

An Optimized Random Forest Model and Its Generalization Ability in Landslide Susceptibility Mapping: Application in Two Areas of Three Gorges Reservoir, China



Deliang Sun¹, Jiahui Xu¹, Haijia Wen^{2,3,4}, Yue Wang¹

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

2. Key Laboratory of New Technology for Construction of Cities in Mountain Area, Ministry of Education, Chongqing 400045, China

3. National Joint Engineering Research Centre of Geohazards Prevention in the Reservoir Areas, Chongqing 400044, China

4. School of Civil Engineering, Chongqing University, Chongqing 400045, China

 Deliang Sun: <https://orcid.org/0000-0003-3153-2811>;  Haijia Wen: <https://orcid.org/0000-0002-2045-729X>

ABSTRACT: Numerous researches have been published on the application of landslide susceptibility assessment models; however, they were only applied in the same areas as the models were originated, the effect of applying the models to other areas than the origin of the models has not been explored. This study is purposed to develop an optimized random forest (RF) model with best ratios of positive-to-negative cells and 10-fold cross-validation for landslide susceptibility mapping (LSM), and then explore its generalization ability not only in the area where the model is originated but also in area other than the origin of the model. Two typical counties (Fengjie County and Wushan County) in the Three Gorges Reservoir area, China, which have the same terrain and geological conditions, were selected as an example. To begin with, landslide inventory was prepared based on field investigations, satellite images, and historical records, and 1 522 landslides were then identified in Fengjie County. 22 landslide-conditioning factors under the influence of topography, geology, environmental conditions, and human activities were prepared. Then, combined with 10-fold cross-validation, three typical ratios of positive-to-negative cells, i.e., 1 : 1, 1 : 5, and 1 : 10, were adopted for comparative analyses. An optimized RF model (Fengjie-based model) with the best ratios of positive-to-negative cells and 10-fold cross-validation was constructed. Finally, the Fengjie-based model was applied to Fengjie County and Wushan County, and the confusion matrix and area under the receiver operating characteristic (ROC) curve value (AUC) were used to estimate the accuracy. The Fengjie-based model delivered high stability and predictive capability in Fengjie County, indicating a great generalization ability of the model to the area where the model is originated. The LSM in Wushan County generated by the Fengjie-based model had a reasonable reference value, indicating the Fengjie-based model had a great generalization ability in area other than the origin of the model. The Fengjie-based model in this study could be applied in other similar areas/countries with the same terrain and geological conditions, and a LSM may be generated without collecting landslide information for modeling, so as to reduce workload and improve efficiency in practice.

KEY WORDS: landslide susceptibility mapping, generalization ability, random forest, Three Gorges Reservoir area, 10-fold cross-validation.

0 INTRODUCTION

Landslides are destructive natural disasters affecting human life, property, and economic development (Wang et al., 2020). According to the Emergency Events Database (EM-DAT), between 2014 and 2018, landslides resulted in 4 914 deaths, with 27 110 people losing their homes and 2.1 billion dollars of assets

destroyed. Therefore, efficient solutions must be developed to reduce and mitigate landslide-related destruction. As a method generated based on local geological environmental factors by assessing the spatial distribution of landslide probabilities in a given area, Landslide Susceptibility Mapping (LSM) is regarded as one of the common counter measures for mitigating the effects of landslides (Silalahi et al., 2019; Hong et al., 2016a).

With the rapid development of data mining technology, many researchers have begun to utilize machine learning methods for LSM (Zhu et al., 2018), and the most popular data mining methods include artificial neural network (Tian et al., 2018; Can et al., 2017), Bayes' net (Chen et al., 2018a, b), support vector machine (Dou et al., 2019a; Pham et al., 2018), decision tree

*Corresponding author: jhw@cqu.edu.cn

© China University of Geosciences (Wuhan) and Springer-Verlag GmbH Germany, Part of Springer Nature 2020

Manuscript received May 20, 2020.

Manuscript accepted August 27, 2020.

(Dou et al., 2019b; Pandey et al., 2018), and random forest (Sun et al., 2020; Chen et al., 2018c). Among these methods, random forest (RF) is a non-linear algorithm, which could deal with large datasets and account for complex interactions, and non-linearity between variables, and has been used widely in recent years (Taalab et al., 2018). A few studies have been made on the application of RF. For example, Sun et al. applied RF model in Fengjie County (China) and achieved the area under the receiver operating characteristic (ROC) curve value (AUC) in test dataset of 0.87, indicating that RF model has great prediction abilities in mountain areas (Sun et al., 2020).

However, although the RF model has been successfully applied in researches worldwide, it is only applied in the same area as its model is located, the effect of applying the model to other areas has not been explored. The generalization ability is a measure of how accurately a model can predict outcome values for previously unseen data, which is the most immediate representation of the model's prediction ability (Jin et al., 2019; Liu et al., 2019). When applied to landslide research, it could be described by the following 3 methods. The first and most intuitive method is to check the accuracy of the model's test dataset. The higher the test dataset's accuracy, the better the model's generalization ability. The second method is to apply the model to the entire study area where the model originated, and to simulate the probability of landslides in such study areas that are not involved in the model training, so as to generate the LSM of the entire study area where the model is originated. If the generated LSM is consistent with the distribution pattern of the actual landslides, the model will be regarded as having good generalization ability; otherwise, the model will have poor generalization ability. Last but not least, what was always overlooked in previous studies is that the generalization ability of a model shall not be limited to the research area that is used to construct the model, but shall be tested in other research areas than the origin of the model. In this way, a LSM may be generated for an area without collecting landslide information for modeling. If the LSM generated in this way is consistent with the LSM generated by the model with this area's own landslides data, it indicates that the model has good generalization ability in the study area. Further, if the model has good generalization ability both for the area where the model is originated and for other similar areas, it indicates that the terrain and geological conditions and the landslide-conditioning factors are identical, and the model could continue to be applied to other similar areas. Lee demonstrates that mountainous areas and areas with more impacting events (e.g., heavy rainfall and earthquakes) are more likely to experience landslides (Lee, 2019). Accordingly, it is desirable to construct a model that can offer both outstanding predictive and generalization abilities for the same terrain and geological conditions, so that efficient performance and reduced workload can be expected in practice.

For any LSM model (such as a RF model), its generalization ability is affected by many factors. Many pre-processing results before constructing a model will have certain impacts on its accuracy, e.g., the ratios of positive (landslide) to negative (non-landslide) cells, the division of the training dataset and the test dataset, and so on (Reichenbach et al., 2018). As for the ratios of positive-to-negative cells, an adequate number of

sample data is extremely important in machine learning to ensure the success and prediction capability of the model. Because the information on positive cells is limited, the way to get an adequate number of sample data is to increase the number of negative cells. The use of additional negative cells could expand the data size for machine learning; however, it will cause a sample imbalance problem, thereby biasing the classifier towards the negative cells and negatively affecting its performance (Hussin et al., 2016; Heckmann et al., 2014). Previous researchers have found no ideal fixed ratio existing between landslide and non-landslide cells (Hussin et al., 2016). As for the division of the training dataset and the test dataset, a ratio of 7 : 3 was used in the previous studies, that is, randomly selecting 70% of the data as the training dataset and remaining 30% as the test dataset (Wang et al., 2020; Dou et al., 2019a). This division method often has certain variability and is not suitable for small size of samples. As a common method to reuse samples and test the algorithm accuracy, *K*-fold cross-validation could reduce the variability and is applicable well in the case of small samples (Jiang and Chen, 2016). Multiple experiments with a large number of datasets and different learning techniques have demonstrated that compared with other cross-validation methods, the 10-fold cross-validation is a suitable choice for obtaining the best error estimate (Han and Kamber, 2006). However, this method has rarely been used in the existing LSM researches for the division between the training and test datasets.

Located in southwestern China, with widespread mountain areas, complex geological environments, and hydro-climatic conditions, the Three Gorges Reservoir area becomes one of the most landslide-prone regions in China. As shown in the previous study (Sun et al., 2020), the application of the RF in one typical county in the Three Gorges Reservoir area—Fengjie County is ideal, so Fengjie County is also selected as the study area in this study. The authors aim to: (1) develop an optimized RF model (Fengjie-based model) based on Fengjie landslides with best ratios of positive-to-negative cells and 10-fold cross-validation; (2) evaluate its generalization ability in Fengjie County where the model is originated; (3) apply this Fengjie-based model in another county, Wushan County, in the Three Gorges Reservoir area, to evaluate generalization ability in the other area than the origin of the model; and (4) explore the reasons for the great generalization ability, and construct a model which cannot only deliver on good accuracy and generalization ability but is also more conducive to wide application.

1 MATERIALS

1.1 Description of the Study Areas

The study areas, Fengjie County and Wushan County, are situated in the east of Chongqing, southwestern China, covering an area of almost 7 045 km² between 109°01'E–109°46'E and 30°29'N–31°28'N (Fig. 1). Both counties are located in the east of the Sichuan Basin, in some mountainous landforms as well as at the junction of the Dabashan arc fold fault zone and the eastern Sichuan fold belt. As for Fengjie County, the elevation of the highest point is 2 125 m a.s.l., the lowest elevation is 87 m a.s.l., and the elevation decreases from the south to the north. As for Wushan County, the elevation of the highest point is 2 688 m

a.s.l., the lowest elevation is 63 m a.s.l., and the elevation decreases from the west to the east. Although the distribution of strata in these two counties is different in each territory, the main types of strata are the same (Fig. 1), that is: Triassic (T), Jurassic (J), Permian (P), Devonian (D), Silurian (S), Quaternary (Q), and Carboniferous (C) (Sun et al., 2020). The Triassic (T) strata are mainly composed of sandstone, limestone, and shale; the Jurassic (J) strata are composed of quartz sandstone, clay rock, and shale; the Permian (P) strata are composed of limestone, shale and a small amount of sandstone; the Devonian (D) strata are composed of quartz sandstone, limestone, and dolomite; the Silurian (S) strata are composed of shale and quartz sandstone; the Quaternary (Q) strata are mainly composed of sediments such as alluvial deposits and slope deposits, including clay and gravel; the Carboniferous (C) strata are composed of limestone, dolomite, and quartz sandstone (Wu et al., 2020; Chen et al., 2018c; Tsangaratos et al., 2016; Fourniadis et al., 2007). The faults are

not distributed in the whole area, but only in some areas of the whole area (Fig. 1).

The two areas have the humid subtropical monsoon climate, with hot and rainy summers and relatively warm and rainless winter, showing four distinct seasons. According to statistics from China Weather Website (<http://www.weather.com.cn/>), the Fengjie area has an annual rainfall of approximately 1 132 mm and the Wushan area has an annual rainfall of approximately 1 049 mm. The monthly average rainfall distribution in Fengjie and Wushan counties is almost the same, mainly occurring from May to September (Fig. 2). The average annual temperature of Fengjie is 8 °C (January) to 28 °C (July and August) with an average annual temperature of approximately 16.5 °C, while the average annual temperature of Wushan is 7 °C (January) to 28 °C (July and August) with an average annual temperature of approximately 18 °C. The monthly average temperature distribution in Fengjie and Wushan counties is also the same (Fig. 2).

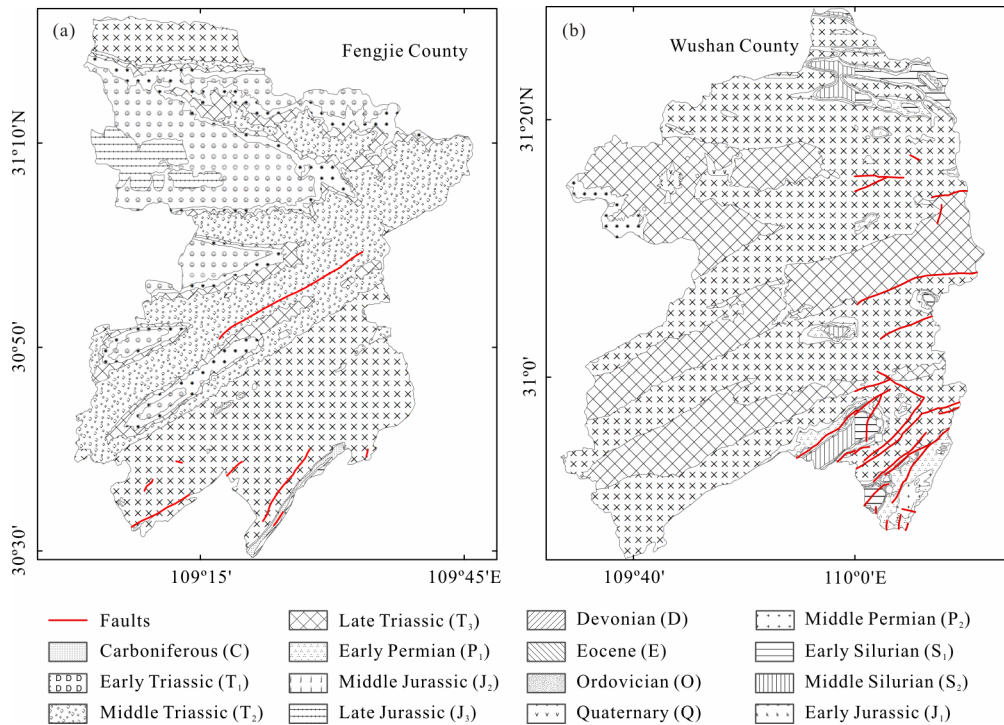


Figure 1. Location and geological map of the study areas.

