<0.01
常量	4.843	0.061	13.732	1	<0.01

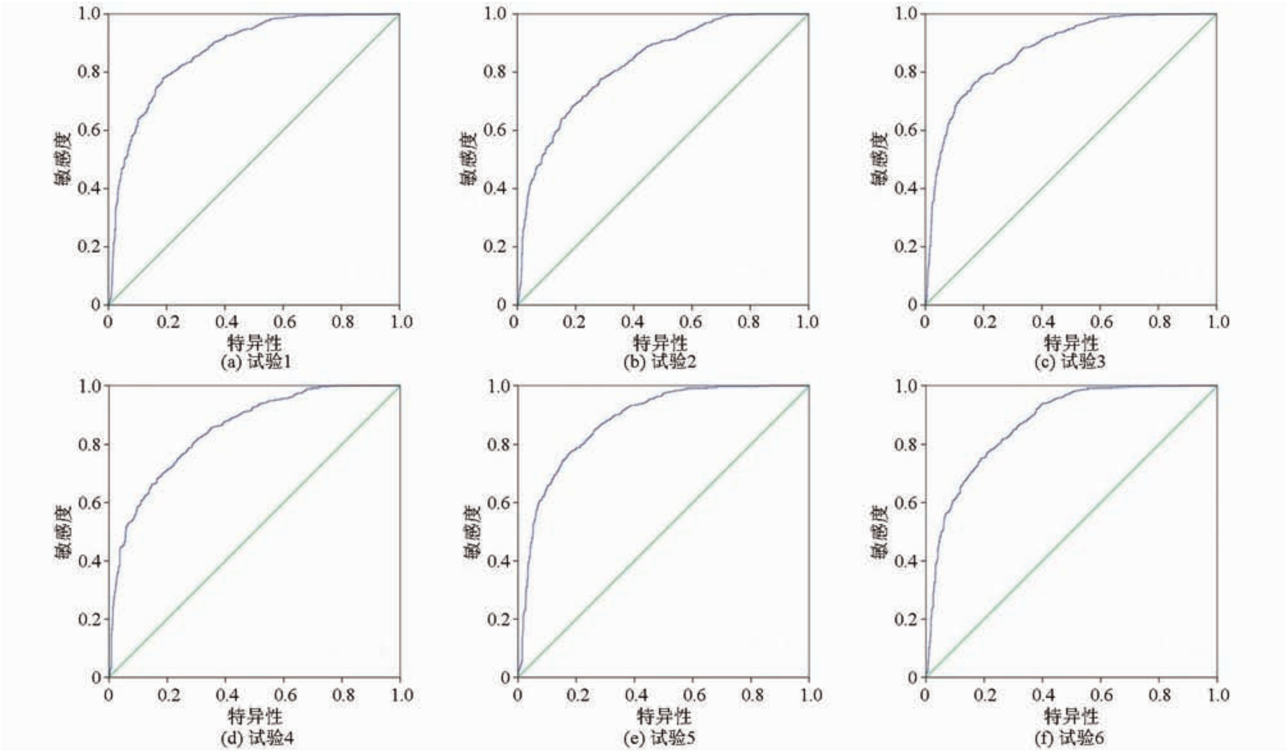
2.1.3　模型检验

图 3 为应用 SPSS 软件作出的各个试验的 ROC 曲线，表 5 为各个试验模型的 AUC 值、临界值和预测率。由表 5 可知，试验 1 ~ 6 的 AUC 值分别为 0.872、0.830、0.876、0.849、0.881、0.873，说明本文建立的逻辑回归模型拟合效果较好，可用于山体滑坡的空间预测。另外，通过建立的模型，结合约登指数得到的最佳临界值，计算各试验组的预测率。结果表明，各试验组的预测率范围为 81.7% ~ 82.8%。

