Comparison of Landslide Susceptibility Mapping Based on Two Machine Learning Algorithm Models

Deliang Sun
Chongqing Normal University
Yue Wang
Southwest University
Jialan Zhang (lanaij@cqu.edu.cn)
Chongqing University
Haijia Wen
Chongqing University https://orcid.org/0000-0002-2045-729X
Jiahui Xu
East China Normal University
Xinzhi Zhou
Chongqing University

Research Article

Keywords: Mountainous region, landslide susceptibility mapping, random forest, artificial neural network, Wushan County

Posted Date: January 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1227768/v1

Comparison of Landslide Susceptibility Mapping Based on Two Machine Learning Algorithm Models

Deliang SUN[1], Yue[2] WANG, Jialan Zhang [3*], Haijia WEN [3*], Jiahui XU [4] and Xinzhi ZHOU [3]

1 Key Laboratory of GIS Application Research, Chongqing Normal University, Chongqing 401331, China

2 Chongqing Engineering Research Center for Application of Remote Sensing Big Data, School of Geographical Sciences, Southwest University, Chongqing 400715, China

3 Key Laboratory of New Technology for Construction of Cities in Mountain Area, Ministry of Education; National Joint Engineering Research Center of Geohazards Prevention in the Reservoir Areas; School of Civil Engineering, Chongqing University, Chongqing 400045, China

4 Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai 200241, China

Abstract: The purpose of this research is to develop and compare two machine learning methods, namely, artificial neural network (ANN) and random forest (RF) used for in landslide susceptibility mapping (LSM) of Wushan County. Firstly, 866 landslides were collected from extensive field investigations, historical records and satellite images from 2001 to 2016. Based on previous literature reviews and the field investigations, a geospatial database was established in geographic information system (GIS) based on topography, geological conditions, environmental conditions and human activities factors, including 22 conditioning factors. Then, based on samples of landslides and non-landslides, the 10-fold cross-validation was used to select the best training and test datasets. Finally, the LSMs of Wushan County were generated. Subsequently, susceptibility maps of the two models were divided into: very low, low, medium, high and very high class using experts' experience method. The two models were compared with area under the receiver operating characteristic (ROC) plot values (AUC) and confusion matrix. The AUC values of the ANN and RF models' test dataset was 0.966 and 1.000, and the accuracy were 0.953 and 1.000, respectively. According to the mean decrease Gini of RF model, the most important conditioning factors were elevation (a mean decrease accuracy of 72.38), followed by annual average rainfall (47.95) and POI kernel density (45.63). Above all, it is concluded that both ANN and RF models were satisfactory for the LSM of Wushan County at the mountainous region. The RF model had an outperformer prediction and is considerably more efficient than the ANN model.

Keywords: Mountainous region; landslide susceptibility mapping; random forest; artificial neural network; Wushan County

1 Introduction

As one of the most important natural disasters, landslides cause heavy loss of human life and property in many countries all over the world yearly (Mondal and Mandal 2017). According to Petley (Petley 2012), 2,620 fatal landslides were recorded worldwide from 2004 to 2010, resulting in the death of at least 32,322 people. The Safe Land-FP7 project reports that landslides

in China are at high risk and widely distributed(Huang and Li 2011), causing more than 700 deaths each year and damages to properties and infrastructures valued at CNY 20 billion. Especially in the southwest of China, the Three Gorges Reservoir area (TGRA) has been having large relief of terrain, coupled with the periodic rise and fall of water level since 2010, resulting in the instability of the slope on both sides. This has led to the frequency of new landslides, putting the life and property of local residents at risk. For example, Shanshucao landslide destroyed 6,140 houses, 40,000 km^2 citrus forest, Daling power station and many roads in 2014, resulting in economic losses of more than 32 million yuan(Zhou et al. 2020). At the same time, the occurrence of these severe events accelerated the process of landslide disaster prevention and prediction research.

Traditional landslide research is limited to the occurrence mechanism of single landslide, and cannot explore the law of macro scale. As an effective tool for detecting the likelihood of landslide occurring in a region (Kanungo et al. 2008), landslide susceptibility mapping (LSM) is a very valuable countermeasure for governments and engineers to mitigate destructions by landslides (Harp et al. 2011; Steger et al. 2017). Therefore, LSM related research has become one of the hottest topics in recent years. Scholars have conducted landslide susceptibility studies in some regions (Pourghasemi et al. 2012; Mondal and Mandal 2017; Silalahi et al. 2019), but few in Wushan County of TGRA. As a server landslide disaster area in TGRA, it is of great significance to explore the landslide susceptibility law and carry out landslide prevention and control work.

In addition, the application of modeling methodologies plays a key role in the effectiveness of LMS. By and large, methods of LSM can be qualitative or quantitative (Reichenbach et al. 2018). Qualitative methods generally need to combine the professional knowledge of engineering geologists and geomorphologists, and have certain subjectivity(Wang et al. 2021). Quantitative methods include traditional statistical methods, physical-based methods, and advanced data mining technologies(Reichenbach et al. 2018) . Although the statistical methods can calculate the spatial probability of landslide occurrence and provide spatial information, it is inevitable to carry out the process of weight assignment, such as frequency ratio (FR) (Silalahi et al. 2019), weights of evidence (Ding et al. 2016), logistic regression (LR) etc (Xie et al. 2018). In physical-based methods, the parameters or weights of some factors can be obtained through experiments, but they are only applicable to a small number of landslides or small-sized areas(Li and Lan 2015). More recently, advanced data mining technologies combined with GIS (Geographic Information Science) and remote-sensing for LSM are more favored by researchers. Data mining technologies can reduce bias and subjectivity in the weight assignment and are suitable for large-scale areas, including support vector machine model (SVM) (Huang and Zhao 2018), artificial neural network model (ANN) (Andang Suryana et al. 2019), random forest model (RF) (Sun et al. 2021), classification and regression trees etc (Chen et al. 2017).

However, data mining method also have different performances due to their different algorithms and study areas. For example, as one of the data mining methods used for spatial modeling of landslides, ANN model is confirmed to be able to deliver high and reliable predictions (Lucchese et al. 2021). RF model is a nonparametric and non-linear algorithm. It can deal with large datasets containing both categorical and numerical data and account for complex interactions and non-linearity between variables (Taalab et al. 2018). Sahin et al. (Sahin et al. 2018) conducted a comparative assessment of different

methods for LSM and the results showed that AUC value of RF model is higher than that of others. Hence, comparing models to find out the most accurate ones has become one of the important research contents. Although researchers have carried out several comparative studies on different models (Hong et al. 2016; Chen et al. 2017; Kalantar et al. 2017), the comparative studies of some advanced methods are still insufficiently applied to Wushan County in TGRA and the relationship between landslides and conditioning factors is not clear. Therefore, two typical and advanced data mining models: ANN and RF models, are selected as examples in this study. According to previous studies (Tian et al. 2018; Harmouzi et al. 2019; Li et al. 2021), there is no doubt that these two methods are effective to analyze the possibility of landslides. In addition, it is of great significance to explore the relationship between landslides and conditioning factors and to excavate the dominant factors for accurate prevention and detection of landslides.

In this study, firstly, based on GIS technology, remote sensing data, topographic data and geological data, four type factors (including 22 secondary factors), 866 historical landslides and non-landslides were used as data sources. Secondly, RF and ANN models were used to construct LSMs in Wushan County of TGRA. ROC (receiver operating characteristic), AUC (the area under the curve) value of ROC curve, precision and other indicators were used for quantitative comparative study. Lastly, we deeply analyzed the importance of landslide conditioning factors and the spatial distribution characteristics of landslides. This article has three objectives: (1) To study the LSM of mountainous county (Wushan County); (2) To compare and analyze two advanced sophisticated data mining techniques; (3) To identify the dominant factors of landslides.

2 Materials

2.1 The Study Area

Wushan County (between 109°33′ - 110°11′E and 30°46′N - 31°28′N) is located in the east of Chongqing, at the heart of the Three Gorges Reservoir area, known as the ‘Northeast Chongqing Portal’ (Fig.1). Its altitude ranges from 63 m to 2,688 m above the sea level, as calculated from a 30 m regular grid digital elevation model (DEM). Wushan County is located at the junction of the three structural systems of the Dabashan arc structure, the Eastern Sichuan fold belt and the Sichuan-Hubei-Hunan-Guizhou uplift fold belt, with sophisticated structural stress fields. The terrain comprises a succession of limestone ridges and gorges, with inter-gorge valleys where interbedded mudstones, shales, and thinly bedded limestones predominate(Liu et al. 2020). Therefore, the stability of the study area is poor, coupled with rainfall softening and reservoir erosion leading to easy formation of landslides. In addition to Triassic and Permian strata, Wushan County is also distributed in Devonian, Cambrian, Eocene, etc. The geology of the area is mainly composed of a Sinian-Jurassic sedimentary, and a pre-Sinian crystalline basement. Under a humid subtropical monsoon climate and obvious three-dimensional climate characteristics, the climate of Wushan County is mild, with abundant rainfall. The annual average precipitation is 1,041 mm mild climate and mean annual temperature is 18.4 °C (Xia et al. 2018). 69 % precipitation mainly occurs in summer and

autumn.

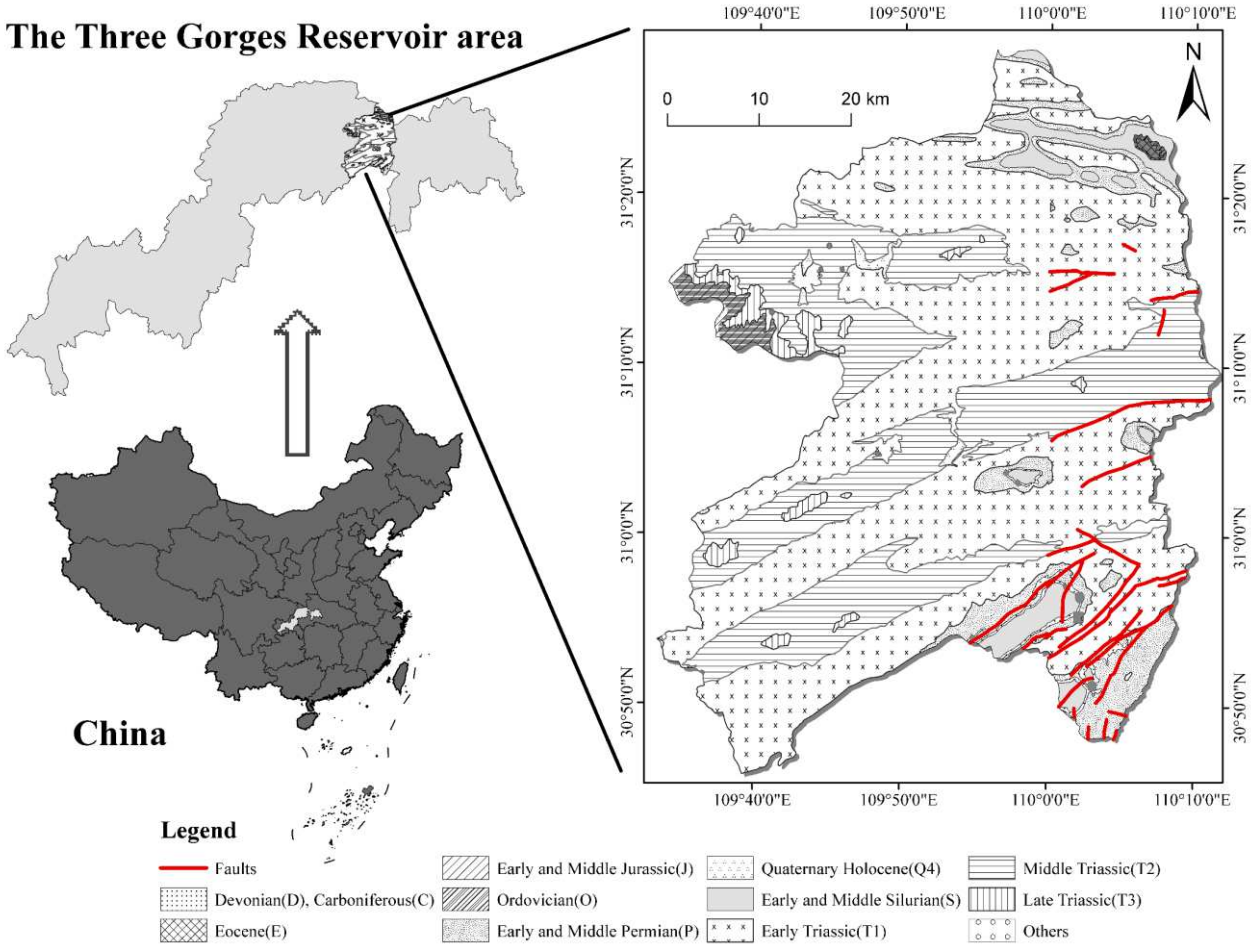


Figure 1 Location and geological map of the study area.

