OPTIMIZATION OF RANDOM FOREST MODEL FOR ASSESSING AND PREDICTING GEOLOGICAL HAZARDS SUSCEPTIBILITY IN LINGYUN COUNTY

Chunfang Kong[1,2,3,4], Kai Xu[1,2,3,*], Junzuo Wang[1], Yiping Tian[1,3,4], Zhiting Zhang[1,3,4] and Zhengping Weng[1,3,4]

[1]School of Computer, China University of Geosciences, Wuhan, China
[2]Hubei Key Laboratory of Intelligent Geo-Information Processing, Wuhan, China
[3]Innovation Center of Mineral Resources Exploration Engineering Technology in Bedrock Area, Ministry of Natural Resources, Guiyang, China
[4]National-Local Joint Engineering Laboratory on Digital Preservation and Innovative Technologies for the Culture of Traditional Villages and Towns, Hengyang, China

ABSTRACT

The random forest (RF) model is improved by the optimization of unbalanced geological hazards dataset, differentiation of continuous geological hazards evaluation factors, sample similarity calculation, and iterative method for finding optimal random characteristics by calculating out-of-bagger errors. The geological hazards susceptibility evaluation model based on optimized RF (OPRF) was established and used to assess the susceptibility for Lingyun County. Then, ROC curve and field investigation were performed to verify the efficiency for different geological hazards susceptibility assessment models. The AUC values for five models were estimated as 0.766, 0.814, 0.842, 0.846 and 0.934, respectively, which indicated that the prediction accuracy of the OPRF model can be as high as 93.4%. This result demonstrated that the geological hazards susceptibility assessment model based on OPRF has the highest prediction accuracy. Furthermore, the OPRF model could be extended to other regions with similar geological environment backgrounds for geological hazards susceptibility assessment and prediction.

KEYWORDS

Geological Hazards, Susceptibility Evaluation, Random Forest (RF), Optimized RF (OPRF), Geographical Information Systems (GIS).

1. INTRODUCTION

The geological hazards system is a nonlinear, dynamic and open complex giant system with multiple levels of structure, multiple control parameters, multiple time scales, and diverse processes [1]. Geological hazards is one of the most serious disasters that can cause not only great economic losses and ecological damage, but can also critically threaten the survival of human beings and the construction of major projects [2-4]. Therefore, the selection of a suitable geological hazards susceptibility assessment method is an important part of geological hazards research, which is of great significance to disaster reduction and prevention [4-6].

David C. Wyld et al. (Eds): CoSIT, AIAPP, SIGL, CRIS, NLPML - 2021
pp. 29-49, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110403

To date, various models and methods have been developed and applied for assessing geological hazards susceptibility in many areas of the world. Among them, the qualitative evaluation method based on expert experience is one of the commonly used methods in the early years. Such as fuzzy comprehensive evaluation model [7-9], analytical hierarchy process [2,3,7,9-12], and weighted linear combination [12], and so on. These methods determine the weight of each evaluation factor through expert scoring, being less time-consuming; However dependence on the subject experience and analysis judgment of the individual experts leads to lack of consistency and portability.

Deterministic model is another commonly used method to evaluate the susceptibility of geological hazards, such as the limit equilibrium method. The mode has high reliability based on the mechanical models of the relationship between relating factors and geological hazards [13], but it requires absolute detailed parameters such as physical, geological environment, tectonic lithology, hydrology and so on. Therefore, the availability of data limits the applicability of this kind of model to the evaluation of local-scale geological hazards.

In addition, quantitative evaluation is the most widely used method in evaluating the susceptibility of geological hazards, such as information value model [2,11,14-16], mathematical statistics method [2,6,10,14,16], certainty factor [17,18], logistic regression (LR) [19,20], artificial neural network (ANN) [3,4,6,21-23], decision tree (DT) [19,24,25], support vector machines (SVM) [6,15,19,26-30], and so on. These methods mainly use mathematical model to establish the quantitative relationship between geological hazards and evaluation factors, which can quantitatively describe the sensitivity of each evaluation factor in different intervals. Meanwhile, the data availability is high and the prediction accuracy is good. However, it is difficult to determine the relationship between each factor and geological hazard point in high dimensional space, and it is easy to overfit. It's also hard for these methods to effectively deal with the multi-source, multi-class, multi-quantity, multi-modal, and multi-temporal geological hazard data accumulated in the long-term geological survey.

Recently, ensemble learning improves the accuracy and generalization ability of the model by integrating multiple weak classifiers into a single strong classifier. As a typical and representative ensemble learning method, random forests (RF) exhibits robust performance in data classification and pattern recognition problems [31]. At the same time, this method does not require the background knowledge of the sample and does not need to choose variables, omitting the tedious work of data pre-processing. Also, it integrates multiple decision trees by random sampling and predicts by majority voting mechanism. Compared with traditional machine learning methods such as ANN and SVM, RF has the advantages of fast classification speed, strong noise resistance and high prediction accuracy. Moreover, the introduction of randomness makes the model not easy to overfit. However, the voting selection mechanism in the RF model will lead to some decision trees with low training accuracy have the same voting ability, reducing the voting accuracy. Moreover, the number of decision trees and other parameters in RF models may also greatly impact the final classification results of RF.

Therefore, the main objective of the current study is to establish a geological hazards susceptibility evaluation model based on optimized random forest (OPRF) with strong processing ability for high-dimensional and large data sets by combining multiple decision trees. For this purpose, the RF model is optimized by optimizing the non-equilibrium data sets, differentiating continuous attributes, and improving the similarity calculation. Next, the out-of-bag (OOB) error estimation is calculated iteratively to find the best random feature and number. After that, taking Lingyun County as the case study, geological hazards susceptibility is divided into four levels for Lingyun County by using the OPRF model. Finally, to evaluate the effectiveness of the proposed OPRF, field investigation and the area under characteristic (AUC) values of the receiver

operating curves (ROC) were used for comparison to the traditional ML classifiers. The evaluation results can provide reference for geological hazard prediction and disaster prevention and mitigation, and also provide decision support for land use development and rational utilization of resources and environment in Lingyun County.

2. STUDY AREA AND DATA TREATMENT

2.1. Study area

Lingyun County is located between longitude 106°23'E to 106°55'E and latitude 24°06'N to 25°37'N in the northwest part of Guangxi, with a total area of about 2048.40km2 and a total population of 193,600, as shown in Figure 1.

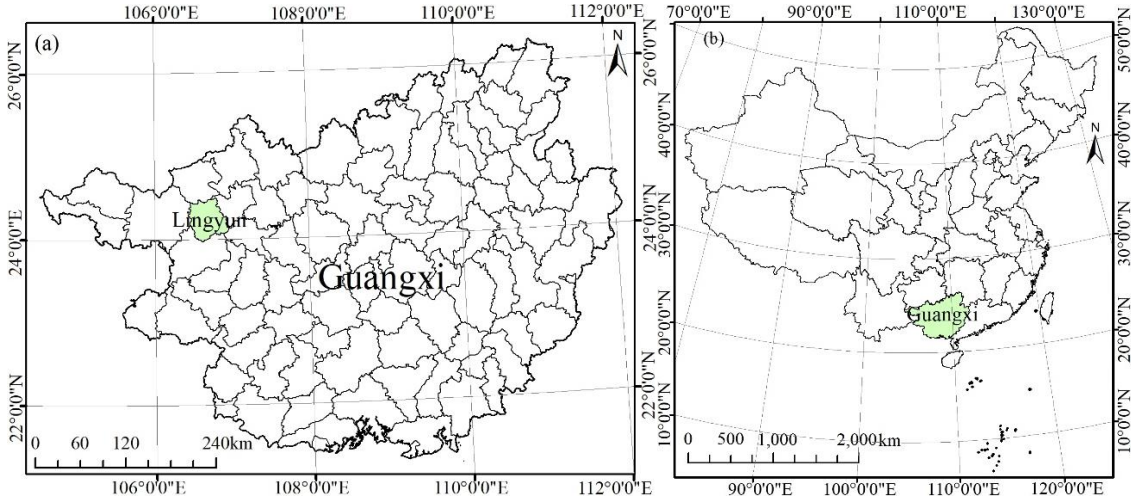


Figure 1. Location of Lingyun County in Guangxi Province (a) and China (b)

It is situated in the transitional zone of the Yungui plateau and the hilly mountainous area of Guangxi. The terrain in the northwest is high and low in the southeast, where in the west it is mostly a clastic rock geomorphology area, and in the east it is mainly a carbonate rock geomorphology. It belongs to a mountainous area with intervening deep valleys, with the mountain area accounting for 93.32% of the total area of the County. There are two main streams and 11 tributaries in the county, which belong to the Youjiang River and Hongshui River. Due to the strong influence of the southern subtropical monsoon and Karst landform, it is under the control of a tropical warm air mass for about half a year. Therefore, heavy rainfalls usually take place during the monsoon season (May to September); it has become one of the heaviest rains centers in Guangxi, and flood disasters occur from time to time in Lingyun County [32].

