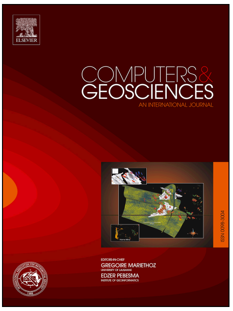

Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China

Chao Zhou, Kunlong Yin, Ying Cao, Bayes Ahmed, Yuanyao Li, Filippo Catani, Hamid Reza Pourghasemi

PII: S0098-3004(17)30484-3

DOI: 10.1016/j.cageo.2017.11.019

Reference: CAGEO 4060

To appear in: *Computers and Geosciences*

Received Date: 30 April 2017

Revised Date: 30 September 2017

Accepted Date: 22 November 2017

Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China

Chao Zhou [a, b], Kunlong Yin [a], Ying Cao [a], Bayes Ahmed [c], Yuanyao Li[d, *], Filippo Catani [b]
Hamid Reza Pourghasemi [e]

[a] Engineering Faculty, China University of Geosciences, Wuhan 430074, China
[b] Department of Earth Sciences, University of Firenze, via La Pira 4, 50121 Firenze, Italy
[c] UCL Institute for Risk and Disaster Reduction, University College London (UCL), London WC1E 6BT, UK
[d] Geological Survey, Institute of China University of Geosciences, Wuhan 430074, China
[e] Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran

Abstract: Landslide is a common natural hazard and responsible for extensive damage and losses in mountainous areas. In this study, Longju in the Three Gorges Reservoir area in China was taken as a case study for landslide susceptibility assessment in order to develop effective risk prevention and mitigation strategies. To begin, 202 landslides were identified, including 95 colluvial landslides and 107 rockfalls. Twelve landslide causal factor maps were prepared initially, and the relationship between these factors and each landslide type was analyzed using the information value model. Later, the unimportant factors were selected and eliminated using the information gain ratio technique. The landslide locations were randomly divided into two groups: 70% for training and 30% for verifying. Two machine learning models: the support vector machine (SVM) and artificial neural network (ANN), and a multivariate statistical model: the logistic regression (LR), were applied for landslide susceptibility modeling (LSM) for each type. The LSM index maps, obtained from combining the assessment results of the two landslide types, were classified into five levels. The performance of the LSMs was evaluated using the receiver operating characteristics curve and Friedman test. Results show that the elimination of noise-generating factors and the separated modeling of each landslide type have significantly increased the prediction accuracy. The machine learning models outperformed the multivariate statistical model and SVM model was found ideal for the case study area.

Keywords: Landslide susceptibility modeling; Machine learning; Support vector machine (SVM); Artificial neural network (ANN); Logistic regression (LR)

1. Introduction

Landslide is a common natural hazard in the mountainous or hilly regions. Every year, extensive economic losses and casualties are caused by landslide disasters (AGU, 2017). The Three Gorges Reservoir Area (TGRA) in

China is highly vulnerable to landslides, and the number of landslides has further increased since the construction of the Three Gorges Dam (Yin et al., 2016). Together, the demand for land is increasing due to rapid urbanization. However, the uncertainty of landslide has restricted the land-use planning in this area. Landslide susceptibility modeling (LSM) is considered as the initial step towards a landslide hazard and risk assessment, and it can also be used for land-use planning and environmental impact assessment (Fell et al., 2008). The decision-makers and engineers value it for developing strategies vis-à-vis landslide disaster risk reduction.

Landslides can be divided into many types according to different deformation mechanisms and failure patterns, and their development laws are often varied (Hungr et al., 2013). Landslide susceptibility assessment is performed based on the assumption that future landslides are more likely to occur under the similar conditions with present landslides. It is obvious that the occurrence conditions of various landslide types are different. For example, the rockfall always occurs in steep rock, while the creep landslide always occurs in soil with a gentle slope. Hereafter, in the area threatened by more than one landslide type, it is essential to conduct landslide susceptibility assessment considering the difference between landslide types.

In recent years, LSM has become a popular research topic. At regional scale, the susceptibility models can be divided into qualitative assessment (inventory-based and knowledge-driven methods) and quantitative assessment (data-driven methods and physically based models). With the improvement of data quality through innovative techniques, the data-driven models are adopted for regional LSM, including the weights-of-evidence (van Westen, 1993; Hussin et al., 2016), artificial neural network (Pradhan and Lee, 2010a; Gorsevski et al., 2016), random forest (Catani et al., 2013; Youssef et al., 2016), support vector machine (Yao et al., 2008; Pradhan, 2013) models and so on. In the data-driven models, the machine learning models performed better, and are considered more efficient than other approaches such as expert opinion based methods and analytic methods (Goetz et al., 2015; Pham et al., 2016a). The support vector machine (SVM) and artificial neural network (ANN) models were widely used in LSM and often achieved high prediction accuracy. However, no general agreement about the landslide susceptibility model exists yet, as the performance of the models requires more comparison in different cases.

Although the machine learning models have shown better performance in mathematics, the occurrence of landslides is considered as an engineering geological problem. Before conducting LSM, it is essential to understand the mechanism of landslides and analyze the relationship between causal factors and landslide occurrences (Guo et al., 2015), especially in an area that is threatened by different landslide types. The bivariate statistical and feature selection methods can quantitatively analyze the relationship between landslide occurrence and causal factors, which provide powerful techniques to analyze the landslide development laws and select the important causal factors for LSM.

In the TGRA, the impoundment and rapid urbanization caused many colluvial landslides and rockfalls (Yin et al., 2016). The previous studies did not consider the landslide types when conducting landslide susceptibility mapping (Bai et al., 2010; Wu et al., 2013). This is the originality and novel approach of this research and the authors hope that it would generate landslide susceptibility map with higher accuracy and better spatial agreement for the study area.

2. Study area

2.1 General characteristics

The study area is located in the southwest of China, the middle reaches of the TGRA, within longitude 108°30′~108°45′ east and latitude 30°30′~30°40′ north. It belongs to Chongqing and Hubei, and the total area is about 440km^2 (Fig. 1). The region is surrounded by middle and low mountains. The average annual rainfall is 1,100~1,400mm, and the monsoon season is from April to September, when the maximum monthly rainfall reaches up to 300 mm.