Affected by complex mountainous landforms, involving lithological folds, human engineering activities, and climatic and hydrological influences, geological disasters occur frequently. The Three Gorges reservoir area has become a serious landslide disaster area. (Li et al. 2019). The frequent landslides of Wushan County have greatly threatened the lives and properties of people. Since 1998, the disasters throughout the county have not diminished: more than 6,000 houses have been damaged and up to 20,000 people have been affected. There has been an economic loss of more than 100 million yuan(Liu et al. 2020).

2.2 Landslide Inventory

The preparation of a landslide inventory is a key for landslide susceptibility modeling. It can assess the relationship between landslide distribution and conditioning factors including landslides of names, geographic position, areas and other basic information. Firstly, the historical landslides come from Chongqing Municipal Geological Environment Agency. Since 2007, the geological disaster management department of Chongqing formed a geological disaster garrison (about 500 members) permanently stationed in geological disaster-prone areas. Once the landslide emergency is found, they will be responsible for the evacuation of personnel and recording basic information on the landslide. In addition, they judge the landslide type, scale, sliding direction and other basic information by landslide geological survey, engineering geophysical exploration and brick

exploration. Ultimately, 1,001 landslides were identified for 2001-2016, including attribute information for landslides of names, geographic position, areas, volumes, and occurrence times. There are 963 small-medium (volume ≤100 × 10^4 m^3) and soil landslides, while there are only 38 large (volume ＞100×10^4 m^3) and rock landslides. Among the small-medium and soil landslides, there are 866 landslides whose movement typologies are slide, 92 movement typologies are flow, and 5 movement typologies are fall. Scholars have suggested that different types of landslides should be dealt with in different manners (Wang et al. 2019), but there are only limited data on large and rock landslides and those movement typologies are flow and fall. Therefore, this research will explore small-medium, soil, and slide landslides to establish the landslide inventory, with the locations of a total of 866 of them mapped in Fig.2. Additionally, the landslide inventory showed most of the landslides that occurred on both sides of the river.

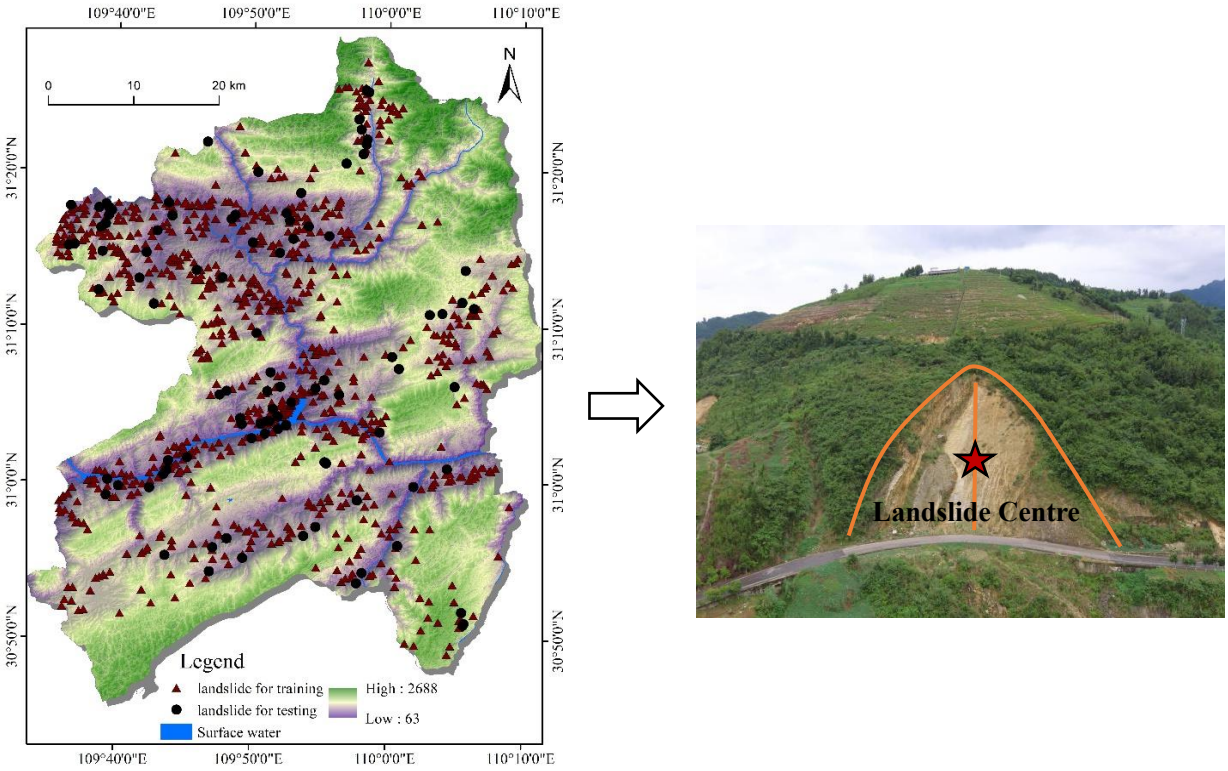


Figure 2 Distribution and on-site investigation of landslides in the study area.

These landslides are typically small. Landslide area varies from 100 m^2 to 96,800 m^2. According to the formation mechanism of the landslide, it is found that the occurrence of a landslide is caused by the movement of a certain position rather than the entire landslide area. In the field survey, we mark the latitude and longitude of the geometric center of the landslide as the location of the landslide (Fig.2). Furthermore, this study uses two different symbols to distinguish landslide for training and testing in landslide inventory map randomly. Consequently, training datasets of landslide were used for modeling, which will make it predictive. The remaining landslides were used to test learning ability of the two models and evaluation processes. In order to obtain good fitting and prediction accuracy, distribution of the training landslide on conditioning factors and selecting a suitable method cannot be ignored(Abedini et al. 2018). Random selection is a common method, which can avoid

subjectivity and enhance the learning ability of the model. All datasets were randomly classified into training (70%) and testing (30%).

2.3 Data

Except the historical landslides, this paper also involves remote sensing data, topographic data, geological data, etc. These data sources, types and accuracy are shown in Table 1. Baidu Map Service (http://map.baidu.com) provides POI (point of interest) data in 2016. These data points include attractions, hospitals, universities etc to characterize the distribution and intensity of human infrastructure.

Table 1 Data and data sources.

Data Name	Data Sources	Type	Scale
Historical landslide	Chongqing Geological monitoring station	Dataset	
Elevation	Aster satellite	Grid	30 m
Geological data	National Geological Data Center	Grid	1:200,000
Land cover	Chongqing Municipal Bureau of land and resources	Vector	1:100,000
Administrative division	Chongqing Municipal Bureau of land and resources	Vector	1:100,000
River network	Chongqing Water Resources Bureau	Vector	1:100,000
Annual rainfall	Chongqing Meteorological Administration	Dataset	90 m
Road	Chongqing Transportation Commission	Vector	1:100,000
Satellite image	Geospatial Data Cloud platform	Grid	30 m
POI of Chongqing	Web Crawler	Dataset	

3 Methods

The study aims to compare the two different machine learning methods (RF and ANN) for LSM in Wushan County. The assessment procedure of this study included four parts: (a) construction of the landslide and non-landslide spatial database; (b) data preparation of landslide conditioning factors and the training and test datasets in ArcGIS 10.4 software and construction; (c) generation of the LSM; and (d) evaluation and comparison of the two models (Fig.3).

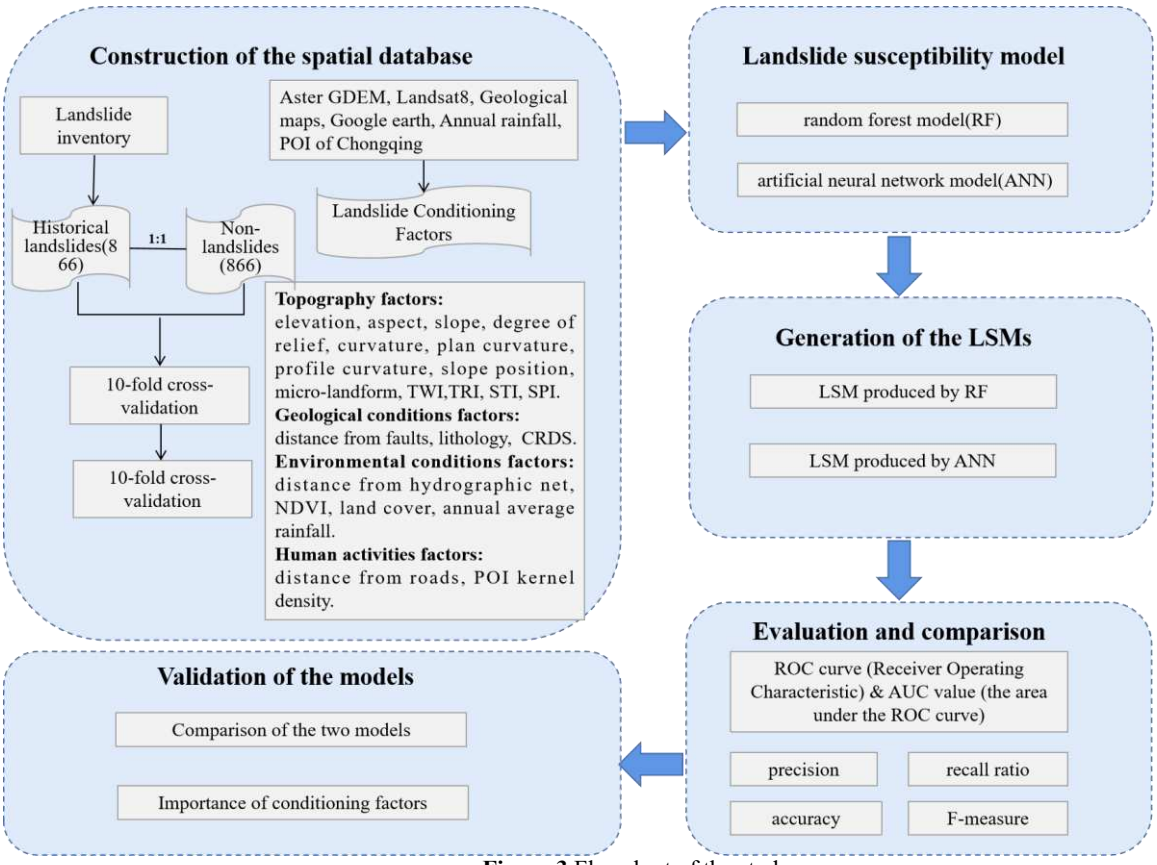


Figure 3 Flowchart of the study.

3.1 Construction of the Spatial Database

The selection of conditioning factors directly affects the accuracy of models and LSM. Conditioning factors for landslide initiation could be classified into inherent and triggering factors. The former refers to the internal factors that play a controlling role in the occurrence of landslides, including mainly topography, geological conditions, and fault structure; the latter refers to the external factors that trigger the occurrence of landslides, such as environmental conditions and human activities. Based on the principles of field investigation, satellite images, the literature review and available data, there were 22 factors selected, including the topography factors: topographic wetness index (TWI), elevation, slope position, plan curvature, aspect, slope, degree of relief, curvature, profile curvature, micro-landform, stream power index (SPI) (Basu and Pal 2017), terrain roughness index (TRI) (Althuwaynee et al. 2014), sediment transport index (STI) (Pourghasemi et al. 2012). Geological conditions factors include: distance from faults, lithology, combination reclassification of stratum dip direction and slope aspect (CRDS) (Sun et al. 2020). Environmental conditions factors are: land cover, annual average rainfall, distance from hydrographic net, normalized vegetation index (NDVI). Human activities factors include POI kernel density and distance from roads. Particularly, point-of-interest (POI) represents the location of shops, hospitals, schools etc. on maps; it can show information on economic society, which is closely linked to human activities (Bakillah et al. 2014). Usually, the higher the POI kernel density, the better would be the urban development in an area. It is noteworthy that previous studies have rarely considered the impact of human activities on landslides, but such activities play a non-negligible role in the occurrence of

landslides (Reichenbach et al. 2018). Therefore, this research selects POI kernel density, a representative factor of the type of human activities, to represent one of the factors in human activities. The description of each factor is shown in Table 2.