2.2　随机森林模型的拟合结果分析

2.2.1　模型特征变量选择

随机森林模型可以进行特征变量的选择，主要



(a) 试验1　(b) 试验2　(c) 试验3　(d) 试验4　(e) 试验5　(f) 试验6

图3　逻辑回归模型的ROC曲线

Fig. 3　ROC curves of logistic regression model

表5　逻辑回归模型的检验结果

Tab. 5　Testing results of logistic regression model

项目	试验1	试验2	试验3	试验4	试验5	试验6
AUC值	0.872	0.830	0.876	0.849	0.881	0.873
临界值	0.818	0.814	0.816	0.820	0.832	0.823
预测率/%	81.7	82.1	82.0	82.8	82.0	81.8

注:预测率是基于40%的检验样本计算所得,下同。

表6　随机森林模型中各试验数据拟合中的显著因子

Tab. 6　Significant factors in random forest model for each experiment data

因子	试验1	试验2	试验3	试验4	试验5	试验6
RA0	Y	Y	Y	Y	Y	Y
RA1	Y	Y	N	Y	N	Y
RA2	N	N	N	N	N	N
ASP	Y	N	N	N	Y	Y
ELE	N	Y	Y	Y	N	Y
DSX	N	N	N	N	Y	N
DXD	N	Y	N	N	N	N
NDVI	Y	Y	Y	Y	Y	Y
PLC	Y	Y	Y	Y	N	Y
PRC	N	N	N	N	N	N
SLO	N	N	Y	Y	N	N
TRLX	N	Y	N	N	N	N

根据袋外数据误差最小进行特征变量的选择。本文对研究区滑坡发生和对应的滑坡因子进行随机森林模型计算,首先采用R统计软件中的程序包varSelRF对5个训练样本进行模型特征变量的选择计算,得到5次试验中不同的显著因子,同样在5次试验中出现3次及以上的因子进入全样本数据进行拟合,各试验中的显著因子见表6。

2.2.2　模型特征变量重要性排序

随机森林模型可以通过平均准确率降低度(Mean decrease accuracy)给出滑坡因子的重要性排序,即通过将某一滑坡因子的取值打乱,然后分析打乱前后随机森林预测准确性的降低程度,其值越大表示该因子的重要性越大。本文在利用随机森林模型进行特征变量选择之后,分别对6次试验组数据进行随机森林模型拟合训练,得到6次随机森林试验中各滑坡因子的重要性排序(图4)。从全样本模型来看(图4f),影响研究区滑坡发生的最重要的因子是归一化植被指数(NDVI),其次是滑坡发生当天降雨量(RA0),滑坡发生前1 d的降雨量(RA1)影响程度最小。从6次试验拟合结果来看,归一化植被指数(NDVI)和滑坡发生当天降雨量(RA0)对山体滑坡发生的影响均高于其他变量。

2.2.3　模型检验

表7为随机森林算法中各个试验模型的Kappa系数和预测率。由表7可知,试验1~6的Kappa系数值分别为0.812 6、0.836 1、0.834 3、0.807 2、0.810 5、0.824 7,说明所建立的随机森林模型拟合效果较好,可用于山体滑坡的空间预测。另外,从检验样本的预测率可以看出,各试验组的预测率范围为86.0%~92.1%,说明建立的随机森林模型有很