The strata in this study area are mainly Mesozoic, the Jurassic red layer covers most of the region, except the Triassic limestone exposed in the anticline core. The main outcropping strata in this area include the Penglaizhen Formation and the Suining Formation of upper Jurassic (J3p and J3s), the upper and lower Shaximiao Formation and the Xintiangou Formation of middle Jurassic (J2s, J2xs, and J2x), and the Badong Formation of middle Triassic (T2b).

The completion of the Three Gorges Dam increased the engineering activities, such as the highway construction and urban reconstruction. The geological environment was seriously damaged by large-scale excavations in the construction site and indiscriminate slope cutting etc. The main human engineering activities in this area contain the new urbanization in rural areas and the construction of traffic routes, such as the G318 national highway and so on.

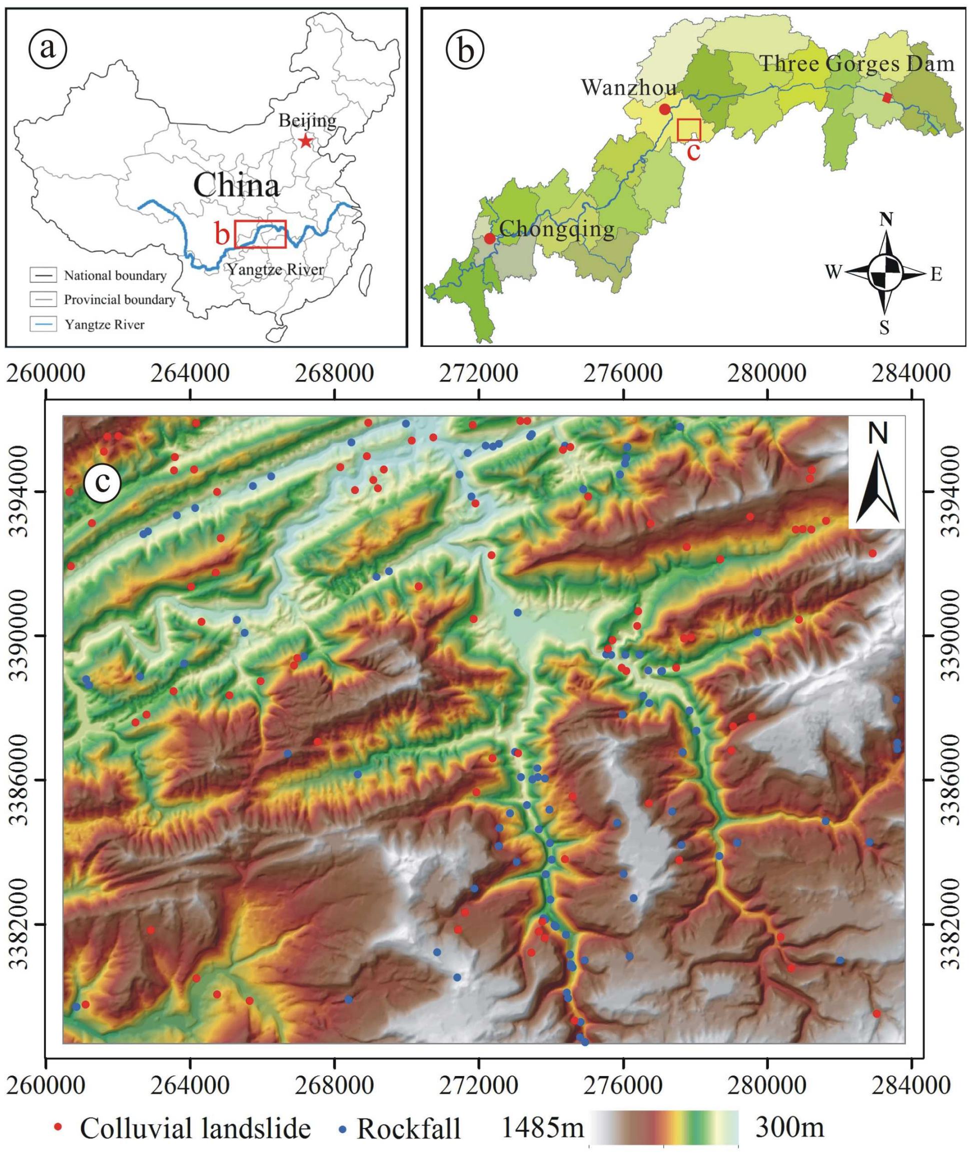

Fig. 1 (a) Site map of the TGRA in China, **(b)** Location of the study area in the TGRA, and **(c)** the digital elevation model (DEM) showing the landslide locations

2.2 Landslide types

The occurrence of landslide is affected by various conditions. Due to regional setup and local context, different landslide types always developed. Two landslide types have been identified in the study area:

Colluvial landslide: The colluvial landslides (Varnes, 1978; Hungr et al., 2014) with small or medium-size developed a lot in the study area (Fig. 2a). The rainfall and reservoir level fluctuation provided external triggering factors for the occurrence of colluvial landslide. The rainfall increases the sliding force of landslide mass, while the reservoir level fluctuation reduces the sliding resistance force, the combined effort of which may decrease landslide stability and improve the occurrence probability.

Rockfall: The rockfall (Varnes, 1978; Hungr et al., 2014) is another main landslide type and often developed in a multi-stage pattern (Fig. 2b). In the abrupt cliff, because of the developed large structural joints, large-scale rockfall often occur. In the gentle slope, there are many human engineering activities, such as road construction. The slope may lose the original equilibrium state under the influence of artificial cutting slope, which could induce the occurrence of small-scale rockfall.