The intricate tectonic framework formed due to the occurrence of three obvious stages of tectonic evolution in Lingyun County, such as the Caledonian, Indosinian-Yanshan, and Himalayan periods. The exposed strata are mainly clastic rocks of Triassic and Cretaceous, carbonate rocks of Devonian, Carboniferous and Permian, accounting for 29.35% and 31.82% of the total area, respectively. In addition, there is also 16.30% clastic rock intercalated with siliceous rock, 11.35% sandstone, shale, conglomerate, 7.35% clastic rock intercalated with limestone. Late Cretaceous feldspar quartz porphyry veins with striped distribution, and the thin thickness of the Quaternary residual layer is distributed in structural erosion middle-low mountain areas, Karst depressions and valleys.

In general, it is a fragile geological environment zone and is prone to geological hazards in Lingyun County. According to inventory data from the Guangxi Geological Survey Bureau, there are 209 geological hazards in Lingyun County (Figure 2), including landslides, unstable slopes, collapses, dangerous rocks, and so on.

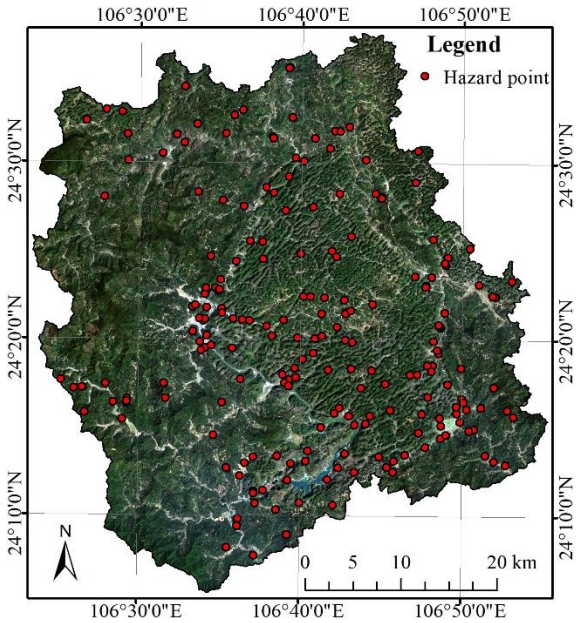


Figure 2. Image of Lingyun County and distribution of geological hazards

2.2. Data source

According to the characteristics of geological hazards and field investigation in Lingyun County, it is found that geological hazards susceptibility is closely related to the characteristics of natural geography, basic geology, ecological environment, human activities, and so on. In the current study, a total of ten geological hazards impacting elements were selected based on the field expedition of Guangxi Geological Survey Bureau as model input variables. They are slope, aspect, topographic curvature, normalized difference vegetation index (NDVI), annual precipitation, strata lithology, tectonic complexity, residential density, road network density, and land use and land cover (LULC). The data adopted in the current study are gathered mainly from the Guangxi Geological Survey Bureau and Guangxi Meteorological Bureau, as shown in Table 1.

Table 1. Data sources of geological hazards impacting elements.

No.	Factors	Data sources and scale
1	Slope	Digital elevation model (DEM) data of 90m
2	Aspect	
3	Topographic curvature	
4	NDVI	Landsat 8 OLI image
5	Annual precipitation	Meteorological data
6	Strata lithology	Geological map with scale 1:50,000
7	Tectonic complexity	
8	Residential density	Topographic map with scale 1:10,000
9	Road network density	
10	LULC	Landsat TM images

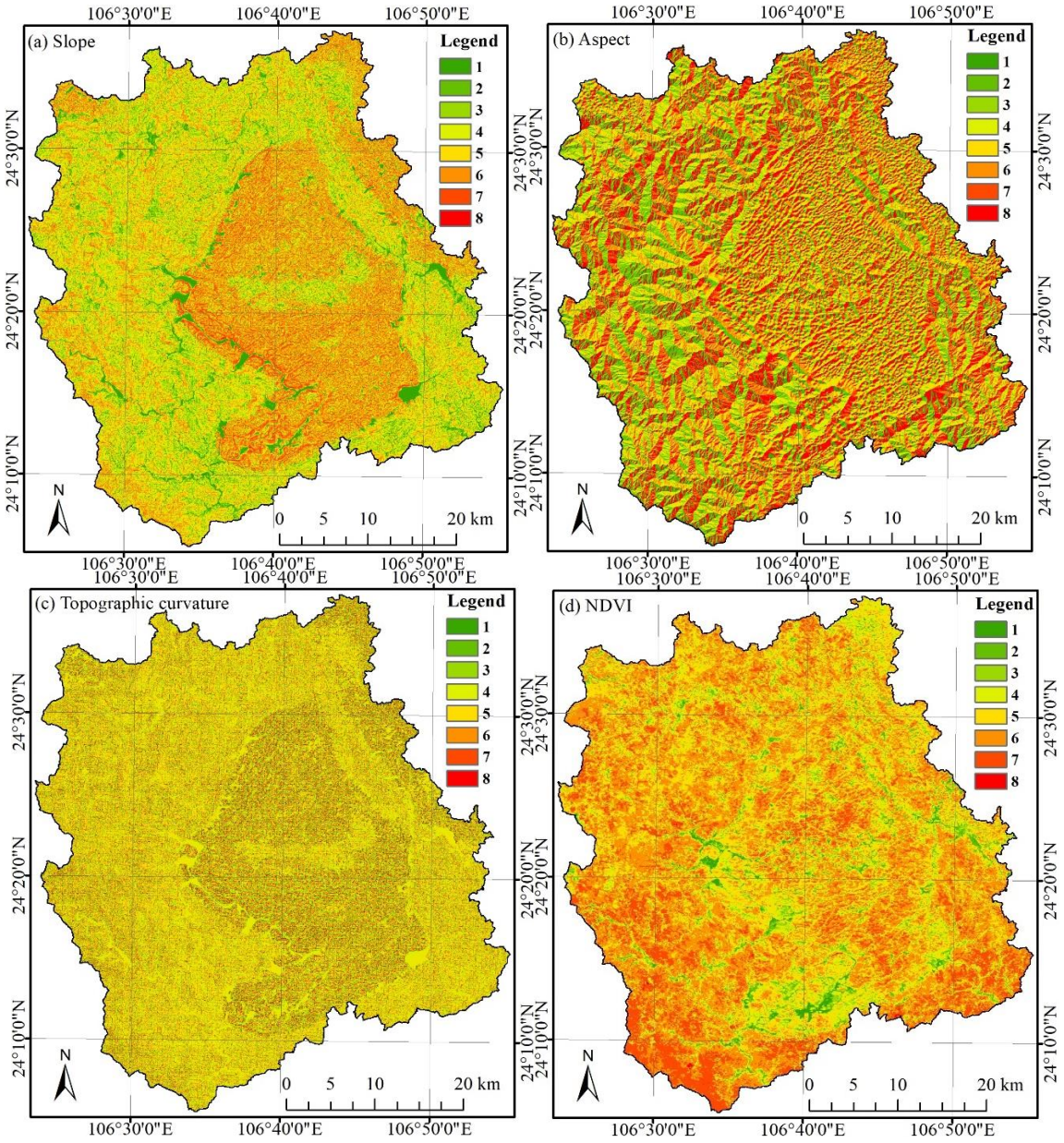
According to the size of geological hazards, this paper adopts a grid with a resolution of 30m×30m as the basic unit for the geological hazards susceptibility assessment, with a total of 2,275,996 evaluation units in Lingyun County.

