

Geological hazard susceptibility assessment based on different mathematical statistical models: a case study of Zhejiang province

Zihang Wang ^a, Jingwei Tong ^b, Yinkai Niu ^c, Xudong Zheng ^d

Faculty of Civil Engineering and Mechanics, Kunming University of Science and Technology,
Kunming 650500, China

^ajknnwzh@163.com, ^b2910287625@qq.com, ^c2035354859@qq.com,

^dzhengxudong0571@outlook.com

Abstract. Geological disaster is a great threat to people's life and property safety, how to predict it reliably and divide it according to the susceptibility of geological disaster is of great significance. Based on the historical collapse data of Zhejiang Province, this paper selects the evaluation indexes of water flow intensity index (SPI), terrain wetness index (TWI), land use type, slope aspect, slope, annual rainfall, river network density, elevation, road density, plane curvature, normalized difference vegetation index (NDVI) and other factors. Based on the frequency ratio model (FR), random forest model (RF) and logistic regression model (LR), the susceptibility zoning of geological disasters in Zhejiang Province is carried out. The results show that: (1) The high and extremely high risk areas in Zhejiang province are mainly distributed near the river system and road network in the high altitude area; (2) In terms of the evaluation effect of comprehensive PR curve, the random forest model has a better evaluation effect than other models; (3) With the increase of the vulnerability level, the proportion of disasters under each model zoning is gradually increasing, indicating that the zoning effect of each model is good; 4) Quzhou city, Jinhua City, Wenzhou city and Shaoxing city in Zhejiang Province account for a large proportion of the high susceptibility areas and above. The results of vulnerability zoning can provide reference for geological hazard prevention and land planning in Zhejiang province.

Keywords: frequency ratio, Logistic regression, Random Forest, Geological hazard susceptibility, Zhejiang Province.

1. Introduction

With the rapid development of China's economy in recent years, the scale of urbanization in China is also growing. By the end of 2021, China's urbanization rate had exceeded 60% [1]. Rapid urbanization has dramatically changed land use patterns. While improving people's life, it has also led to the surge of urban built-up area in China, which has brought many environmental problems. With the continuous development of urbanization, China's urban development has expanded from scale gradually entered the stage of high-quality development [2 3].

Zhejiang Province is located in the eastern coastal area of China, the terrain is mainly mountainous and hilly, affected by typhoons and other climate, geological disasters frequent. According to public data from the National Bureau of Statistics [4], nearly 3,700 geological disasters occurred in Zhejiang Province from 2010 to 2020, causing a total of 120 casualties and direct economic losses of 4.9484 billion yuan. Geological disasters cause huge casualties and property losses every year. Different scholars have studied it from the aspects of vulnerability [5], harmfulness [6], and disaster prevention and mitigation measures [7], but its content mainly focuses on the geological disaster object itself. There are few studies on the evaluation methods. The use of susceptibility evaluation research based on different mathematical statistical models can make cities more effectively predict the frequency of geological disasters and effectively prevent them, and find the most suitable evaluation method for regional disaster prevention.

Under the background of continuous urbanization, how to prevent geological disasters and reduce the loss of national property is of great significance. It is of great academic and social value to establish a research framework of geological hazard vulnerability based on different mathematical statistical

models. In view of this, this study plans to select Zhejiang Province as the research area to study the local geological hazard susceptibility based on the single-factor model (frequency ratio [8]), multi-factor model (logistic regression [9]) and machine learning model (random forest [10]), and comprehensively analyze the calculation results according to certain division rules.

2. Geological overview of the study area

The climate of Zhejiang Province is diverse, with most areas having a subtropical monsoon climate [11]. Monsoon and typhoon have a significant impact, and rainfall is mainly distributed in summer, with the average annual rainfall in the province ranging from 1200mm to 2200mm [12]. The terrain of Zhejiang Province is complex, with mountains, plains and coastal zones interlocking. In the province, the western Fujian Mountain belt and the central Fujian Mountain belt dominate the terrain trend, and the mountains mainly spread south-north and northeast-southwest trends. The valley plain is distributed among mountains and coastal zones, and the terrain slopes from inland mountains to coastal zones, with large topographic fluctuations and complex geomorphic types [13]. Affected by topography and precipitation, the river network in Zhejiang Province is developed and the river density is large. There are a considerable number of rivers with a basin area of more than 50km², forming a rich river system. Zhejiang Province is located in the junction area of Eurasian plate and Pacific plate, which is affected by geological tectonic movement. There are multi-stage structures in the province, and faults and folds are widely distributed [14]. The strata in the interior of Zhejiang Province are rich and have different degrees of outcrops from Proterozoic to Quaternary. Rocks of different periods are widely distributed, including metamorphic, volcanic and sedimentary rocks. The stratigraphic characteristics of different regions show diversity, for example, the northwest is dominated by Proterozoic metamorphic rocks, the central and southwestern are distributed by Sinian to Cretaceous metamorphic rocks, volcanic rocks and sedimentary rocks, and the eastern is dominated by Jurassic to Cretaceous volcanic rocks. The topographic distribution in Zhejiang Province is shown in Fig.1.

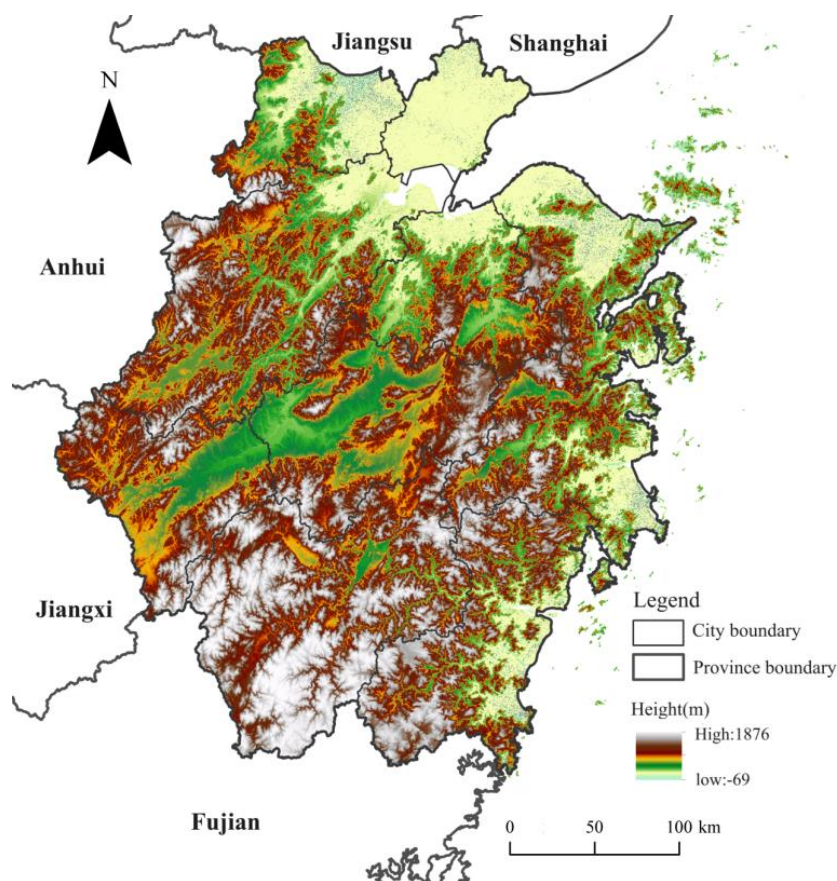


Figure 1. Topographic distribution in Zhejiang Province

3. Data sources and research methods

3.1. Data sources

The data used in this paper include: (1) The data of historical geological disasters in Zhejiang Province are from the open data of historical disasters and engineering geological exploration data of Resources and Environmental Science and Data Center of Chinese Academy of Sciences (<https://www.resdc.cn/>); (2) Data elevation model (DEM), spatial distribution data of road, river, land use type and ormalized difference vegetation index (NDVI) were obtained from Resources and Environmental Sciences and Data Center, Chinese Academy of Sciences. Slope and plane curvature were extracted from DEM data. Annual rainfall data was obtained from the National Earth System Science Data Cloud Platform [15].