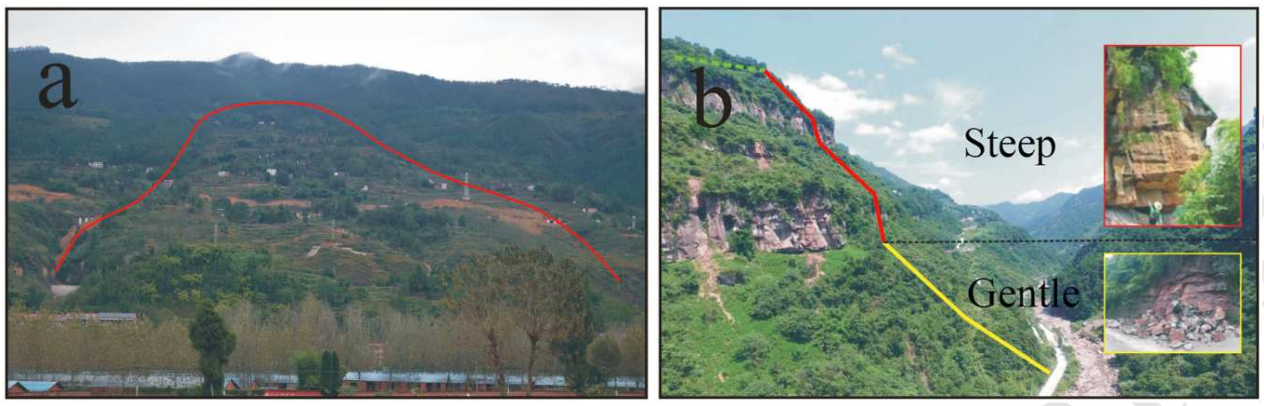

Fig. 2 Landslide types: **(a)** Colluvial landslide, **(b)** Multi-stage rockfall

3. Methodology

3.1 Landslide causal factors analysis

3.1.1 Information value model

The information value model (Yin and Yan, 1988) is based on the concept that landslide occurrence (y) is affected by various factors (x_i), and their influences to landslides are different. According to a conditional probability, the formula for the information value can be written as:

$$\mathrm{I}(y,x_i) = \mathrm{Log}_2 \frac{\mathrm{P}(y,x_i)}{\mathrm{P}(y)} \tag{1}$$

Where $\mathrm{I}(y,x_i)$ is the information value under the causal factors x_i; $\mathrm{P}(y)$ is the probability of landslide occurrence; $\mathrm{P}(y,x_i)$ is the probability of the occurrence of landslide under the causal factor x_i. The probability can be calculated using the area ratio as well. The formula (1) can be expressed as:

$$\mathrm{I}(y,x_i) = \mathrm{Log}_2 \frac{\mathrm{S}_0^{\mathrm{i}} / \mathrm{S}}{\mathrm{A}_0^{\mathrm{i}} / \mathrm{A}} \tag{2}$$

Where S is the total area of the landslide; S_0^i is the landslide area under the factor x_i; A is the total area of the study area; A_0^i is the area under the factor x_i. It is worth to highlight that a positive value of $\mathrm{I}(y,x_i)$

indicates factor x_i plays a promotion influence for landslide occurrence. In contrast, a negative value of $I(y, x_i)$ indicates factor x_i plays an inhibitive effect on landslide occurrence.

3.1.2 Information gain ratio

Information gain ratio (IGR) is one of the most efficient feature selection methods (Quinlan, 1993; Tien Bui et al., 2016). The factors with a higher value of IGR indicate a higher prediction ability of the models. Assuming that the training data T consists of n samples, and belongs to the class Ci (landslide, non landslide). Then, the information entropy can be calculated as:

$$Info(T) = -\sum_{i=1}^{2} \frac{n(Ci,T)}{|T|} \log_2 \frac{n(Ci,T)}{|T|} \tag{3}$$

The amount of information $(T_1, T_2 \cdots T_m)$ split from T regarding the causal factor F is estimated as:

$$Info(T,F) = -\sum_{j=1}^{m} \frac{T_j}{|T|} \log_2 Info(T) \tag{4}$$

Then, the IGR of the landslide causal factor F can be written as follows:

$$IGR(T,F) = \frac{Info(T) - Info(T,F)}{SplitInfo(T,F)} \tag{5}$$

Where $SplitInfo$ represents the potential information generated by dividing the training data T into m subsets. The formula of $SplitInfo$ was shown as follows:

$$SplitInfo(T,F) = -\sum_{j=1}^{m} \frac{|Tj|}{|T|} \log 2 \frac{|Tj|}{|T|} \tag{6}$$

3.2 Landslide susceptibility modeling

3.2.1 Support vector machine

Support vector machine (Vapnik, 1995) is a nonlinear classification method, which is based on the principle of Vapnik-Chervonenkis Dimension and structural risk minimization. The input variables in the original space are mapped into a high-dimensional linear feature space by nonlinear transformation. Then, in order to split the positive from the negative, SVM model operates by attempting to find an optimal surface in the feature space between the two types (Zhou et al., 2016). Assuming samples $(x_i, y_i): i = 1, 2 \cdots n$, the optimal hyperplane can be solved by the following function:

$$\begin{cases} \boldsymbol{Min}(\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i) \\ y_i(\vec{w}\cdot\vec{x}_i + b) - 1 + \xi_i \geq 0 \\ \xi_i \geq 0, i = 1, 2\cdots, n \end{cases} \tag{7}$$

Where w is the weight vector that determines the orientation of the hyperplane, b is the bias, ξ_i is the positive slack variables for the data points that allow for penalized constraint violation, C is the penalty parameter that controls the trade-off between the complexity of the decision function and the number of training examples misclassified. The function can be converted into an equivalent dual problem based on the Wolf duality theory:

$$\begin{cases} \boldsymbol{Max}(\sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j(\vec{x_i}\cdot\vec{x_j})) \\ \sum_i \alpha_i y_i = 0, 0 \leq \alpha_i \leq C \end{cases} \tag{8}$$

Where α_i are Lagrange multipliers, C is the penalty. Then, the decision function, which will be used for the classification of new data, can be written:

$$f(x) = \boldsymbol{sgn}(\sum_{i=1}^{n} y_i\alpha_i\boldsymbol{K}(x_i, x_j) + b) \tag{9}$$

Where $\boldsymbol{K}(x_i, x_j)$ is the kernel function. The radial basis kernel was adopted as kernel function for SVM model in this study.

3.2.2 Artificial neural networks

Artificial neural network is a reasoning model established on the imitation of human brain function and nervous system. Back propagation neural network (BPNN) (Hecht-Nielsen, 1988) is one of the most effective ANNs, it is a multilayer neural network consisting of an input layer, hidden layers, and an output layer (Fig. 3). In signal propagation, the input signal is processed layer by layer from the input to the output. If the result of the output layer is not expected, it would be transferred to the reverse propagation, and adjust to the network weights and thresholds according to the prediction error to approximate the desired output.