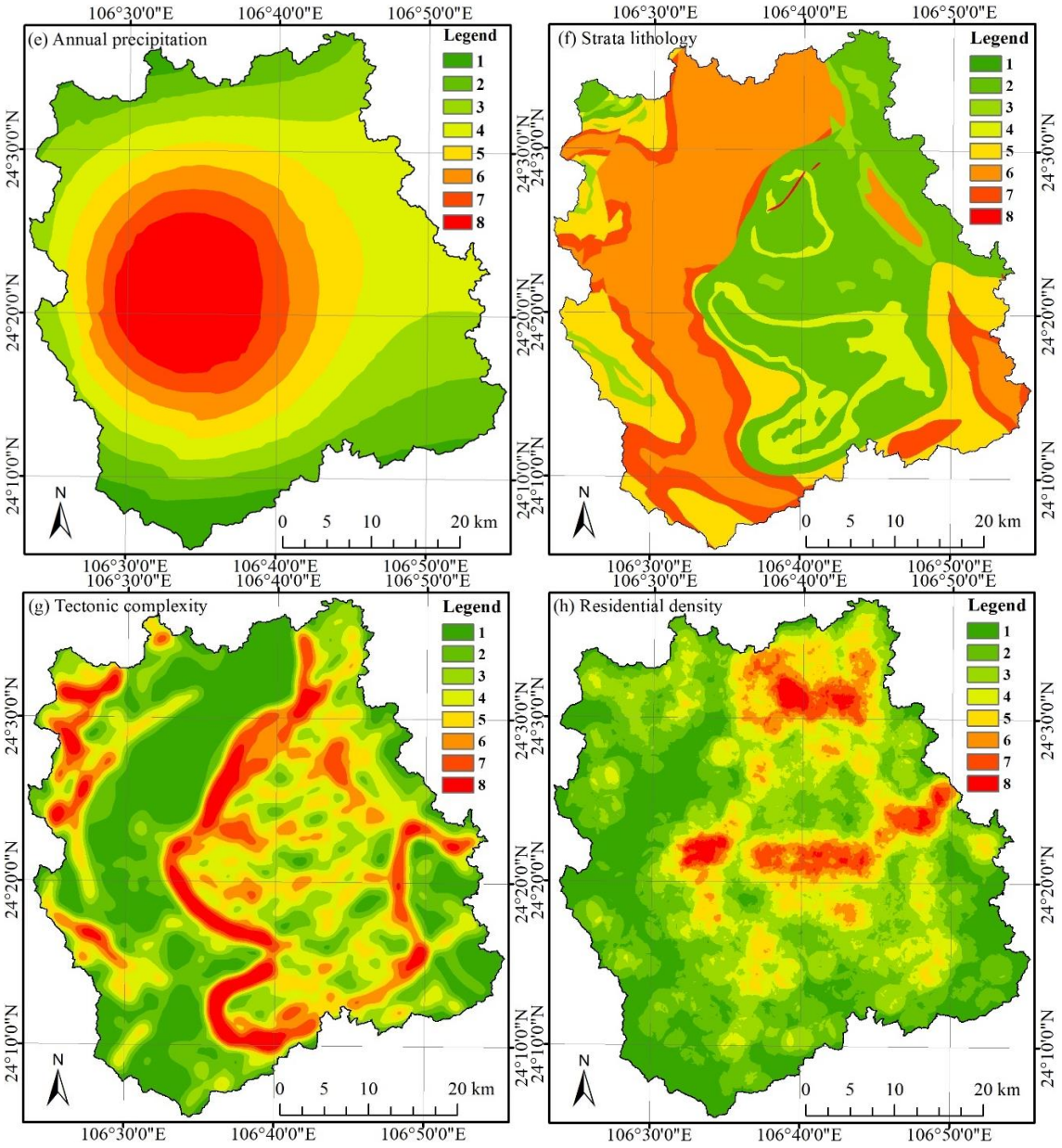
2.3. Treatment and analysis of geological hazards assessment factors

The classification of geological hazards impacting elements is closely related to the evaluation results of geological hazards susceptibility grade. In order to more objectively evaluate the susceptibility of geological hazards, the geological hazards impacting elements have been classified into different levels (Table 2) according to geological hazards characteristic and evaluation criterion developed by Guangxi Geological Survey Bureau for Lingyun County. At the same time, the geological hazards impacting elements of Lingyun County were differentiated, and the distinct effect is shown in Figure 3(a)-(j).

Table 2. Geological hazards impacting elements and their Classification.

No.	Evaluation factor	Classification
(a)	Slope (°)	1-[0,7); 2-[7,13); 3-[13,19); 4-[19,25); 5-[25,34); 6-[34,50); 7-[50,70); 8-[70,79)
(b)	Aspect (°)	1-[337.5,22.5); 2-[22.5,67.5); 3-[67.5,112.5); 4-[112.5,157.5); 5-[157.5,202.5); 6-[205.2,247.5); 7-[247.5,292.5); 8-[292.5,337.5)
(c)	Topographic curvature	1-[-25,-5); 2-[-5,-2.5); 3-[-2.5,-1); 4-[-1,0); 5-[0,1); 6-[1,2.5); 7-[2.5,5); 8-[5,25)
(d)	NDVI	1-[0,0.01); 2-[0.01,0.09); 3-[0.09,0.17); 4-[0.17,0.25); 5-[0.25,0.33); 6-[0.33,0.4); 7-[0.4,0.5); 8-[0.5,0.57)
(e)	Annual precipitation	1-[0,1930); 2-[1930,1990); 3-[1990,2050); 4-[2050,2110); 5-[2110,2170); 6-[2170,2230); 7-[2230,2290); 8-[2290,2350)
(f)	Strata lithology	1-Quaternary; 2-carbonate rock; 3-carbonatite with clastic rock; 4-clastic rock intercalated limestone; 5-clasolite intercalated with siliceous rocks; 6-clastic rock; 7-sandstone, shale, conglomerate; 8-granite or basal rocks
(g)	Tectonic complexity	1-[0,1.4); 2-[1.4,2.7); 3-[2.7,3.8); 4-[3.8,4.9); 5-[4.9,6); 6-[6,7.3); 7-[7.3,8.9); 8-[8.9,14.4)
(h)	Residential density	1-[0,1.2); 2-[1.2,2.4); 3-[2.4,3.5); 4-[3.5,4.5); 5-[4.5,5.8); 6-[5.8,7.1); 7-[7.1,8.6); 8-[8.6,12)
(i)	Road network density (km/km^2)	1-[0,3.2); 2-[3.2,4.7); 3-[4.7,6.1); 4-[6.1,7.8); 5-[7.8,9.7); 6-[9.7,11.7); 7-[11.7,13.9); 8-[13.9,15.3)
(j)	LULC	1-cultivated land; 2-woodland; 3-grassland; 4-river and lake; 5-construction land


(a) Slope
(b) Aspect
(c) Topographic curvature
(d) NDVI
Legend
1
2
3
4
5
6
7
8
106°30'0"E
106°40'0"E
106°50'0"E
24°30'0"N
24°20'0"N
24°10'0"N
N
0 5 10 20 km


(e) Annual precipitation
(f) Strata lithology
(g) Tectonic complexity
(h) Residential density
Legend
1
2
3
4
5
6
7
8
106°30'0"E
106°40'0"E
106°50'0"E
24°30'0"N
24°20'0"N
24°10'0"N
N
0 5 10 20 km

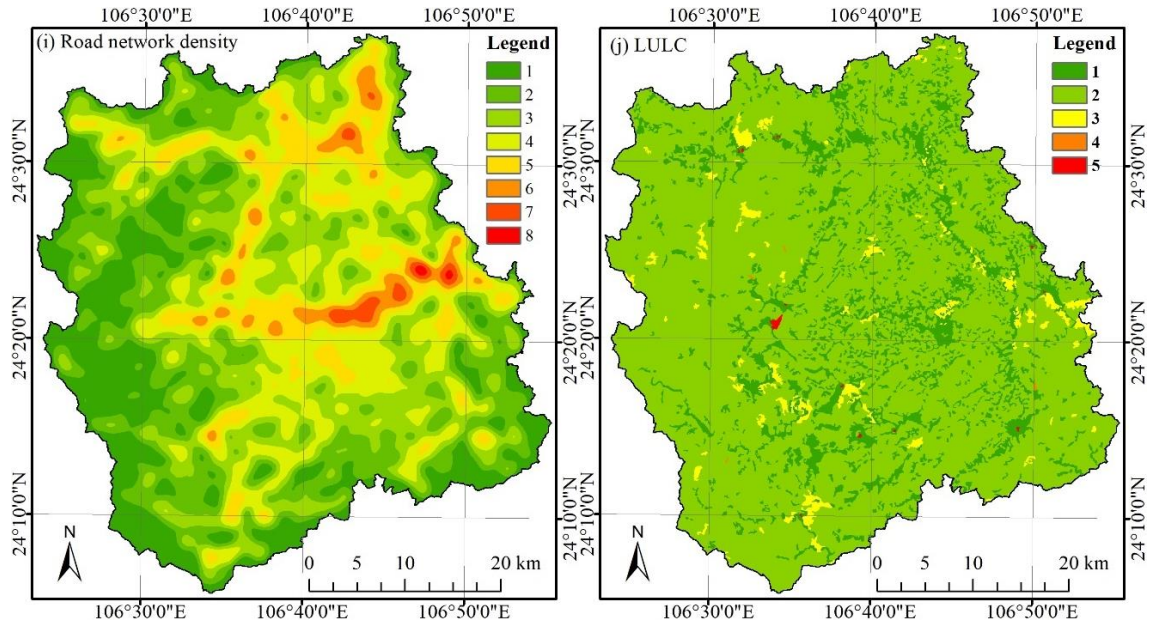


Figure 3. Attribute value of geological hazards evaluation factors [(a) slope, (b) Aspect, (c) Topographic curvature, (d) NDVI, (e) Annual precipitation, (f) Strata lithology, (g) Tectonic complexity, (h) Residential density, (i) Road network density, (j) LULC]

Meanwhile, the information values of each geological hazards impacting element was used to measure the impact of each element on geological hazards; the greater the information value, the greater the impact on geological hazards, which indicates that the higher the probability of occurrence of geological hazards in the region, the higher the susceptibility level [33,34]. The information values of each geological hazards impacting element in Lingyun County are shown in Figure 4.