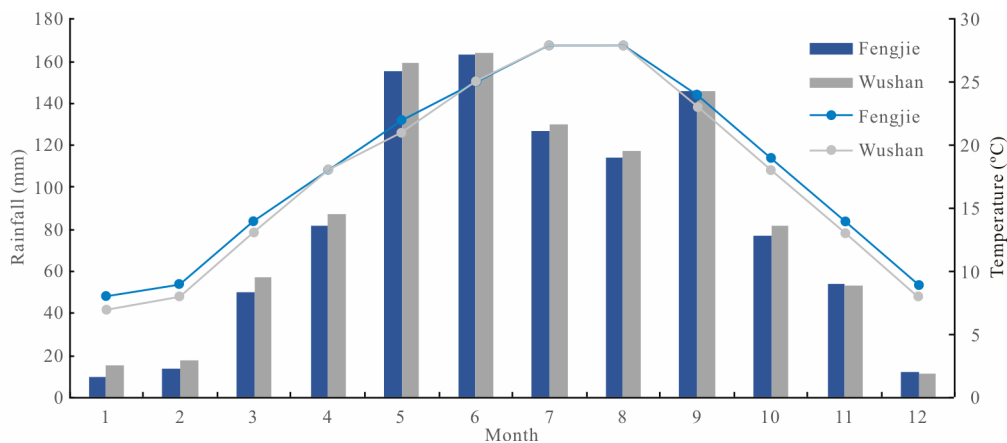


Figure 2. Monthly average rainfall and temperature distributions in the study areas (2009–2018).

1.2 Landslide Inventory

Landslide inventory is used to record the location, the occurrence, and other messages about landslides (if known) in the study area (Tian et al., 2019; Shirzadi et al., 2017; Zêzere et al., 2017). As an essential basis for assessing the landslide susceptibility, based on field investigations, satellite images, and historical records, preparation of landslide inventory identified and mapped a total of 1 522 landslides (2001–2016) in Fengjie County (Fig. 3). Meanwhile, the preparation of the landslide inventory also identified and mapped 952 landslides (2001–2016) in Wushan County to explore and verify the generalization ability of the model (Fig. 4). These landslides are typically small and medium: the smallest ones have an area of 115 m², largest ones of 106 743 m², averaged at 60 326 m² in Fengjie County (Fig. 3); the smallest ones have an area of 100 m², largest ones of 96 800 m², averaged at 56 495 m² in Wushan County (Fig. 4). These landslides were classified by two factors, type and trigger. In terms of type, most of the landslides were shallow/soil ones (82%) and only 18% were deep/bedrock landslides in Fengjie County; most of the landslides were shallow/soil ones (89%) and only 11% were deep/bedrock landslides in Wushan County (Fig. 5a). In terms of trigger, most of its landslides were caused by rainfall (75.88%), 10.05% by groundwater (pore water), 2.01% by human construction activities and 12.06% by others in Fengjie County; most of its landslides were caused by rainfall (82.33%), 6.72% by groundwater, 3.24% by human construction activities and 7.71% by others in Wushan County (Fig. 5b). Distributed together with slope soils, residuals and shale, or mudstones without active faults, most of these landslides are developed in low mountain-landforms with a 20–40 slope degree of poor stability. Jurassic (J), Triassic (T), and Permian (P) strata are the main bedrock of these landslides.

2 METHODS

The assessment procedure included four phases: (a) selection of landslide-conditioning factors; (b) construction and comparison of the RF models based on different ratios of positive-to-negative cells; (c) application of the optimized model with best ratios and 10-fold cross-validation to Fengjie County and Wushan County; and (d) evaluation the generalization ability of the optimized model (Fig. 6).

2.1 Landslide-Conditioning Factors

In the LSM research, a crucial step is to select landslide-conditioning factors to be included in the universal model as input variables. According to Ayalew and Yamagishi (2005), the selected conditioning factors for GIS-based LSM should be measurable, operational, non-uniform, complete, and non-redundant. Based on the field investigation, the literatures and the data available in the study area, 22 landslide-conditioning factors were identified, including topography elements (elevation, degree of relief, aspect, slope, curvature, plan curvature, profile curvature, slope position, micro-landform, topographic wetness index (TWI), terrain roughness index (TRI), sediment transport index (STI), and stream power index (SPI)), geological conditions (distance from faults, lithology, combination reclassification of stratum dip direction, and slope aspect (CRDS)), environmental conditions (distance from hydrographic net, normalized vegetation index (NDVI), land cover, and annual average rainfall), and human activities (distance from roads, and POI kernel density). Table 1 shows the description of each factor.

Slope, degree of relief, aspect, slope position, micro-landform, curvature, profile curvature, plan curvature, TRI, TWI, STI, and SPI were all acquired by processing the digital elevation model

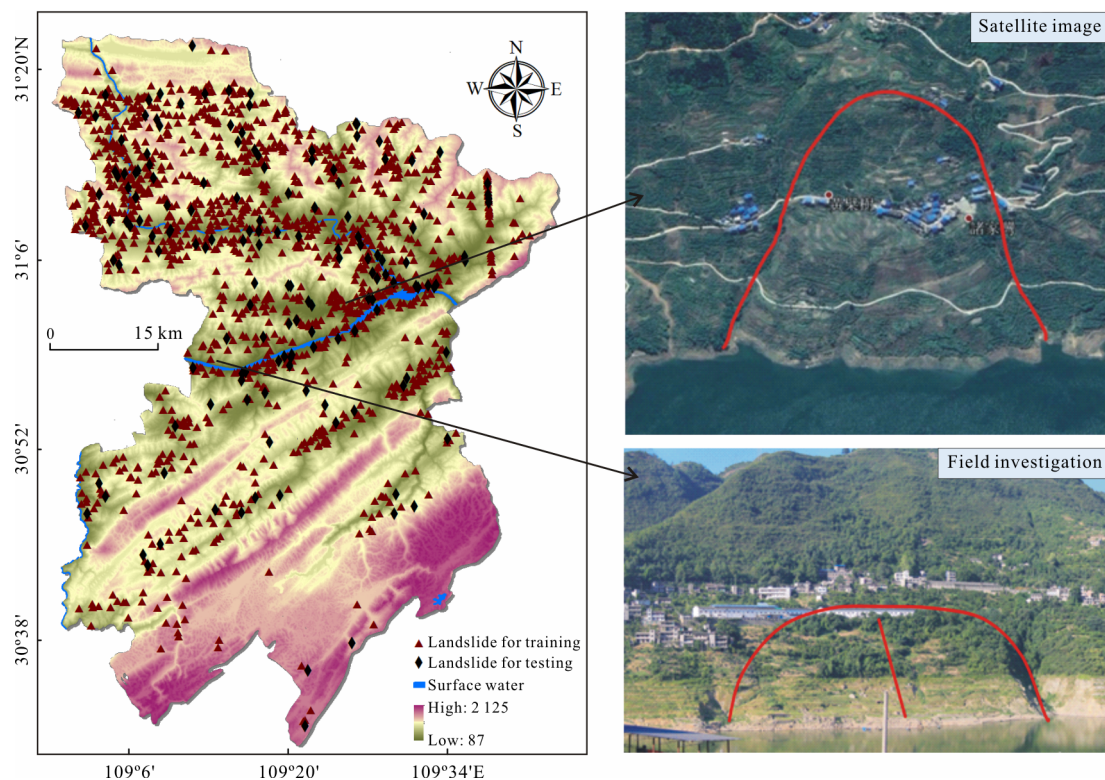


Figure 3. Landslide inventory map of Fengjie County.

Table 1 Description of landslide-conditioning factors

Category	Conditioning factor	Description
Topography	Elevation (m)	Elevation is the basic data to extract terrain data and usually represents vegetation patterns and climate
	Degree of relief (m)	Degree of relief refers to the maximum elevation difference in a certain area, which can be used to describe the terrain in a macroscopic way
	Slope (°)	Slope is the degree of steepness and slowness of the surface unit, usually indicated by the ratio of the vertical height of the slope to the distance in the horizontal direction
	Aspect	Aspect is the projection direction of slope, normally on the horizontal plane. The orientation of mountain effects on sunshine hours and solar radiation intensity
	Curvature	Curvature could be seen as slope of a slope, which affects soil erosion and affects the movement of water on slope surface
	Profile curvature	Profile curvature is a measure of the change rate of the ground elevation along the orientation of the maximum gradient of the ground slope. Its essence is the change rate of the ground slope. A value greater than 0 means the grid is part of a concave up slope. A value less than 0 means the slope is raised
	Plan curvature	Plan curvature is the slope perpendicular to the orientation of the maximum slope. A value greater than 0 means the unit is part of the lateral concave slope. A value less than 0 means the slope is laterally convex
	Slope position	Slope position refers to the geomorphic element of the slope
	Micro-landform	Micro-landform is a small-scale geomorphic form
	TWI (Hong et al., 2016b)	TWI can quantitatively describe the spatial distribution of soil moisture in an area. The larger the value, the higher the soil water content, and the more likely the soil will reach saturation state to generate runoff
	TRI	TRI is a surface roughness index, referring to the ratio of the surface area of the earth to the projected area in a certain region. It can describe the surface morphology macroscopically
	STI (Pourghasemi et al., 2012; Moore and Wilson, 1992)	STI is a comprehensive index used to characterize the transport and deposition of surface materials with water flows
	SPI (Sestraş et al., 2019; Yu et al., 2019)	SPI can quantitatively describe the erosion capacity of surface water. The results include the path formed by the flow convergence and the possible erosion gully points. The larger the SPI value, the stronger the erosion ability of surface water flow
Geological conditions	Distance from faults (m)	Within a certain range, the closer to the main fault, the looser the soil quality of the slope body, thus prone to landslides
	Lithology	Due to the different formation time and weathering degrees and various weakness degrees of formation lithology, cultivated soil is different and the ability to maintain soil water content is different
	CRDS (Wen et al., 2017)	It relates to the combination of the rock inclination and the slope aspect, as well as the sedimentary stack direction of the slope
Environmental conditions	Distance from hydrographic net (m)	The closer the water area is to rivers and lakes, the lower the altitude is, and the higher the soil moisture content is. To a certain extent, it can reflect the soil moisture in a small area, and can also reflect the terrain and topography of this area
	NDVI	NDVI can describe vegetation cover on the surface
	Land cover	Land cover is the result of reclassification and merging based on the data of Chongqing's Second Survey in 2015. Land cover represents vegetation and land use patterns, both of which may affect susceptibility
	Annual average rainfall (mm)	Its time scale is 15 years from 2000 to 2014, and it can roughly reflect the rainfall in the study area
Human activities	Distance from roads (m)	The construction of roads will wreck the stability of slopes and let it prone to landslides
	POI kernel density	POI, a feature that specifies coordinates of specific points (e.g., schools, restaurants, churches, etc.) on maps, can be used to identify information on urban structures and economic vitalities, which are related to various human activities at a micro-scale (Yao et al., 2017; Bakillah et al., 2014). POI kernel density based on POI data could effectively reflect the patterns of population distributions and human activities, which will also make the soils around the slope loose and affect the stability of strata to a certain extent

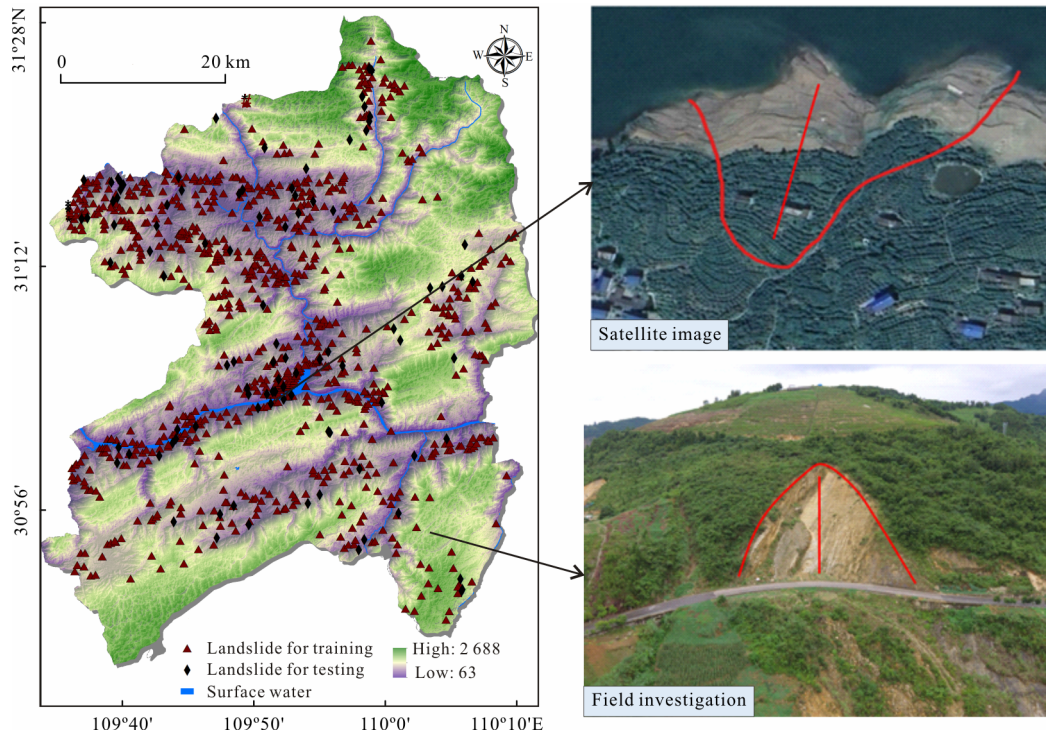


Figure 4. Landslide inventory map of Wushan County.