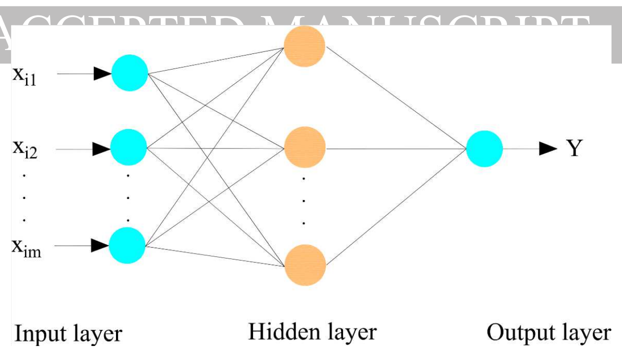

Fig. 3 The architecture of a three layers BPNN

The learning rate is an important parameter of ANN model, which may affect its performance. In this study, the learning rate will be automatically calculated using the following formula:

$$\eta(n) = \eta(n-1) * \exp(\log(\eta_{\min} / \eta_{\max}) / d) \tag{10}$$

Where $\eta(n)$ is the learning rate in the n*th* times training; $\eta_{\min}$ is the minimum value of the learning rate; $\eta_{\max}$ is the maximum value of the learning rate, and d is the delay rate. In this study, the initial rate, the maximum and minimum learning rate, and the delay rate are 0.3, 0.1, 0.01 and 30, respectively.

3.2.3 Logistic regression

Logistic regression (LR) (Cox, 1958) is a multivariate statistical method for landslide susceptibility mapping (Budimir et al., 2015). LR can reveal the relationship between a target variable and multiple predictor variables, and predict the occurring probability of a certain event. In a statistical analysis of LR, the predictor variables can be either continuous or discrete, and there is no need to meet the normal distribution. The formula of LR is as follows:

$$y = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \tag{11}$$

Where α is a constant, n is the number of independent variables, $x_i (i=1,2,\cdots,n)$ is the predictor variables and $\beta_i (i=1,2,\cdots,n)$ is the coefficient of the LR.

4. Data preparation and analysis

4.1 Landslide inventory

Landslide inventory is the basis for landslide susceptibility mapping. An accurate and reliable landslide inventory data is crucial for LSM (Corominas et al., 2013; Zhu et al., 2014). According to the Chinese National Standard of Specification for landslide survey and risk assessment (http://www.caghp.org/standard.php), the landslide inventory map was prepared by incorporating and analyzing high-resolution remote sensing images of Pleiades-1 (9/22/2014) and GF-1 (3/30/2015), through field investigation, and the historical landslide data. A total of 202 landslides were identified which contains 95 colluvial landslides and 107 rockfalls (Fig. 1c). The total area of the colluvial landslides was calculated as 3.35km^2, while the area of individual colluvial landslide ranges from 7.1m^2 to 0.24km^2. The total area of rockfall is 0.28 km^2 and the area of individual rockfall ranges from 1.93m^2 to 0.04km^2. The colluvial landslide and rockfall are dominant in the study area. Both types are sensitive to different engineering geology conditions, which causes the differences of their development laws. In this study, the colluvial landslide and rockfall were analyzed and assessed separately, and the final landslide susceptibility map was obtained by combining them (Fig. 4).