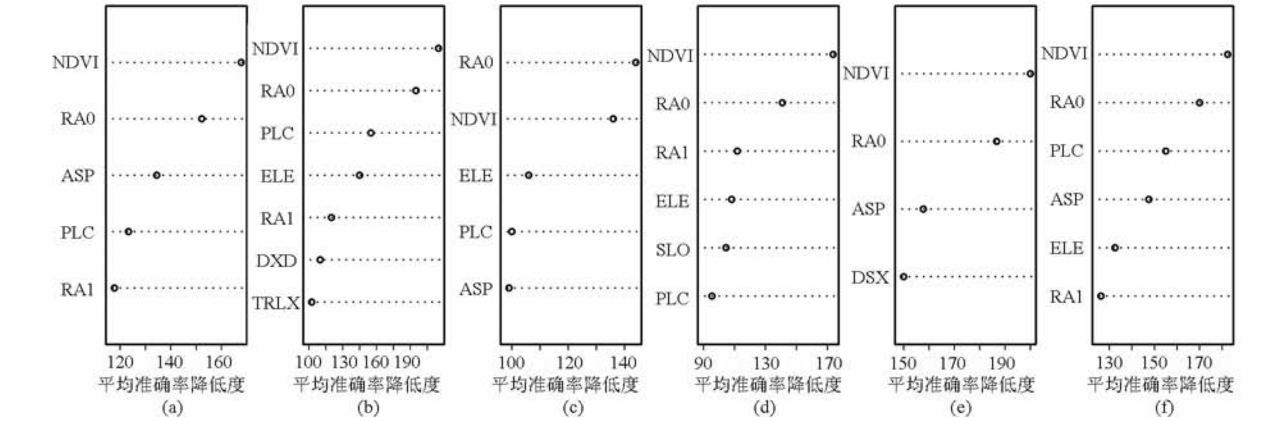


图 4　随机森林模型中滑坡因子的重要性排序

Fig. 4　Importance sorting of landslide factors in random forest model

表 7　随机森林模型的检验结果

Tab. 7　Testing results of random forest model

项目	试验 1	试验 2	试验 3	试验 4	试验 5	试验 6
Kappa 系数	0. 812 6	0. 836 1	0. 834 3	0. 807 2	0. 810 5	0. 824 7
预测率/%	88. 7	92. 1	91. 5	86. 0	87. 9	90. 8

好的泛化能力。

2. 3　逻辑回归模型与随机森林模型拟合结果对比分析

2. 3. 1　模型预测率对比分析

根据模型显著变量选择结果，分别进行逻辑回归模型和随机森林模型预测准确率的计算（表 8）。研究结果显示，6 次试验中，随机森林模型的正确判别率均比逻辑回归模型高，在 5 个子样本的训练和测试样本中（试验 1 ~5），随机森林模型分别高于逻辑回归模型 3. 2% ~10. 9% 之间。在全样本的训练和测试样本中（试验 6），随机森林预测率较逻辑回归模型高约 7. 7% 和 9. 0%。说明随机森林算法对福建顺昌地区山体滑坡发生的拟合效果优于传统的逻辑回归模型，可用于该地区的山体滑坡发生预测。

表 8　逻辑回归模型和随机森林模型的预测精度

Tab. 8　Prediction accuracy of logistic regression model and random forest model　　%

试验组	逻辑回归模型		随机森林模型	
	训练样本	检验样本	训练样本	检验样本
试验 1	80. 2	81. 7	91. 1	88. 7
试验 2	81. 9	82. 1	90. 4	92. 1
试验 3	80. 8	82. 0	89. 7	91. 5
试验 4	83. 4	82. 8	88. 2	86. 0
试验 5	81. 3	82. 0	84. 9	87. 9
试验 6	82. 6	81. 8	90. 3	90. 8

2. 3. 2　模型泛化能力对比分析

建模效果检验和评价后，将基于全样本数据的逻辑回归模型和随机森林模型应用于整个研究区域形成滑坡灾害的空间分布预测图，在此基础上通过分级体系得到 2 种模型的滑坡灾害危险性预测图（图 5、图 6）。对于每种模型，4 个危险性分区内滑坡比例及贡献值如表 9 所示。