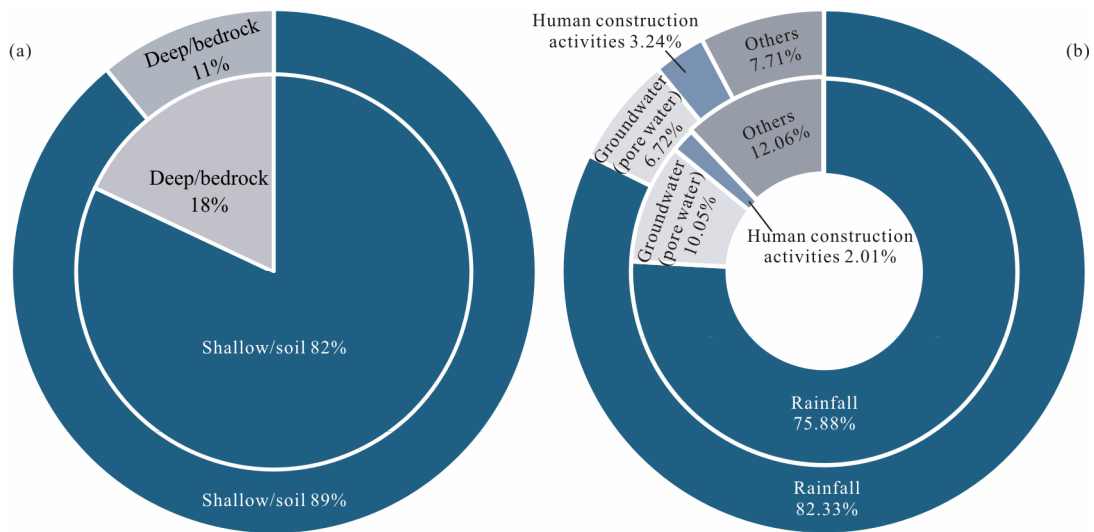


Figure 5. Statistics of landslide type and trigger (inner ring: Fengjie County; outer ring: Wushan County). (a) In terms of type; (b) in terms of trigger.

(DEM) from ASTER GDEM with 30 m resolution. Landsat 8 OLI satellite images were used to extract NDVI and land cover data. Lithologies and faults were extracted by vectorizing geological maps at a scale of 1 : 200 000. Google Earth was used to obtain the hydrographic nets and roads. The distances from faults, hydrographic nets, and roads were generated after buffering the faults, river networks and roads. CRDS was extracted by subtraction and reclassification of aspects and tendencies. Annual average rainfall was derived from information of local climate stations by using the spatial interpolation method. The POIs in this research were provided by Baidu Map Service (<http://map.baidu.com>), the largest and most-widely-used web map service provider in China. With the help of the application programming interfaces (API) provided by Baidu Map, the population-related POIs were ex-

tracted from the Fengjie and Wushan area to get a total of about 17 324 and 12 688 records, respectively, including commercial sites, business establishments, residential communities, educational facilities (e.g., kindergartens, primary schools, and middle schools), clinical facilities and scenic locations. POI kernel density was based on POI data generated by using the kernel density tool of ArcGIS software.

Since all the 22 landslide-conditioning factors were at different intervals or scales, all grid data were transformed to raster grids with 30 m resolution corresponding to DEM. According to relevant literature and the areas of the landslides in our study area, 30 m×30 m cells are detailed enough to catch the spatial characteristics of landslides and also comprehensive enough to lessen the computational complexity (Huang et al.,

2017; He et al., 2012). Meanwhile, classification was performed for the continuous factors, i.e., elevation, slope, degree of relief, curvature, profile curvature, plan curvature, TRI, TWI, STI, SPI, NDVI, annual average rainfall, and POI kernel density, so as to facilitate further analyses. Based on field investigations, experts' experiences and several examples in the literature, the threshold value of each category was first determined by the natural breakpoint method, and then slightly adjusted by calculating the landslide density under each category, so as to make it more in line with the actual situation (Wu et al., 2020; Li et al., 2019; Xie et al., 2018; Wen et al., 2016; Xu et

al., 2012) (Table 2). To reduce the data discreteness, the 22 landslide-conditioning factors after reclassification were normalized. The classification index values of these factors were then transformed linearly so that the values were reduced to [0, 1] interval. Since RF is a non-linear model, there is no need to check the multicollinearity between the factors. The thematic layers of these factors are shown in Fig. 7.

Moreover, in order to explore the generalization ability of the model in Wushan County, the 22 landslide-conditioning factors of Wushan County were also collected, and the reclassification and normalization method during data processing

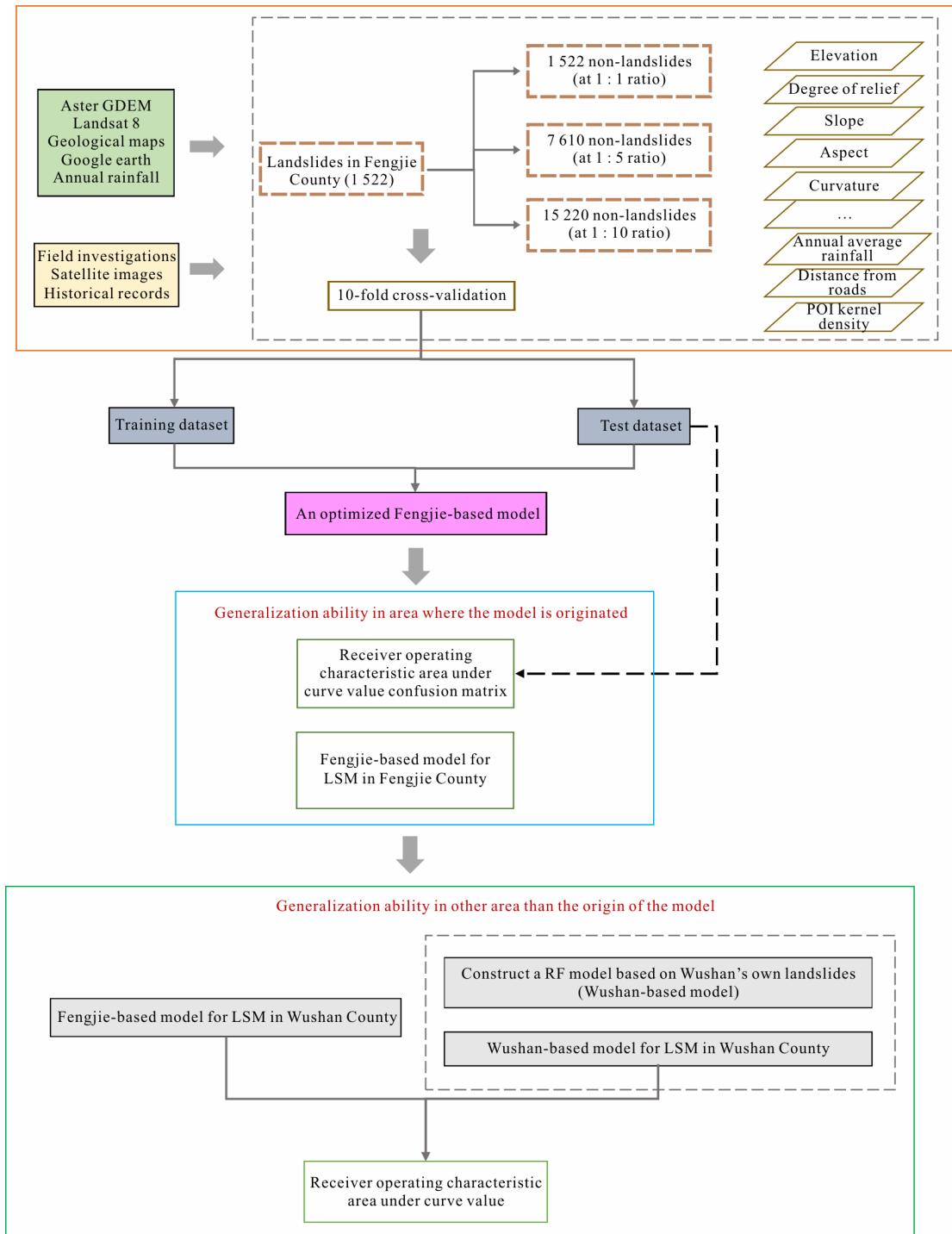


Figure 6. Flowchart of the study.

Table 2 Categories of landslide-conditioning factors

Conditioning factor	Class	Classification standard
Elevation (m)	11	1. < 340; 2. 340–543; 3. 543–690; 4. 690–832; 5. 832–951; 6. 951–1 053; 7. 1 053–1 144; 8. 1 144–1 302; 9. 1 302–1 556; 10. 1 556–1 654; 11. >1 654
Degree of relief (m)	7	1. < 20; 2. 20–30; 3. 30–40; 4. 40–50; 5. 50–80; 6. 80–120; 7. >120
Slope (°)	9	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
Aspect	9	1. Flat; 2. north; 3. northeast; 4. east; 5. southeast; 6. south; 7. southwest; 8. west; 9. northwest
Curvature	6	1. < -1; 2. -1– -0.5; 3. -0.5–0; 4. 0–0.5; 5. 0.5–1; 6. >1
Profile curvature	6	1. < -1; 2. -1– -0.5; 3. -0.5–0; 4. 0–0.5; 5. 0.5–1; 6. >1
Plan curvature	6	1. < -1; 2. -1– -0.5; 3. -0.5–0; 4. 0–0.5; 5. 0.5–1; 6. >1
Slope position	6	1. Ridge; 2. upper slope; 3. middle slope; 4. flats slope; 5. lower slope; 6. valleys
Micro-landform	10	1. Canyons, deeply incised streams; 2. midslope drainages, shallow valleys; 3. upland drainages, headwaters; 4. U-shape valleys; 5. plains; 6. open slopes; 7. upper slopes, mesas; 8. local ridges hills in valleys; 9. midslope ridges, small hills in plains; 10. mountain tops, high narrow ridges
TWI	5	1. < 4; 2. 4–6; 3. 6–8; 4. 8–10; 5. >10
TRI	5	1. <1.05; 2. 1.05–1.1; 3. 1.1–1.15; 4. 1.15–1.2; 5. >1.2
STI	6	1. < 20; 2. 20–40; 3. 40–70; 4. 70–100; 5. 100–200; 6. >200
SPI	7	1. <15; 2. 15–30; 3. 30–45; 4. 45–60; 5. 60–100; 6. 100–1 000; 7. >1 000
Distance from faults (m)	7	1. < 500; 2. 50–1 000; 3. 1 000–1 500; 4. 1 500–2 000; 5. 2 000–2 500; 6. 2 500–3 000; 7. >3 000
Lithology	12	1. J _{3p} /J _{3s} ; 2. J _{2s} /J _{2xs} ; 3. J _{1-z} /J _{1z} ; 4. T _{3xj} ; 5. T _{3b1} ; 6. T _{2b2} ; 7. T _{1d} ; 8. T _{1j} ; 9. P ₂ ; 10. P ₁ ; 11. D ₃ /D ₂ ; 12. S ₁₋₂
CRDS	7	1. Dip-slope I; 2. dip-slope II; 3. outward slope; 4. oblique slope; 5. tangential slope; 6. reverse slope; 7. flat
Distance from hydrographic net (m)	7	1. <100; 2. 100–200; 3. 200–300; 4. 300–400; 5. 400–500; 6. 500–600; 7. >600
NDVI	7	1. <0.10; 2. 0.10–0.15; 3. 0.15–0.20; 4. 0.20–0.25; 5. >0.25
Land cover	8	1. Meadow; 2. farmland; 3. water area; 4. forest; 5. garden plot; 6. transportation; 7. residential land; 8. others
Annual average rainfall (mm)	8	1. <1 221; 2. 1 221–1 251; 3. 1 251–1 276; 4. 1 276–1 308; 5. 1 308–1 343; 6. 1 343–1 389; 7. 1 389–1 440; 8. >1 440
Distance from roads (m)	7	1. <100; 2. 100–200; 3. 200–300; 4. 300–400; 5. 400–500; 6. 500–600; 7. >600
POI kernel density	7	1. <1; 2. 1–2; 3. 2–3; 4. 3–4; 5. 4–5; 6. 5–10; 7. >10

was consistent with the corresponding factors of Fengjie County.