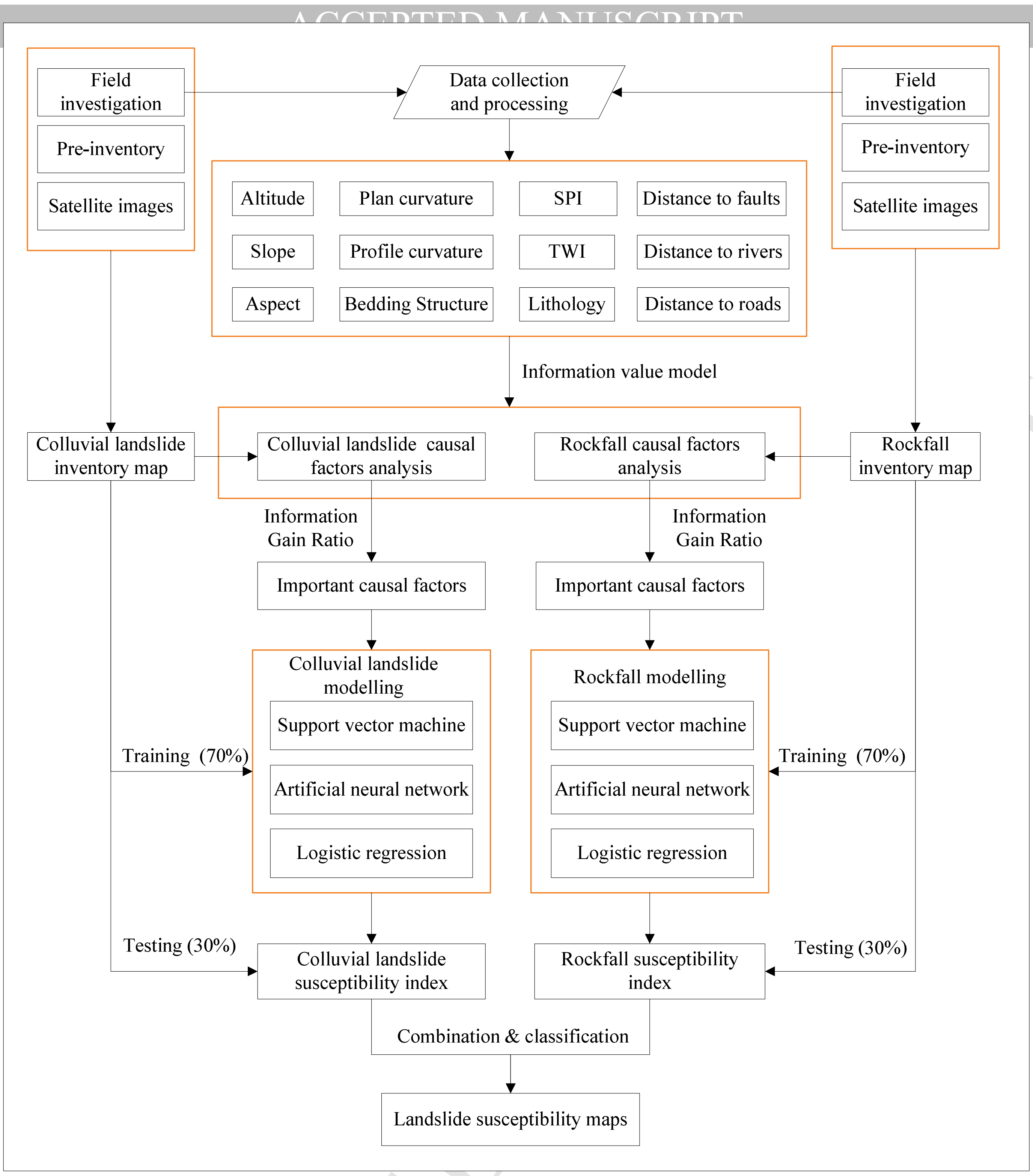

Fig. 4 The flowchart of the landslide susceptibility assessment

4.2 Landslide causal factors

Landslide hazard is caused by the interaction between the internal geological conditions of slope and the external environmental factors. Based on field investigation, data analysis, and previous researches (Wu et al., 2013; Peng et al., 2014), twelve factors were prepared initially for landslide susceptibility assessment: altitude, slope, aspect, plan curvature, profile curvature, stream power index (SPI), topographic wetness index (TWI), lithology, bedding structure, and distance to faults, rivers, and roads. The relationship between landslide occurrences and causal factors was analyzed quantitatively using the information value model. Moreover, in order to classify the continuous causal factors (altitude, slope, and so on) reasonably, they were discretized into small intervals first, and then three kinds of curves were obtained by statistics, namely the distribution curve of the whole area, the distribution curve of the landslide area, and the curve of information value. Finally, the continuous causal factors were classified by the breakpoints of the three kinds of curves (Zhou et al., 2015).

Topographic factors

The topographic factors used in this study were prepared using a digital elevation model (DEM) with a spatial resolution of 25m, which was collected from China Geological Survey. Subsequently, six topographic factors (altitude, slope, aspect, plan curvature, profile curvature, SPI and TWI) were extracted in ArcGIS 10.0 using the mentioned DEM.

Altitude

The altitude range in this area is 300m~1,300m, which was divided into five classes: [300~450), [450~700), [700~950), [950~1,100), [1,100~1,500) (Fig.5a). The colluvial landslide mainly occurred in the altitude range of 450~700m and 700m~950m, and their information values are 0.086 and 0.303, respectively (Table 1). The rockfall mainly occurred in the altitude from 300m to 950m, the altitude ranges of [300, 450) and [450, 750) have the largest information values of 1.196 and 0.741, respectively.

Slope

The slope was divided into five classes: very gentle (0~6°), gentle (6~18°), moderate (18~30°), steep (30~39°), and very steep (>39°) (Fig. 5b). The colluvial landslide mainly occur in the gentle and moderate slope, and the moderate slope shows the highest promotion influence on it, whose information value is 0.911 (Table 1). Different effect shows on rockfall, which is more prone to occur in the steep and very steep slopes, and their information values are 0.970 and 1.432, respectively.

Table 1 Spatial relationship between causal factors and landslides

Causal factor	Category	Percentage of domain	Percentage of CL	IV of CL	Normalized class of CL	Percentage of rockfall	IV of rockfall	Normalized class of rockfall
Altitude (m)	<450	6.39	4.03	-1.231	0.010	14.64	1.196	0.990
	450-700	23.38	28.11	0.086	0.745	39.07	0.741	0.750
	700-950	30.33	48.11	0.303	0.990	23.40	-0.374	0.500
	950-1100	20.24	16.03	-0.214	0.500	15.15	-0.418	0.260
	>1100	19.66	3.73	-0.799	0.255	7.73	-1.346	0.010
Slope (°)	< 6	9.46	10.14	0.666	0.745	2.07	-2.190	0.010
	6 - 18	26.47	43.95	0.911	0.990	7.37	-1.844	0.255
	18 - 30	38.99	23.25	-0.383	0.255	36.64	-0.090	0.500
	30 - 39	18.59	20.36	0.008	0.500	36.41	0.970	0.745
	> 39	6.49	2.30	-3.097	0.010	17.51	1.432	0.990
Aspect	Flat	3.10	0.56	-2.463	0.010	0.23	-3.777	0.010
	N	11.86	13.08	0.142	0.745	14.93	0.333	0.623
	NE	8.38	3.57	-1.231	0.133	18.78	1.164	0.990
	E	9.25	5.55	-0.735	0.255	6.33	-0.546	0.255
	SE	14.90	21.76	0.547	0.990	3.17	-2.234	0.133
	S	11.22	8.21	-0.451	0.500	14.48	0.368	0.745
	SW	9.22	5.57	-0.727	0.378	14.03	0.605	0.868
	W	16.91	19.34	0.193	0.868	14.93	-0.179	0.500
	NW	15.17	15.68	0.048	0.623	13.12	-0.209	0.378
Plan curvature	Concave	26.71	21.08	-0.342	0.010	32.58	0.287	0.500
	Flat	45.60	44.07	-0.049	0.500	29.64	-0.622	0.010
	Convex	27.69	28.18	0.025	0.990	37.78	0.448	0.990
Profile curvature	Concave	26.42	26.24	-0.010	0.500	31.22	0.241	0.500
	Flat	41.82	43.00	0.040	0.990	31.00	-0.432	0.010
	Convex	31.76	24.08	-0.399	0.010	37.78	0.251	0.990
SPI	0 - 2	32.40	22.99	-0.495	0.010	38.01	0.230	0.990
	2 - 4	42.81	41.84	-0.033	0.663	40.27	-0.088	0.337
	4 - 8	12.39	19.18	0.631	0.990	12.22	-0.020	0.663
	> 8	12.41	9.32	-0.414	0.337	9.50	-0.385	0.010
TWI	0 - 4.5	61.17	42.89	-0.512	0.010	42.76	-0.517	0.010
	4.5 - 6.5	14.62	17.72	0.277	0.337	21.72	0.571	0.663
	6.5 - 8	10.88	15.26	0.488	0.990	10.41	-0.064	0.337
	> 8	13.32	17.45	0.390	0.663	25.11	0.914	0.990
Distance to rivers/m	0 - 200	27.55	38.19	0.470	0.990	31.24	0.182	0.990
	200 - 500	32.20	29.47	-0.130	0.663	32.28	0.003	0.663
	500-1000	35.29	30.00	-0.240	0.337	33.88	-0.059	0.337
	> 1100	4.97	2.34	-1.090	0.010	2.60	-0.934	0.010
Distance to roads/m	0 - 50	30.91	43.14	0.480	0.990	67.32	1.123	0.990
	50 - 150	35.90	34.92	-0.040	0.663	27.42	-0.388	0.663
	150 - 400	25.27	19.04	-0.410	0.337	2.68	-3.237	0.010
	> 400	7.93	2.90	-1.450	0.010	2.58	-1.622	0.337