3.2. Evaluation factors and models

3.2.1. Evaluation factors

The State Council's "Regulations on the Prevention and Control of Geological Hazards"[16] classify geological disasters into six types: collapse, landslide, debris flow, ground subsidence, ground fissures, and ground subsidence, Collapse, landslides, and mudslides are the main types of disasters that cause casualties and material losses [17]. Based on the data availability and the characteristics of the study area, this paper selected 12 factors as evaluation factors: elevation, slope, aspect, river density, road density, annual rainfall, normalized ormalized difference vegetation index (NDVI), plane curvature, land use type, terrain wetness index (TWI), water flow intensity index (SPI) and terrain relief. The basic evaluation unit with 50m×50m resolution is adopted. The distribution of geological disaster points in Zhejiang Province is shown in Fig.2.

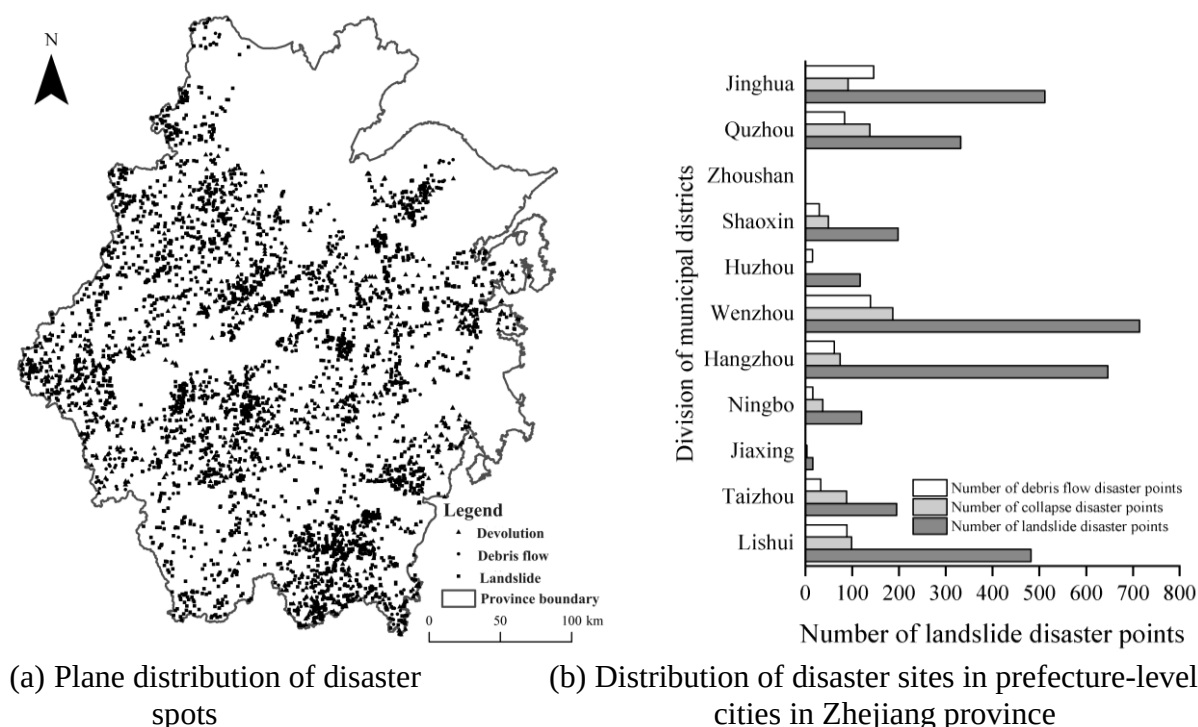


Figure 2. Distribution of disaster sites in Zhejiang Province

3.2.2. Level of elevation

Elevation, that is, the absolute distance from a certain point along the vertical direction to a certain point, is one of the important factors that affect the occurrence, development and morphological characteristics of geological disasters such as collapse. In different elevation ranges, vegetation coverage, rainfall, hydrothermal conditions, temperature difference, and intensity of human engineering activities are obviously different, and these factors will directly or indirectly affect the

occurrence of geological disasters. The topography of Zhejiang Province is dominated by hilly plain topography. Considering the elevation distribution in Zhejiang Province, the elevation data of the study area are divided into <100m, 100M-200M, 200m-300m, 300m-400m and >400m. There are 5 types of 400m. The elevation data distribution in Zhejiang Province is shown in Fig.3(a), and the data elevation model (DEM) is from the public data of the Center for Resources, Environmental Sciences and Data of the Chinese Academy of Sciences [18].

3.2.3. Degree of slope

Slope represents the steepness and gentleness of the surface of the disaster point and potentially dangerous slope, which is generally expressed as the ratio of the vertical distance and horizontal distance from a certain point on the slope to the foot of the slope. The slope directly affects the stability of the slope and plays an important role in controlling the occurrence of geological disasters. The steeper the slope is, the water and soluble salt in the slope will gradually accumulate downward, which will directly affect the vegetation distribution on the slope. With different slope positions, the movement of water in the slope, the denudation and material distribution of rock and soil on the slope, and the stress characteristics inside the slope will also be different, and the probability of geological disasters such as collapse will also change accordingly. In this section, the slope data of the study area are divided into <5°, 5°-10°, 10°-15°, 15°-20° and >20°; There are 5 types of 20°. The distribution of slope data in Zhejiang Province is shown in Fig.3(b), and the slope data are extracted from DEM data.

3.2.4. Aspect of slope

The aspect is expressed as the Angle between the projection of the normal direction of the slope on the plane and the true north direction. The slope aspect has an effect on the number of days of sunshine and the total amount of radiation, especially in areas with strong subtropical monsoon climate such as Zhejiang. Due to the differences in the amount of solar radiation obtained, water evaporation, plant distribution, pore water pressure and dry-wet cycle of slope surface will also be different in different slope directions, and the risk of collapse disaster will also be different. According to the public cognition, this paper divides the aspect data of the study area into eight types: north, northeast, east, southeast, south, southwest, west and northwest. The spatial distribution of slope aspect in Zhejiang is shown in Fig.3(c), and the aspect data are extracted from DEM data.

3.2.5. Density of rivers

The presence of water can cause considerable erosion of the coastal rock and soil. The higher the moisture content of the soil, the lower the shear strength, the worse the stability, and the more prone to slope disasters. The density of the river system indicates the number of river networks in a unit area of the region, which is an important representation of the development degree of the local river system. The higher the density of the water system is, the greater the distribution density of the local surface runoff is, and the greater the erosion effect of the water body on the adjacent rock and soil is. The formula for calculating the density of the stream system is expressed as follows:

$$\rho_{w=\frac{l_{wi}}{A_i}} \quad (1)$$

Where, ρ_w represents the density of the water system, l_{wi} represents the total number of river network distribution lengths of the i th computing unit, and A_i represents the area of the i th computing unit. In this section, the water system density data of the study area are divided into <0.5km-1, 0.5km-1-1.0km-1, 1.0km-1-1.5km-1, 1.5km-1-2.0km-1 and >2.0km-1. There are 5 types in total. The distribution of water system density data in Zhejiang Province is shown in Fig.3(d). River data were obtained from the Center for Resources, Environmental Sciences and Data of the Chinese Academy of Sciences [18].