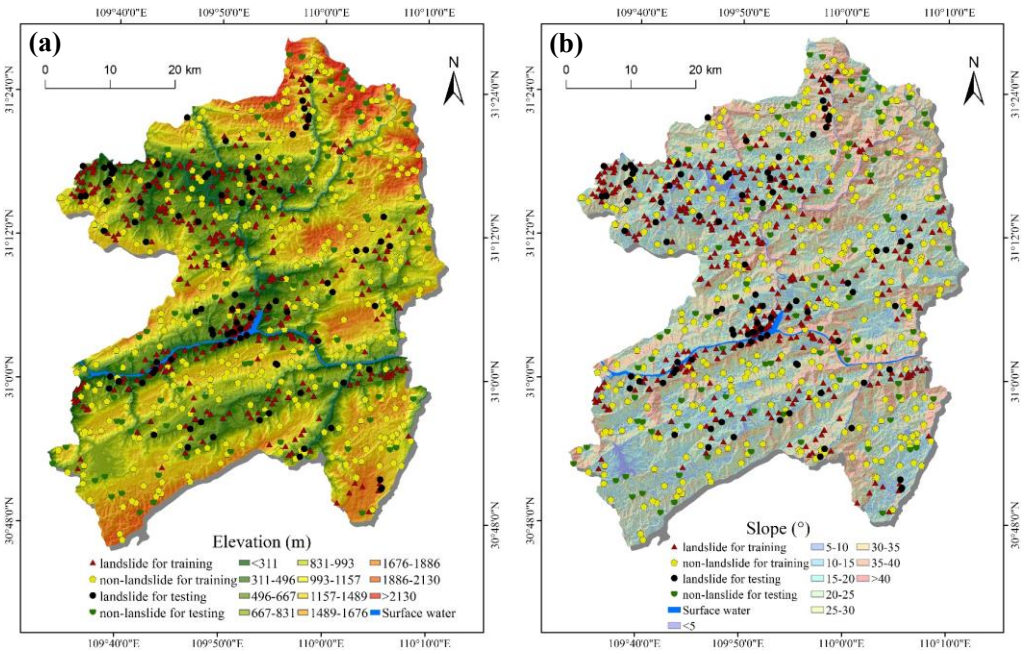
Table 2 Description of landslides conditioning factors.

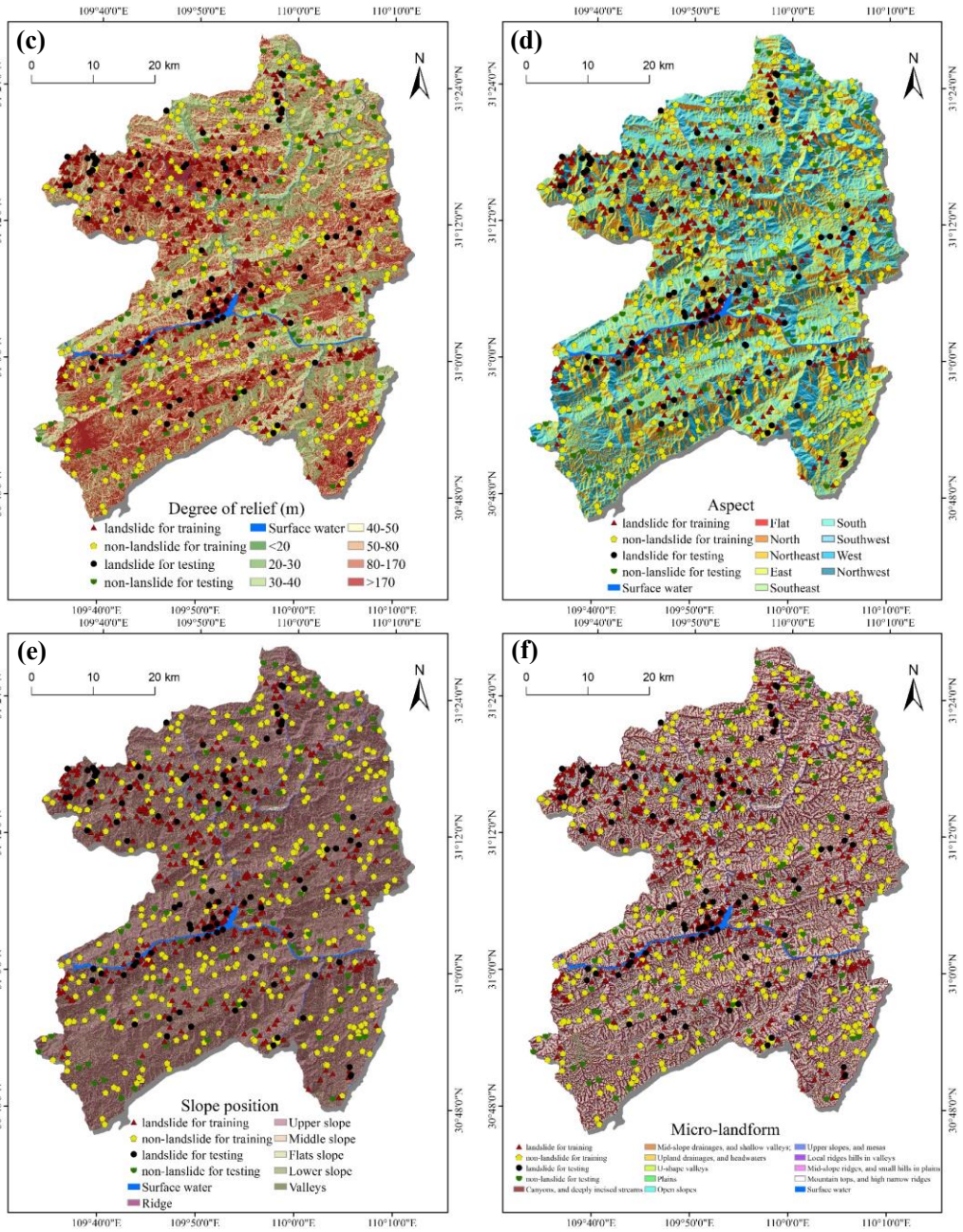
Influence	Conditioning factor	Description
Topography	Elevation (m)	Elevation is the basic data for extracting terrain data, usually reflecting climate and vegetation patterns.
	Aspect	Aspect is defined as the direction of projection of slope normal on horizontal plane. It plays an important role in mountain ecology. The orientation of mountain has an effect on sunshine hours and solar radiation intensity.
	Slope (°)	Slope refers to the degree of surface unit steepness, equal to the vertical height of the slope divided by horizontal distance.
	degree of relief (m)	It is a macroscopic index that describes the topographic characteristics of a region and characterizes the potential energy of surface erosion and material movement. In a certain area, the degree of relief is equal to the maximum value of elevation minus the minimum value. Different degree of relief regions have different influence on the occurrence of landslides.
	Curvature	Curvature can be seen as the slope of a slope. It stresses the materials on the slope and affects the movement of water on the slope surface.
	Plan Curvature	It is the slope perpendicular to the direction of the maximum slope. A positive value indicates that the cell is part of the lateral concave slope. A negative value indicates that the slope is laterally convex.
	Profile Curvature	The change rate of elevation along the maximum direction of slope becomes profile curvature, that is, the change rate of ground slope. A positive value indicates that the cell is part of a concave up slope. A negative value indicates that the slope is raised.
	Slope position	The core work of slope position extraction is the division of landform parts and landform types.
	Micro-landform	Relatively small-scale geomorphologic form and the smallest geomorphologic unit.
	TWI	Based on DEM data, TWI comprehensively considers the influence of topography and soil on soil moisture distribution, and has important theoretical and application value in the study of spatial distribution of soil moisture in the basin.
	TRI	TRI is the surface roughness index. Based on DEM, it is an index to measure surface fluctuation and erosion degree.
	STI	STI is a sediment transport index, which is used to characterize the surface sediment transport and deposition.
	SPI	The stream power index (SPI) is used to calculate the erosion value of surface water based on DEM data. The greater the value, the more obvious the erosion of surface water flow.
Geological conditions	Distance from faults(m)	Setting a buffer zone to the fault, the closer to the main fault area, the looser the slope soil, more likely to cause landslides.
	Lithology	There are abundant outcropped strata and many types of lithology in the study area. Due to the different degrees of weakness of various strata lithology, their shear strength is different, and the difficulty of landslide is different.
	CRDS	CRDS refers to the relationship between the dip angle of rock and the aspect, and the direction of the sedimentary heap.
Environmental Conditions	Distance from hydrographic net (m)	The study area water system to establish buffers to determine the different scope of influence. Because the bank slope is cut and eroded by water flow, the slope closer to the river is more unstable, which leads to the occurrence of landslides.
	NDVI	Vegetation growth is another key factor reflecting slope stability. Therefore, the normalized difference vegetation index (NDVI) was used to detect the vegetation growth status, vegetation coverage and eliminate some radiation errors.
	Land cover	Land cover refers to the surface complex covered by natural and artificial buildings, including vegetation, soil, lakes and various buildings, with different land stability.
	Annual average rainfall (mm)	The data are derived from the average rainfall from 2000 to 2014 to reflect the long-term rainfall and distribution characteristics in the study area.
Human Activities	Distance from roads (m)	The construction of a large number of roads inevitably poses a threat to the stability of the slope and is easy to induce landslides.
	POI kernel density	POI nuclear density covers almost all types of human engineering buildings. Similar to the construction of highways, these buildings will also loose the soil around the slope, thus affecting the slope stability.

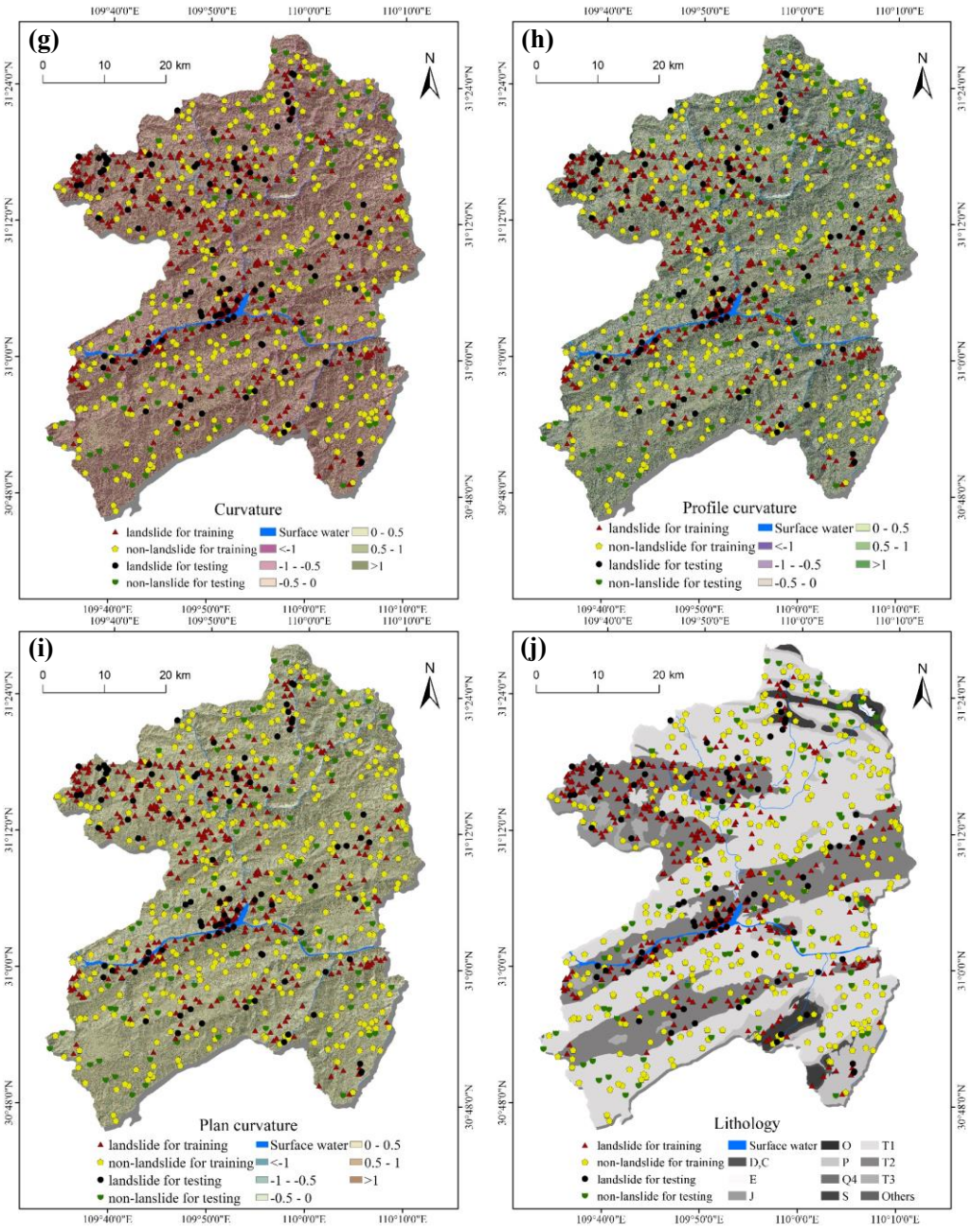
DEM from ASTER GDEM was used to extract the topography factors. Land cover and NDVI come from Landsat 8 OLI satellite images. We vectorized the 1: 200,000 geological maps to obtain lithologic and faults factors. Similarly, the hydro-

graphic nets and roads factors were vectorized by Google Earth images. CRDS was extracted by subtraction and reclassification of aspects and tendencies. The multi-level buffer zones are established for faults, hydrographic nets and roads to generate distances from faults, hydrographic nets and roads. Rainfall data were derived from the local climate stations and using Kriging spatial interpolation method to generate annual average rainfall factor. Lastly, Baidu Map Service provides POI records and a total of about 12,688 records in the study area, including shops, hospitals, schools, etc. POI kernel density was generated by Kernel Density tool.