Causal factor	Category	Percentage of domain	Percentage of CL	IV of CL	Normalized class of CL	Percentage of rockfall	IV of rockfall	Normalized class of rockfall
Distance to faults/m	0 - 200	5.81	11.40	0.970	0.990	12.47	1.102	0.990
	200 - 400	5.71	7.99	0.480	0.663	6.70	0.230	0.663
	400 - 800	11.33	12.81	0.180	0.337	7.53	-0.590	0.337
	> 800	77.15	67.79	-0.190	0.010	73.30	-0.074	0.010
Lithology	A	9.78	15.49	0.663	0.794	8.80	-0.151	0.598
	B	6.62	4.43	-0.581	0.206	0.23	-4.875	0.010
	C	14.22	17.21	0.275	0.598	2.93	-2.277	0.206
	D	24.36	25.65	0.074	0.402	6.32	-1.946	0.402
	E	39.01	19.34	-1.012	0.010	73.14	0.907	0.990
	F	6.01	11.21	0.899	0.990	8.58	0.513	0.794
Bedding structure	BS1	57.14	27.07	-1.080	0.010	83.61	0.549	0.990
	BS2	0.71	0.95	0.420	0.500	0.01	-6.105	0.010
	BS3	7.93	26.98	1.770	0.990	1.13	-2.806	0.173
	BS4	5.86	7.49	0.360	0.337	1.13	-2.369	0.337
	BS5	9.31	9.52	0.030	0.173	4.23	-1.139	0.663
	BS6	6.90	9.90	0.520	0.663	1.44	-2.258	0.500
	BS7	12.16	18.09	0.570	0.827	8.45	-0.524	0.827

Note: CL means Colluvial landslide, and IV means Information value.

Aspect

The aspect was divided into nine categories (Fig. 5c). The colluvial landslides on the southeast aspect represent the highest occurrence probability with an information value of 0.547. The rockfalls on the northeast aspect are the easiest to occur, its information value is the highest of 1.164. Because of the inhibition effect on slope movement, the information value of flat terrain are the least in both the landslide types (Table 1).

Plan curvature

The plan curvature varies within the range of -14.0~7.9, and the slope pattern was divided into convex, flat, and concave (Fig. 5d). The convex slope has slightly promotion effect on colluvial landslide; its information value is 0.025 (Table 1). For rockfall, the flat curvature shows slightly inhibition effect and the information value is -0.662. The information values of concave and convex curvature are 0.287 and 0.448, respectively.

Profile curvature

The profile curvature varies within the range of -12.9~13.3. The slope pattern was divided into convex, flat, and concave as well (Fig. 5e). As shown in Table 1, the profile curvature has slight influence on the occurrence of both colluvial landslide and rockfall. The flat slope has the highest information value of 0.004 for the colluvial landslide, while the convex slope has the highest information value of 0.251 for rockfall.