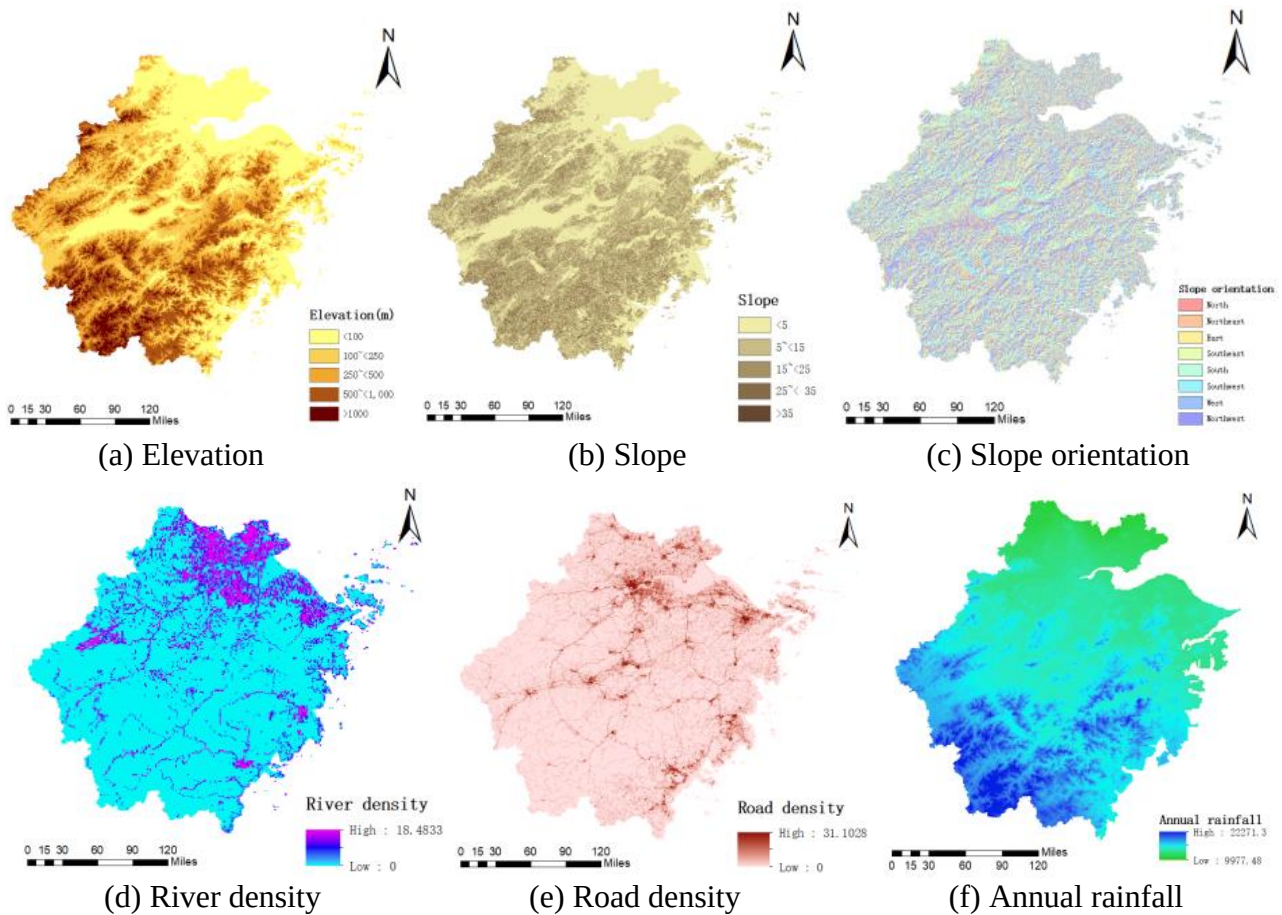
3.2.6. Density of roads

The stability of rock and soil mass will be affected by human engineering activities. On the other hand, large-scale construction will also change the original balance condition of rock and soil mass,

and increase the risk of geological disasters. Road density represents the mileage of roads (including national roads, township roads, urban expressways, subways, etc.) within a unit area of a region, which is an important representation of local human activities. Road density represents the mileage of roads (including national roads, township roads, urban expressways, subways, etc.) within a unit area of a region, which is an important representation of local human activities. The calculation formula is as follows:

$$\rho_r = \frac{l_{r_i}}{A_i} \quad (2)$$

Where, ρ_r represents the road density, l_{r_i} represents the total number of road network distribution length of the i th computing unit, A_i represents the area of the i th unit of computation. In this section, the road density data of the study area are divided into $<0.5\text{km-1}$, 0.5km-1-1.0km-1 , 1.0km-1-1.5km-1 , 1.5km-1-2.0km-1 and $>2.0\text{km-1}$. There are 5 types in total. The distribution of water system density data in Zhejiang Province is shown in Fig.3(e). It was obtained from the Center for Resources, Environmental Sciences and Data of the Chinese Academy of Sciences [18].



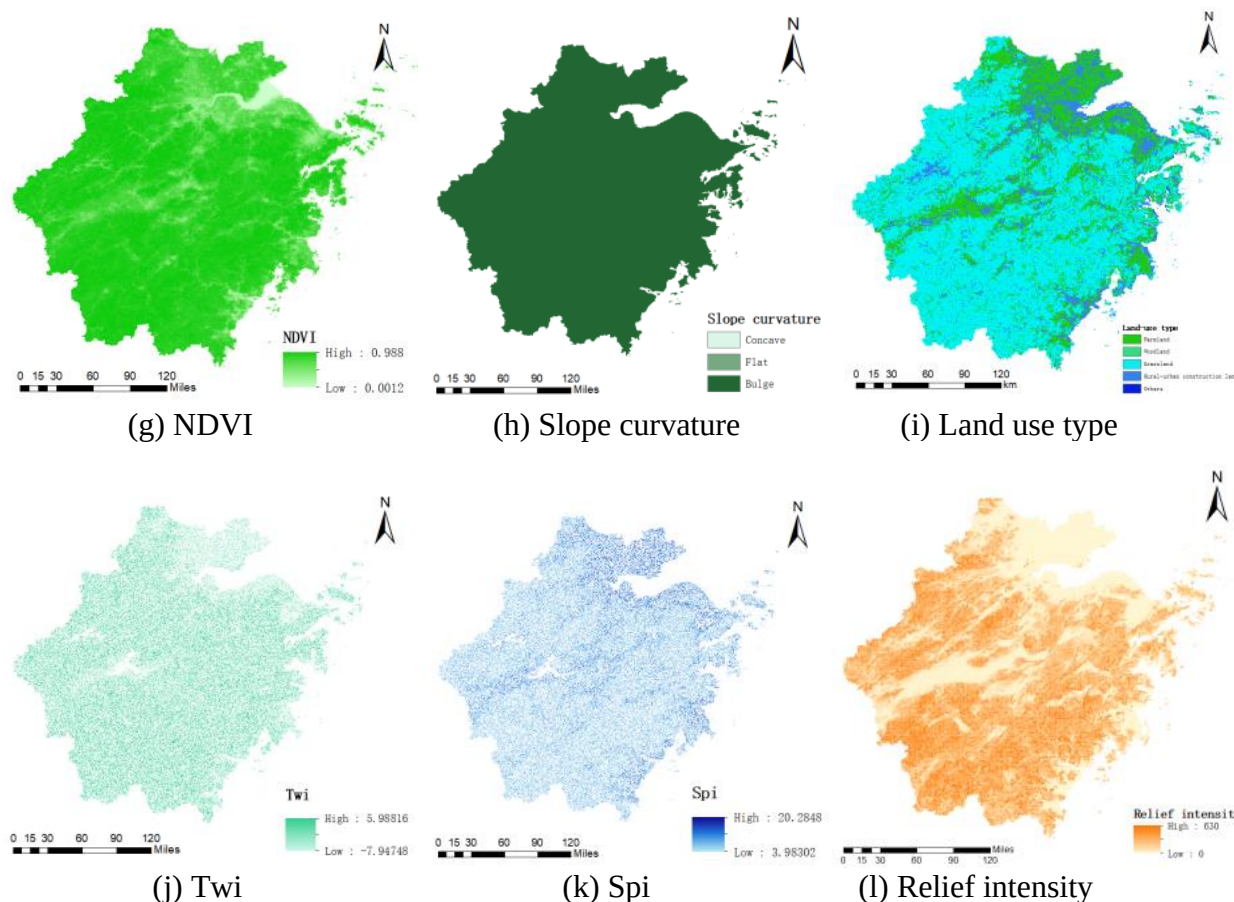


Figure 3. Evaluation factors

3.2.7. Annual rainfall

Rainfall is one of the important causes of geological disasters. Frequent rainfall can create objective conditions for the occurrence of geological disasters. During summer, abundant precipitation can bring a large amount of water supply to the surface runoff, improve the moisture content of the slope, and deteriorate its physical and mechanical properties. The increase of slope weight and the resulting permeability will also increase the sliding force of the slope on the most dangerous sliding surface, which will greatly increase the probability of geological disasters. Most areas in Zhejiang province belong to the subtropical monsoon climate, the average annual rainfall distribution in the province is between 1000mm and 2000mm, the precipitation has obvious monsoon climate characteristics, mostly concentrated in July to September in summer, at the same time due to the location factors of the territorial sea, Zhejiang spring and summer precipitation is vulnerable to typhoons, heavy rain, floods and other natural disasters often break out during the rainy season. In this section, the road density data of the study area are divided into <1000mm, 1000mm-1200mm, 1200mm-1400mm, 1400mm-1600mm and >1600mm. There are 5 types in total. The distribution of annual rainfall data in Zhejiang Province is shown in Fig.3(f), and the annual rainfall data are from the public data of the National Earth System Science Data Cloud Platform.