Slope is an important indicator in the geological hazards survey process to measure the probability of movement of the slope deposits or Quaternary cover [21]. In the current study, the slope, aspect and topographic curvature was extracted from the digital elevation model (DEM) with 30m resolution by ArcGIS, as shown in Figures 3(a)-(c). At the same time, their information value is calculated, as shown in Figures 4(a)-(c). Figure 4(a) shows that the information value of the slope decreases first and then increases with the increase of the slope. This indicates that the impact of the slope with the occurrence of geological hazards also decreases first and then increases, and the impact of the slope with the occurrence of geological hazards is the most significant in the range of 50-70 degrees, followed by 35-50 degrees. Figure 4(b) shows that the information value of the aspect decreases first and then increases and then decreases and then increases with the increase of the aspect, which presents that the impact of the aspect on the occurrence of geological hazards is relatively complex, with the least impact in the range of 67.5-112.5, and the most significant in the range of 292.5-337.5. Figure 4(c) shows that the information value of topographic curvature decreases first and then increases with the increase of the topographic curvature, which states that the effect of the topographic curvature on the occurrence of geological hazards also decreases first and then increases, with the least effect in the range of -1 to 0, and the most significant in the range of 5 to 25.

The vegetation types are diverse, and the forest coverage rate is 71% in Lingyun County because it is a subtropical monsoon forest vegetation area, where the climate is mild, it is wet and rainy, and natural soil fertility is good. The NDVI of Lingyun County were extracted by Landsat 8 OLI image (2017/5/3, 127/043) and ArcGIS, as shown in Figure 3(d). At the same time, its

information value is calculated, as shown in Figure 4(d). Figure 4(d) shows that the information values of NDVI decrease with the increase of NDVI, indicating that the effect of NDVI with the occurrence of geological hazards decrease with the increase of NDVI. That is to say, the better the vegetation cover, the less likely geological hazards will occur.

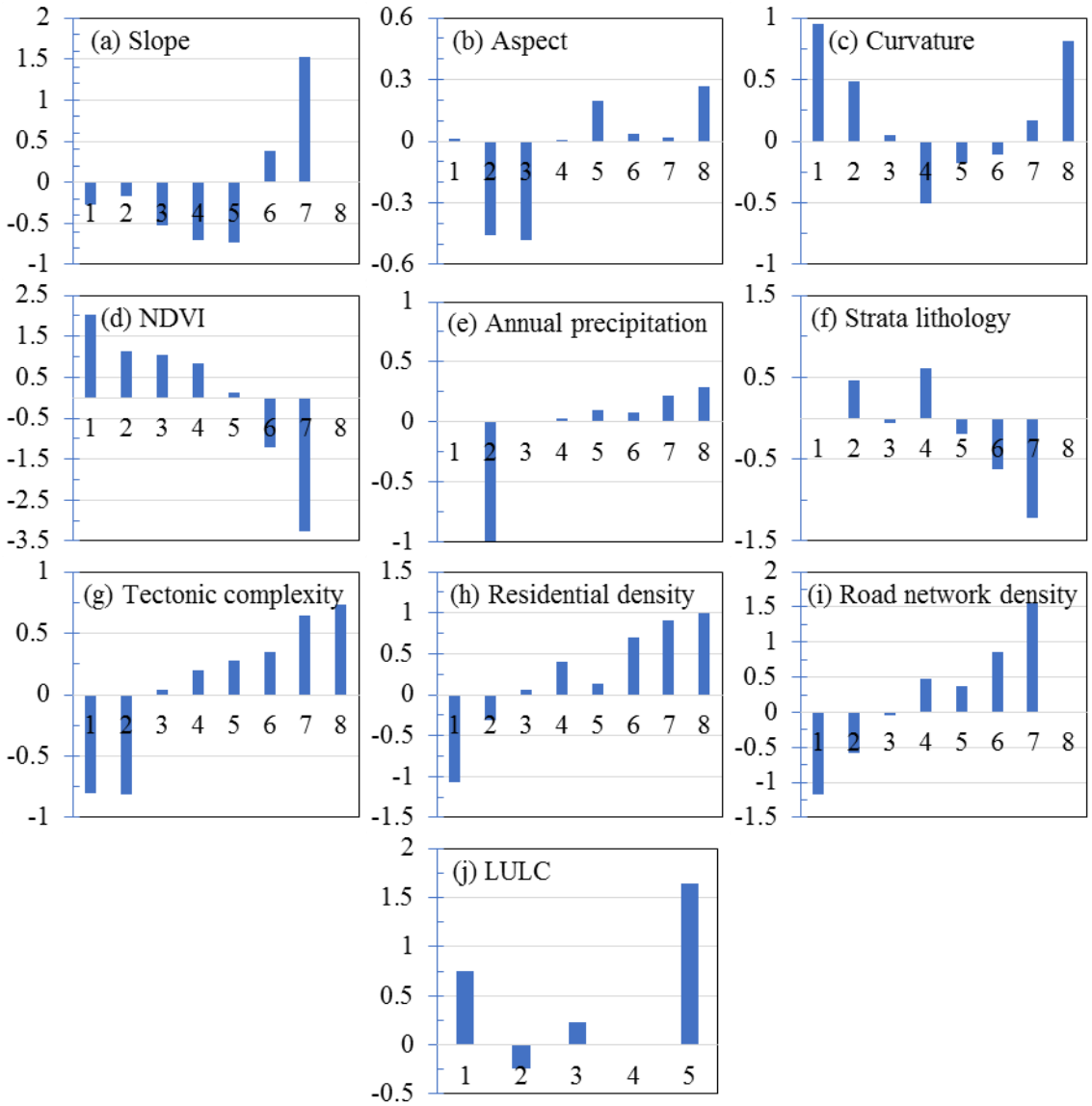


Figure 4. Information values distribution of main geological hazards impacting elements [(a) slope, (b) Aspect, (c) Topographic curvature, (d) NDVI, (e) Annual precipitation, (f) Strata lithology, (g) Tectonic complexity, (h) Residential density, (i) Road network density, (j) LULC]

Precipitation, especially heavy rain or continuous precipitation is the external dynamic factor that induces geological hazards [32]. There is plenty of precipitation in Lingyun County, and the average annual precipitation is 1235 mm. Under the action of precipitation infiltration, scour, and erosion, geological hazards occur from time to time. Meanwhile, the geological hazards and frequent periods of heavy rain are basically the same, indicating that the formation of geological hazards is closely related to heavy rain in Lingyun County [32]. Figure 3(e) is the annual precipitation map and Figure 4(e) is the information value of annual precipitation. Figure 4(e) indicates that the information value of precipitation increases with the increase of the

precipitation, illustrating that the greater the precipitation, the greater the information value, and the greater the impact on the occurrence of geological hazards.

The strata exposed in Lingyun County are mainly carbonate rock and clastic rocks; also there is the clasolite intercalated with siliceous rocks, sandstone, shale, conglomerate, and clastic rock intercalated limestone, and so on. In the carbonate geomorphology area, rock joints and fissures developed, coupled with long-term weathering and dissolution, and rock collapse is easy to occur. In Karst depressions and valleys, it is easy to produce collapse under the action of groundwater [3], because shallow Karst develops and the thin Quaternary is overburdened. The landform of clastic rock is mainly composed of soft mud and shale, alternating with hard sandstone and siltstone. The mud shale is easy to weather and soften when it meets water, so it is easy to form a weak structural surface, resulting in geological hazards such as landslide, collapse and debris flow which are easy to occur. The strata and lithology of Lingyun County is exhibited in Figure 3(f), and the information value is expressed in Figure 4(f). Figure 4(f) indicates that the information value of clastic rock intercalated limestone is the largest, followed by carbonate rock, indicating that clastic rock intercalated limestone and carbonate rock are the most advantageous for the occurrence of geological hazards.

Fault is a zone with fragile structure and is prone to geological hazards [21]. Different periods and different forms of folds and faults with different properties have been formed after the occurrence of three strong crustal movements. At the same time, the later crustal rise suffered erosion and denudation, which caused some early-formed faults to reoccur, resulting in more complex geological structures in Lingyun County. Figure 3(g) states the tectonic complexity and Figure 4(g) states the information value of tectonic complexity in Lingyun County. Figure 4(g) also states that the information value of tectonic complexity increases with the increase of the tectonic complexity, illustrating that the greater the tectonic complexity, the greater the information value, and the greater the impact on the occurrence of geological hazards.