2.2 Preparation of the Dataset

The most important problem in pre-processing works before constructing a model for LSM is to ensure that there are a sufficient number of samples, including positive and negative cells. As a rule, relatively few positive cells are used because landslides are relatively rare, whereas non-landslide samples are much more numerous than landslide samples and are more easily collected (Hussin et al., 2016). The use of additional negative cells could expand the data size for machine learning, but it will cause a sample imbalance problem, thereby biasing the classifier towards the negative cells and negatively affecting its performance (Wang et al., 2019). Heckmann et al. (2014) summarized the types of regression analyses, finding that the ratios of positive-to-negative cells often range between 1 : 1 and 1 : 10. In this study, positive cells consisted of 1 522 historical landslides, each of which was regarded as a single grid cell (30 m×30 m). Three typical ratios of positive-to-negative cells, i.e., 1 : 1, 1 : 5, and 1 : 10, are adopted for comparative analyses to construct a model with great prediction capability. Accordingly, the number of non-landslides corresponding to each ratio (1 522, 7 610, and 15 220) was randomly extracted from the non-landslide area in Fengjie County.

2.3 Random Forest Model

As an ensemble of individually trained binary decision trees, RF is one of the most widely used classifier methods, with successful results for regression, classification and feature selection (Chen et al., 2018c). Unlike other machine learning models, RF returns several important measures of a variable; and then the most reliable measure is used to reduce the classification accuracy when the values of the variables in the nodes of the tree are randomly arranged (Breiman, 2001).

The key point of RF is that it combines n independent decisions $[y(X, \theta_k; k=1, 2, \dots, n)]$ to construct a model. Each decision tree in the model judges or predicts the samples. Different classification models $y_1(X), y_2(X), \dots, y_k(X)$ are obtained after sample training. Then, these classification models are used to build RF models. Last, voting is conducted

$$Y(X) = \arg \max_Z \sum_{i=1}^k I(y_i(X) = Z) \quad (1)$$

where $Y(X)$ represents a RF model; $y_i(X)$ denotes a single decision tree model; Z means output variable; and $I(\cdot)$ is an explicit function.

The generalization error of RF depends on the accuracy of a single tree and the correlation between the trees. The final

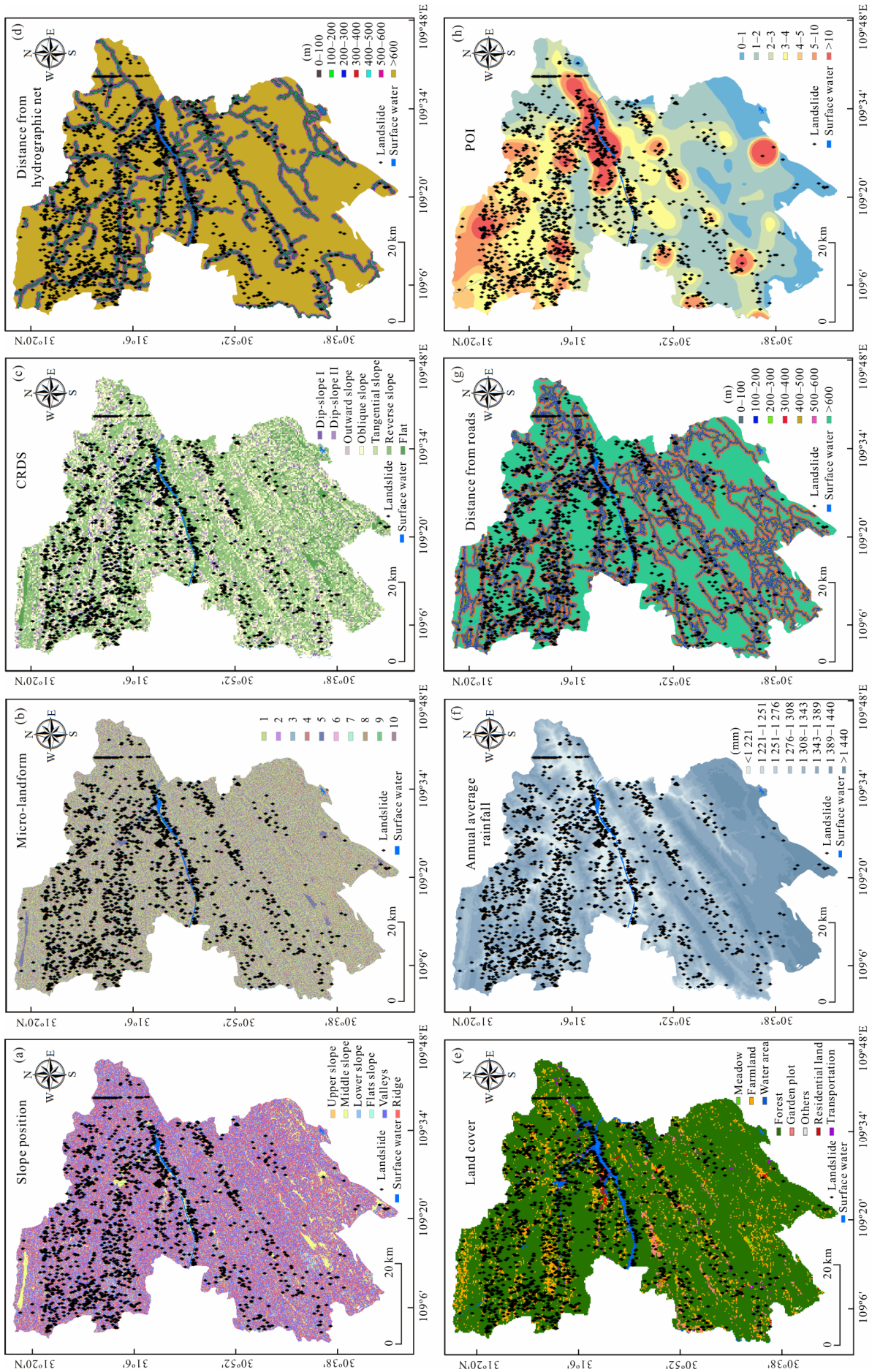


Figure 7. Partial thematic maps of landslide-conditioning factors. (a) Slope position; (b) micro-landform; (c) CRDS; (d) distance from hydrographic net; (e) land cover; (f) annual average rainfall; (g) distance from roads; (h) POI kernel density.

prediction result is achieved by a majority vote of the decision trees (Sahin et al., 2018). Figure 8 shows the steps of the RF algorithm.

In this study, the RF for LSM consists of two trees (i.e., positive and negative cells), each of which is constructed with 22 random features (i.e., the 22 landslide-conditioning factors).

2.4 Validation Method

2.4.1 10-fold cross-validation

When utilized to validate the accuracy of algorithms, the 10-fold cross-validation could reduce the variability and is well applicable in the case of small samples. The general process of 10-fold cross-validation is: first, to randomly divide the dataset into ten subsets, with one of them used as the test dataset and the other nine as the training dataset each time. Then under a certain model, each operation of fitting will get certain results and a correct rate (or an error rate). A total of ten such operations are performed. Finally, the average value of the accuracies (or error rates) for the 10 subsets is utilized to evaluate the accuracy of the algorithm (Jiang and Chen, 2016). After this, among these subsets, the subset with the highest test accuracy is utilized to construct a model related to LSM.

2.4.2 Confusion matrix and ROC curve

Any landslide susceptibility assessment will have no scientific significance without validation, and there is no exception for

this study. The confusion matrix and ROC have been widely used to evaluate the performance of landslide susceptibility assessment models (Huang and Zhao, 2018; Kalantar et al., 2017; Shirzadi et al., 2017). ROC is a kind of curves based on a confusion matrix, which takes sensitivity and specificity as horizontal and vertical axes, respectively. The value of AUC under the ROC curve of the training dataset is a symptom of the success power of a model, while the AUC of the test dataset is a symptom of the predictive capabilities of a model (Tsangaratos et al., 2016). Such statistical measures based on the confusion matrix as accuracy and precision were also utilized to evaluate the model, and these measures could be calculated as shown in Table 3. For the value of AUC, accuracy, and precision, a value of 1 indicates a perfect model, a value of 0 indicates an information-less model, and higher value indicates a better model (Zhang et al., 2019; Das et al., 2012).

3 ANALYSES AND RESULTS

3.1 An Optimized RF Model with Best Ratios and 10-Fold Cross-Validation

Table 4 shows the accuracy of the 10-fold cross-validation of the RF models at different ratios of positive-to-negative cells. As can be seen, the average accuracy of the test dataset of the RF model at ratios of 1 : 1, 1 : 5, and 1 : 10 were 0.769, 0.851, and 0.914, respectively. Subset 3 with the ratio of 1 : 1, subset 1 with

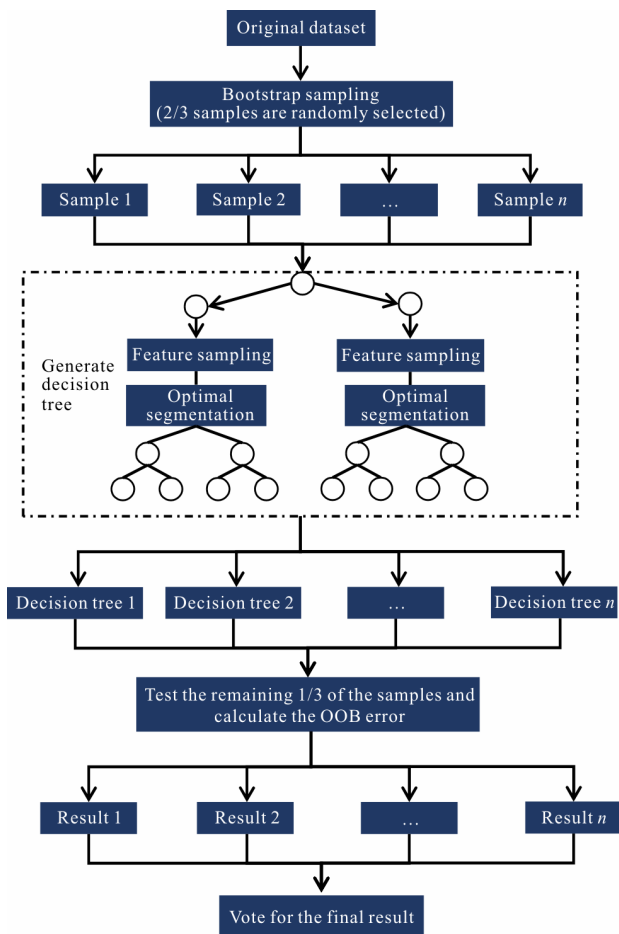


Figure 8. The schematic diagram of the RF algorithm.