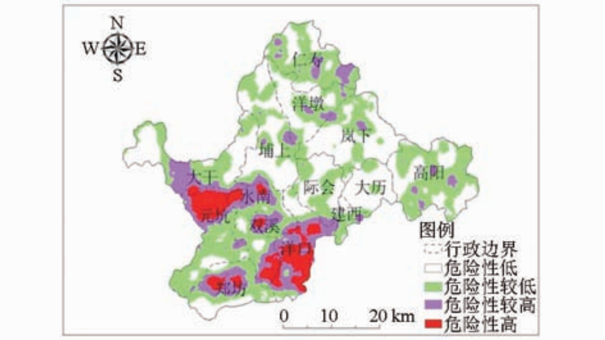


图 5　基于逻辑回归模型得到的预测结果图

Fig. 5　Landslide susceptibility map based on logistic regression model

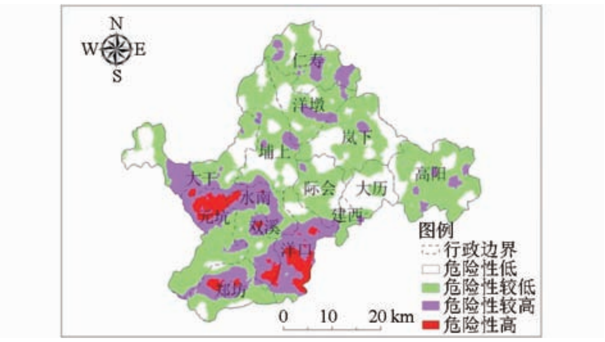


图 6　基于随机森林模型得到的预测结果图

Fig. 6　Landslide susceptibility map based on random forest model

由表 9 可以看出，2 种模型所预测出来的危险性分区贡献值随着危险性等级的升高而迅速的升高，说明模型划分的危险性预测图符合历史滑坡数据，模型结果与研究区实际滑坡分布情况相吻合。逻辑回归模型预测出的高危险区和较高危险区面积总和占研究区总面积的 21. 11%，高于随机森林模型的 21. 02%；对于随机森林模型，其预测出的高危险区和较高危险区所包含的滑坡比占总和的

表9　2种模型结果图中危险性分区内的滑坡比例和贡献值

Tab.9　Proportions of landslide and contribution values of different susceptibility classes for two models

等级	逻辑回归模型					随机森林模型				
	面积/km^2	面积比/%	滑坡数量/个	滑坡比/%	贡献值	面积/km^2	面积比/%	滑坡数量/个	滑坡比/%	贡献值
危险性低	775.71	38.94	129	8.74	0.22	447.01	22.44	45	3.05	0.14
危险性较低	795.77	39.95	412	27.91	0.70	1126.28	56.54	456	30.89	0.55
危险性较高	284.30	14.27	323	21.88	1.53	234.26	16.78	439	29.74	1.77
危险性高	136.22	6.84	612	41.46	6.06	84.45	4.24	536	36.31	8.57

注:为消除面积差异,引入贡献值,贡献值为滑坡比与面积比的比值。

66.05%,高于逻辑回归模型的63.34%;高危险区和较高危险区的贡献值均是随机森林模型高于逻辑回归模型。根据前文述及的一个良好的滑坡空间预测模型既要求滑坡灾害尽量多的落在滑坡高危险区域内,又要求预测图上滑坡高危险区域尽可能小这2个准则,说明随机森林模型的泛化能力优于逻辑回归模型。此结果和2种模型的预测率结果相一致,进一步说明随机森林模型的预测性能较逻辑回归模型好,可用于研究区的山体滑坡预测。

3　结论

(1)从模型的变量选择结果可以看出,归一化植被指数、滑坡发生当天降雨量、平面曲率、坡向、高程5个滑坡因子均进入了逻辑回归和随机森林2种模型的最终全样本中(试验6),这5个滑坡因子是滑坡发生的重要因子。

(2)模型的预测结果显示,2种模型的AUC值和Kappa系数均有统计学意义,在对2种模型的5个样本组预测中,随机森林模型分别高于逻辑回归模型3.2%~10.9%之间,在全样本的训练和测试样本中(试验6),随机森林预测率较逻辑回归模型高约7.7%和9.0%,模型的泛化能力进一步证明了随机森林模型较逻辑回归模型有更好的预测效果。