The geological engineering conditions of Lingyun are more complex because they are in the geological engineering environment composed of carbonate rocks, clastic rocks and loose accumulated rocks. As the scope of human activities continues to expand and strengthen, human activities such as steep slope cultivation and engineering construction strongly disturbed the topography and geomorphology in Lingyun County, which led to the occurrence of geological hazards, such as landslide, collapse, collapse, ground fissure, flood, water inrush, leakage, soil erosion, and so on. Residential density and road density of Lingyun were calculated, as exhibited in Figures 3(h)-(i). Meanwhile, their information values were also calculated, as exhibited in Figures 4(h)-(i). Figures 4(h)-(i) exhibit that the greater the density of settlements and roads, the greater the information value, the greater the impact on the occurrence of geological hazards.

In addition, there is 303.83km^2 of arable land and 1687.95km^2 of forest in Lingyun County. Figure 3(j) reveals the LULC map of Lingyun County, and its information value is exhibited in Figure 4(j). Figure 4(j) reveals that the information value of woodland is the smallest, while that of the construction land is the largest, indicating that woodland has the least influence on the occurrence of geological hazards, while construction land has the greatest influence on the occurrence of geological disasters; this further illustrates that the impact of human activities on the occurrence of geological disasters is relatively far-reaching.

2.4. Set up the geological hazards susceptibility assessment database

On the basis of the above, the database of the geological hazards susceptibility evaluation factors in Lingyun County was established, with a total of 2,275,996 grid evaluation units and 209 geological hazards points. Among them, 70% of the geological hazards points (146) were

randomly selected as the geological hazards training samples, the rest 30% of the geological hazards points (63) were selected as the geological hazards testing samples. Accordingly, the non-hazards sample points of 10 times the number of geological hazards points (1460) were randomly selected as the geological hazards training samples, and 630 non-hazards sample points were selected as the geological hazards testing samples. The aim is to reduce the imbalance and spatial autocorrelation between the data of geological hazards points and non-hazards points.

3. METHODS

3.1. RF model

RF is an ensemble learning method that generates a large number of independent training sets and multiple classification and regression trees (CART) by combining bagging [35,36]. The expression of the model is:

$$\{h(X, \theta_k), k = 1,2,3, \dots\} \quad (1)$$

where $h(X, \theta_k)$ is the classification regression tree without pruning generated by the CART algorithm; X is the input vector; $\{\theta_k\}$ is the random vector of independent distribution.

In geological hazards susceptibility assessment, first, 146 geological hazards and 1460 non-hazards sites samples were randomly selected by the bagging method as independent spatial training sets. Secondly, 10 geological hazards impacting elements were randomly selected for internal node branching without pruning to separately set up the CART tree for each training set [35, 36]. Thirdly, the other unselected 63 geological hazards and 630 non-hazards sites data as OOB data were to estimate the OOB error for each tree, and the OOB error for all trees is averaged to the RF [37]. Finally, the class with the most votes is taken as the geological hazards assessment result by synthesizing all decision trees. The specific implementation process is shown in Figure 5.

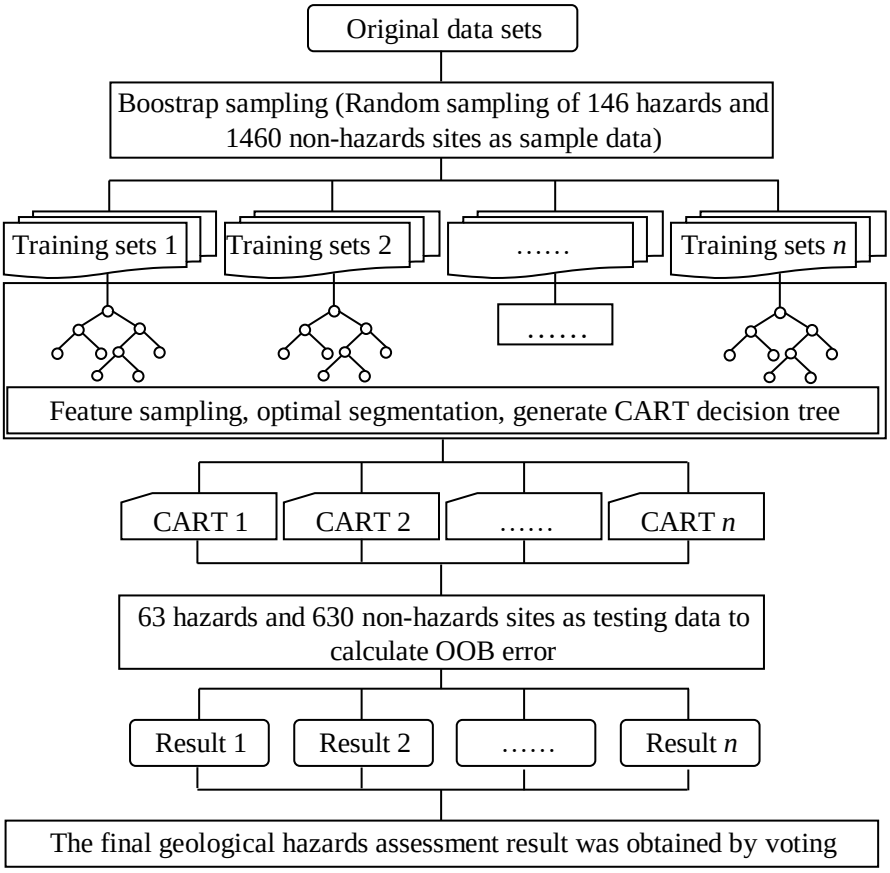


Figure 5. Diagram of RF Algorithm

OOB error consists of unbiased estimates, approximated by cross-validation errors, and is bounded by generalization errors in RF [38]:

$$P^* \leq \bar{\rho}\frac{(1-s^2)}{s^2} \quad (2)$$

where P* is the generalization error of the RF; $\bar{\rho}$ is the average of the correlation between CART trees; s is the average intensity of the decision tree.

Formula (2) illustrates that to enhance the generalization ability of RF, it can weaken the correlation between decision trees or increase the intensity of decision trees. For this purpose, this study introduces randomness to the feature selection of CART trees to weaken the correlation between decision trees.

The specific steps are as follows: (1) m features were randomly selected, ($m \leq 10$); (2) according to the principle of minimum non-purity of nodes, the optimal features are selected from these m characteristics to split the nodes; (3) the intensity and correlation of the CART tree are affected by m [38]. When m is too small, the intensity of the CART tree is weak; when m is too large, the intensity of CART tree increases, but the correlation between CART trees also increases.

In addition, this study further optimizes the RF model by optimizing the non-equilibrium data sets, differentiation of continuous attributes, and improving the similarity calculation of RF samples.

3.2. Optimization of non-equilibrium data sets

The sample data in the geological hazards susceptibility evaluation of Lingyun County are typically unbalanced data, because the number of geological hazards sites is far less than that of non-hazards sites, based on field investigations by the Guangxi Geological Survey Bureau.
In order to improve the evaluation accuracy, the C_SMOTE algorithm was applied to solve the non-equilibrium problem of the sample data. The steps are as follows:

(1) Calculate the central of the hazards sites, recorded as X_{center}. The formula is as follows:

$$X_{center} = \left(\frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1}, \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i2}, \ldots\ldots, \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ir} \right) \quad (3)$$

(2) Synthesis of "artificial" samples. The formula is as follows:

$$p_j = X_i + rand(0,1) \times (X_{center} - X_i) \quad (4)$$

where n_1 is the total sample number of the hazards sites; r is the attribute of each sample; $X_i\ (i = 1, 2, \ldots\ldots, n_1)$ is the hazards sites sample; X_{center} is the center of the hazards sites sample; $p_j\ (j = 1, 2, \ldots\ldots, m)$ is the synthetic "artificial" sample; and $rand(0,1)$ is a random number within the interval (0, 1).