3.2.8. Normalized vegetation index

The Normalized difference vegetation index (NDVI) is an important measure of land surface vegetation cover. Vegetation has anchorage effect on soil through the growth of roots, which can affect the permeability of surface water to soil, and inhibit the erosion of surface runoff on slope surface, which can effectively prevent the occurrence of geological disasters. In this section, NDVI data in the study area are divided into <0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8 and >0.8. There are 5 types in total. The distribution of NDVI data in Zhejiang Province is shown in Fig.3(g).

3.2.9. Curvature of plane

The curvature shows the concave and convex condition of the surface of the evaluation element. When the curvature is positive, the evaluation element is convex slope. When is negative, it is a concave slope; When it is 0, it is a plane. Generally speaking, a flat slope can effectively prevent the occurrence of geological disasters. Curvature itself is also the most intuitive manifestation of the slope surface, which has an important impact on the weathering and denudation of the slope surface and surface runoff. The local topography of Zhejiang province is mainly hilly and mountainous, and the surface fluctuation is large. According to the different concave and convex situations, the curvature data of the study area are divided into three types in this section: concave, plane and convex. The distribution of curvature data in Zhejiang Province is shown in Fig.3(h), and the curvature data are extracted from DEM data.

3.2.10. Type of land use

Land use type is an important indicator reflecting land use, use nature and economic characteristics of users, which can characterize the impact of urbanization on land to a certain extent. Different land use types have great differences in the corresponding human engineering activities, vegetation distribution, roads, houses and other infrastructure construction degrees, which will have an impact on the stability of local geotechnical masses. In this section, based on the reference to relevant literature, the data of land use types in the study area are divided into five types: urban and rural residential land, forest land, cultivated land, grassland, water area and others. The data distribution of land use types in Zhejiang Province is shown in Fig.3(i).

3.2.11. Topographic moisture index and water flow capacity index

The presence of water can cause considerable erosion of the coastal rock and soil. The higher the moisture content of the soil, the lower the shear strength, the worse the stability under the action of natural forces, and the more prone to geological disasters. Topographic wetness index (TWI) is a commonly used index to measure the control effect of topography on the flow and accumulation of water bodies. Areas with higher TWI values will collect more standing water during rainfall. Stream power index (SPI) is a commonly used quantitative index to quantitatively describe the flow capacity of surface flowing water and its ability to scour coastal areas. TWI and SPI are important factors to measure the extent to which geotechnical bodies are affected by water bodies, and their calculation formulas are as follows:

$$TWI = \ln\left(\frac{A}{\tan \beta}\right) \quad (3)$$

$$SPI = \ln(A \cdot \tan \beta) \quad (4)$$

Where, β is the slope of the location where the unit is evaluated; A Represents the area of the upstream catchment area that may be formed after precipitation or surface runoff passes through. This section divides the SPI data in the study area into <2, 2-5, 5-8, 8-11 and >11. There are 5 types in total, and the TWI data in the study area are divided into <1, 1-2, 2-3, 3-4 and >4. There are 5 types in total. The distribution of TWI and SPI data in Zhejiang Province is shown in Figure 3(j) and Figure 3(k), respectively

3.2.12. Degree of relief

Topographic relief refers to the difference between the elevation of the highest point and the elevation of the lowest point in a specific area. It is a macro indicator to describe the topographic characteristics of a region. Topographic relief can directly reflect the topographic relief characteristics, is an important index to classify geomorphic types, is the most direct factor leading to soil erosion. In this section, the topographic relief data of the study area are divided into <50m, 50m-100m, 100m-150m, 150m-200m and >200m, 5 types in total. Distribution of topographic relief data in Zhejiang Province is shown in Fig.3(l).

3.3. Analysis of correlation

The evaluation factors in the evaluation model of geological disaster susceptibility are likely to have certain correlation, and these related factors may interfere with each other and overlap repeatedly during the evaluation, and the results will seriously affect the accuracy of the model. Therefore, in order to ensure the independence of the influencing factors, it is necessary to conduct collinearity analysis on the elevation and other factors before the collapse susceptibility study. In this paper, Pearson correlation coefficient is used to analyze the correlation between factors, and the calculation is completed by importing the original data matrix into the Pandas library correlation function in Python. Assuming factor $(X, Y) = (x_1, y_1), \dots, (x_n, y_n)$, the calculation formula of correlation coefficient between evaluation factors is as follows:

$$PCC = \frac{\sum_{i=1}^n (x_i - \bar{x}) \sum_{j=1}^n (y_j - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{j=1}^n (y_j - \bar{y})^2}} \quad (5)$$

Where, (X, Y) is the data set for comparison, (x_i, y_i) is the sample of the (X, Y) , $(\bar{x}, \bar{y})$ is the average value of all samples in dataset (X, Y) . Pearson correlation coefficient is between $[-1, 1]$, when its absolute value is 0, it means there is no relationship between the evaluation factors, when its absolute value is 1, it means that the evaluation factors are completely linear. When $0 \leq |PCC| \leq 0.3$, it shows that the correlation between evaluation factors is not significant; When $|PCC| > 0.3$, it indicates that there is a strong correlation between the evaluation factors. The calculation results are shown in Table.1. It can be seen from Table 1 that the pcc between spi and twi is 0.75, and the pcc of slope and topographic relief is 0.338, indicating a strong correlation. The data set satisfies the requirements in collinearity. Therefore, ten factors including spi, soil benefit, aspect, slope, precipitation, river network density, elevation, road density, plane curvature, and vegetation were selected as evaluation factors.

Table 1. Correlation coefficients of evaluation indicators

Index of evaluation	NDVI	SPI	Degree of slope	precipitation	Density of river network	Level of elevation	Degree of relief	Density of roads	TWI	Curvature of plane
NDVI	1.000									
SPI	0.067	1.000								
Degree of slope	0.148	-0.238	1.000							
precipitation	0.281	0.086	0.162	1.000						
Density of river network	-0.138	-0.004	-0.073	-0.182	1.000					
Level of elevation	0.173	-0.039	0.240	0.275	-0.143	1.000				
Degree of relief	0.139	-0.173	0.339	0.116	-0.080	0.142	1.000			
Density of roads	-0.122	0.066	-0.110	-0.079	0.203	-0.135	-0.109	1.000		
TWI	0.099	0.755	-0.131	0.138	-0.059	0.036	0.080	-0.099	1.000	
Curvature of plane	0.001	-0.043	0.007	0.004	-0.004	0.060	0.003	0.007	-0.099	1.000

3.4. Methods of research

3.4.1. Frequency ratio model

The frequency ratio model (FR) is a quantitative analysis method to describe the correlation degree between the target variable and a single influencing factor. The frequency ratio of each grading interval of the factor is determined based on the percentage of the number of geological disasters occurring and the percentage of the area of the corresponding grading interval.

The frequency ratio method is used to reveal the correlation between historical geological disasters and various factors according to the distribution degree of the collected disaster points in the grading interval of each influencing factor. The frequency ratio of geological disasters is determined by analyzing the relationship between geological disasters and causative factors, and then the possibility of geological disasters (DSI) in each evaluation unit can be calculated. The specific calculation formula is as follows:

$$FR_{ij} = \frac{num_{ij} / Num}{A_{ij} / A} \quad (6)$$

$$DSI = \sum_{i=1}^N FR_i \quad (7)$$

Where, A_{ij} is the area of the j th grading interval in the i th factor; A is the total area of the study area; num_{ij} is the number of samples of geological disasters occurring in the j th grading interval of the i th factor; Num is the total number of geological hazard samples in the study area; FR_{ij} is the frequency ratio corresponding to the j th interval in the i th factor; N is the number of influencing factors; DSI is the disaster vulnerability index, which can be obtained by summing the FR_i superposition corresponding to each factor of the evaluation unit.

3.4.2. Logistic regression model

The logistic regression model (LR for short) because the occurrence of geological disasters is non-linearly related to a variety of factors, the selection of evaluation factors is also random, all these factors increase the difficulty of analyzing the trend of disaster, and logistic regression model can analyze the specific numerical relationship between the target variable and multiple influencing factors, and then evaluate the susceptibility of geological disasters. It solves this problem well. Logistic regression analysis is used to characterize the specific numerical relationship between multiple influencing factors. After determining the weight, the susceptibility of geological disasters in the study area can be quantitatively represented by vector superposition.