(3)随机森林模型预测出的高危险区和较高危险区所包含的滑坡个数仅占总和的66.05%,精度相对较低,主要原因在于:①本文的危险性分级体系设置的阈值较高。②地形地貌、气象水文、土壤植被3个方面是本文选择的主要滑坡因子,而地质状况、人类活动、社会经济条件等滑坡潜在因子及这些因子对滑坡的影响未考虑,这可能对模型的预测结果造成一定的误差。

参考文献

1　张帆宇,刘高,谌文武,等.基于多变量统计分析的大型滑坡敏感性评价:以汶川地震影响的陇南地区为例[J].中南大学学报:自然科学版,2012, 43(9): 3595-3600.
ZHANG Fanyu, LIU Gao, CHEN Wenwu, et al. Large landslide susceptibility assessment by multivariate statistical analysis in the Longnan area affected by the Wenchuan earthquake [J]. Journal of Central South University: Science and Technology, 2012, 43(9):3595-3600. (in Chinese)

2　陈展鹏,雷廷武,晏清洪,等.汶川震区滑坡堆积体坡面侵蚀量测算方法[J].农业机械学报,2014, 45(4): 195-200.
CHEN Zhanpeng, LEI Tingwu, YAN Qinghong, et al. Estimation of erosion from earthquake landslides in Wenchuan area [J]. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(4): 195-200. (in Chinese)

3　刘光旭,席建超,戴尔阜,等.中国滑坡灾害承灾体损失风险定量评估[J].自然灾害学报,2014, 23(2): 39-46.
LIU Guangxu, XI Jianchao, DAI Erfu, et al. Loss risk assessment of the hazard-affectted body of landslides in China [J]. Journal of Natural Disasters, 2014, 23(2): 39-46. (in Chinese)

4　谭龙,陈冠,曾润强,等.人工神经网络在滑坡敏感性评价中的应用[J].兰州大学学报:自然科学版,2014, 50(1): 15-20.
TAN Long, CHEN Guan, ZENG Runqiang, et al. Application of artificial neural network in landslide susceptibility assessment [J]. Journal of Lanzhou University: Natural Sciences, 2014, 50(1): 15-20. (in Chinese)

5　宋盛渊,王清,潘玉珍,等.基于突变理论的滑坡危险性评价[J].岩土力学,2014, 35(增刊2): 422-428.
SONG Shengyuan, WANG Qing, PAN Yuzhen, et al. Evaluation of landslide susceptibility degree based on catastrophe theory [J]. Rock and Soil Mechanics, 2014, 35(Supp.2): 422-428. (in Chinese)

6　许冲,戴福初,徐素宁,等.基于逻辑回归模型的汶川地震滑坡危险性评价与检验[J].水文地质工程地质, 2013, 40(3): 98-104.
XU Chong, DAI Fuchu, XU Suning, et al. Application of logistic regression model on the Wenchuan earthquake triggered landslide hazard mapping and its validation [J]. Hydrogeology & Engineering Geology, 2013, 40(3): 98-104. (in Chinese)

7　HONG Haoyuan, PRADHAN B, XU Chong, et al. Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines [J]. CATENA, 2015, 133: 266-281.

8　OZDEMIR A, ALTURAL T. A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping：Sultan Mountains，SW Turkey［J］. Journal of Asian Earth Sciences，2013，64：180－197.

9　WANG L J，SAWADA K，MORIGUCHI S. Landslide susceptibility analysis with logistic regression model based on FCM sampling strategy［J］. Computers & Geosciences，2013，57：81－92.