In order to unify the spatial resolution of 22 landslide conditioning factors, all factors were converted to a raster grid (30 × 30 m cells) that corresponds to the DEM resolution. According to relevant literature (Huang et al. 2017), a pixel of 30 × 30 m can not only capture the spatial characteristics of a landslide in detail, but also reduce the workload and time. The acquisition path and classification of conditioning factor is directly related to their different natures. Therefore, according to references, field investigation and experts' experience, the corresponding classification scheme for each continuous factor was established in Table 3. In short, the spatial thematic maps of landslide conditioning factors after reclassification were constructed with the grid unit of 30 m spatial resolution. (Fig.4).

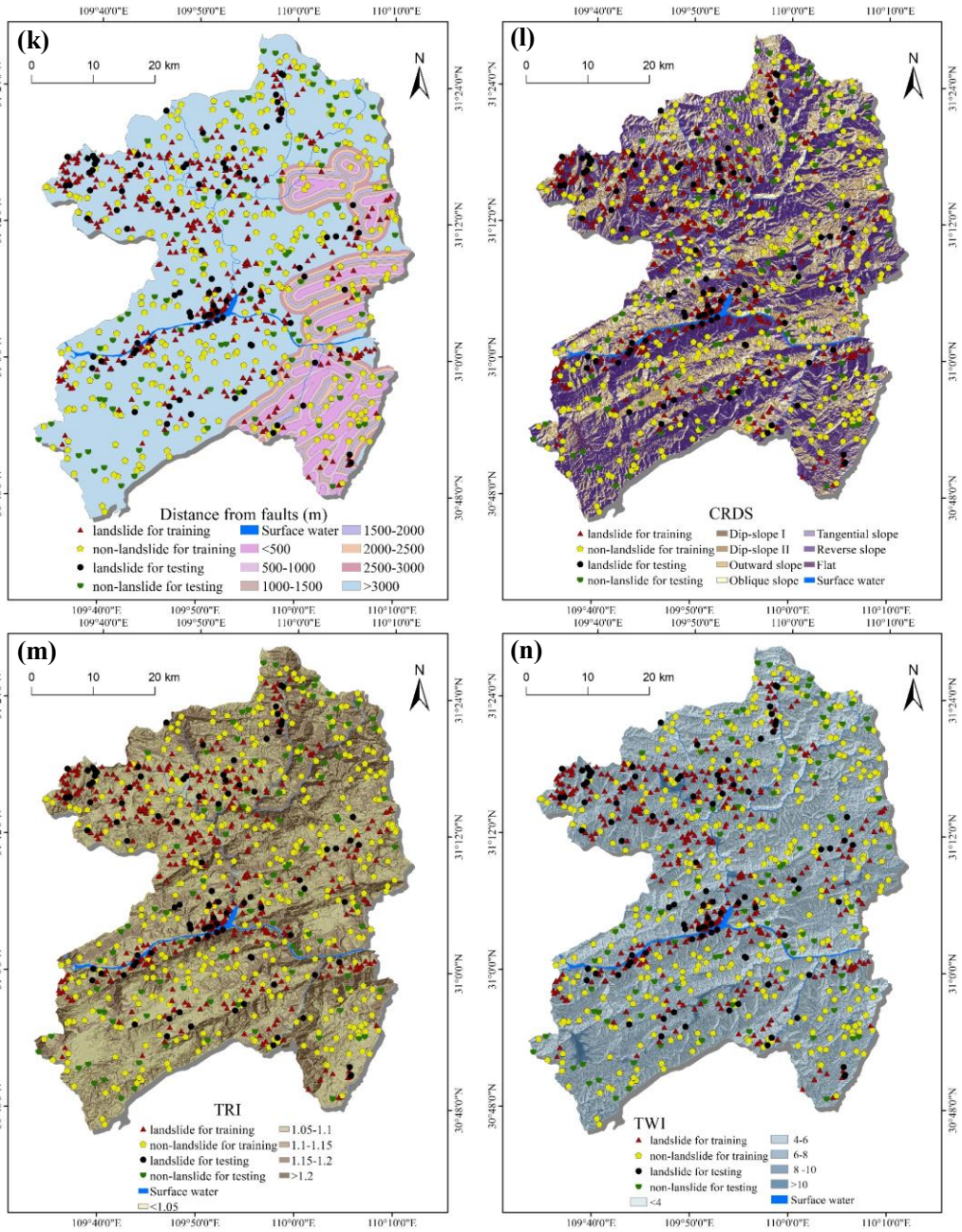


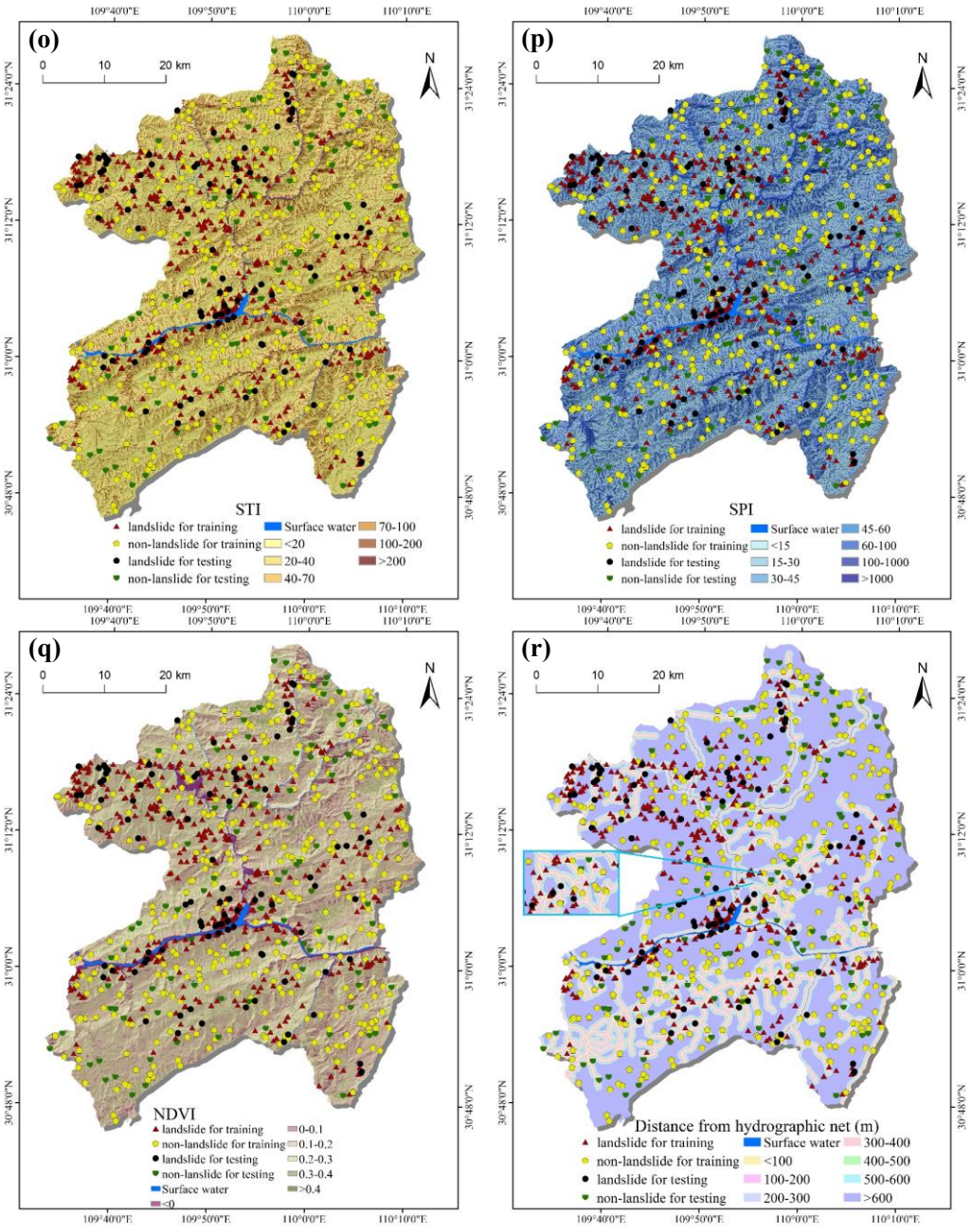

(c)
0 10 20 km
N
Degree of relief (m)
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
<20
20-30
30-40
40-50
50-80
80-170
>170
(d)
Aspect
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
Flat
North
Northeast
East
Southeast
South
Southwest
West
Northwest
(e)
Slope position
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
Ridge
Upper slope
Middle slope
Flats slope
Lower slope
Valleys
(f)
Micro-landform
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Canyons, and deeply incised streams
Mid-slope drainages, and shallow valleys;
Upland drainages, and headwaters
U-shape valleys
Plains
Open slopes
Upper slopes, and mesas
Local ridges hills in valleys
Mid-slope ridges, and small hills in plains
Mountain tops, and high narrow ridges
Surface water
109°40'0"E
109°50'0"E
110°0'0"E
110°10'0"E
31°24'0"N
31°12'0"N
31°0'0"N
30°48'0"N


(g)
Curvature
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
<-1
-1 - -0.5
-0.5 - 0
0 - 0.5
0.5 - 1
>1
(h)
Profile curvature
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
<-1
-1 - -0.5
-0.5 - 0
0 - 0.5
0.5 - 1
>1
(i)
Plan curvature
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
<-1
-1 - -0.5
-0.5 - 0
0 - 0.5
0.5 - 1
>1
(j)
Lithology
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
D,C
E
J
O
P
Q4
S
T1
T2
T3
Others
0
10
20 km
N

178

179


(k)
0 10 20 km
N
Distance from faults (m)
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
<500
500-1000
1000-1500
1500-2000
2000-2500
2500-3000
>3000
(l)
CRDS
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Dip-slope I
Dip-slope II
Outward slope
Oblique slope
Tangential slope
Reverse slope
Flat
Surface water
(m)
TRI
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
<1.05
1.05-1.1
1.1-1.15
1.15-1.2
>1.2
(n)
TWI
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
<4
4-6
6-8
8 -10
>10
Surface water


(o)
0 10 20 km
N
STI
landslide for training
non-landslide for training
landslide for testing
non-lanslide for testing
Surface water
<20
20-40
40-70
70-100
100-200
>200
(p)
SPI
<15
15-30
30-45
45-60
60-100
100-1000
>1000
(q)
NDVI
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
>0.4
<0
(r)
Distance from hydrographic net (m)
<100
100-200
200-300
300-400
400-500
500-600
>600
109°40'0"E
109°50'0"E
110°0'0"E
110°10'0"E
31°24'0"N
31°12'0"N
31°0'0"N
30°48'0"N