As the target variable, the geological hazard is used as 1 and 0 to represent the two cases of occurrence and non-occurrence, which can correspond to the dual variables in the data grid. When establishing a disaster assessment model, the LR model can perform hypothesis testing and parameter estimation on qualitative data and continuous quantitative data that meet the independence requirements, and on this basis determine the occurrence probability of geological disasters in the evaluation unit. The weight of each factor in the LR model is mainly determined by its influence on the target variable, which has the advantages of fast training speed and good interpretability, and is very suitable for binary classification problems such as susceptibility evaluation. The expression of the LR model is as follows:

$$DSI = 1 / [1 + e^{-(\alpha + \beta_1 x_1 + \dots + \beta_N x_N)}] \quad (8)$$

Where, DSI is the disaster susceptibility index, and in the logistic regression model is the probability of occurrence of geological disasters in the evaluation unit; β_i is the weight of the i th

factor; α is the corresponding constant term and x_i is the corresponding value of the i th influencing factor. After taking the logarithm of both sides, the above equation can be transformed into:

$$\ln\left(\frac{DSI}{1-DSI}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_N x_N \quad (9)$$

3.4.3. Random forest model

The random forest model (RF) is a classification model developed from the generalized portrait algorithm in pattern recognition, which is essentially a collection of multiple decision tree models. During each evaluation, the RF model uses a random method to build decision trees, while there is no correlation between the decision trees. For binary classification problems such as geological disaster susceptibility assessment, a single decision tree model can classify different factors according to the principle of "minimum information entropy" based on the data distribution characteristics of different factors. The calculation method is as follows:

$$Ent(D) = -\sum_{k=1}^{|Y|} p_k \log_2 p_k \quad (10)$$

Where, $Ent(D)$ is the information entropy of set D , which is a common factor to measure the purity of the sample set. The smaller its value is, the more balanced the partition result is; p_k is the proportion of the k th sample in the sample set D ; Y is the total number of samples for evaluation. After each training, a single decision tree model will calculate the difference between the predicted value and the target value, and iteratively adjust the partition rule of each leaf node accordingly until convergence is reached. RF usually adopts Bootstrap method for resampling, which can extract a specific proportion of data from the original data set to form a new sample set, so as to generate different decision trees specifically. Since the evaluation result of RF model is decided by voting according to the evaluation result of decision tree, and the selection of child node splitting characteristics and training samples between different decision trees are random, RF model can avoid the problem of overfitting to a certain extent. Based on the historical geological disaster records of Zhejiang Province, this section uses RF model to evaluate the susceptibility of geological disasters in Zhejiang Province, and analyzes the evaluation effect from the perspectives of accuracy and AUC value. The basic evaluation steps of RF model for a single time are shown in Fig.4.

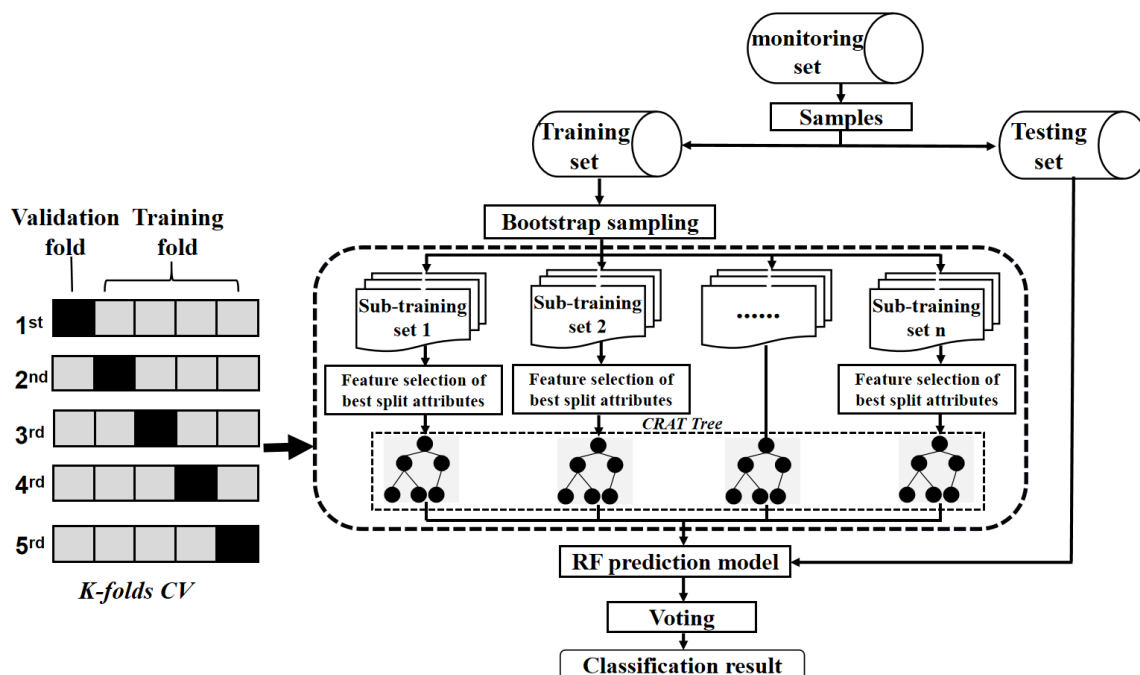


Figure 4. basic evaluation steps of RF model

4. Results and accuracy analysis

4.1. Analysis of results

The frequency ratio can represent the influence degree of different grading intervals of evaluation factors on geological disasters. The larger the value is, the higher the probability of geological disasters occurring in this interval is. If the corresponding frequency value of a grading interval is 0, it can be indicated that there is no disaster record in this interval, the occurrence of geological disasters has nothing to do with this grading interval, and this interval has no influence on the occurrence of geological disasters. Table 2 shows that the frequency ratio of NDVI is high at 1.2 when NDVI is more than 0.8, and the frequency ratio is increasing with the increase of zoning level, indicating that NDVI factors have a positive correlation with the occurrence of geological disasters in the study area. The influence of plane curvature on geological disasters is mainly reflected in the convexity of curvature. When the land use type is grassland, it has the greatest impact on geological disasters. The same when annual rainfall is between 1400mm and 1600mm, spi is between 5-8, elevation is between 200m-300m, and road density is between 1.0-1.5. Although its influence on geological disasters is only within a certain range, it also maintains a certain degree of positive correlation. Among other factors, the frequency ratio is highest when the slope is between 5 and 10, and when the river network density is 0.5 the subcategory frequency ratio is the highest, which is contrary to the conventional understanding of the occurrence of geological disasters, and it is likely that human activities are more frequent in areas with fewer rivers and low slopes. The logistic regression coefficient represents the degree of influence of a factor on the occurrence of geological disasters; the larger the value is, the greater the correlation of the factor on the hazard of geological disasters is. It can be seen from Table.3 that factors such as annual rainfall (0.61) and elevation (-0.410) have a greater impact on the occurrence of geological disasters in Zhejiang Province (weight>0.4).

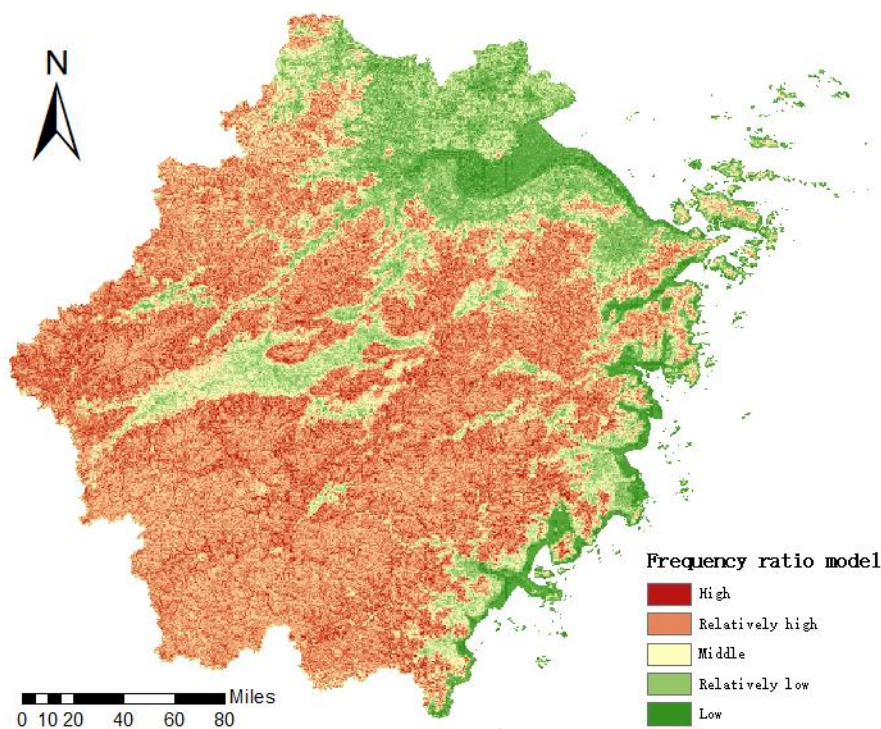
Table 2. Calculation results of frequency ratio method