10　许冲，徐锡伟. 基于不同核函数的 2010 年玉树地震滑坡空间预测模型研究［J］. 地球物理学报，2012，55（9）：2994－3005.
XU Chong，XU Xiwei. The 2010 Yushu earthquake triggered landslides spatial prediction models based on several kernel function types［J］. Chinese Journal of Geophysics，2012，55（9）：2994－3005.（in Chinese）

11　BREIMAN L. Random forest［J］. Machine Learning，2001，45（1）：5－32.

12　CHEN Weitao，LI Xianju，WANG Yanxin，et al. Forested landslide detection using LiDAR data and the random forest algorithm：a case study of the Three Gorges，China［J］. Remote Sensing of Environment，2014，152：291－301.

13　STUMPF A，KERLE N. Obeject-oriented mapping of landslides using random forests［J］. Remote Sensing of Environment，2011（115）：2564－2577.

14　方匡南，吴见彬，朱建平，等. 随机森林方法研究综述［J］. 统计与信息论坛，2011，26（3）：32－38.
FANG Kuangnan，WU Jianbin，ZHU Jianping，et al. A review of technologies on random forests［J］. Statistics & Information Forum，2011，26（3）：32－38.（in Chinese）

15　李粉玲，王力，刘京，等. 基于高分一号卫星数据的冬小麦叶片 SPAD 值遥感估算［J］. 农业机械学报，2015，46（9）：273－281.
LI Fenling，WANG Li，LIU Jing，et al. Remote sensing estimation of SPAD value for wheat leaf based on GF－1 data［J］. Transactions of the Chinese Society for Agricultural Machinery，2015，46（9）：273－281.（in Chinese）

16　张雷，王琳琳，张旭东，等. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例［J］. 生态学报，2014，34（3）：650－659.
ZHANG Lei，WANG Linlin，ZHANG Xudong，et al. The basic principle of random forest and its applications in ecology：a case study of *Pinus yunnanensis*［J］. Acta Ecologica Sinica，2014，34（3）：650－659.（in Chinese）

17　马玥，姜琦刚，孟治国，等. 基于随机森林算法的农耕区土地利用分类研究［J］. 农业机械学报，2016，47（1）：297－303.
MA Yue，JIANG Qigang，MENG Zhiguo，et al. Classification of land use in farming area based on random forest algorithm［J］. Transactions of the Chinese Society for Agricultural Machinery，2016，47（1）：297－303.（in Chinese）

18　姚雄，余坤勇，刘健，等. 基于随机森林模型的降水诱发山体滑坡空间预测技术［J］. 福建农林大学学报：自然科学版，2016，45（2）：219－227.
YAO Xiong，YU Kunyong，LIU Jian，et al. Application of random forest model on the landslide spatial prediction caused by precipitation［J］. Journal of Fujian Agriculture and Forestry University：Natural Sciences Edition，2016，45（2）：219－227.（in Chinese）

19　李欣海. 随机森林模型在分类与回归分析中的应用［J］. 应用昆虫学报，2013，50（4）：1190－1197.
LI Xinhai. Using “random forest” for classification and regression［J］. Chinese Journal of Applied Entomology，2013，50（4）：1190－1197.（in Chinese）

20　DONG J J，TUNG Y H，CHEN C C，et al. Logistic regression model for predicting the failure probability of a landslide dam［J］. Engineering Geology，2011，117（1－2）：52－61.

21　司康平，田原，汪大明，等. 滑坡灾害危险性评价的 3 种统计方法比较——以深圳市为例［J］. 北京大学学报：自然科学版，2009，45（4）：19－26.
SI Kangping，TIAN Yuan，WANG Daming，et al. The comparison of three statistical methods on landslide susceptibility analysis：a case study of Shenzhen City［J］. Acta Scientiarum Naturalium Universitatis Pekinensis，2009，45（4）：19－26.（in Chinese）

22　CAN T，NEFESLIOGLU H A，GOKCEOGLU C，et al. Susceptibility assessments of shallow earthflows triggered by heavy rainfall at three catchments by logistic regression analyses［J］. Geomorphology，2005，72（1－4）：250－271.