182

183

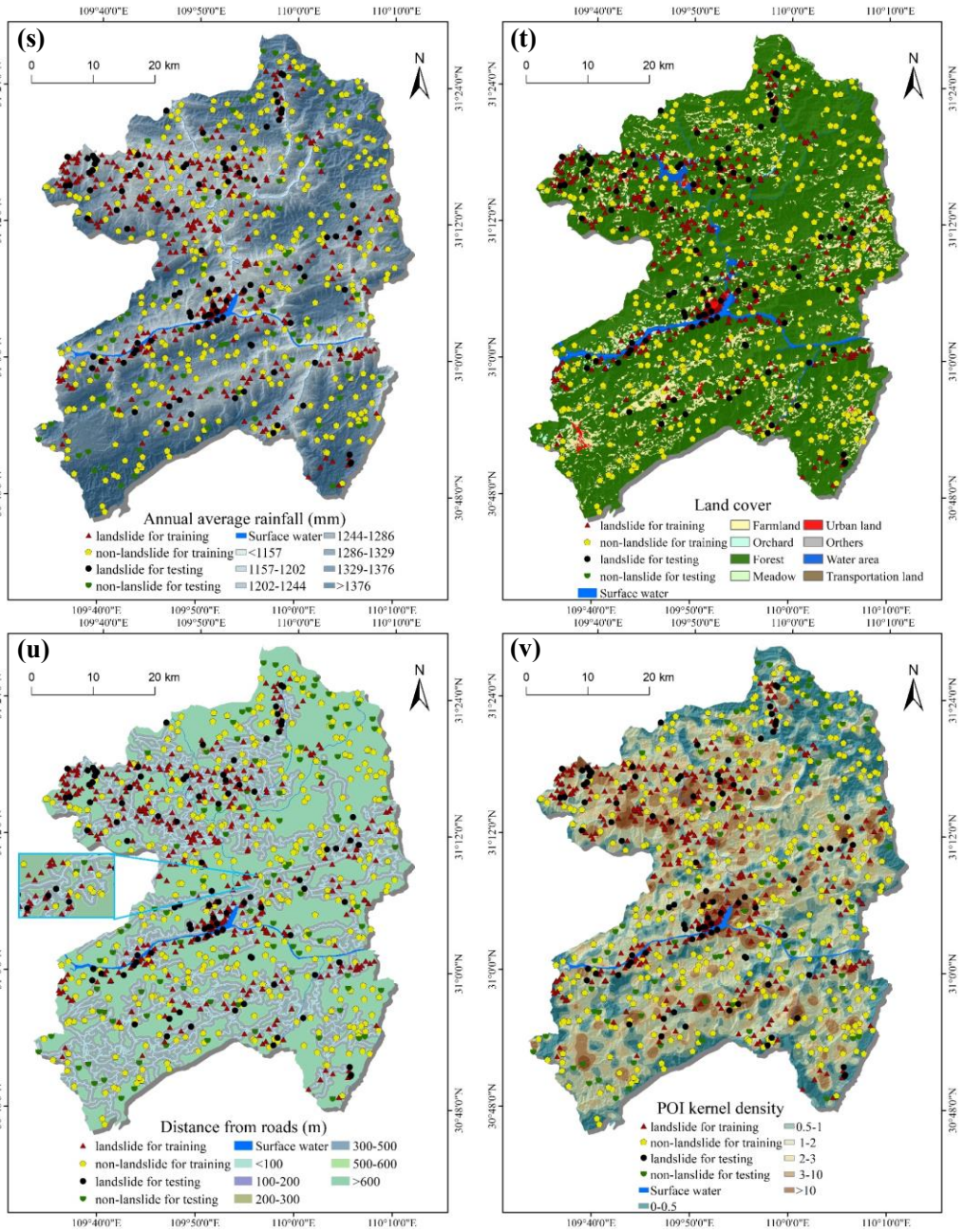


Figure 4 Thematic maps of landslide conditioning factors: (a) elevation; (b) slope; (c) degree of relief; (d) aspect; (e) slope position; (f) micro-landform; (g) curvature; (h) profile curvature; (i) plan curvature; (j) lithology; (k) distance from faults; (l) CRDS; (m) TRI; (n) TWI; (o) STI; (p) SPI; (q) NDVI; (r) distance from hydrographic net; (s) annual average rainfall; (t) land cover; (u) distance from roads; and (v) POI kernel density.

Note: The landslide and non-landslides for testing are all displayed, while a part of landslide and non-landslides for training are selected randomly for display.

To lower the data discreteness, the classification factors were normalized(Wen 2015). Therefore, the values of each factor in different dimensions are limited to the interval [0, 1]. The normalization formula is as follows:

$$X^* = (X - X_{\min})/(X_{\max} - X_{\min}) \quad (1)$$

Where, X^* represents the normalized data; X represents the data before normalization; $X_{\min}$ represents the minimum value of each factor; and $X_{\max}$ represents the maximum value of each factor.

Table 3 Classification of landslide-conditioning factors.

Conditioning Factor	Classification Standard
Elevation (m)	1. <311; 2. 311 ~ 496; 3. 496 ~ 667; 4. 667 ~ 831; 5. 831 ~ 993; 6. 993~1157; 7. 1157 ~ 1489; 8. 1489 ~ 1676; 9. 1676 ~ 1886; 10. 1886 ~ 2130; 11. >2130
Slope (°)	1. < 5; 2. 5 ~ 10; 3. 10 ~ 15; 4. 15 ~ 20; 5. 20 ~ 25; 6. 25 ~ 30; 7. 30 ~ 35; 8. 35 ~ 40; 9. >40
Degree of relief (m)	1. <20; 2. 20~30; 3. 30 ~ 40; 4. 40 ~ 50; 5. 50 ~ 80; 6. 80 ~ 170; 7. >170
Aspect	1. Flat; 2. North; 3. Northeast; 4. East; 5. Southeast; 6. South; 7. Southwest; 8. West; 9. Northwest
Slope position	1. Ridge; 2. Upper slope; 3. Middle slope; 4. Flats slope; 5. Lower slope; 6. Valleys;
Micro-landform	1. Canyons, and deeply incised streams; 2. Mid-slope drainages, and shallow valleys; 3. Upland drainages, and headwaters; 4. U-shape valleys; 5. Plains; 6. Open slopes; 7. Upper slopes, and mesas; 8. Local ridges hills in valleys; 9. Mid-slope ridges, and small hills in plains; 10. Mountain tops, and high narrow ridges
Curvature	1. <-1; 2. -1 ~ -0.5; 3. -0.5 ~ 0; 4. 0 ~ 0.5; 5. 0.5 ~ 1; 6. > 1
Profile curvature	1. <-1; 2. -1 ~ -0.5; 3. -0.5 ~ 0; 4. 0 ~ 0.5; 5. 0.5 ~ 1; 6. > 1
Plan curvature	1. <-1; 2. -1 ~ -0.5; 3. -0.5 ~ 0; 4. 0 ~ 0.5; 5. 0.5 ~ 1; 6. > 1
Lithology	1. D, C; 2. E; 3. J; 4. O; 5. P; 6. Q_4; 7. S; 8. T_1; 9. T_2; 10. T_3; 11. Others
Distance from faults (m)	1. < 500; 2. 50 ~ 1000; 3.1000 ~ 1500; 4. 1500 ~ 2000; 5. 2000 ~ 2500; 6. 2500 ~ 3000; 7. > 3000
CRDS	1. Dip-slope I; 2. Dip-slope II; 3. Outward slope; 4. Oblique slope; 5. Tangential slope; 6. Reverse slope; 7. Flat
TRI	1. < 1.05; 2. 1.05 ~ 1.1; 3. 1.1 ~ 1.15; 4. 1.15 ~ 1.2; 5. > 1.2
TWI	1. < 4; 2. 4 ~ 6; 3. 6 ~ 8; 4. 8 ~ 10; 5. >10
STI	1. < 20; 2. 20 ~ 40; 3. 40 ~ 70; 4. 70 ~ 100; 5. 100 ~ 200; 6. > 200
SPI	1. < 15; 2. 15 ~ 30; 3. 30 ~ 45; 4. 45 ~ 60; 5. 60 ~ 100; 6. 100 ~ 1000; 7. > 1000
NDVI	1. <0; 2. 0 ~ 0.1; 3. 0.1 ~ 0.2; 4. 0.2 ~ 0.3; 5. 0.3 ~ 0.4; 6. >0.4
Distance from hydrographic net (m)	1. < 100; 2. 100 ~ 200; 3. 200 ~ 300; 4. 300 ~ 500; 5. 500 ~ 600; 6. > 600
Annual average rainfall (mm)	1. <1157; 2. 1157 ~ 1202; 3. 1202 ~ 1244; 4. 1244 ~ 1286; 5. 1286 ~ 1329; 6. 1329 ~ 1376; 7. >1376
Land cover	1. Meadow; 2. Farmland; 3. Water area; 4. Forest; 5. Orchard; 6. Transportation land; 7. Urban land; 8. Others
Distance from roads (m)	1. <100; 2. 100 ~ 200; 3. 200 ~ 300; 4. 300 ~ 500; 5. 500 ~ 600; 6. >600
POI kernel density	1. 0 ~ 0.5; 2. 0.5 ~ 1; 3. 1 ~ 2; 4. 2 ~ 3; 5. 3 ~ 10; 6. >10

3.2 Preparation of the Training and Testing Datasets

In our study, landslide and non-landslide formed all datasets. 866 historical landslides consisted of positive cells and non-landslide served as negative cells. Additional non-landslide cells could expand the data size for machine learning, but it will cause a sample imbalance problem, thereby biasing the classifier towards the non-landslide data and negatively affecting its performance (Wang et al. 2020). After many attempts at the ratios of positive and negative cells, a balanced (1:1) positive and negative dataset has been proposed, and 866 non-landslides were randomly extracted from the non-landslide area. Then, all datasets were randomly divided into 70 % (the training dataset) and 30 % (the testing dataset). Specially, RF and ANN models do not need to consider the problem of multivariate collinearity in general regression analysis and the predictive ability of the two models is not affected by multicollinearity.

3.3 Landslide Susceptibility Models

3.3.1 Artificial neural network

Artificial Neural Network (ANN) (Pavel et al. 2011; Chakraborty and Goswami 2017) is a mathematical model used for distributed parallel information processing with capability to abstracts human brain neurons from the perspective of

information processing, acquire, present, and calculate mapping from data multivariate space to another. ANN is an operation model composed of a large number of interconnected nodes. Each node represents a neuron. The conversion function in the neuron is called activation function (Melchiorre et al. 2008). Without the limitations of statistical techniques of oversimplification, prohibitive data requirements and ineffective input data acquisition, ANN has a remarkable ability for handling imperfect or incomplete data and the nonlinear and complex problems.

The most prominently employed neural network method is the Back Propagation (BP) algorithm. BP neural network is a multi-layer artificial neural network, which consists of information forward propagation and error back propagation (Arora et al. 2003). The output result can be expressed by Eq. (2):

$$y_n = \sum_{j=1}^{N_2} W_{kj} f\left(\sum_{i=1}^{N_1} W_{ij}\, x_i + b_k\right) \tag{2}$$

where, y_n is the nth output, W_{kj} is weight from the k neuron in the hidden layer to the j neuron in the output layer. $f(.)$ is the transfer function of the hidden layer neuron. $b_k$is the bias value of the k neuron in the hidden layer. $N_1$is the number of neurons in the input layer, N_2 is the number of neurons in hidden layer. As long as the number of hidden layer neurons is sufficient, the BP network with a hidden layer can approximate any complex nonlinear function with arbitrary accuracy.

As shown in Fig.5, the BP neural network model comprises three layers, namely, input layer (i.e., landslide conditioning factors), hidden layers and output layer (i.e., landslide susceptibility). There are many neurons in the model which are connected in layers. ANN model is trained in R studio software.