Table 3 Statistical measures based on the confusion matrix

Statistical measure	Equation	Definition
Accuracy	$ACC = \frac{TP + TN}{TP + FP + TN + FN}$	The proportion of positive and negative cells which are correctly identified
Precision	$PRE = \frac{TP}{TP + FP}$	The proportion of positive cells in the positive sample identified
Sensitivity	$SST = \frac{TP}{TP + FN}$	The percentage of positive cells that are correctly identified
Specificity	$SPF = \frac{TN}{TN + FP}$	The percentage of negative cells that are correctly identified

Table 4 The accuracy of 10-fold cross-validation at different ratios

Subset	Accuracy of the test dataset		
	Ratio of 1 : 1	Ratio of 1 : 5	Ratio of 1 : 10
1	0.796	0.838	0.916
2	0.779	0.884	0.913
3	0.800	0.847	0.914
4	0.786	0.837	0.918
5	0.737	0.859	0.912
6	0.734	0.864	0.922
7	0.757	0.843	0.918
8	0.763	0.852	0.904
9	0.787	0.847	0.921
10	0.760	0.841	0.905
Mean	0.769	0.851	0.914

Table 5 Confusion matrix of the modelsat different ratios

Ratios		Actual value			
		Landslide (1)	Non-landslide (0)		
1 : 1	Predicted	Landslide (1)	1 473	49	ACC: 0.969
		Non-landslide (0)	49	1 473	PRE: 0.969
1 : 5	Predicted	Landslide (1)	1 346	28	ACC: 0.978
		Non-landslide (0)	176	7 582	PRE: 0.980
1 : 10	Predicted	Landslide (1)	1 229	7	ACC: 0.982
		Non-landslide (0)	293	15 213	PRE: 0.994

the ratio of 1 : 5, and subset 6 with the ratio of 1 : 10 had a relatively higher accuracy (0.800, 0.884, and 0.922). Therefore, the RF models based on different ratios of positive-to-negative cells were constructed by using the training dataset of these subsets. Table 5 presents the confusion matrix of all datasets of these RF models. As for the statistical measure of accuracy, the values of it at ratios of 1 : 1, 1 : 5, and 1 : 10 were 0.969, 0.978, and 0.982, respectively. As for the statistical measure of precision, the values of it at ratios of 1 : 1, 1 : 5, and 1 : 10 were 0.969, 0.980, and 0.994, respectively. Figure 9 shows the ROC curves of the RF models at different ratios, where the AUC value of the test dataset at ratios of 1 : 1, 1 : 5, and 1 : 10 were 0.851, 0.862, and 0.876, respectively. Although all models at different ratios showed reasonable goodness of fit for the test datasets, however, the RF model at the ratio of 1 : 10 performed better, whether by measuring of confusion matrix or AUC value. The RF model at a ratio of 1 : 10 had a better prediction than ratios of 1 : 1 and 1 : 5 in this case. Therefore, at last, the optimized RF model (Fengjie-based model) based on Fengjie landslides with best ratios of positive-to-negative cells and 10-fold cross-validation was adopted to evaluate its generalization ability in Fengjie County and Wushan County.

3.2 Generalization Ability in Fengjie County

After construction, the optimized model was applied to the whole Fengjie area to evaluate its generalization ability in the area where the model is originated. Based on expert experience method, 4 categorized breakpoints (at 0.05, 0.20, 0.30, and 0.51) were extracted to divide the susceptibility values in Fengjie area into 5 classes, corresponding to very low, low, medium, high, and very high susceptibility regions, respectively. The selection of the categorized breakpoints should be able to minimize the density of historical landslides in the low susceptibility region and maximize that in the high susceptibility region, which is as consistent as possible with the actual situation (Sun et al., 2020). Figure 10 is the LSM of Fengjie County generated by the Fengjie-based model, which shows that most areas of Fengjie County were located in the low susceptibility regions, concentrated in the east and southwest. The high susceptibility regions were concentrated along the Yangtze River and its tributaries, mainly located in the northwest of Fengjie, a result that is well consistent with the distribution pattern of actual landslides.

Moreover, statistical analysis was used to quantitatively evaluate the model generalization ability via the following metrics: the percentage of each susceptibility classification, the number of landslides, and the proportion and density of each class' region. As shown in Table 6, more than 78.65% of the

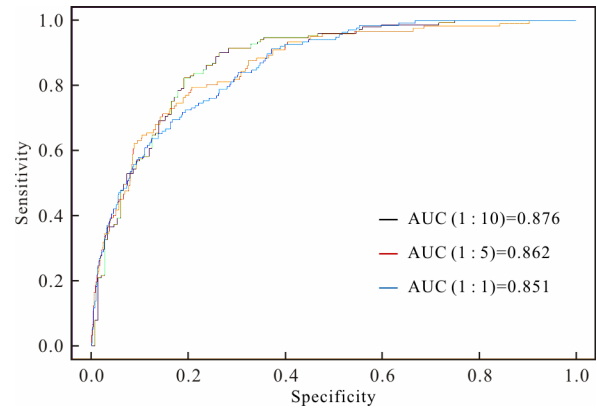


Figure 9. ROC curve of the Fengjie-based model.

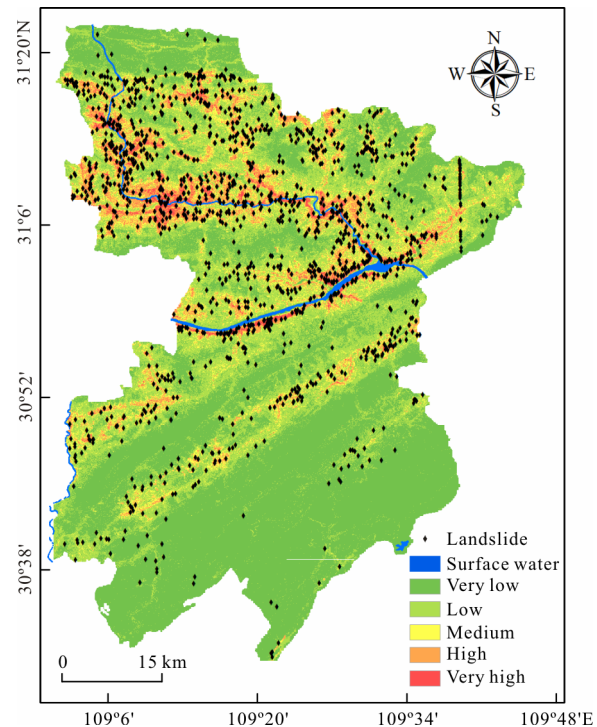


Figure 10. LSM of Fengjie County generated by Fengjie-based model.

landslides were located in 2.11% of the very high susceptibility regions, and only 0.46% of the landslides were located in 45.97% of very low susceptibility region. The landslide density was increased by approximately 53 times (from 0.070 to 3.743) when the susceptibility level varied from “very low” to “very high”. This result showed that the LSM generated by the Fengjie-based model is well consistent with the distribution of actual

Table 6 Statistics of the susceptibility classes for Fengjie County

Susceptibility class	Coverage of area	Landslide (point)	Landslide (%)	Landslide density (Pcs/km ²)
Very low	45.97%	7	0.46%	0.070
Low	32.68%	45	2.96%	0.143
Medium	10.21%	30	1.97%	0.326
High	9.04%	242	15.90%	0.483
Very high	2.11%	1 197	78.65%	3.743

landslides and delivered high stability and predictive capability in Fengjie County, indicating a great generalization ability of the model to the area where the model is originated.

3.3 Generalization Ability in Wushan County

The LSM for Wushan County generated by the Fengjie-based model (LSM1) is shown in Fig. 11. Based on expert experience method, 4 categorized breakpoints (at 0.08, 0.18, 0.25, and 0.38) were extracted to divide the susceptibility values in the Wushan area into 5 classes, corresponding to very low, low, medium, high, and very high susceptibility regions, respectively. According to the LSM1, most areas of Wushan County were located in regions with low susceptibility to landslides and were concentrated in relatively flat areas, such as the upper east and southwest regions. The high susceptibility regions were concentrated along the Yangtze River and its tributaries, mainly located in the northwest of Wushan.

To evaluate the generalization ability of the Fengjie-based model in the Wushan area, we collected the Wushan area landslides and generated the ROC curve based on these actual landslides in Wushan County and the predicted value of the Fengjie-based model (Fig. 12). As shown in Fig. 12, the AUC value was 0.813 in this case, indicating the efficient application performance and great generalization ability of the Fengjie-based model.

Additionally, a RF model (Wushan-based model) based on Wushan’s own landslides data and conditioning factors was constructed by following the same steps as in the Fengjie-based model in this study. Figure 13 shows the ROC curves of this Wushan-based model; and the AUC values of the training dataset and test dataset were 1.00 and 0.823, respectively. The LSM for Wushan County generated by the Wushan-based model (LSM2) is shown in Fig. 14. Based on expert experience method, 4 categorized breakpoints (at 0.13, 0.22, 0.33, and 0.5) were extracted to divide the susceptibility values in Wushan area into 5 classes, corresponding to very low, low, medium, high, and very high susceptibility regions, respectively.

To compare the results of different models, statistical analysis was made to quantitatively evaluate the validity of the models. As shown in Table 7, in the very low susceptibility region, the landslide density of LSM1 is 0.055 Pcs/km² and the landslide density of LSM2 is 0.034 Pcs/km². In the very high susceptibility region, the landslide density of LSM1 is 1.595 Pcs/km² and the landslide density of LSM2 is 1.987 Pcs/km². As for LSM1, when the susceptibility level varied from “very low” to “very high”, the landslide density would be increased by approximately 29 times. As for LSM2, under the same change, the landslide density would be increased by approximately 58 times. As the susceptibility level increases, the

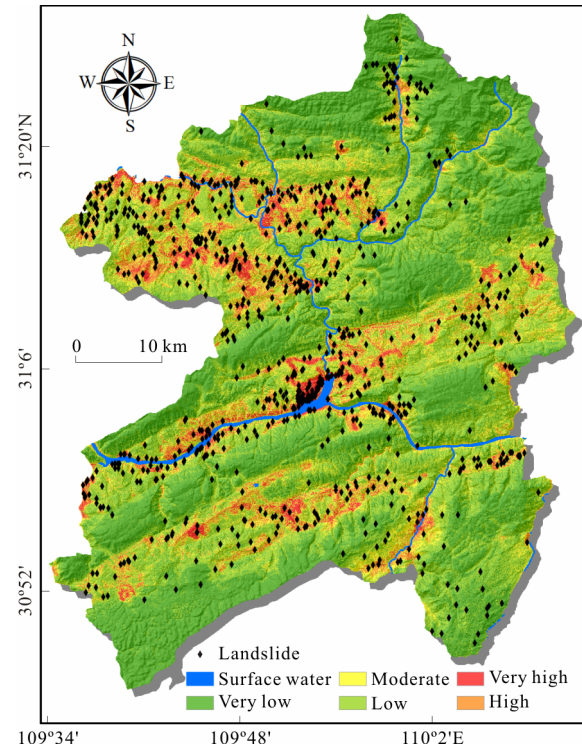


Figure 11. LSM of Wushan County generated by the Fengjie-based model.

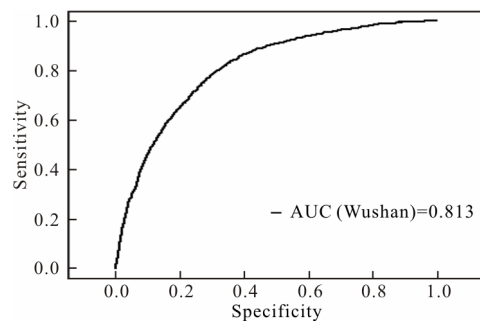


Figure 12. ROC curve of Wushan County generated by the Fengjie-based model.

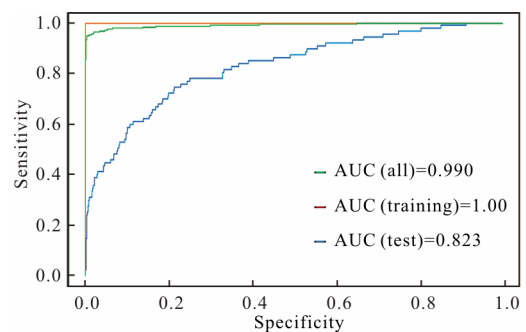


Figure 13. ROC curve of the Wushan-based model.

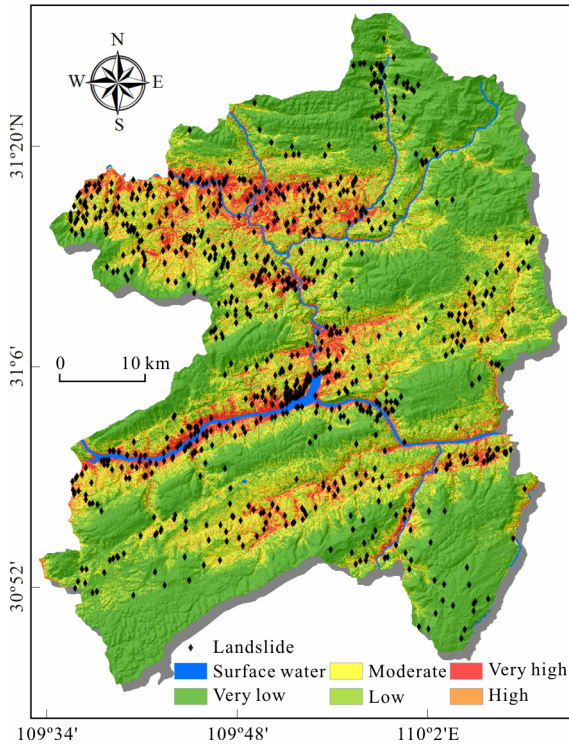


Figure 14. LSM of Wushan County generated by the Wushan-based model.