Factors	Division of areas	Frequency ratio	Factors	Division of areas	Frequency ratio
NDVI	<0.2	0.141	SPI	<2	0.889
	0.2-0.4	0.215		2-5	0.826
	0.4-0.6	0.195		5-8	1.438
	0.6-0.8	0.401		8-11	1.278
	>0.8	1.272		>11	0.172
Density of roads(km-1)	<0.5	0.752	Level of elevation(m)	<100	0.352
	0.5-1.0	1.489		100-200	1.564
	1.0-1.5	1.729		200-300	1.928
	1.5-2.0	1.656		300-400	1.530
	>2.0	1.070		>400	1.073
Annual rainfall(mm)	<1000	0.072	Degree of slope(°)	<5	0.589
	1000-1200	0.277		5-10	1.665
	1200-1400	0.826		10-15	1.452
	1400-1600	1.415		15-20	1.080
	>1600	1.245		>20	0.754
Aspect of slope	North	0.815	Density of river network(km ⁻¹)	<0.5	1.131
	Northeast	0.607		0.5-1.0	0.901
	East	1.064		1.0-1.5	0.728
	Southeast	1.183		1.5-2.0	0.474
	South	1.304		>2.0	0.385
	Southwest	1.309	Type of land use	Land under cultivation	0.695
	West	1.134		Forest land	1.234
	Northwest	0.965		Grass land	1.340
Curvature of plane	Flat	0.823		Urban and rural construction land	0.396
	Concave	1.129		Waters and others	0.443
	Convex	1.071			

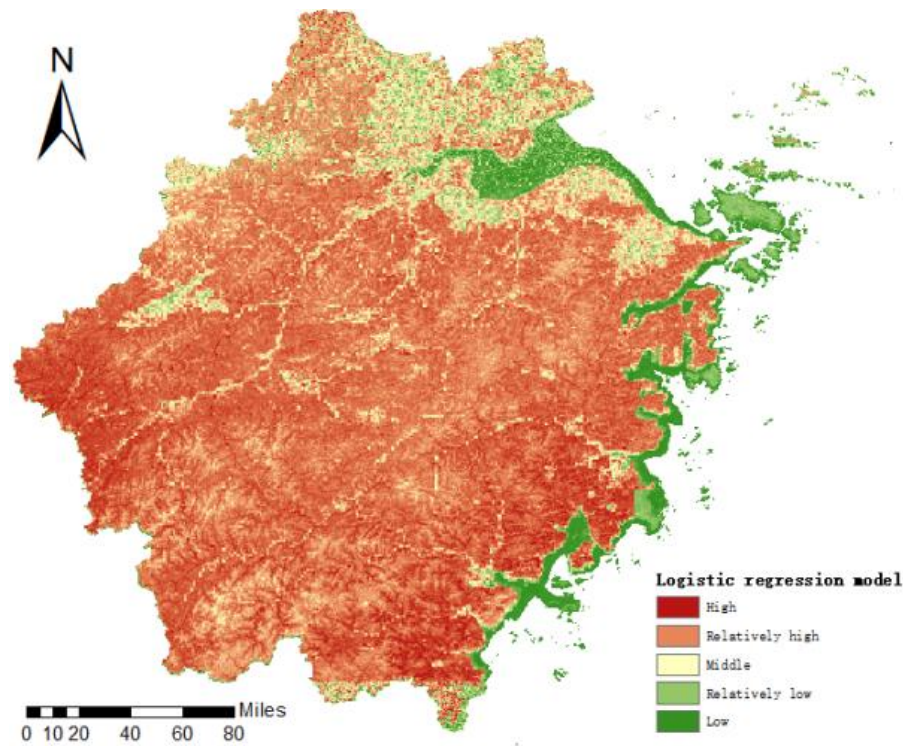
Table 3. Logistic regression model to calculate the weights

Factor	Weight	Factor	Weight
spi	0.047	Land utilization rate	-0.076
Aspect of slope	0.014	Degree of slope	0.086
Rainfall	0.615	Density of river network	-0.249
Level of elevation	-0.410	Density of roads	0.128
Curvature of plane	-9.531e-05	NDVI	0.449
Constant term	0.0001		

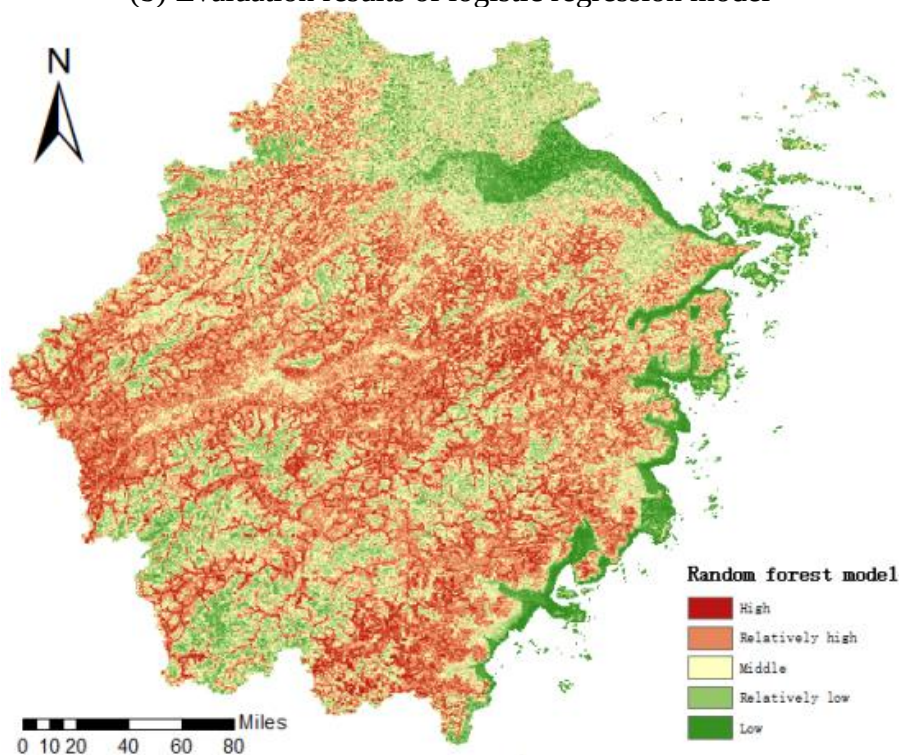
In the process of establishing the evaluation model of geological hazard susceptibility in Zhejiang Province based on the theory of different mathematical calculation models, 70% of the samples in the data set were randomly selected as the training set, and the remaining 30% were selected as the validation set. It is imported into the corresponding Python API in Arcgis, and finally superimposed to obtain the geological disaster prone regionalization map of Zhejiang Province calculated separately by different models, as shown in Fig.5. In the zoning results calculated by the frequency ratio model, the high and extremely high susceptibility areas are mainly concentrated in the hilly area of western Zhejiang province and the mountainous area of southern Zhejiang Province, which are similar to the topographic distribution of Zhejiang Province. The results of logistic regression model showed that the high and extremely high risk areas were mainly concentrated in the west and southeast of Zhejiang. The high and extremely high risk areas in the random forest model were mainly distributed near the river system and road network in the high altitude area, and the transition part from the outside of the mountain to the coastal plain was the concentration zone of the middle and low risk areas. The northern, southern and some valley plain areas are dominated by extremely low risk areas.



(a) Evaluation results of frequency ratio model



(b) Evaluation results of logistic regression model



(c) Evaluation results of random forest model

Figure 5. Calculation results of geological hazard model

The area distribution and the proportion of the number of disasters in each partition were counted, and the results are shown in Fig.6. It can be seen from the analysis in Figure 6 that the disaster proportion of each model is still increasing with the increase of the level of the prone area in general, and its characteristics are in line with the objective law of geological disasters, indicating that the model is effective. However, compared with pure objective methods such as frequency ratio method (2.01) and logistic regression method (1.63), random forest method has a more obvious enrichment degree of disaster points (2.40) in extremely high susceptibility.