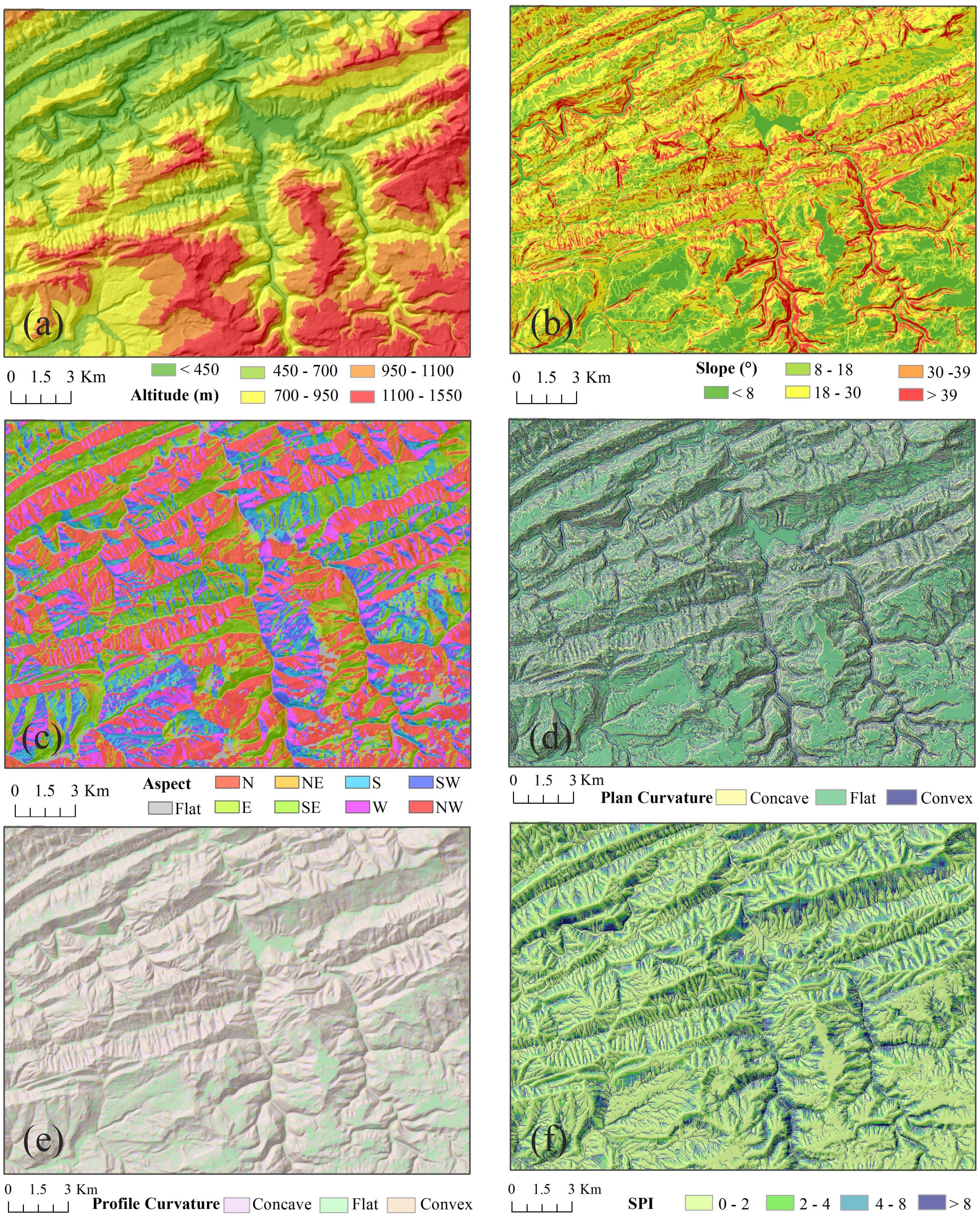

Fig. 5 Landslide causal factors of the study area: **a** altitude, **b** slope, **c** aspect, **d** plan curvature, **e** profile curvature, **f** SPI, **g** TWI, **h** lithology, **i** bedding structure, **j** distance to faults, **k** distance to rivers and **l** distance to roads.

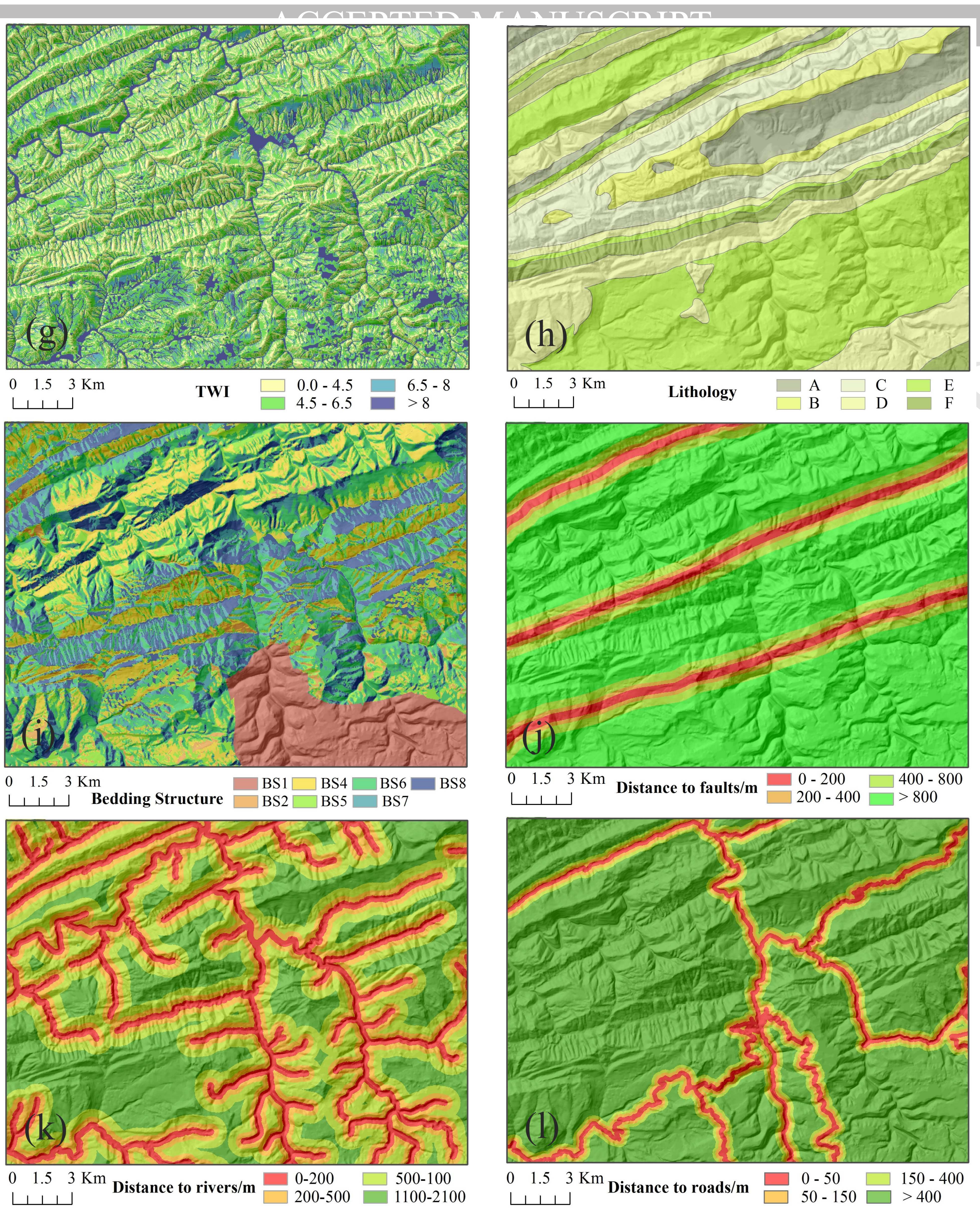

Fig. 5 (continued).

SPI and TWI

The SPI and TWI are commonly used to quantify topographic influence on hydrological processes (Moore et al., 1991). In this study, the value of SPI was classified into four categories: 0-2, 2-4, 4-8, and >8 (Fig. 5f), while the value of TWI was divided into four classes of 0-4.5, 4.5-6.5, 6.5-8, and >8 (Fig. 5g). The positive and negative influence of SPI and TWI are slight, all the information values are relatively smaller (Table 1).

Lithology

The main outcropping strata of the study area include Badong Formation, the upper and lower Shaximiao Formation of middle Jurassic and so on. The lithology was extracted from the geological map (Fig. 5h) and grouped into six categories (Table 2). The category F shows the strongest positive influence on colluvial landslide with the largest information value of 0.899. More than 70% of rockfalls occurred in category E, and its information value is the largest of 0.907 (Table 1).