Table 7 Statistics of the susceptibility classes for Wushan County

Susceptibility class	Landslide density (Pcs/km ²)	
	LSM1	LSM2
Very low	0.055	0.034
Low	0.250	0.163
Medium	0.483	0.347
High	0.841	0.900
Very high	1.595	1.987

landslide density also gradually increases, which indicates that the LSMs generated by the Fengjie-based and Wushan-based models were consistent with the distribution pattern of the actual historical landslides, and both models had certain reference value.

Moreover, two typical landslides are selected to demonstrate the model performance at a micro-scale. Located in the southwest of Wushan County (Fig. 15), the Gongjiapo Landslide occurred in September 2014, mainly triggered by rainfall. The main type of the landslide strata is T_{2b}, dominated by sandstone and limestone, with an elevation of 433 m a.s.l. and a slope of 44°. The landslide area covers about 7.4×10⁴ m², with a volume of about 18.5×10⁴ m³, so it is categorized as a medium landslide, burying more than 70 people of 16 families in Gongjiapo Village and leading to an asset loss of CNY 5.4 million yuan. Also triggered by rainfall, another typical landslide occurred in July 2016 in Liujiapo, also located in the southwest of Wushan County (Fig. 15). The landslide materials are mainly composed of sandstone, limestone, and shale. At an elevation of 1 025 m a.s.l. and a slope of 33°, the landslide covers an area of about 0.32×10⁴ m², with a volume of about 0.8×10⁴ m³, so it can also be categorized as a medium landslide, claiming the life of more than 122 people of 132 families in Liujiapo Village and leading to an asset loss of CNY 200 million yuan. As shown in Fig. 15, Gongjiapo Landslide was in very high susceptibility regions whether in the Fengjie-based or Wushan-based model, while Liujiapo Landslide was in high susceptibility region in Wushan-based model, and in medium susceptibility region in Fengjie-based model.

It is concluded that, although the LSM generated by the Fengjie-based model was not as good as the LSM generated by Wushan-based model in Wushan area, it also had a reasonable reference value. In other words, the great generalization ability of the Fengjie-based model was demonstrated in the area other than the origin of the model.

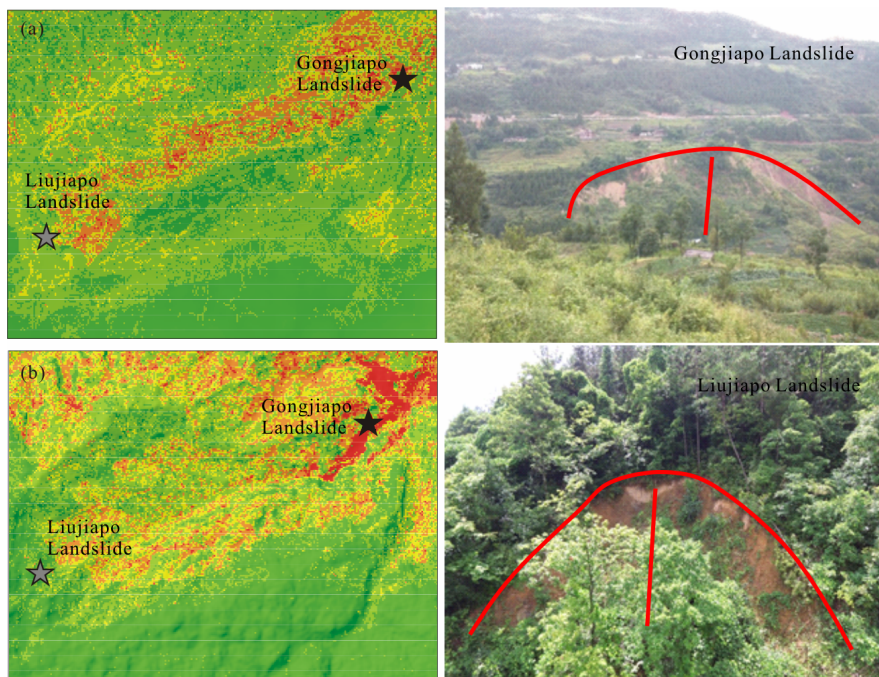


Figure 15. Typical landslides in Wushan County.

4 DISCUSSION

4.1 Ratio of Positive-to-Negative Cells

It is an essential procedure to define the sampling size that will be exploited to train and test the susceptibility model. The modeler needs to decide not only the number of positive cells but also the number of negative cells to be used in assessing the success and prediction capability of the model. The use of additional negative cells could expand the data size for machine learning, so a sample has to be large enough to cover the variability of geo-factors within the study area, and to yield stable and reproducible results (Heckmann et al., 2014). However, it will cause a sample imbalance problem, thereby biasing the classifier towards the negative cells and negatively affecting its performance (Wang et al., 2019). Moreover, a large sample is likely to violate the assumption of independent observations due to spatial autocorrelation. Therefore, in the actual process, it is necessary to constantly try different ratios of positive-to-negative cells to construct a model with better prediction capability and performance. In this study, 3 typical ratios of positive-to-negative cells, i.e., 1 : 1, 1 : 5, and 1 : 10, are adopted for comparative analyses. As shown in Fig. 5 and Table 9, as the proportion of negative cells increases, the accuracy of the model becomes increasingly higher and there is no imbalance problem that the classifier is biased towards the negative cells. However, only these 3 most typical ratios have been tried in this study, and higher ratios have not been tried. How to improve the accuracy and avoid imbalance problems in which classifier is biased towards the negative cells is a question worthy of further discussion in the subsequent research.

4.2 Generalization Ability in Area other than the Origin of the Model

In the current study, the great generalization ability of the

Fengjie-based model was demonstrated in the Wushan area, which is an area other than the origin of the model. The same terrain, geological and hydrological conditions in both Wushan County and Fengjie County could facilitate the explanation of the generalization ability to some extent. As can be seen from Section 2.1, both counties are located in the east of the Sichuan Basin, in some mountainous landforms as well as at the junction of the Dabashan arc fold fault zone and the eastern Sichuan fold belt, with the same elevation range. The main types of strata are the same, including Triassic (T), Jurassic (J), Permian (P), Devonian (D), Silurian (S), Quaternary (Q), and Carboniferous (C) (Fig. 1). The monthly average rainfall and temperature distribution in Fengjie County and Wushan County are also almost the same (Fig. 2). Moreover, the types of landslides in these two counties are identical, that is, most of them are small-medium, shallow/soil landslides, and triggered by rainfall. Distributed together with slope soils, residuals, and shale, or mudstones without active faults, most of these landslides are developed in low mountain-landforms with a 20°–40° slope degree of poor stability. Hence, it was concluded that the same elevation range, annual temperature and rainfall, and identical landslides in both counties are the most basic elements to achieve the great generalization ability of the model.

Additionally, the prediction capability of landslide assessment models is closely related to the landslide-conditioning factors selected. Therefore, analyzing the importance of the factors for landslide occurrence in the different regions could also facilitate the explanation of the generalization ability. The mean decrease Gini, a metric in the RF model, can measure how each variable promotes the homogeneity of the nodes and leaves in the resulting RF and has particular reference value for the critical order of factors (Hong et al., 2016b). Figures 16 and 17 illustrate the 22 factors ordered by the mean decrease Gini of the Fengjie-

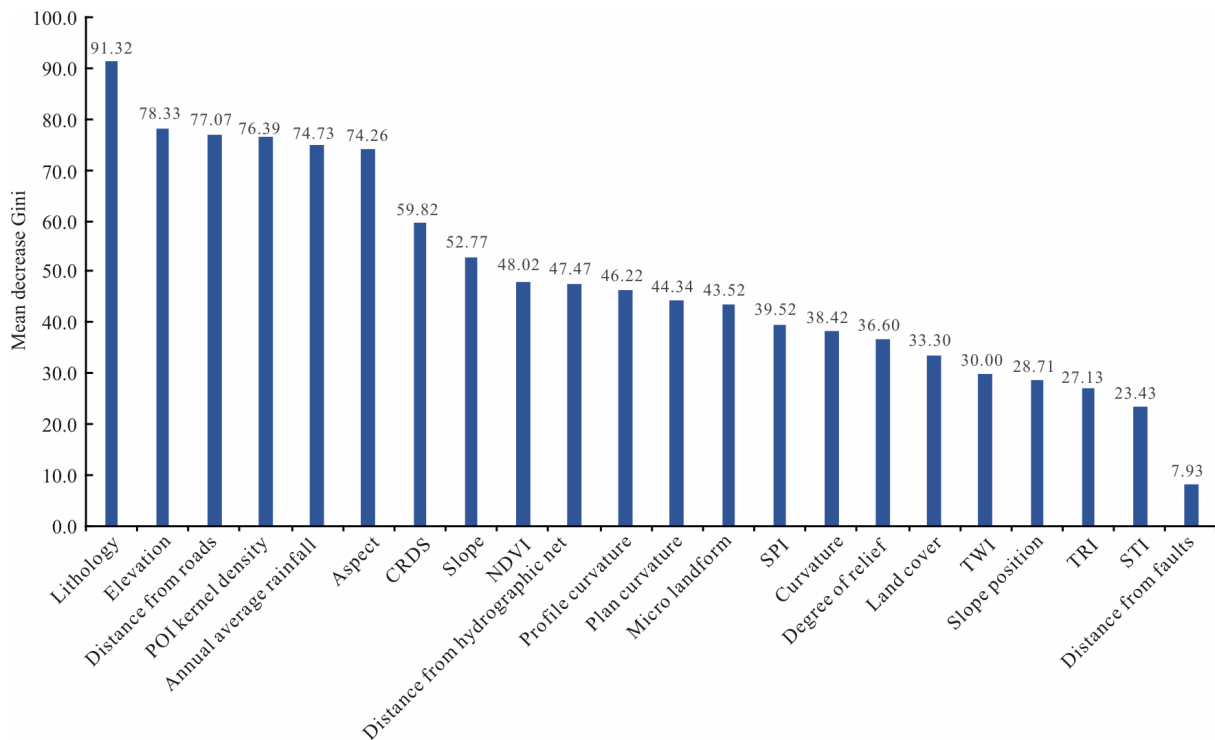


Figure 16. The importance of the factors of the Fengjie-based model.

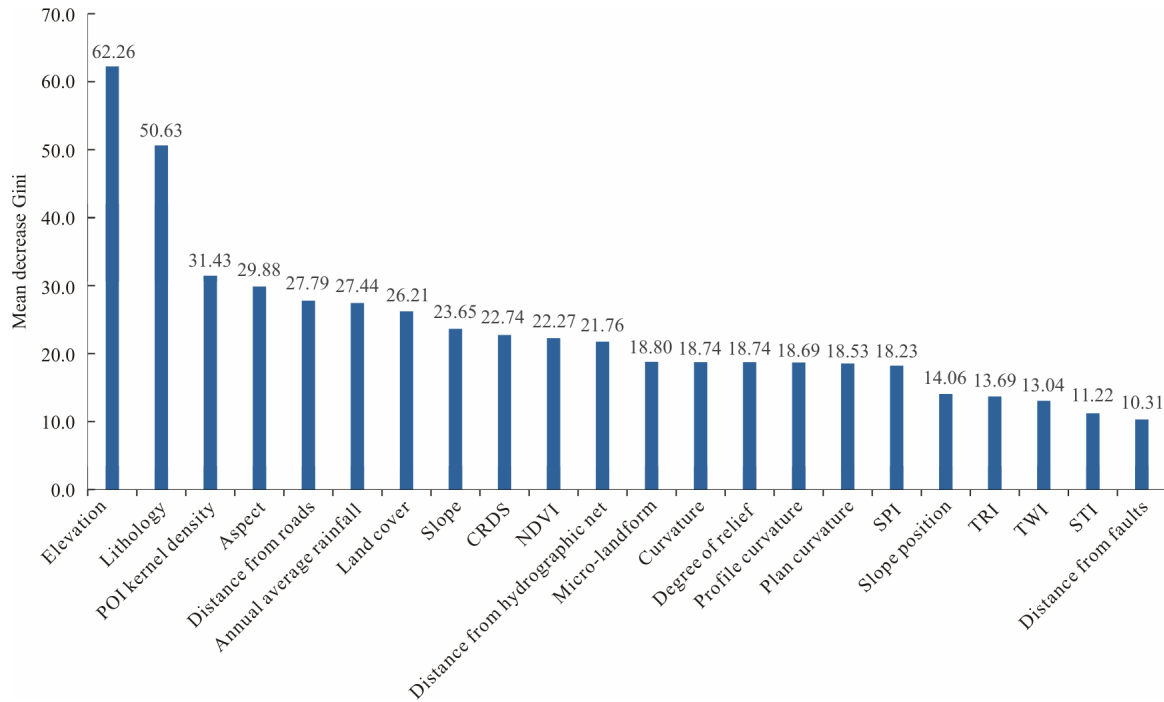


Figure 17. The importance of the factors of the Wushan-based model.

based and Wushan-based models, showing that the 10 most important conditioning factors for Fengjie-based model were: lithology, elevation, distance from roads, POI kernel density, annual average rainfall, aspect, CRDS, slope, NDVI, and distance from hydrographic net. On the other hand, the most 10 important factors of the Wushan-based model were: elevation, lithology, POI kernel density, aspect, distance from roads, annual average rainfall, land cover, slope, CRDS, and NDVI. Although there were slight differences in the ranking of the factors of these two models, the top 10 factors of them were the same, except for distance from hydrographic net and land cover. However, although distance from hydrographic net was not ranked in the top 10 factors for the Wushan-based model, it was still ranked in 11, which means this factor was also very important. In terms of the last factors in the ranking, the last 5 ones for the two counties were the same, i.e., TWI, slope position, TRI, STI, and distance from faults.