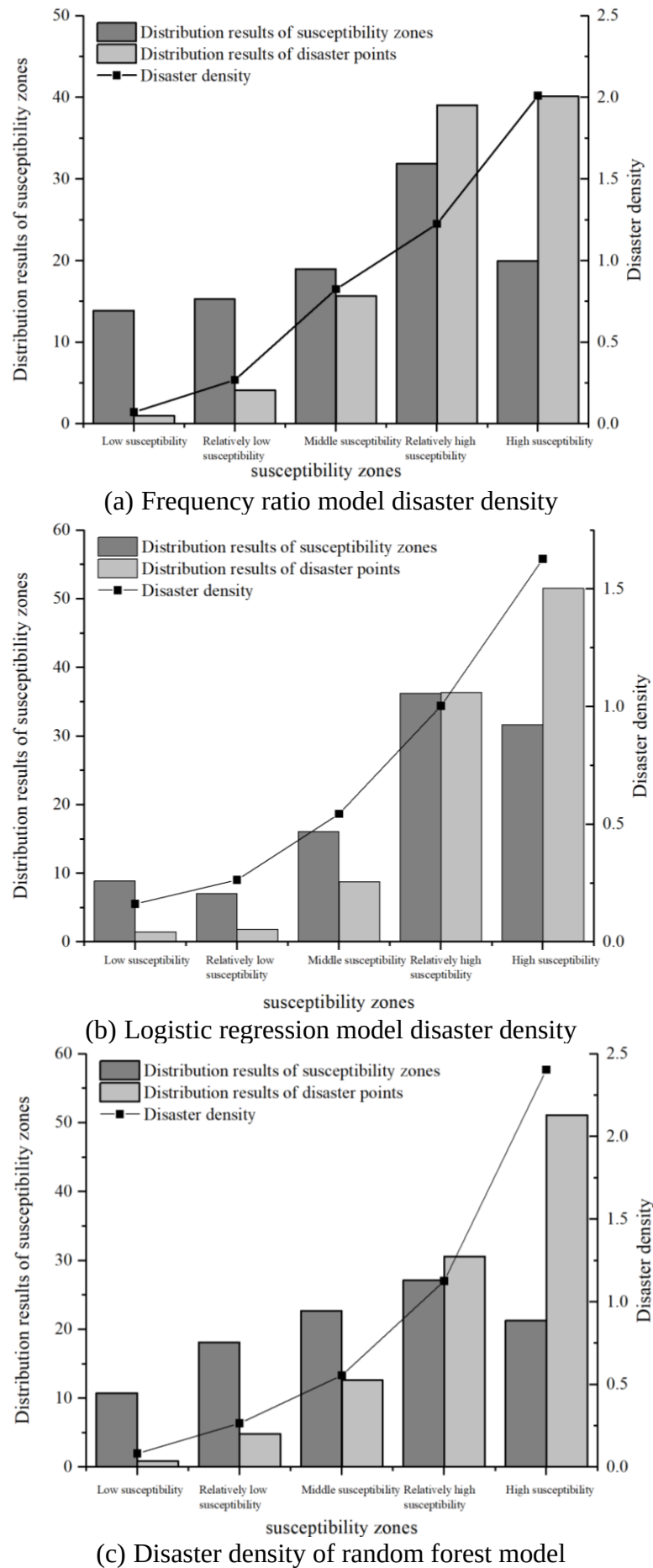


Figure 6. Calculation results of disaster density

4.2. Analysis of accuracy

In order to illustrate the differences in evaluation effects among different mathematical calculation models, this section combines the Receiver operating characteristic curve (ROC) to verify the accuracy of the calculation results of each model. The ROC curve uses False positive rate (FPR, also known as specificity) as the horizontal axis and True positive rate (TPR, also known as sensitivity) as the vertical axis, where TPR measures the proportion of actual positive samples that are correctly identified, while FPR measures the proportion of actual negative samples that are incorrectly identified as positive, which is calculated as follows:

$$TPR = \frac{TP}{TP + FN} \quad (11)$$

$$FPR = \frac{FP}{FP + TN} \quad (12)$$

Where, TPR is the true positive rate, that is, the rate of correct prediction in positive samples; FPR is the false positive rate, that is, the rate of prediction errors in negative samples; TP is the number of samples predicted to be positive but actually positive; FN is the number of samples predicted to be negative but actually positive; FP is the number of samples that are predicted to be positive but actually negative; TN is the number of samples predicted to be negative but actually negative. The Area enclosed by the ROC curve and the X-axis is the Area under curve (AUC) value. The higher the AUC value is, the more deviated the ROC curve is from the diagonal, and the better the performance of the classification model is. The ROC curves of each model are shown in Fig.7.

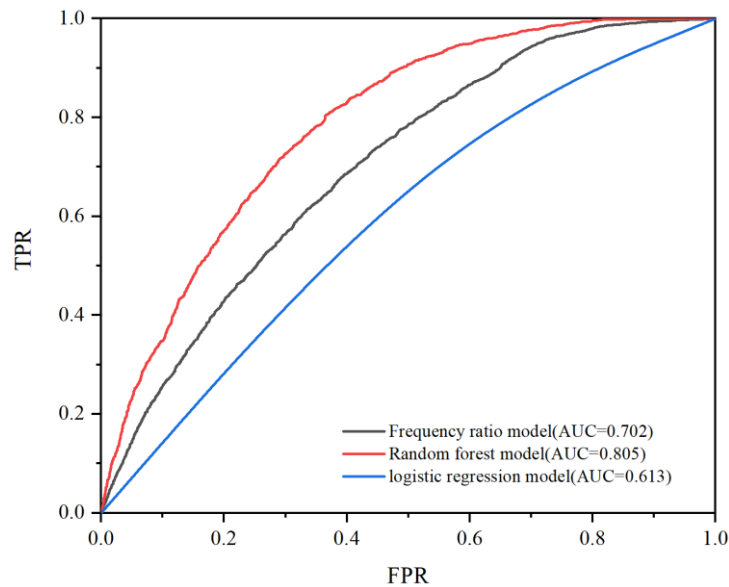


Figure 7. ROC curve

It can be seen from Figure 3 that the AUC areas under different models are as follows: random forest model (RF, AUC=0.805), frequency ratio method (FR, AUC=0.702), and logistic regression model (LR, AUC=0.613). Except that the AUC value of the random forest model is above 0.8, the AUC value of the models of other methods is mostly distributed between 0.6 and 0.7, and the evaluation results are effective, but the random forest model has higher accuracy. The PR curve takes the Precision (Pre, which is calculated in the same way as the true positive rate) as the horizontal axis, and the Recall (Rec) as the vertical axis. The calculation method is as follows:

$$Pre = \frac{TP}{TP + FP} \quad (13)$$

$$Rec = \frac{FP}{TP + FN} \quad (14)$$

Where, Pre is the accuracy rate, that is, the proportion of true disaster points in the evaluation unit that predicts the occurrence of disasters; Is the recall rate, which is the proportion of disaster points that are correctly predicted. Generally speaking, the accuracy rate can reflect the prediction accuracy of the evaluation model for the positive samples in the validation set. The recall rate represents the generalization degree of the model to the evaluation object. Generally speaking, the higher the precision of the model is, the higher the corresponding precision under the condition of the same recall rate will be, and the PR curve of the model will shift to the upper right corner. The PR curves of the different models are shown in Fig.8.

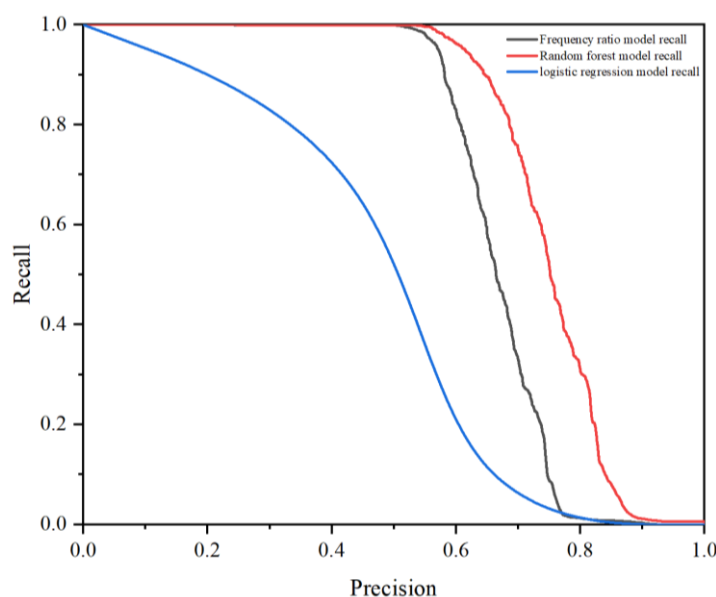


Figure 8. PR curve

It can be seen from Figure 8 that the PR curves of each model are close to the upper right corner, with a trend of decreasing with the increase of recall rate, and the accuracy of the model is decreasing with the increase of recall rate, which conforms to the objective law of statistics. However, except for the evaluation results of the random forest model, the PR curves of the other models are relatively rugged and have different degrees of bending. To sum up, the ranking of evaluation accuracy among models is random forest model (RF); Frequency ratio (FR); Logistic regression model (LR). The precision and accuracy of random forest model (RF) are at a high level, and it has better evaluation effect than other models.