(3) If the synthetic hazards sites sample number exceeds the actual required sample number, then use the under-sampling method to remove some samples far away from the center, finally, make the synthesized sample number reach the required equilibrium rate.The flow chart is shown in Figure 6:

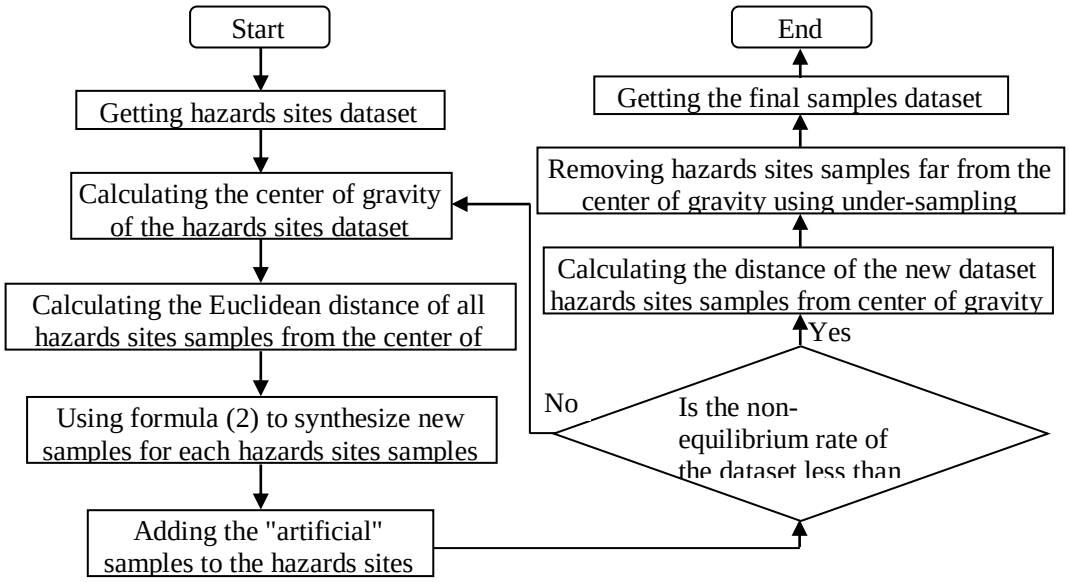


Figure 6. Flow chart of the C_SMOTE algorithm

3.3. Differentiation of continuous attributes

There are 2 discontinuous attribute elements and 8 continuous attribute elements for geological hazards susceptibility evaluation in Lingyun County. To improve the accuracy of the RF model, this study adopts the entropy based on minimal description length principle (Ent-MDLP) to differentiate the attribute values of continuous evaluation factors. The steps are as follows:

(1) Dichotomy recursion to find breakpoints. First of all, find the adjacent points of different classes, and takes the midpoint between them as the candidate breakpoint T; secondly, each candidate breakpoint T can divide the sample set R into two subsets, calculate the information entropy of the two subsets respectively, then weight the summation to obtain the classification information entropy E(A,T,R); Finally, take the breakpoint T that makes the classification information entropy minimum as the final selected breakpoint.

(2) Determine the recursive downtime condition. The minimal description length principle (MDLP) is introduced here [38], and the downtime condition is that the information gain G should be satisfied:

$$\begin{aligned} G(A,T,R) = E(R) - E(A,T,R) &= E(R) - \frac{|R_1|}{N \times E(R_1)} - \frac{|R_2|}{N \times E(R_2)} \\ &> \frac{\log_2(N-1)}{N} + \log_2(3^k - 2) - [k \times E(R) - k_1 \times E(R_1) - k_2 \times E(R_2)] \end{aligned} \quad (5)$$

where A is an input variable, T is a breakpoint, R is a sample set, N is the total sample size, k is the number of categories; E(R) is the entropy of the sample set R; $E(R_1)$ and $E(R_2)$ are the entropy of the instance set R_1 and R_2 in each subinterval; and k_1 and k_2 are the number of categories in each subinterval.

3.4. Improving the similarity calculation of RF samples

It is an outstanding advantage of RF over other classifiers that RF can calculate the degree of similarity between samples and obtain the similarity matrix between samples. The similarity between the two samples can be measured by the frequency at which the two samples appear on the same leaf node on each tree, or by the probability that the two samples belong to the same class.

Assuming that the number of samples is N, the calculation process of the similarity matrix is as follows: First, the sample similarity matrix prox(i,j) is initialized as the all-zero matrix of N row N column. Then, all samples are discriminated with each tree generated, and each sample falls on one of the leaf nodes of the tree. Finally, for samples i and j, if they all land on the same leaf node of the tree, add 1 at row i and column j corresponding to the prox(i,j) matrix. Meanwhile, for the similarity between samples falling on different leaf nodes, the prox(i,j) matrix is improved by calculating the distance (d) between different leaves. The formula is as follows:

$$prox(i,j) = prox(i,j) + \frac{1}{d^m} \quad (6)$$

where d is the distance between the leaf nodes of sample i and j, and m is any positive real number.

The above procedure is repeated for each tree in the RF, traverse each tree to get a total addition value. Then each element is divided in the prox(i,j) matrix by the total number of trees to get the final prox(i,j) matrix. It is a symmetric matrix of N row N columns for prox(i,j) matrix, in which the diagonal elements are all 1, and the element of prox(i,j) in line i column j is defined as the similarity between sample i and sample j.

3.5. Set up the geological hazards susceptibility assessment model based on OPRF

In order to find the optimal random feature number, the OOB error of OPRF with a different random feature number was calculated by the cyclic iterative method, as shown in Figure 7.

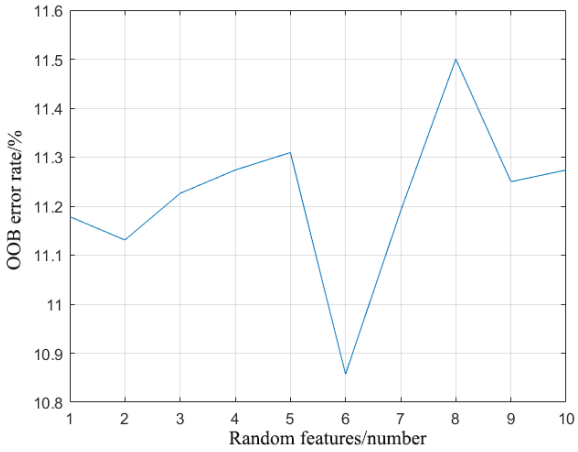


Figure 7. OOB Error distribution of OPRF with different numbers of random feature

Figure 7 indicates the variation characteristic of the OOB error with the increase of random feature number. When the number of random features is 6, the OOB error is the smallest, which indicates that the prediction accuracy of the geological hazards susceptibility evaluation model based on the OPRF established in the present study is the highest. Currently the number of decision trees in this OPRF is 81 and the maximum depth is 20.

4. RESULTS AND DISCUSSIONS

4.1. Model evaluation metrics

Model precision and validation analysis is one of the essential steps for geological hazards susceptibility assessment and prediction [39]. Here, to test and verify the improvements and scientific significance of the proposed method in the current study, the proposed OPRF model was recommended and compared for comprehensive performance comparison with the RF and three other models, including LR, ANN, and SVM. The remaining 30% testing samples were used to test the five models, and the receiver operating characteristics (ROC) curves and the area under the curve (AUC) of each model prediction result were calculated (Figure 8), which is a widely used independent performance valuator [40-42]. The prediction performance is assessed by the AUC compared with the total plot area. If the AUC is equal to 1, it represents excellent prediction capability, while the AUC close to 0.5 represents a poor prediction capability [6, 8, 24, 30, 43, 44]. Figure 8 exhibits the ROC curves of the LR, ANN, SVM, RF and OPRF models in the current study.

Figure 8 shows that the AUC values of the LR, ANN, SVM, RF and OPRF models are 0.766, 0.814, 0.842, 0.846 and 0.934, respectively, which states that the prediction accuracy of five models for geological hazards susceptibility assessment in Lingyun County are 76.6%, 81.4%, 84.2%, 84.6%, and 93.4%, respectively. This result demonstrates that the geological hazards susceptibility assessment model based on OPRF has the highest prediction accuracy. which is mainly owing to the large number of elements selected in present study, the OPRF model, a type of ensemble learning, presented superiorities over a traditional method by not only accounting for different types of elements but also assessing the relative importance of the elements in terms of geological hazards stability [25]. At the same time, the result also demonstrates that the improvements proposed in the current study increase the performance of the RF model in evaluating and predicting the geological hazards susceptibility. Consequently, the OPRF model can be applied to the geological hazards susceptibility assessment under the same natural ecological environment.

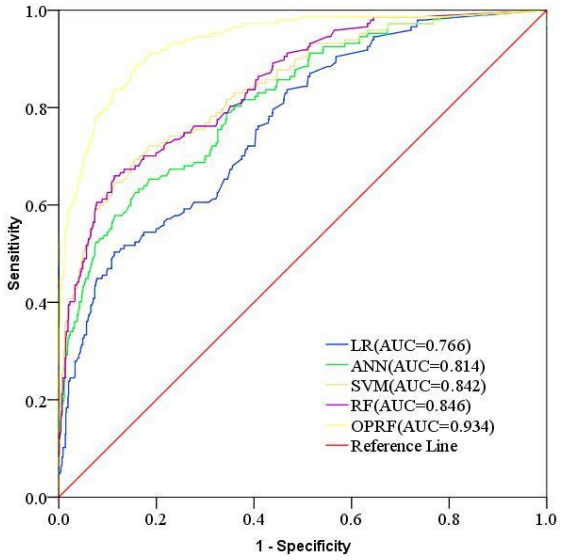


Figure 8. ROC curves and AUC values of test set for LR, ANN, SVM, RF and OPRF models

4.2. Evaluation results

The geological hazards susceptibility index of Lingyun County is calculated between 0 and 1, using the OPRF model, corresponding to the geological hazards susceptibility from low to high. At the same time, the Ent-MDLP method was used for the grading treatment, which was divided into four grades: [0-0.6776], (0.6776-0.7074], (0.7074-0.7372], and (0.7372-1], corresponding to the non-prone region, low-prone region, middle-prone region, and high-prone region, as shown in Figure 9.