Among the most important conditioning factors, the lithology, elevation, distance from roads, POI kernel density, and annual average rainfall had higher values of mean decrease Gini in both models. This means that, if removing them, the error rate of the model will increase. Therefore, the landslide density maps of the Fengjie-based model and Wushan-based model were generated for these most important factors as an example to explore the influence of the factors for landslide occurrence (Fig. 18). As for continuous conditioning factors (including elevation, distance from roads, POI kernel density, and annual average rainfall), they were analyzed from a quantitative perspective. Obviously, although the values of the landslide-conditioning factor densities in the two counties were different in each class, the overall trend was the same. For example, the landslide density of the two counties were both negatively correlated with the elevation, distance from roads, and annual average rainfall, but positively with POI kernel density (Fig. 18). As for the qualitative factor (lithology), analyses found that although the lithology of the two coun-

ties is slightly different, the density of landslides was relatively high when the lithology was Jurassic (J), Triassic (T), and Permian (P).

Therefore, it was concluded that the Fengjie-based model had such good predictive power in Wushan County (i.e., the generalization ability), partly because the conditioning factors have the same effect on the occurrence of landslides in both Fengjie and Wushan counties. Combined with the same terrain, geological and hydrological conditions in both Wushan and Fengjie counties, it can be concluded that the optimized Fengjie-based model in this study could be applied in other similar areas/countries with the same terrain and geological conditions and small-medium, shallow/soil landslides. The result of this study is a good contribution to the landslide assessment research at the regional scale, as it could generate a LSM without collecting landslide information for modeling, so as to reduce workload and improve efficiency in practice, which has not been reported in previous studies.

4.3 A more Generalizable Model after Factor Selection

In order to avoid the disadvantages of the subjective selection of factors by expert experience and missing important factors, this study comprehensively weighs four categories of factors that affect landslides including topography elements, geological conditions, environmental conditions, and human activities, and finally selects 22 landslide-conditioning factors under these four categories. Selecting a large quantity of landslide-conditioning factors has both advantages and disadvantages. On the one hand, because the choice is subject to the subjective influence, the selection is less likely to miss those factors that are important for landslides, resulting in poor performance of the constructed models and difficulties in obtaining an ideal LSM. On the other hand, selecting too many landslide-conditioning factors will bring in the following shortcomings: (1) first, the workload of

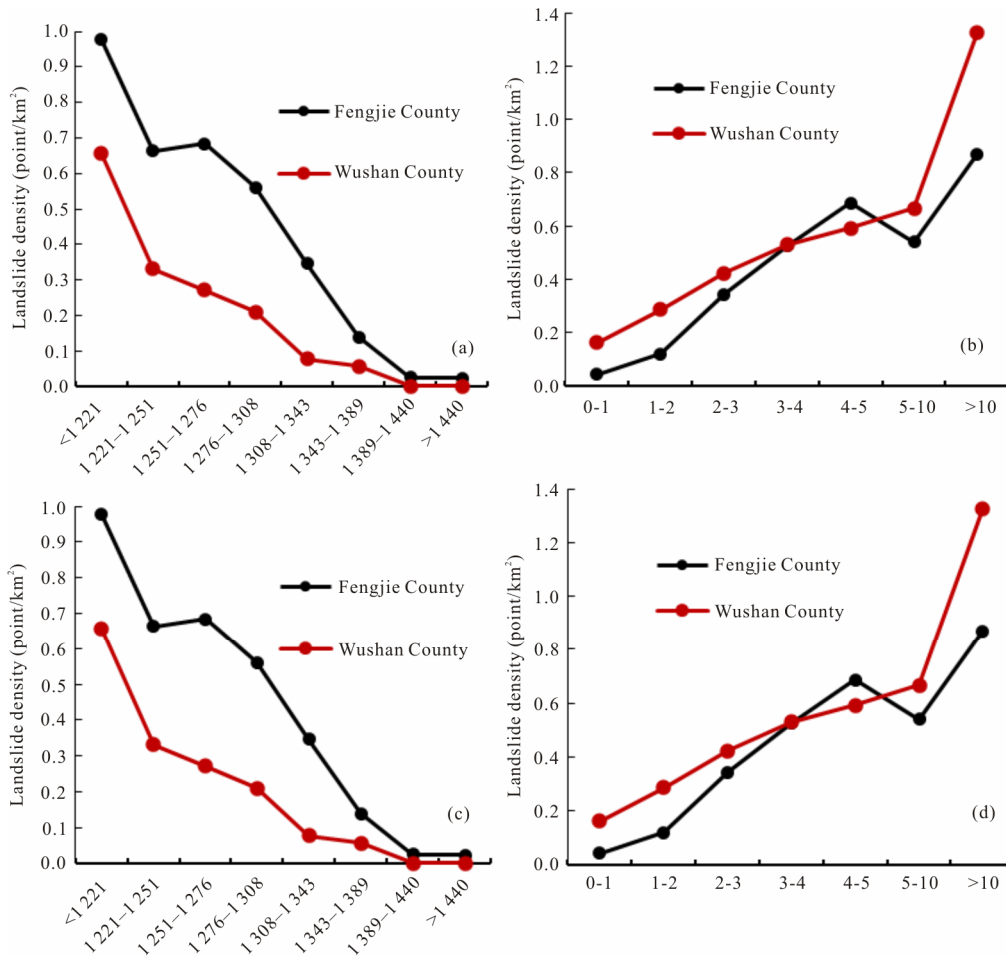


Figure 18. Landslide density charts. (a) Elevation; (b) distance from roads; (c) annual average rainfall; (d) POI kernel density.

obtaining data is increased, and the difficulty of obtaining data varies in different regions. Too many landslide-conditioning factors are not conducive to applying the constructed models to other regions; (2) the influence of some landslide-conditioning factors on landslides is not so significant, and (3) many landslide-conditioning factors have correlations with each other, so some factors could be replaced by others with the same influence.

In the current work, to facilitate the application of the model to other areas than its origin area more easily, we selected the top 10 most important factors ordered by the mean decrease Gini in the Fengjie-based and Wushan-based models (i.e., elevation, lithology, POI kernel density, aspect, distance from roads, annual average rainfall, slope, CRDS, distance from hydrographic net, and NDVI), and reconstructed a Fengjie RF model (Fengjie-based-generalizable model) by following the same steps as in the original Fengjie-based model. Figure 19 shows the AUC values of this model, it can be found that compared with the original Fengjie-based model, the AUC value of the Fengjie-based generalizable model on the test dataset is decreased only by 0.022, and the AUC value applied to Wushan County is decreased only by 0.013. Using just these 10 factors has achieved almost the same effect as using the 22 factors, but fewer factors are more conducive to the wide application of the model. Therefore, in the future research, more objective methods will be adopted to select the factors that may have a dominate role in the occurrence of landslides, such as eliminating later factors in the primary selection

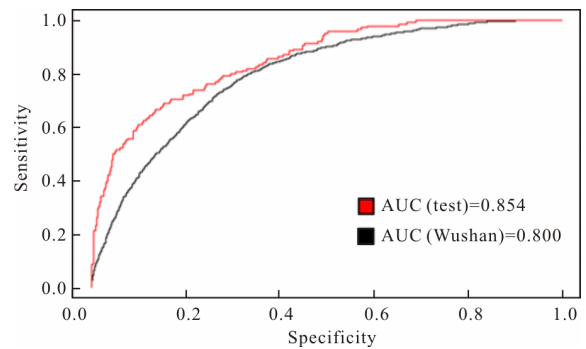


Figure 19. ROC curve of the Fengjie-based-generalizable model.

factor ranking, recursive feature elimination (Zhou et al., 2014) and so on, and a model will be constructed that cannot only deliver on good accuracy and generalization ability but is also more conducive to wide application.

5 CONCLUSIONS

This study developed an optimized RF model and explored its generalization ability not only in the area where the model is originated but also in area other than the origin of the model, with the following conclusions reached.

(1) An optimized RF model based on Fengjie landslides with best ratios of positive-to-negative cells and 10-fold cross-validation was constructed. The ratio of positive-to-negative cells

of 1 : 10 had the best prediction capability and performance, and there is no imbalance problem that the classifier is biased towards the negative cells.

(2) The accuracy, precision, and AUC value of the test dataset of the Fengjie-based model were 0.982, 0.994, and 0.876, respectively, and the LSM generated by the Fengjie-based model in Fengjie County is well consistent with the distribution pattern of actual landslides. The Fengjie-based model delivered high stability and predictive capability in Fengjie County, indicating a great generalization ability of the model to the area where the model is originated.

(3) The AUC value formed by the actual landslide value of Wushan and the predicted value of the Fengjie-based model was 0.813. The LSM in Wushan County generated by the Fengjie-based model had a reasonable reference value, indicating the Fengjie-based model had a great generalization ability in area than the origin of the model.

(4) The Fengjie-based model had such a generalization ability in area other than the origin of the model, partly because the selected counties have the same terrain, geological and hydrological conditions, and the conditioning factors have the same effect on the occurrence of landslides in both counties. Thus, it can be inferred that the Fengjie-based model in this study could be applied in other similar areas/counties with the same terrain and geological conditions.

(5) Based on Conclusion (4), a model after factor selection is reconstructed, which cannot only deliver on good accuracy and generalization ability but is also more conducive to wide application.

The result of this study is a good contribution to the landslide assessment research at the regional scale, as it could generate a LSM without collecting landslide information for modeling, so as to reduce workload and improve efficiency in practice, which has not been reported in previous studies.

ACKNOWLEDGMENTS

We want to express our gratitude to Chongqing Meteorological Administration for providing essential meteorological data and also to Chongqing Institute of Geology and Mineral Resources for offering valuable research data on historical landslides. We are also grateful to the editors and anonymous reviewers for their valuable comments on this manuscript. This study was supported by the National Natural Science Foundation of China (No. 41807498), the National Key Research and Development Program of China (No. 2018YFC1505501), and the Humanities and Social Sciences Foundation of the Ministry of Education of China (No. 20XJAZH002). The final publication is available at Springer via <https://doi.org/10.1007/s12583-020-1072-9>.