The evaluation results are further divided according to the administrative division of prefecture-level cities in Zhejiang Province, and the results are shown in Figure 9. It can be seen from Fig.9 that 71.69% and 69.70% of the areas of Quzhou and Jinhua are located at or above the high susceptibility zone, followed by Wenzhou (58.57%) and Shaoxing (57.51%), and the proportions of other cities are below 40%. Based on the evaluation results, it is suggested that the relevant departments should take Quzhou, Jinhua, Wenzhou and Shaoxing as the focus of geological disaster prevention and control in Zhejiang province.

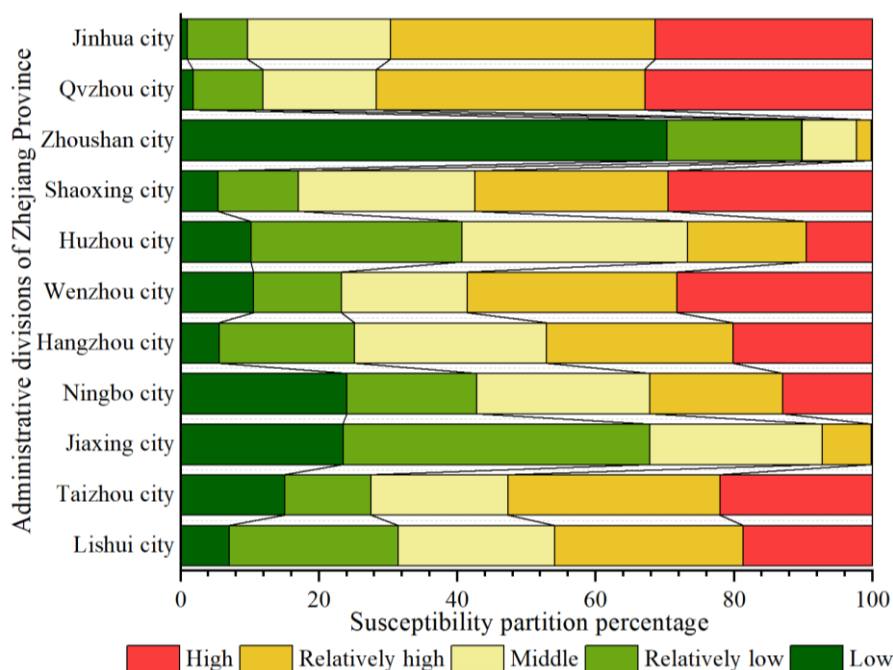


Figure 9. Classification of prefecture-level cities

5. Conclusion

In this study, 12 evaluation indices including elevation, slope, aspect, river density, road density, annual rainfall, normalized vegetation index (NDVI), plane curvature, land use type, terrain moisture index (TMI), water flow capacity index (SPI), and terrain relief were selected based on the historical geological disaster data from Zhejiang Province. This paper assesses the susceptibility of geological disasters in Zhejiang Province using various mathematical statistical models, yielding the following key findings:

(1) The high and extremely high-risk areas in Zhejiang Province are mainly distributed near the river system and road network in the high-altitude areas, and the transition part from the outside of the mountains to the coastal plain is the concentration zone of the middle and low risk areas; The northern, southern and some valley plain areas are mainly very low risk areas;

(2) The AUC area of each model was random forest model (RF, AUC=0.805), frequency ratio method (FR, AUC=0.702), logistic regression model (LR, AUC=0.613). In terms of the evaluation effect of comprehensive PR curve, the evaluation effect of random forest model was better than that of other models. With the increase of the vulnerability level, the proportion of disasters under each model zoning is gradually increasing, indicating that the zoning effect of each model is good.

(3) From the perspective of disaster point change analysis and ROC curve, the evaluation effect of each model is good and conforms to the objective law;

(4) According to the administrative divisions of each prefecture-level city in Zhejiang province, the distribution of the susceptibility of geological disasters obtained under the RF model is statistically divided, and it is found that Quzhou city, Jinhua City, Wenzhou city and Shaoxing city account for a large proportion of the high susceptibility areas and above.

References

- [1] Information on: www.stats.gov.cn.
- [2] Zhong X, Xie R, Chen P, et al. Internet Development and Environmental Quality—Evidence from the Development of Chinese Cities [J]. Sustainability, 2021, 13 (20): 11308.
- [3] Luo J, Wang W, Wu Y, et al. Analysis of an Urban Development Boundary Policy in China Based on the IAD Framework [J]. Land, 2021, 10 (8): 855.

- [4] Information on: www.mnr.gov.cn.
- [5] Zhou S H, Zhou S K, Tan X. Nationwide susceptibility mapping of landslides in Kenya using the fuzzy analytic hierarchy process model [J]. *Land*, 2020, 9 (12): 535 - 557.
- [6] He B, Bai M, Shi H, et al. Risk Assessment of Pipeline Engineering Geological Disaster Based on GIS and WOE-GA-BP Models [J]. *Applied Sciences*, 2021, 11 (21): 9919.
- [7] Bhat D R, Osawa S, Wakai A, et al. Rigorous Analysis of Stress-Dependent Landslide Movements with Groundwater Fluctuations Applicable to Disaster Prevention in Monsoon Asia[J]. *Journal of Disaster Research*, 2021, 16 (4): 658 - 673.
- [8] GUO Z Z, YING K LM, HUANG F M, et al. Evaluation of landslide susceptibility based on landslide classification and weighted frequency ratio model [J]. *Chinese Journal of Rock Mechanics and Engineering*, 2019, 38 (02): 287 - 300.
- [9] WU C R, JIAO Y M, WANG J L, et al. Evaluation of landslide susceptibility in Shuangbai County based on frequency ratio logistic regression coupling model [J]. *Journal of Natural Disasters*, 2021, 30 (4): 213 – 224.
- [10] ZHAI W H, WANG X D, WU M T, etc Geological hazard susceptibility evaluation based on the coupling of frequency ratio model and random forest model [J] *Journal of Natural Disasters*, 2023, 32 (06): 74 - 82 DOI: 10.13577/j.jnd.2023.0607.
- [11] BAO X Y, ZENG Y, QIU X F, etc Climate livability evaluation of Zhejiang Province based on the GridMet model [J] *Meteorological Science*, 2022, 42 (02): 244 - 253.
- [12] TAO X W, ZHU X H, KONG J W, etc Research on the Characteristics of Geological Disasters in Zhejiang Province and Countermeasures for Disaster Prevention and Reduction [J] *China Safety Production Science and Technology*, 2022, 18 (S1): 11 - 17.
- [13] LIU Y L, WANG Y, LOU S P. Formation conditions and influencing factors of debris flows in Zhejiang Province[J]. *Technology Bulletin*, 2015 (1): 37 - 40.
- [14] ZHU Y F, LIANG Q N, LIN D G. Comprehensive Study on Landslide Geological Hazard Risk Mapping in Zhejiang Province [J] *Land and Natural Resources Research*, 2019, (03): 34 - 41 DOI: 10.16202/j.cnki.tnrs. 2019. 03. 010.
- [15] Information on: <http://auth.geodata.cn/>
- [16] State Council of the People's Republic of China. Regulations on the Prevention and Control of Geological Disasters [EB/OL]. [2006-7-27]. <https://www.resdc.cn/>.
- [17] WNAG Y, LIN Q G, SHI P J. Spatial pattern and influencing factors of casualty events caused by landslides [J]. *ACTAGEOGRAPHICA SINICA*, 2017, 72 (05): 906 - 917.
- [18] Information on: www.resdc.cn.