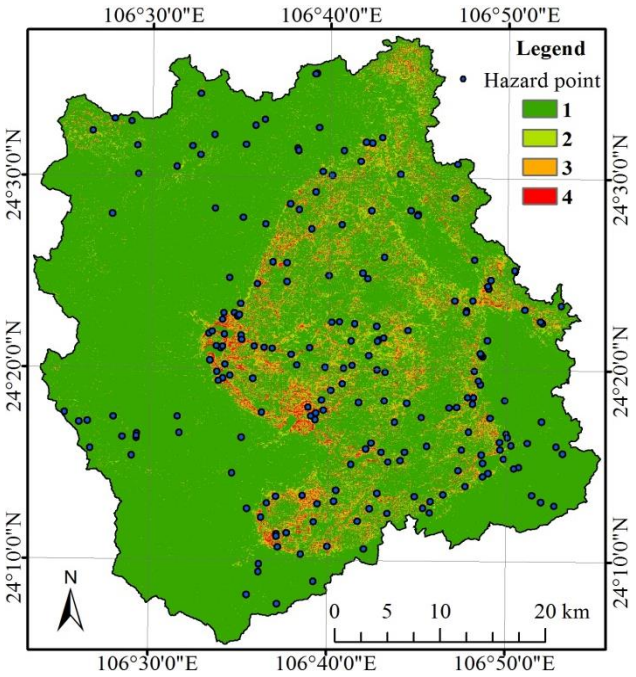


Figure 9. Evaluation results of geological hazards susceptibility in Lingyun County

Figure 9 shows that the high-prone region of geological hazards is 59.93 km^2, accounts for 2.93% of the total area in Lingyun County, mainly distributed in the regions of the carbonate rocks, where the slope is steep at 34-70 degrees, the aspect is between 292.5-337.5 and 157.5-202.5 degrees, the topographic curvature is between ±5-25°, vegetation coverage is low, the geological tectonic is complex, and the density of residents and road network is large. These regions are affected by multi-stage tectonic movement, which makes the joint fissure of rock mass develop, and the rock differentiation is strong, causes the frequent occurrence of disasters such as dangerous rocks, unstable slope, landslide, and collapse, indicating that carbonate rocks have a profound influence on the stability of geological hazards in the region. At the same time, there are many towns and traffic lines in these regions, indicating that these regions are strongly influenced by human activities.

The middle-prone region of geological hazards is 93.44 km^2, accounts for 4.56% of the total area in Lingyun County, mainly distributed in the regions of clastic rocks, clastic rock intercalated with limestone, and clastic rock intercalated with siliceous rock. Here the slope is from 7 to 34 degrees, vegetation coverage is low, and moderate density of population and road network. These regions have poor rock stability and strong weathering erosion, which provide a good material basis for the development of geological hazards.

The low-prone region of geological hazards is 139.31km^2, accounts for 6.8% of the total area in Lingyun County, mainly distributed near rural settlements where the rock mass is stable, the vegetation covers well, and is less disturbed by human activities.

The remaining region is the non-prone area of geological hazards, accounts for 85.71% of the total area in Lingyun County, where the rock mass is stable, the vegetation coverage is high, and is rarely affected by human activities to maintain its original natural ecological environment.

Figure 9 also indicates that the occurrence of geological hazards has a strong correlation with the vegetation index, road network density, and residential density, indicating the far-reaching impact

of human activities on the occurrence of geological hazards in Lingyun County. It also indirectly illustrates that the construction of human engineering strongly interferes with the natural ecological environment of the region and leads to the frequent occurrence of geological hazards. Therefore, the research results of the current study also suggest that the stability and carrying capacity of the regional natural environment system should be fully considered in human engineering construction.

5. CONCLUSIONS

Geological hazards susceptibility evaluation is considered as an important task of geological hazards survey and is also the first important step in geological hazards risk assessments. Therefore, it is essential to accurately assess and predict geological hazards susceptibility regions with high performance-based models. Since performance of all kinds of proposed methods and techniques for simulating geological hazards is still being discussed, explorations of new methods for the evaluation of geological hazards are highly essential. These explorations will help obtain enough background knowledge to achieve some rational conclusions. The rapid development of advanced machine-learning allows for systems such as RF with high accuracy and better overall performance; use of these is recommended in disaster assessment and prediction. In this current study, the geological hazards susceptibility evaluation model based on OPRF was set up to assess and divide the hazards levels for Lingyun County. Meanwhile, field investigation and ROC curve were used to verify the evaluation results. The following conclusions have been reached in this study:

(1) The C_SMOTE algorithm is re-sampled on the line between the negative sample of the geological hazards point and the gravity center of the data set, so that the newly generated "artificial" sample is always between the center point and the negative sample of the geological hazards point; its position is determined by a random number, so it will not deviate from the geometric space of the negative sample set of the geological hazards point, and so it will not produce the tendency of marginalization, but will be directed towards the center point, thus reducing the randomness and blindness.

(2) The Ent-MDLP can better solve the differentiation problem when continuous geological hazards factors are increased and there is a lack of enough experience in the geological hazards susceptibility evaluation. At the same time, the discrete results show obvious trend characteristics and avoid the inconvenience of RF randomness to continuous factor analysis.

(3) When calculating the similarity between samples, for the similarity between samples falling at different leaf nodes, the loss of the sample similarity measure caused by “one-size-fits-all” is avoided by calculating the path distance d between different leaves to improve similarity matrix prox.

(4) The optimal random characteristic number is determined by finding the smallest OOB error of OPRF under different random characteristic numbers, which is calculated by iterative method.

(5) AUC values of the ROC curves and field investigation proved that the prediction accuracy of the geological hazards susceptibility evaluation model based on OPRF is higher than the original RF and the other three models.

In general, the improvements proposed in the current study aim to improve the accuracy and overall performance of the RF model for the geological hazards susceptibility evaluation. The RF model is improved in three aspects: optimization of unbalanced geological hazards data sets, differentiation of continuous geological hazards evaluation factor and the sample similarity

calculation. On this basis, the geological hazards susceptibility evaluation model was set up based on OPRF. At the same time, the geological hazards susceptibility evaluation model was optimized by iteratively calculating the OOB error to find the best number of random features. Finally, geological hazards susceptibility is assessed by using the OPRF model, and the geological hazards susceptibility levels of Lingyun County are divided. Meanwhile, the accuracy and overall performance of evaluation results is verified by field investigation and ROC curves. The results indicate that the optimization strategies proposed in the current study are effective for the RF model. Furthermore, the OPRF can be expanded to the geological hazards susceptibility evaluation under the same natural ecological environment.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (Grant No.41201193), Guizhou Science and Technology Planning Project: Research and Application of Three-dimensional Prediction System for Gold Deposit in Southwest Guizhou Based on Geological Big Data ([2020]4Y039); Research and Development of Big Data Management and Intelligent Processing System for Manganese Ore Exploration ([2017]2951); Evaluation and Demonstration of Three-Dimensional Geological Survey of Deep Gold Mine in Guizhou Province (2020196700); Open fund project of National-Local Joint Engineering Laboratory on Digital Preservation and Innovative Technologies for the Culture of Traditional Villages and Towns (CTCZ19K01), and Open research project of key laboratory of Tectonics and Petroleum Resources (China University of Geosciences), Ministry of Education (No.TPR-2019-11). The authors would like to thank the anonymous reviewers for providing valuable comments on the manuscript.

REFERENCES

[1] Huang, R., Xu, X., Tang, C., & Xiang, X. (2008) Geological Environmental Assessment and Geological Hazard Management, Beijing, Science press.

[2] Sharma, S., & Mahajan, A. K., (2019) “A comparative assessment of information value, frequency ratio and analytical hierarchy process models for landslide susceptibility mapping of a Himalayan watershed, India”, B. Eng. Geol. Environ, Vol.78, pp2431–2448.

[3] Sun, P., Cai, R., Xie, C., & Yi, Z. (2019) “Slope stability evaluation based on genetic optimization neural network”, Mod. Electron. Tech, Vol.42, pp75–78.

[4] Wang, Y., Fang, Z., Wang, M., Peng, L., & Hong, H. (2020) “Comparative study of landslide susceptibility mapping with different recurrent neural networks”, Comput. Geosci, Vol.138, pp104445.

[5] Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S., & Al-Katheeri, M.M. (2016) “Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at wadi Tayyah Basin, Asir region, Saudi Arabia”, Landslides No.13, pp839–856.

[6] Tien Bui, D., Tuan, T. A., Klempe, H., Pradhan, B., & Revhaug, I. (2016) “Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree”, Landslides, No.13, pp361–378.