REFERENCES CITED

- Ayalew, L., Yamagishi, H., 2005. The Application of GIS-Based Logistic Regression for Landslide Susceptibility Mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology*, 65(1/2): 15–31. <https://doi.org/10.1016/j.geomorph.2004.06.010>
- Bakillah, M., Liang, S., Mobasheri, A., et al., 2014. Fine-Resolution Population Mapping Using OpenStreetMap Points-of-Interest. *International Journal of Geographical Information Science*, 28(9): 1940–1963. <https://doi.org/10.1080/13658816.2014.909045>
- Breiman, L., 2001. Random Forests. *Machine Learning*, 45(1): 5–32. <https://doi.org/10.1023/a:1010933404324>
- Can, A., Dagdelenler, G., Ercanoglu, M., et al., 2017. Landslide Susceptibility Mapping at Ovacik-Karabük (Turkey) Using Different Artificial Neural Network Models: Comparison of Training Algorithms. *Bulletin of Engineering Geology and the Environment*, 78(1): 89–102. <https://doi.org/10.1007/s10064-017-1034-3>
- Chen, W., Peng, J. B., Hong, H. Y., et al., 2018a. Landslide Susceptibility Modelling Using GIS-Based Machine Learning Techniques for Chongren County, Jiangxi Province, China. *Science of the Total Environment*, 626: 1121–1135. <https://doi.org/10.1016/j.scitotenv.2018.01.124>
- Chen, W., Yan, X. S., Zhao, Z., et al., 2018b. Spatial Prediction of Landslide Susceptibility Using Data Mining-Based Kernel Logistic Regression, Naive Bayes and RBFNetwork Models for the Long County Area (China). *Bulletin of Engineering Geology and the Environment*, 78(1): 247–266. <https://doi.org/10.1007/s10064-018-1256-z>
- Chen, W., Zhang, S., Li, R. W., et al., 2018c. Performance Evaluation of the GIS-Based Data Mining Techniques of Best-First Decision Tree, Random Forest, and Naïve Bayes Tree for Landslide Susceptibility Modeling. *Science of the Total Environment*, 644: 1006–1018. <https://doi.org/10.1016/j.scitotenv.2018.06.389>
- Das, I., Stein, A., Kerle, N., et al., 2012. Landslide Susceptibility Mapping along Road Corridors in the Indian Himalayas Using Bayesian Logistic Regression Models. *Geomorphology*, 179: 116–125. <https://doi.org/10.1016/j.geomorph.2012.08.004>
- Dou, J., Yunus, A. P., Bui, D. T., et al., 2019a. Improved Landslide Assessment Using Support Vector Machine with Bagging, Boosting, and Stacking Ensemble Machine Learning Framework in a Mountainous Watershed, Japan. *Landslides*, 17(3): 641–658. <https://doi.org/10.1007/s10346-019-01286-5>
- Dou, J., Yunus, A. P., Bui, D. T., et al., 2019b. Assessment of Advanced Random Forest and Decision Tree Algorithms for Modeling Rainfall-Induced Landslide Susceptibility in the Izu-Oshima Volcanic Island, Japan. *Science of the Total Environment*, 662: 332–346. <https://doi.org/10.1016/j.scitotenv.2019.01.221>
- Fourniadis, I. G., Liu, J. G., Mason, P. J., 2007. Landslide Hazard Assessment in the Three Gorges Area, China, Using ASTER Imagery: Wushan-Badong. *Geomorphology*, 84(1/2): 126–144. <https://doi.org/10.1016/j.geomorph.2006.07.020>
- Han, J. W., Kamber, M., 2006. Data Mining: Concepts and Techniques. *Data Mining Concepts Models Methods & Algorithms Second Edition*, 5(4): 1–18. <https://doi.org/10.1002/9781118029145.ch1>
- He, S. W., Pan, P., Dai, L., et al., 2012. Application of Kernel-Based Fisher Discriminant Analysis to Map Landslide Susceptibility in the Qinggan River Delta, Three Gorges, China. *Geomorphology*, 171/172: 30–41. <https://doi.org/10.1016/j.geomorph.2012.04.024>
- Heckmann, T., Gegg, K., Gegg, A., et al., 2014. Sample Size Matters: Investigating the Effect of Sample Size on a Logistic Regression Susceptibility Model for Debris Flows. *Natural Hazards and Earth System Sciences*, 14(2): 259–278. <https://doi.org/10.5194/nhess-14-259-2014>
- Hong, H. Y., Naghibi, S. A., Pourghasemi, H. R., et al., 2016a. GIS-Based Landslide Spatial Modeling in Ganzhou City, China. *Arabian Journal of Geosciences*, 9(2): 1–26. <https://doi.org/10.1007/s12517-015-2094-y>
- Hong, H. Y., Pourghasemi, H. R., Pourtaghi, Z. S., 2016b. Landslide Susceptibility Assessment in Lianhua County (China): A Comparison between a Random Forest Data Mining Technique and Bivariate and Multivariate Statistical Models. *Geomorphology*, 259: 105–118. <https://doi.org/10.1016/j.geomorph.2016.02.012>

- Huang, F. M., Yin, K. L., Huang, J. S., et al., 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, Y., Zhao, L., 2018. Review on Landslide Susceptibility Mapping Using Support Vector Machines. *CATENA*, 165: 520–529. <https://doi.org/10.1016/j.catena.2018.03.003>
- Hussin, H. Y., Zumpano, V., Reichenbach, P., et al., 2016. Different Landslide Sampling Strategies in a Grid-Based Bi-Variate Statistical Susceptibility Model. *Geomorphology*, 253: 508–523. <https://doi.org/10.1016/j.geomorph.2015.10.030>
- Jiang, P., Chen, J. J., 2016. Displacement Prediction of Landslide Based on Generalized Regression Neural Networks with K-Fold Cross-Validation. *Neurocomputing*, 198: 40–47. <https://doi.org/10.1016/j.neucom.2015.08.118>
- Jin, Y. F., Yin, Z. Y., Zhou, W. H., et al., 2019. Bayesian Model Selection for Sand with Generalization Ability Evaluation. *International Journal for Numerical and Analytical Methods in Geomechanics*, 43(14): 2305–2327. <https://doi.org/10.1002/nag.2979>
- Kalantar, B., Pradhan, B., Naghibi, S. A., et al., 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://doi.org/10.1080/19475705.2017.1407368>
- Lee, S., 2019. Current and Future Status of GIS-Based Landslide Susceptibility Mapping: A Literature Review. *Korean Journal of Remote Sensing*, 35: 179–193. <https://doi.org/10.7780/kjrs.2019.35.1.12>
- Li, C. D., Fu, Z. Y., Wang, Y., et al., 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: 105279. <https://doi.org/10.1016/j.enggeo.2019.105279>
- Liu, Z. H., Zhan, W. F., Lai, J. M., et al., 2019. Balancing Prediction Accuracy and Generalization Ability: A Hybrid Framework for Modelling the Annual Dynamics of Satellite-Derived Land Surface Temperatures. *ISPRS Journal of Photogrammetry and Remote Sensing*, 151: 189–206. <https://doi.org/10.1016/j.isprsjprs.2019.03.013>
- Moore, I. D., Wilson, J. P., 1992. Length-Slope Factors for the Revised Universal Soil Loss Equation: Simplified Method of Estimation. *Journal of Soil and Water Conservation*, 47(5): 423–428. <https://doi.org/10.1073/pnas.91.1.271>
- Pandey, V. K., Pourghasemi, H. R., Sharma, M. C., 2018. Landslide Susceptibility Mapping Using Maximum Entropy and Support Vector Machine Models along the Highway Corridor, Garhwal Himalaya. *Geocarto International*, 35(2): 168–187. <https://doi.org/10.1080/10106049.2018.1510038>
- Pham, B. T., Prakash, I., Khosravi, K., et al., 2018. A Comparison of Support Vector Machines and Bayesian Algorithms for Landslide Susceptibility Modelling. *Geocarto International*, 34(13): 1385–1407. <https://doi.org/10.1080/10106049.2018.1489422>
- Pourghasemi, H. R., Mohammady, M., Pradhan, B., 2012. Landslide Susceptibility Mapping Using Index of Entropy and Conditional Probability Models in GIS: Safarood Basin, Iran. *CATENA*, 97: 71–84. <https://doi.org/10.1016/j.catena.2012.05.005>
- Reichenbach, P., Rossi, M., Malamud, B. D., et al., 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., Colkesen, I., Kavzoglu, T., 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>
- Sestras, P., Bilaşco, Ş., Roşca, S., et al., 2019. Landslides Susceptibility Assessment Based on GIS Statistical Bivariate Analysis in the Hills Surrounding a Metropolitan Area. *Sustainability*, 11(5): 1362. <https://doi.org/10.3390/su11051362>
- Shirzadi, A., Bui, D. T., Pham, B. T., et al., 2017. Shallow Landslide Susceptibility Assessment Using a Novel Hybrid Intelligence Approach. *Environmental Earth Sciences*, 76(2). <https://doi.org/10.1007/s12665-016-6374-y>
- Silalahi, F. E. S., Pamela, Arifianti, Y., et al., 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>
- Sun, D. L., Wen, H. J., Wang, D. Z., et al., 2020. A Random Forest Model of Landslide Susceptibility Mapping Based on Hyperparameter Optimization Using Bayes Algorithm. *Geomorphology*, 362: 107201. <https://doi.org/10.1016/j.geomorph.2020.107201>
- Taalab, K., Cheng, T., Zhang, Y., 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. Y., Xu, C., Hong, H. Y., et al., 2018. Mapping Earthquake-Triggered Landslide Susceptibility by Use of Artificial Neural Network (ANN) Models: An Example of the 2013 Minxian (China) M_w 5.9 Event. *Geomatics, Natural Hazards and Risk*, 10(1): 1–25. <https://doi.org/10.1080/19475705.2018.1487471>
- Tian, Y. Y., Xu, C., Ma, S. Y., et al., 2019. Inventory and Spatial Distribution of Landslides Triggered by the 8th August 2017 M_w 6.5 Jiuzhaigou Earthquake, China. *Journal of Earth Science*, 30(1): 206–217. <https://doi.org/10.1007/s12583-018-0869-2>
- Tsangaratos, P., Ilia, I., Hong, H. Y., et al., 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., Sun, D. L., Wen, H. J., et al., 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(12): 4206. <https://doi.org/10.3390/ijerph17124206>
- Wang, Y. M., Wu, X. L., Chen, Z. J., et al., 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. J., Xie, P., Xiao, P., et al., 2016. Rapid Susceptibility Mapping of Earthquake-Triggered Slope Geohazards in Lushan County by Combining Remote Sensing with the AHP Model Developed for the Wenchuan Earthquake. *Bulletin of Engineering Geology and the Environment*, 76(3): 909–921. <https://doi.org/10.1007/s10064-016-0957-4>
- Wen, H. J., Wang, G. L., Huang, X. L., 2017. A Preliminary Evaluation Method of Slope Stability Based on Topographic Map and Geological Map. Chinese patent No 2017105719823 (in Chinese)
- Wu, W. Y., Xu, C., Wang, X. Q., et al., 2020. Landslides Triggered by the 3 August 2014 Ludian (China) M_w 6.2 Earthquake: An Updated Inventory and Analysis of Their Spatial Distribution. *Journal of Earth Science*, 31(4): 853–866. <https://doi.org/10.1007/s12583-020-1297-7>
- Xie, P., Wen, H. J., Ma, C., et al., 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>
- Xu, C., Xu, X. W., Dai, F. C., et al., 2012. Landslide Hazard Mapping Using GIS and Weight of Evidence Model in Qingshui River Watershed of 2008 Wenchuan Earthquake Struck Region. *Journal of Earth Science*, 23(1): 97–120. <https://doi.org/10.1007/s12583-012-0236-7>
- Yao, Y., Liu, X. P., Li, X., et al., 2017. Mapping Fine-Scale Population Distributions at the Building Level by Integrating Multisource Geospatial Big Data. *International Journal of Geographical Information Science*, 13(1): 1–25. <https://doi.org/10.1080/13658816.2017.1290252>
- Yu, L. B., Cao, Y., Zhou, C., et al., 2019. Landslide Susceptibility Mapping Combining Information Gain Ratio and Support Vector Machines: A Case Study from Wushan Segment in the Three Gorges Reservoir Area, China. *Applied Sciences*, 9(22): 4756. <https://doi.org/10.3390/app9224756>
- Zêzere, J. L., Pereira, S., Melo, R., et al., 2017. Mapping Landslide Susceptibility Using Data-Driven Methods. *Science of the Total Environment*, 589: 250–267. <https://doi.org/10.1016/j.scitotenv.2017.02.188>
- Zhang, T. Y., Han, L., Zhang, H., et al., 2019. GIS-Based Landslide Susceptibility Mapping Using Hybrid Integration Approaches of Fractal Dimension with Index of Entropy and Support Vector Machine. *Journal of Mountain Science*, 16(6): 1275–1288. <https://doi.org/10.1007/s11629-018-5337-z>
- Zhou, Q. F., Zhou, H., Zhou, Q. Q., et al., 2014. Structure Damage Detection Based on Random Forest Recursive Feature Elimination. *Mechanical Systems and Signal Processing*, 46(1): 82–90. <https://doi.org/10.1016/j.ymsp.2013.12.013>
- Zhu, A. X., Miao, Y. M., Wang, R. X., et al., 2018. A Comparative Study of an Expert Knowledge-Based Model and Two Data-Driven Models for Landslide Susceptibility Mapping. *CATENA*, 166: 317–327. <https://doi.org/10.1016/j.catena.2018.04.003>