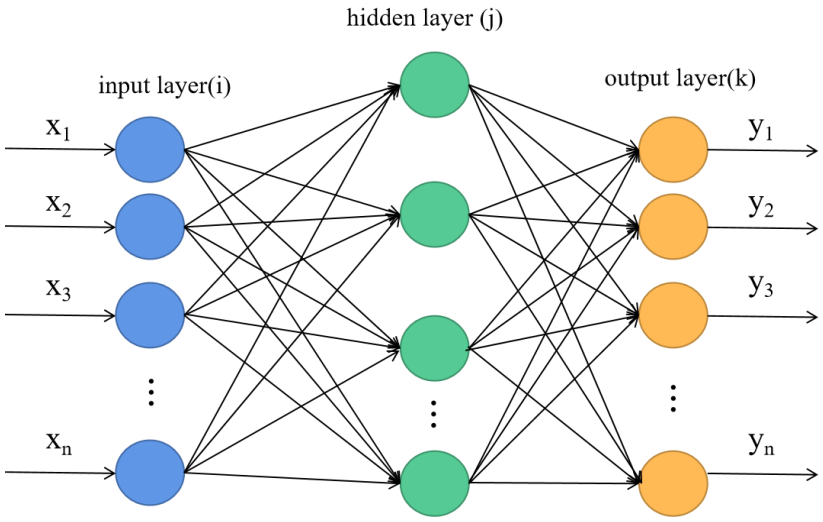


Figure 5 BP neural network model structure

3.3.2 Random Forest

First proposed by Breiman, Random Forest (RF) is an ensemble method of separately trained binary decision trees. Multiple decision trees are established through different data subsets, and their judgment results are voted to obtain the output results of the RF. Numerous researches indicate that the random forest has a relatively high tolerance to outliers and noises, and it is not easy to produce over-fitting (Sahin et al. 2018).

Combining *N* unrelated decision tree models $[h(X, \theta_k); k = 1, \ldots\ldots]$ through classification model is the core of RF. In

the model, each decision tree judges and predicts the classification of the sample (via classification algorithm). In order to improve the performance of random forest models, an irrelevant training set is established to provide differences between models. Through sample training, different classification models $h_1(X)$ ... ... $h_k(X)$ are obtained, and then the random forest models are combined by using these classifications. Then to vote:

$$H(x) = arg_z^{max} \Sigma_{i=1}^{k} I(h_i(x) = Z) \quad (3)$$

Where, *H(x)* denotes a random forest model, h_i is a single decision tree model, *Z* denotes an output variable, and *I(.)* is an explicit function.

RF model is trained in R studio software. Fig. 6 shows the steps of the RF algorithm.

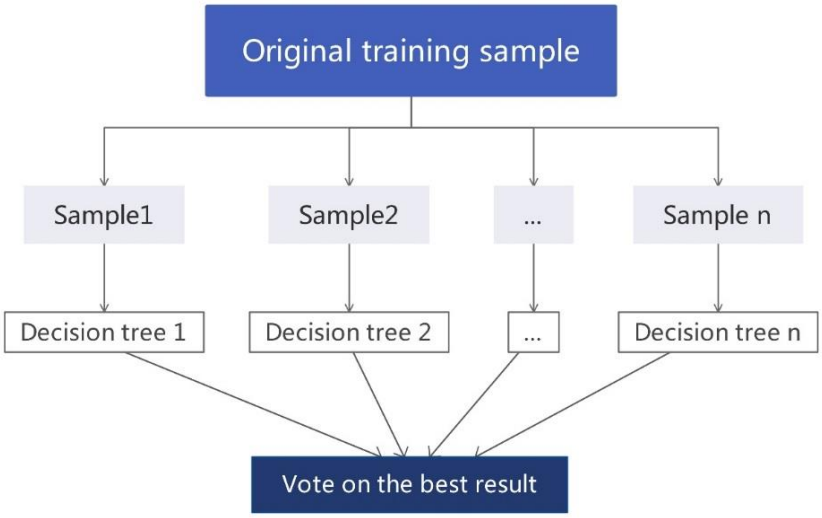


Figure 6 The schematic diagram of the random forest (RF) algorithm.

3.4 Performance and Validation of Two Models

Validation of the model is a key step to test whether the results are scientific and reasonable. Therefore, it is necessary to use different indicators and methods. In particular, the confusion matrix, receiver operating characteristic curve (ROC) and area under ROC curve (AUC) are used to evaluate the accuracy commonly, as study on landslide susceptibility is prone to be a typical binary classification problem (0,1) (Kalantar et al. 2017). Secondly, based on the confusion matrix, statistical measures such as accuracy, precision, recall, and F-measure were also used to evaluate the predictive capabilities of the models, and these measures were calculated by the following formulas:

$$AUC = \frac{(\sum TP + \sum TN)}{(P + N)} \quad (4)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (5)$$

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F\text{-}measure = \frac{2 \times TP}{2 \times TP+FP+FN} \quad (10)$$

Where, P is the sum of landslides, N is the sum of non-landslides, TP (True Positive) and TN (True Negative) are the numbers of correctly classified cells, and FP (False Positive) and FN (False Negative) are the numbers of incorrectly classified cells. AUC between 1.0 and 0.5, in the case of AUC > 0.5, the AUC is closer to 1, indicating that the model prediction effect is better. For accuracy, precision, recall, and F-measure metrics between 0 and 1, 0 and 1 represent ideal and invalid models, respectively, and a higher value represents a better model.

4 Results and Analyses

4.1 LSM Acquired by Artificial Neural Network (ANN) Model

In order to reduce the variability and avoid over-fitting, the 10-fold cross-validation method was used to select the training and testing dataset. Through multiple experiments using a lot of datasets and different techniques, it is shown that 10-fold cross-validation is a suitable choice for obtaining the best precision sample. In the 10-fold cross-validation method, all datasets (866 positive cells and 866 negative cells) were divided into ten individual subsets randomly and averagely. When one subset was tested, other subsets were used for model training over 10 rounds. The results of the 10-fold cross-validation of the ANN model are shown in Table 4. Averagely, the accuracy of the test dataset was 0.888, and the highest accuracy (0.948) is Subset 3. Therefore, Subset 3 was used to construct the ANN model. After it was constructed, we applied the established model to the entire study area to obtain landslide susceptibility value. As a result, in order to facilitate the comparison of the evaluation results of the different models, the LSM is divided into five classes (highest 10 %, second 10 %, third 10 %, fourth 20 % and remaining 50 %) by two method of specified area ratio (Pradhan and Lee 2010), corresponding to very high, high, moderate, low and very low susceptibility regions, respectively. The LSM of Wushan County was generated by using the ANN model in Fig. 7a. It shows that low class of landslide susceptibility in most regions of Wushan County is mainly concentrated in the east and southwest. The high susceptibility regions were mainly concentrated in Yangtze River and its tributaries, which matched well with the distribution of actual historical landslides.

Table 4 The accuracy of 10-fold cross-validation.

Subset (ANN)	Accuracy		Subset (RF)	Accuracy	
	Training dataset	Test dataset		Training dataset	Test dataset
1	0.967	0.884	1	1.000	1.000
2	0.949	0.920	2	1.000	0.995
3	0.945	0.948	3	1.000	0.994
4	0.972	0.855	4	1.000	1.000
5	0.951	0.914	5	1.000	0.990
6	0.970	0.839	6	1.000	0.995
7	0.990	0.902	7	1.000	0.990
8	0.941	0.827	8	1.000	1.000
9	0.972	0.919	9	1.000	1.000
10	0.942	0.873	10	1.000	1.000

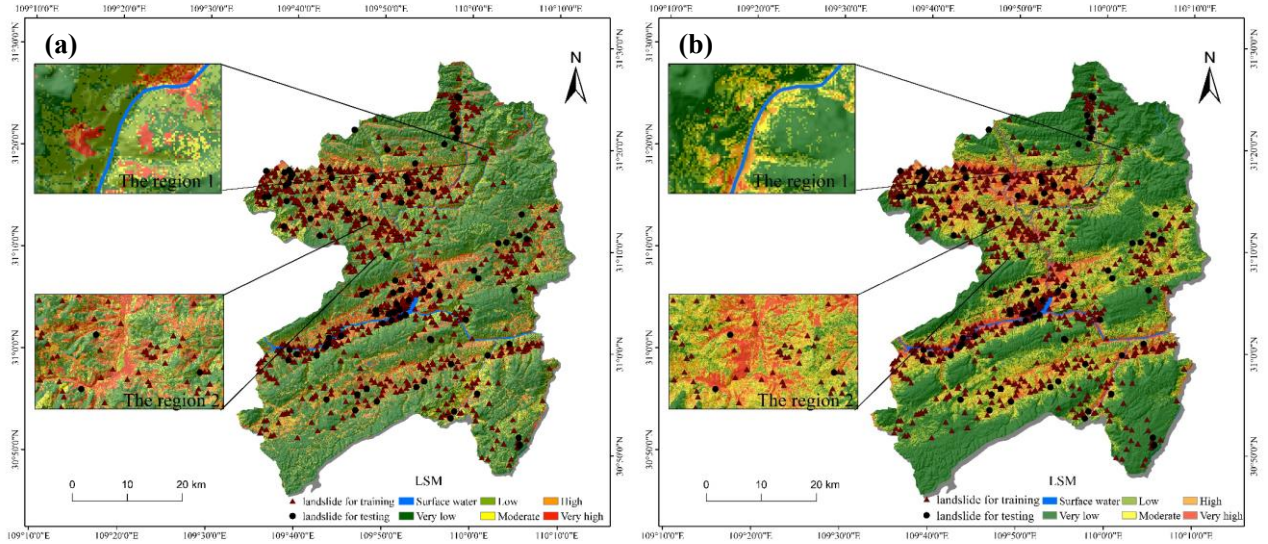


Figure 7 Landslide susceptibility mappings: (a) ANN model; (b) RF model.

Moreover, in order to quantitatively evaluate the effectiveness of the two models, the percentage of each susceptibility classification, the number of landslides, the proportion and density of each class's region were statistically analyzed (Table 5). The results of the ANN model indicated 47 % of the landslides was located in 20 % of the high and very high susceptibility regions, but only 26 % of the landslides was located in 50 % of very low susceptibility region. The evaluation shows that the landslide density increased by approximately 5 times (from 0.145 to 0.700) when the susceptibility class varied from very low to very high. This which indicates that landslides are concentrated in high and very high susceptibility regions.

Table 5 Statistics of the susceptibility classes.

	Class:	very low:	Low:	Moderate:	High:	very high:
	Coverage (%)	50 %	20 %	10 %	10 %	10 %
ANN	Landslide (number)	223	124	113	173	233
	Landslide (%)	26 %	14 %	13 %	20 %	27 %
	Landslide density (Pcs/km^2)	0.145	0.234	0.461	0.605	0.700
RF	Landslide (number)	51	98	86	175	456
	Landslide (%)	6 %	11 %	10 %	20 %	53 %
	Landslide density (Pcs/km^2)	0.035	0.167	0.293	0.596	1.552

4.2 LSM Acquired by Random Forest (RF) Model

The 10-fold cross-validation of the RF model is also shown in Table 4. Averagely, accuracy of the test dataset of the RF model was 0.996, and the highest accuracy is 1.000 in Subset 1. Therefore, we used Subset 1 dataset to construct the RF model. Fig. 7b shows the LSM produced by the RF model for Wushan County, which was covered with landslides. The LSM fits well with the distribution of historical landslides.

Furthermore, all statistical data acquired in statistical analysis are listed in Table 5. As shown in Table 5, most areas of Wushan County were located in low susceptibility class regions; more than 73 % of landslides points were located in 20 % of high and very high susceptibility regions; only 6 % of landslides was located in 50 % of low and very low susceptibility regions. The landslide density from 0.035 to 1.552 increased by about 44 times when the susceptibility class varied from very low to very high.

4.3 Validation and Comparison

To further compare the performance of the two models, the results of confusion matrix, the ROC plots and AUC values of the two models are shown in Table 6 and Fig.8, respectively. Furthermore, we used the precision, AUC, accuracy, recall rate and F-measure to evaluate the models. The AUC of the training data represents the success rate, and the AUC of the test data represents the predictive ability of the model. (Tsangaratos et al. 2016). Figs. 8a and 8b show that The AUC values of ANN and RF model training data were 0.990 and 1.00, respectively, and the AUC values of test data were 0.966 and 1.00, respectively. Besides, in Table 6, we can calculate precision, overall accuracy, recall rate, and F-measure of the ANN and RF models simply. For all the datasets, the precision, overall accuracy, recall rate, and F-measure of ANN model were 0.950, 0.953, 0.957, and 0.953, respectively, while the counterparts of RF model were all 1.00. All indicators show that although the two models show reasonable goodness of fit, RF model shows better performance both in training and test data sets. In this case, RF model has better prediction results than ANN model.