[7] Myronidis, D., Papageorgiou, C., & Theophanous, S. (2016) “Landslide susceptibility mapping based on landslide history and analytic hierarchy process (AHP)”, Nat. Hazards Vol.81, pp245–263.

[8] Ciurleo, M., Mandaglio, M. C., & Moraci, N. (2019) “Landslide susceptibility assessment by TRIGRS in a frequently affected shallow instability area”, Landslides No.16, pp175–188.

[9] Sezer, E. A., Nefeslioglu, H. A., & Osna, T. (2017) “An expert-based landslide susceptibility mapping (LSM) module developed for Netcad Architect Software”, Comput. Geosci, Vol.98, pp26–37.

[10] Bourenane, H., Guettouche, M. S., Bouhadad, Y., & Braham, M. (2016) “Landslide hazard mapping in the Constantine city, Northeast Algeria using frequency ratio, weighting factor, logistic regression, weights of evidence, and analytical hierarchy process methods”, Arab. J. Geosci, No.9, pp1–24.

[11] Achour, Y., Boumezbeur, A., Hadji, R., Chouabbi, A., Cavaleiro, V., & Bendaoud, E.A. (2017) "Landslide susceptibility mapping using analytic hierarchy process and information value methods along a highway road section in Constantine, Algeria", Arab. J. Geosci, No.10, pp194–209.

[12] Hung, L.Q., Van, N.T.H., Duc, D.M., Ha, L.T.C., Son, P.V., Khanh, N.H., & Binh, L.T. (2016). "Landslide susceptibility mapping by combining the analytical hierarchy process and weighted linear combination methods: a case study in the upper lo river catchment (vietnam)", Landslides, Vol.13 No.5, pp1285-1301.

[13] Wang, X., Zhang, L., Wang, S., & Lari, S. (2014). "Regional landslide susceptibility zoning with considering the aggregation of landslide points and the weights of factors", Landslides, Vol.11, No.3, pp399-409.

[14] Liao, L., Zhu, Y., Zhao, Y., Wen, H., Yang. Y., Chen. L., Ma. S., & Xu. Y. (2019) "Landslide integrated characteristics and susceptibility assessment in Rongxian county of Guangxi, China", J. Mt. Sci, No16, pp657–676.

[15] Mokhtari, M., & Abedian, S. (2019) "Spatial prediction of landslide susceptibility in Taleghan basin, Iran", Stoch. Environ. Res. Risk Assess, Vol.33, pp1297–1325.

[16] Chen, W., Fan, L., Li, C., & Pham, B. T. (2020) "Spatial prediction of landslides using hybrid integration of artificial intelligence algorithms with frequency ratio and index of entropy in Nanzheng county, China", Appl. Sci, No10, pp29.

[17] Li, Y., Mei, H., Ren, X., Hu, X., & Li, M. (2018) "Geological disaster susceptibility evaluation based on certainty factor and support vector machine", J. Geo-info. Sci., Vol.20 No12, pp1699-1709.

[18] Zheng, Y., Chen, J., Wang, C., & Cheng, T. (2020) "Application of certainty factor and random forests model in landslide susceptibility evaluation in Mangshi City, Yunnan Province", B. Geol. Sci. Tech., Vol.39, No.6, pp131-144.

[19] Wang, F., Yin, K., Gui, L., & Chen L. (2018) "Landslide hazard analysis under different daily rainfall conditions in Wanzhou District", J. Geo-info. Sci., Vol.37 No.1, pp190-195.

[20] Hu, T., Fan, X., Wang, S., Guo, Z., Liu, A., & Huang, F. (2020) "Landslide susceptibility evaluation of Sinan County using logistics regression model and 3S technology", B. Geol. Sci. Tech., Vol.39, No2, pp113-121.

[21] Xu, K. Guo, Q., Li, Z., Xiao, J., Qin, Y., Chen, D., & Kong, C. (2015) "Landslide susceptibility evaluation based on BPNN and GIS: a case of Guojiaba in the Three Gorges Reservoir Area", Int. J. Geogr. Inf. Sci, Vol.29, pp1111–1124.

[22] Zhou, C., Yin, K., Cao, Y., Ahmed, B., Li, Y., Catani, F., & Pourghasemi, H. R. (2018) "Landslide susceptibility modeling applying machine learning methods: a case study from Longju in the Three Gorges Reservoir area, China", Comput. Geosci, Vol.112, pp23–37.

[23] Lee, D.H., Kim, Y.T., & Lee, S.R. (2020) "Shallow landslide susceptibility models based on artificial neural networks considering the factor selection method and various non-linear activation functions", Remote Sens, No12, pp1194.

[24] Hong, H. Liu, J., Tien Bui, D., Pradhan, B., Acharya, T.D., Pham, B.T., Zhu, A.X., Chen, W., & Ahma, B.B. (2018) "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)", Catena Vol.163, pp399–413.

[25] Zhang, K., Wu, X., Niu, R., Yang, Y., & Zhao, L. (2017) "The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China", Environ. Earth Sci, Vol.76, pp405.

[26] Li, X., Cheng, X., & Chen, W. (2015) "Identification of forested landslides using LiDar data, object-based image analysis, and machine learning algorithms", Remote sens., Vol.7, No.8, pp9705-9726.

[27] Chen, Q., Liu, G., Ma, X., Zhang, J., & Zhang, X. (2019) "Conditional multiple-point geostatistical simulation for unevenly distributed sample data", Stoch. Env. Res. Risk A, Vol.33, pp973–987.

[28] Nguyen, H., Bui, X.N., Choi, Y., Lee, C.W., & Armaghani, D.J. (2020) "A novel combination of whale optimization algorithm and support vector machine with different kernel functions for prediction of blasting-induced fly-rock in quarry mines", Nat. Resour. Res, https://doi.org/10.1007/s11053-020-09710-7.

[29] Yu, X., & Gao, H. (2020) "A landslide susceptibility map based on spatial scale segmentation: A case study at Zigui-Badong in the Three Gorges Reservoir Area, China", PLOS ONE Vol.15, ppe0229818.

[30] Pham, B. T., Pradhan, B., Tien Bui, D., Prakash, I., & Dholakia, M. B. (2016) "A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India)", Environ. Modell. Softw, Vol.84, pp240–250.

[31] Chen, W., Li, X., Wang,Y., Chen, G., & Liu, S. (2014) "Forested landslide detection using LiDAR data and the random forest algorithm: A case study of the Three Gorges, China", Remote Sen Environ., Vol.152, pp291-301.
[32] Zhang, L., Shi, S., & Liu, Q. (2016) "Spatial-temporal distribution characteristics and genetic analysis of geological disasters in Guangxi", Guangxi Water Resour. Hydropower. Eng, No.6, pp64–67.
[33] Murat, E., & Candan, G. (2004) "Use of fuzzy relations to produce landslide susceptibility map of landslide prone area (West Black Sea Region, Turkey)", Eng. Geol, Vol.75, pp229–250.
[34] Chen, L., Ye, J., Wei, C., & Xu, Y. (2016) "Application of ArcGIS and information method to landslide susceptibility evaluation", J. Guangxi Univ, Vol.41, pp141–148.
[35] Breiman, L. (1996) "Bagging predictors", Mach. Learn, Vol.24, pp123–140.
[36] Breiman, L. (2001) "Random forests", Mach. Learn, Vol.45, pp5–32.
[37] Catani, F., Lagomarsino, D., Segoni, S., & Tofani, V. (2013) "Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues", Nat. Hazards Earth Syst. Sci, Vol.13, pp2815–2831.
[38] Liu, J., Li, S., & Chen, T. (2018) "Landslide susceptibility assessment based on optimized random forest model", Geomat. Inf. Sci. Wuhan Univ, Vol.43, pp1085–1091.
[38] Frattini, P., Crosta, G., & Carrara, A. (2010) "Techniques for evaluating the performance of landslide susceptibility models", Eng. Geol, Vol.111, p 62–72.
[40] Hanley, J.A., & McNeil, B.J. (1983) "A method of comparing the areas under receiver operating characteristic curves derived from the same cases", Radiology, Vol.148, pp839–843.
[41] Swets, J. A. (1988) "Measuring the accuracy of diagnostic systems", Science, Vol.240, pp1285–1293.
[42] Fielding, A.H., & Bell, J.F. (1997) "A review of methods for the assessment of prediction errors in conservation presence/absence models", Environ. Conserv, Vol.24, pp38–49.
[43] Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A.C., & Peruccacci, S. (2010) "Optimal landslide susceptibility zonation based on multiple forecasts", Geomorphology, Vol.114, pp129–142.
[44] Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Tien Bui, D., Duan, Z., & Ma, J. (2017) "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility", Catena, Vol.151, pp147–160.

AUTHORS

Chunfang Kong, Ph.D., Associate Professor, My current research interests in remote sensing of the resource environment, data mining and processing, machine learning, and GIS applications.