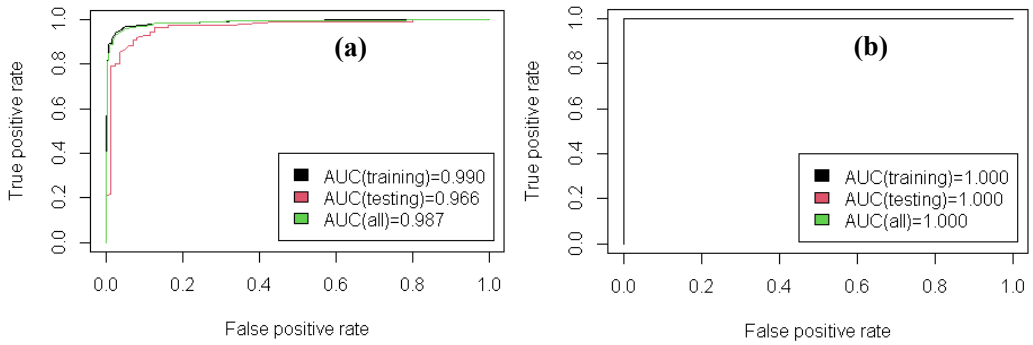


Figure 8 ROC curve of the ANN and RF models: (a) ANN model; (b) RF model.

We also compared the two models from the results of LSMs. On the one hand, for the generated LSMs (Fig. 7), in order to display the difference between the two LSMs more intuitively, two regions are selected for comparison. As shown in the region 1, compared with the LSM generated by the ANN model, the accuracy of landslide falling into high and very high

susceptibility region is higher in LSM of RF model. For the region 2, although the LSMs generated by both models almost include most of the landslides, the high-very high susceptibility regions of LSM generated by RF occupy less area. On the other hand, in LSMs, 26 % of landslides are located in very low susceptibility region by ANN, while only 6 % are located in the same very low susceptibility region by RF; 27 % of landslides are located in very high susceptibility region by ANN, while 53 % are located in the same region by RF. In general, in the same area, landslides are more distributed in the high and very high susceptibility region in RF, while in the LSM generated by ANN, the distribution law of landslides is not obvious. This shows that the RF model is better than the ANN model in the landslide susceptibility of the whole area.

Table 6 Confusion matrix of the ANN and RF models.

Model			Actual value	
			Non-landslide (0)	Landslide (1)
ANN	Predicted value	Non-landslide (0)	822	37
		Landslide (1)	44	829
RF			Non-landslide (0)	Landslide (1)
	Predicted value	Non-landslide (0)	866	0
		Landslide (1)	0	866

5 Discussion

5.1 Mapping Units

Selection of a mapping unit is a basic step of any landslide susceptibility model that affects significantly the susceptibility terrain zonation. Grid cell is very popular among landslide susceptibility researchers, because it is easy to process, at all resolutions and geographical scales. However, grid cell has clear drawbacks for susceptibility modelling. Firstly, there is no physical relationship between a grid cell or a group of grid cells, and landslides. Landslides are the result of slope processes that act on different spatial and temporal scales. These slope processes result in geomorphological forms of very different shapes and sizes, which is difficult to capture by grid cells accurately. Secondly, Reichenbach (Reichenbach et al. 2018) demonstrated that the majority of the models for LSM that adopt grid cell as the mapping unit use the cell of the same resolution of the DEM. Using of the same resolution as DEM for the landslide has clear practical advantages, but it may introduce biases or lead to potentially misleading results. Another unit type that is used commonly is the slope unit, which is hydro-logical terrain unit bounded by drainage and divide lines. The size of the slope unit can be customized according to the type and size of the landslide, allowing the use of a geo environmental information that is best suited for the specific type of landslide. Despite its conceptual and operational advantage over grid cells, only a few authors have used slope unit for landslide susceptibility. The reason is that slope unit is difficult to obtain manually, particularly for large areas. In future, we will continue to try and compare the grid and slope units to improve our scientific research.

5.2 Importance of Conditioning Factors

Although the occurrence of landslide is caused by multiple factors, the influence of each factor is not the same. Therefore, quantitative analysis of the importance of each factor helps to identify the dominant factors, so as to carry out targeted landslide research and prevention.

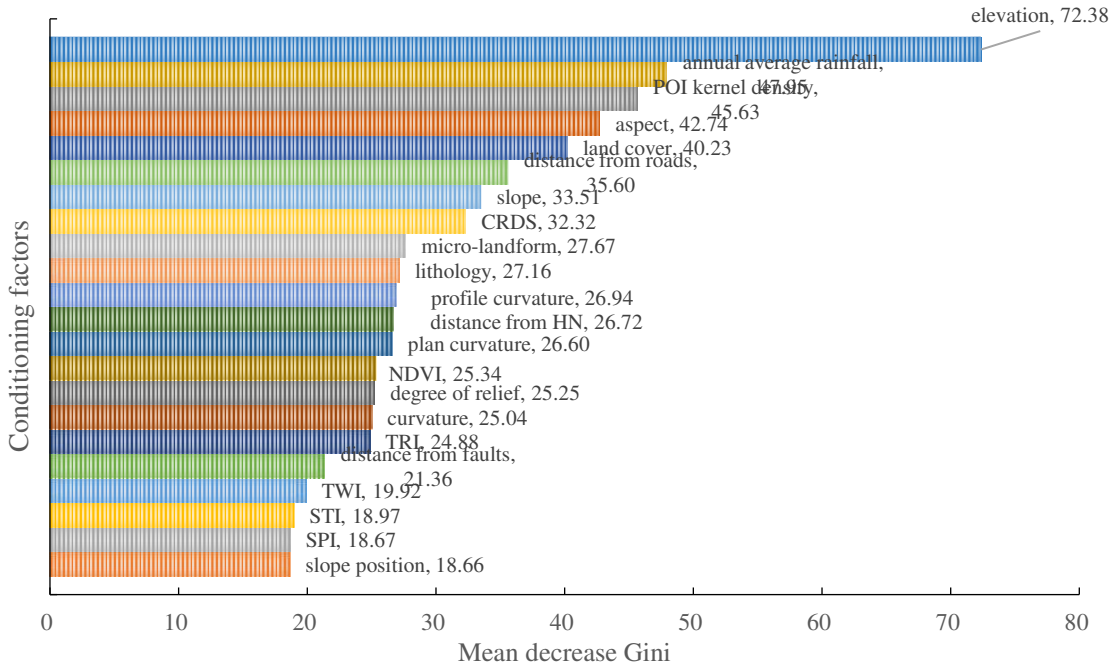
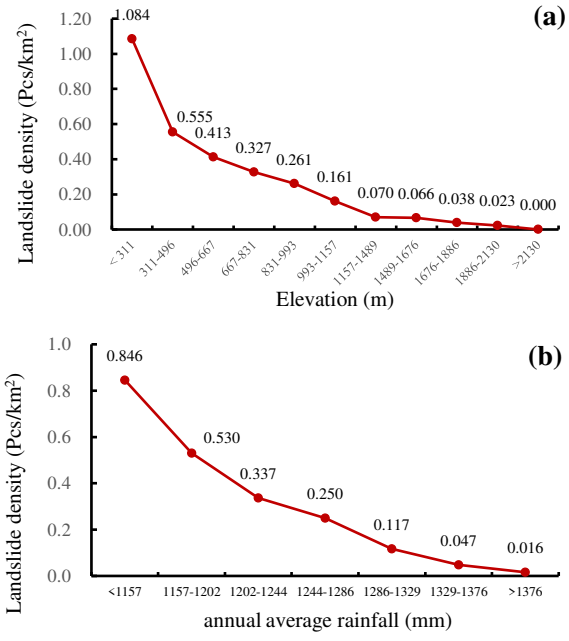


Figure 9 Constructing importance of factors based on mean decrease Gini.

The mean decrease Gini of RF model can rank the conditioning factors and has high reference value (Tsangaratos et al. 2016). Fig. 9 shows 22 conditioning factors ordered by the mean decrease Gini, and the first three important conditioning factors were elevation, followed by annual average rainfall and POI kernel density, with a mean decrease accuracy of 72.38, 47.95, and 45.63, respectively. Therefore, we generated a landslide density map of these three factors (Fig. 10).



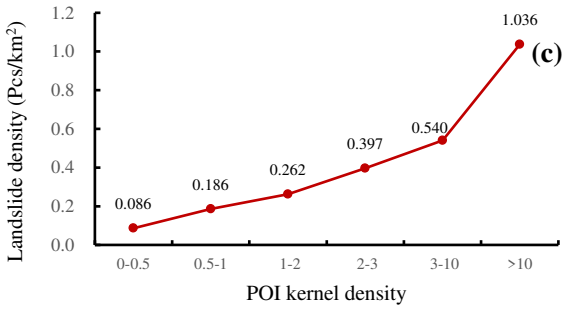


Figure 10 Landslide density charts: (a) elevation; (b) annual average rainfall; (c) POI kernel density.

Firstly, interpreting the spatial variability of different geomorphological phenomena (such as the vegetation distribution), the elevation has a critical effect on terrain attributes (Fig.10a). Landslide density is negatively correlated with elevation. As a typical mountainous area with complex terrain, low-elevation areas in Wushan County are often with looser soil bed and more human engineering activities. Therefore, the most intensive elevation range of landslide distribution was the low-elevation range with frequent human activities and low vegetation coverage.

Secondly, Fig. 10b shows that the landslide density was negatively correlated with the annual average rainfall. Annual average rainfall is long-term observation data, including rainfall in the early years and rainfall on the day of landslide. Especially in the rainy season, the erosion of rainfall on the slope makes unstable rock and soil to be taken away by surface runoff formed by rainfall. Annual average rainfall not only erodes the slope, but also affects the development of vegetation on the slope, thereby affecting the occurrence of landslides.

Previous studies rarely consider the impact of human activities on landslides, but the impact of such activities on landslides cannot be ignored; all excavation and backfill sections have varying degrees of deformation and damage (Bourenane et al. 2016). As the value of POI kernel density increases, the frequency of landslides increases (Fig.10c), indicating frequent human activities affect the occurrence of landslides to a certain extent. This factor has rarely been taken into account by previous studies. POI is a new factor that represents the impact of human activities on landslides.

5.3 Comparison of the Two Models

The confusion matrix and ROC curve are shown in Table 6 and Fig. 8, respectively. The AUC values of the ANN and RF models' test dataset were 0.966 and 1.000, respectively, and other indexes of ANN model were lower than the counterparts of RF model. This shows the interpretation and generalization ability of the RF model is higher than ANN model. As for the data that builds the model, only dozens or hundreds of sample data may not prove the full power of machine learning, especially the RF model. For example, Hong (Hong et al. 2016) only used 163 positive cells to build the model, which may lead to inaccurate conclusions. Sun (Sun et al. 2020) used a large number of dataset (1,520 positive cells and 15,200 negative cells) to build the models and compared the LR and RF models in Fengjie County (China), whose terrain and geological conditions are almost the same as our study area. The result is that the performance of RF is better than LR model.

Comprehension of the strength and weakness of each method is made possible by provided for selecting the most accurate model in any case. On the one hand, the main difference is that RF is the set of decision trees, and the decision trees in the set are independent. Each decision tree can independently predict landslide susceptibility. ANN is a network model composed of neurons that are connected. These neurons are grouped by layers and process the data of each layer, and then are passed to the next layer. The prediction results are only determined by the last layer of neurons. Therefore, RF may have better prediction performance than ANN. On the other hand, RF has less pretreatment and simpler training process, which makes forecasting easier. Therefore, RF has better application in this study. In addition, RF not only has achieved good results in the LSM, but also has been widely used in researches works of medicine, city, agriculture etc.

6 Conclusion

LSM shows the spatial location of landslide in an intuitive way, which is an important tool for preventing and evaluating landslides. In general, LSMs of ANN and RF models were generated based on landslides in Wushan County. The conclusions are as follows:

(1) When the two models were compared, the AUC values of the test data of ANN and RF models were 0.966 and 1.00, respectively. When the LSM is made by the ANN model, 47 % landslides are located in 20 % of high and very high susceptibility regions, but only 26 % landslides are located in 50 % of very low susceptibility regions. On the other hand, when the LSM is made by the RF model, more than 73 % of landslides are located in 20 % of high and very high susceptibility regions, but 6% landslides in 50 % of low and very low susceptibility regions. Therefore, we conclude that both models show excellent performances, but we have to admit the RF model has better stability and robustness in this study.

(2) In this study region, based on the importance ranking of factors obtained by the mean decrease Gini, the results quantify and highlight the dominant factors, in line with the basic laws of geology. The conditioning factors of elevation, annual average rainfall play an important role in landslide occurrence. It is worth mentioning that although POI kernel density is a new landslide influencing factor, its importance ranking ranks third, confirming its existence value.

(3) Comparison of the strength and weakness of all the methods indicates that the ANN and RF models are completely comparable in terms of predictive capabilities. Whether modeling set or prediction set, RF model has better prediction accuracy and stronger generalization ability. In general, the ANN and RF models can be applied to other areas with similar topography, as well as small and medium-sized soil landslides.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research reported in this manuscript is funded by National Key R&D Program of China (Grants No. 2018YFC1505504) and Humanities and Social Sciences Research Project of the Ministry of Education (Grants No. 16YJCZH061).

References

Abedini M, B Ghasemian, A Shirzadi, H Shahabi, K Chapi, B T Pham, B Bin Ahmad, D Tien Bui (2018) A novel hybrid approach of Bayesian Logistic Regression and its ensembles for landslide susceptibility assessment. Geocarto International 34(13): 1427-1457. 10.1080/10106049.2018.1499820

Althuwaynee O F, B Pradhan, H-J Park, J H Lee (2014) A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping. Catena 114: 21-36. https://doi.org/10.1016/j.catena.2013.10.011

Andang Suryana S, K Tetsuya, M Hideaki (2019) Optimization of causative factors using logistic regression and artificial neural network models for landslide susceptibility assessment in Ujung Loe Watershed, South Sulawesi Indonesia. Journal of Mountain Science 16: 383-401. https://doi.org/10.1007/s11629-018-4884-7

Arora M K, A S Das Gupta, R P Gupta (2003) An artificial neural network approach for landslide hazard zonation in the Bhagirathi (Ganga) Valley, Himalayas. International Journal of Remote Sensing 25(3): 559-572. http://dx.doi.org/10.1080/0143116031000156819

Bakillah M, S Liang, A Mobasheri, J J Arsanjani, A Zipf (2014) Fine-resolution population mapping using OpenStreetMap points-of-interest. International Journal of Geographical Information Science 28: 1940-1963. http://dx.doi.org/10.1080/13658816.2014.909045

Basu T, S Pal (2017) Identification of landslide susceptibility zones in Gish River basin, West Bengal, India. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 12(1): 14-28. http://dx.doi.org/10.1080/17499518.2017.1343482

Bourenane H, M S Guettouche, Y Bouhadad, M Braham (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. Arabian Journal of Geosciences 9(2). https://doi.org/10.1007/s12517-015-2222-8

Chakraborty A, D Goswami (2017) Prediction of slope stability using multiple linear regression (MLR) and artificial neural network (ANN). Arabian Journal of Geosciences 10(17). https://doi.org/10.1007/s12517-017-3167-x

Chen W, X Xie, J Wang, B Pradhan, H Hong, D T Bui, Z Duan, J Ma (2017) A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena 151: 147-160. http://dx.doi.org/10.1016/j.catena.2016.11.032

Ding Q, W Chen, H Hong (2016) Application of frequency ratio, weights of evidence and evidential belief function models in landslide susceptibility mapping. Geocarto International: 1-21. https://dx.doi.org/10.1080/10106049.2016.1165294

Harmouzi H, H A Nefeslioglu, M Rouai, E A Sezer, A Dekayir, C Gokceoglu (2019) Landslide susceptibility mapping of the Mediterranean coastal zone of Morocco between Oued Laou and El Jebha using artificial neural networks (ANN). Arabian Journal of Geosciences 12(22). 10.1007/s12517-019-4892-0

Harp E L, D K Keefer, H P Sato, H Yagi (2011) Landslide inventories: The essential part of seismic landslide hazard analyses. Engineering Geology 122(1-2): 9-21. https://dx.doi.org/10.1016/j.enggeo.2010.06.013

Hong H, S A Naghibi, H R Pourghasemi, B Pradhan (2016) GIS-based landslide spatial modeling in Ganzhou City, China. Arabian Journal of Geosciences 9(2). https://dx.doi.org/10.1007/s12517-015-2094-y

Huang F, K Yin, J Huang, L Gui, P Wang (2017) Landslide susceptibility mapping based on self-organizing-map network and extreme learning machine. Engineering Geology 223: 11-22. https://doi.org/10.1016/j.enggeo.2017.04.013

Huang R, W Li (2011) Formation, distribution and risk control of landslides in China. Journal of Rock Mechanics and Geotechnical Engineering 3(2): 97-116.

10.3724/sp.J.1235.2011.00097
Huang Y, L Zhao (2018) Review on landslide susceptibility mapping using support vector machines. Catena 165: 520-529. https://doi.org/10.1016/j.catena.2018.03.003
Kalantar B, B Pradhan, S A Naghibi, A Motevalli, S Mansor (2017) Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomatics, Natural Hazards and Risk 9(1): 49-69. https://dx.doi.org/10.1080/19475705.2017.1407368
Kanungo D P, M K Arora, R P Gupta, S Sarkar (2008) Landslide risk assessment using concepts of danger pixels and fuzzy set theory in Darjeeling Himalayas. Landslides 5(4): 407-416. https://dx.doi.org/10.1007/s10346-008-0134-3
Li C, Z Fu, Y Wang, H Tang, J Yan, W Gong, W Yao, R E Criss (2019) Susceptibility of reservoir-induced landslides and strategies for increasing the slope stability in the Three Gorges Reservoir Area: Zigui Basin as an example. Engineering Geology 261. https://doi.org/10.1016/j.enggeo.2019.105279
Li L, H Lan (2015) Probabilistic modeling of rockfall trajectories: a review. Bulletin of Engineering Geology and the Environment 74(4): 1163-1176. 10.1007/s10064-015-0718-9
Li S H, X H Luo, L Z Wu (2021) A new method for calculating failure probability of landslide based on ANN and a convex set model. Landslides 18(8): 2855-2867. 10.1007/s10346-021-01652-2
Liu R, S Shi, D Sun, J Xu (2020) Based on GIS and random forest model for landslide susceptibility mapping in Wushan county. Journal of Chongqing Normal University(Natural Science) 37(3): 86-96. http://dx.chinadoi.cn/10.11721/cqnuj20200306
Liu X, R Jing, L Miao, Y Han, Z Ddeng, C Xiong (2020) Reconstruction models and typical case analysis of the fluctuation belt of reservoir bank slopes in Wushan. Chinese Journal of Rock Mechanics and Engineering 39(7): 1321-1332. http://dx.chinadoi.cn/10.13722/j.cnki.jrme.2019.0887
Lucchese L V, G G de Oliveira, O C Pedrollo (2021) Investigation of the influence of nonoccurrence sampling on landslide susceptibility assessment using Artificial Neural Networks. Catena 198. https://doi.org/10.1016/j.catena.2020.105067
Melchiorre C, M Matteucci, A Azzoni, A Zanchi (2008) Artificial neural networks and cluster analysis in landslide susceptibility zonation. Geomorphology 94(3-4): 379-400. http://dx.doi.org/10.1016/j.geomorph.2006.10.035
Mondal S, S Mandal (2017) RS & GIS-based landslide susceptibility mapping of the Balason River basin, Darjeeling Himalaya, using logistic regression (LR) model. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 12(1): 29-44. https://doi.org/10.1080/17499518.2017.1347949
Pavel M, J D Nelson, R Jonathan Fannin (2011) An analysis of landslide susceptibility zonation using a subjective geomorphic mapping and existing landslides. Computers & Geosciences 37(4): 554-566. https://dx.doi.org/10.1016/j.cageo.2010.10.006
Petley D (2012) Global patterns of loss of life from landslides. Geology 40: 927-930. https://doi.org/10.1130/G33217.1
Pourghasemi H R, M Mohammady, B Pradhan (2012) Landslide susceptibility mapping using index of entropy and conditional probability models in GIS: Safarood Basin, Iran. Catena 97(none): 71-84. https://doi.org/10.1016/j.catena.2012.05.005
Pradhan B, S Lee (2010) Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environmental Modelling & Software 25(6): 747-759. https://dx.doi.org/10.1016/j.envsoft.2009.10.016
Reichenbach P, M Rossi, B D Malamud, M Mihir, F Guzzetti (2018) A review of statistically-based landslide susceptibility models. Earth-Science Reviews 180: 60-91. https://doi.org/10.1016/j.earscirev.2018.03.001
Sahin E K, I Colkesen, T Kavzoglu (2018) A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping. Geocarto International 35(4): 341-363. https://doi.org/10.1080/10106049.2018.1516248
Silalahi F E S, Pamela, Y Arifianti, F Hidayat (2019) Landslide susceptibility assessment using frequency ratio model in Bogor, West Java, Indonesia. Geoscience Letters 6(1). https://doi.org/10.1186/s40562-019-0140-4
Steger S, A Brenning, R Bell, T Glade (2017) The influence of systematically incomplete shallow landslide inventories on statistical susceptibility models and suggestions for improvements. Landslides 14: 1767–1781. https://doi.org/10.1007/s10346-017-0820-0
Sun D, S Shi, H Wen, J Xu, X Zhou, J Wu (2021) A hybrid optimization method of factor screening predicated on GeoDetector and Random Forest for Landslide Susceptibility Mapping. Geomorphology 379. https://doi.org/10.1016/j.geomorph.2021.107623
Sun D, H Wen, D Wang, J Xu (2020) A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. Geomorphology 362. https://doi.org/10.1016/j.geomorph.2020.107201
Sun D, H Wen, Y Zhang, M Xue (2020) An optimal sample selection-based logistic regression model of slope physical resistance against rainfall-induced landslide. Natural Hazards. https://doi.org/10.1007/s11069-020-04353-6
Taalab K, T Cheng, Y Zhang (2018) Mapping landslide susceptibility and types using Random Forest. Big Earth Data 2(2): 159-178. https://doi.org/10.1080/20964471.2018.1472392
Tian Y, C Xu, H Hong, Q Zhou, D Wang (2018) Mapping earthquake-triggered landslide susceptibility by use of artificial neural network (ANN) models: an example of the 2013 Minxian (China) Mw 5.9 event. Geomatics, Natural Hazards and Risk 10(1): 1-25. 10.1080/19475705.2018.1487471
Tsangaratos P, I Ilia, H Hong, W Chen, C Xu (2016) Applying Information Theory and GIS-based

quantitative methods to produce landslide susceptibility maps in Nancheng County, China. Landslides 14(3): 1091-1111. https://doi.org/10.1007/s10346-016-0769-4
Wang Y, D L Sun, H J Wen, H Zhang, F T Zhang (2020) Comparison of Random Forest Model and Frequency Ratio Model for Landslide Susceptibility Mapping (LSM) in Yunyang County (Chongqing, China). International Journal of Environmental Research and Public Health 17: 4206. https://doi.org/10.3390/ijerph17124206
Wang Y, H Wen, D Sun, Y Li (2021) Quantitative Assessment of Landslide Risk Based on Susceptibility Mapping Using Random Forest and GeoDetector. Remote Sensing 13(13). 10.3390/rs13132625
Wang Y, X Wu, Z Chen, F Ren, L Feng, Q Du (2019) Optimizing the Predictive Ability of Machine Learning Methods for Landslide Susceptibility Mapping Using SMOTE for Lishui City in Zhejiang Province, China. International Journal of Environmental Research and Public Health 16(3): 368. https://doi.org/10.3390/ijerph16030368
Wen H (2015) A susceptibility mapping model of earthquake-triggered slope geohazards based on geo-spatial data in mountainous regions. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 9(1): 25-36. http://dx.doi.org/10.1080/17499518.2015.1005634
Xia H, K Yin, X Liang, F Ma (2018) Landslide susceptibility assessment based on SVM-ANN Models: a case stualy for Wushan County in the Three Gorges Reservoir. The Chinese Journal of Geological Hazard and Control 29(5): 13-19. http://dx.chinadoi.cn/10.16031/j.cnki.issn.1003-8035.2018.05.03
Xie P, H Wen, C Ma, L G Baise, J Zhang (2018) Application and comparison of logistic regression model and neural network model in earthquake-induced landslides susceptibility mapping at mountainous region, China. Geomatics, Natural Hazards and Risk 9(1): 501-523. https://doi.org/10.1080/19475705.2018.1451399
Zhou J, M Deng, Z Li, F Zhang, Y Lin (2020) Analysis on the Formation and Apparent-dip Lateral Sliding Mechanism of Shanshucao Landslide in the Three Gorges Reservoir Area. Journal of Disaster Prevention and Mitigation Engineering 40(6): 860-866+883. http://dx.chinadoi.cn/10.13409/j.cnki.jdpme.2020.06.003