Table 2 Lithological classification in the study area

Category	Main lithology	Geologic group
A	Muddy limestone	T_2b^1, T_2b^3
B	Lithic sandstone	T_3xj
C	Sandstone, mudstone (shale)	J_1z, J_2x
D	Mudstone, pelitic siltstone with sandstone	J_2xs^1, J_2s^2, J_3s
E	Lithic sandstone with mudstone	J_2xs^2, J_2s^3
F	Interbeds of mudstone and sandstone	J_2s^1

Bedding structure

Bedding structure indicates the intersection relationship between strata and slope, its classification is shown in Table 3, In this study area, the colluvial landslide mostly occurred in the under-dip slope and horizontal strata slope (Fig. 5i), and the under-dip slope got the maximum information values of 1.770. Because of rock outcropping and its developed vertical fissure (Fig. 2b), more than 80% of rockfalls are distributed in the horizontal strata slope, whose information value is the highest of 0.549 (Table 1).

Table 3 Classification of bedding structure

Category	Type of Bedding Structure	Definition(Slope: θ ,Aspect: σ ,bed dip angle: α ,bed dip direction: β)
BD1	Horizontal strata slope	$\alpha \leq 10°$
BD2	Over-dip slope	$((\lvert\sigma - \beta\rvert \in (0, 30°]) \square (\lvert\sigma - \beta\rvert \in [330°, 360°))) \& \& (\alpha > 10°) \& \& (\theta > \alpha)$
BD3	Under-dip slope	$((\lvert\sigma - \beta\rvert \in (0, 30°]) \square (\lvert\sigma - \beta\rvert \in [330°, 360°))) \& \& (\alpha > 10°) \& \& (\theta < \alpha)$
BD4	Dip-oblique slope	$(\lvert\sigma - \beta\rvert \in [30°, 60°)) \square (\lvert\sigma - \beta\rvert \in [300°, 330°))$
BD5	Transverse slope	$(\lvert\sigma - \beta\rvert \in [60°, 120°)) \square (\lvert\sigma - \beta\rvert \in [240°, 300°))$
BD6	Anaclinal oblique slope	$(\lvert\sigma - \beta\rvert \in [120°, 150°)) \square (\lvert\sigma - \beta\rvert \in [210°, 240°))$
BD7	Anaclinal slope	$(\lvert\sigma - \beta\rvert \in [150°, 180°)) \square (\lvert\sigma - \beta\rvert \in [180°, 210°))$

Distance to faults

The proximity parameters (distance to faults, rivers and roads) were calculated using geological and geomorphology maps based on the Euclidean distance method in ArcGIS 10.0. The faults in the study area is relatively simple (Fig. 5j), most of the landslides occurred far away from the faults. Within the influence area, the faults show a more positive effect on landslide occurrence. When the distance to faults is smaller than 200m, the information values for colluvial landslide and rockfall are the maximum of 0.970 and 1.102, respectively (Table 1).

Distance to rivers

The distance to rivers was divided into four classes, namely 0~200m, 200~500m, 500~1,100m, and >1,100m (Fig. 5k). In the study area, 38% of the colluvial landslides are distributed within the range of 200m from rivers, its information value is the maximum of 0.471. There are few colluvial landslides when the distance is greater than 1,100m, whose information value is the minimum of -1.090. The rivers show a slight effect on rockfall, when the distance to rivers is less than 200m, the information value is the highest of 0.182 (Table 1).

Distance to roads

The distance to roads was classified into four categories, namely 0~50m, 50m~150m, 150m~300m, and >300m (Fig. 5l). In the study area, 43% of colluvial landslides are distributed within the range of 50m from roads and the information value is the highest of 0.480. The road has a strong influence on rockfall, because the cutting slope was caused by road construction (Fig. 2b), 67% of rockfalls are distributed within the range of 50m from roads and the information value is the maximum of 1.123. Only 2.58% of rockfalls occurred when the distance to roads is more than 400m, its information value is the minimum of -1.451.

5. Results and analysis

5.1 Landslides susceptibility mapping

5.1.1 Data preparation and multicollinearity analysis

The machine learning models are more sensitive to data in their desired range. Consequently, the landslide causal factors were normalized into the range of [0.01, 0.99] according to the information values (Table 1). The normalized data of the factors were taken as input data, and the landslide susceptibility index (landslide:1, non landslide:0) was taken as output data. 70% of colluvial landslide and rockfall locations were randomly selected as the training samples, and the remaining 30% were used to evaluate the performance of the models. Furthermore, the negative data (non colluvial landslide, non rockfall) and positive data (colluvial landslide, rockfall) were considered equally important in LSM. The same number of negative data was randomly selected from the landslide free area (Felicísimo et al., 2013), its distribution is shown in Fig. 6.

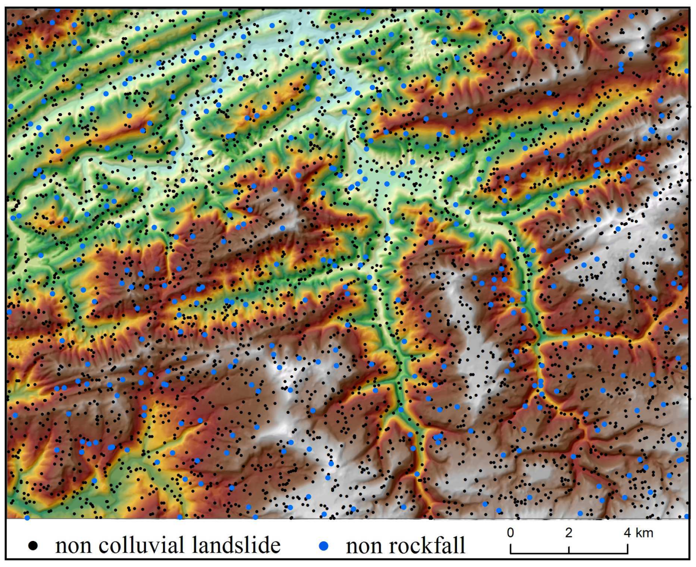

Fig. 6 The distribution of non landslide samples

Multicollinearity among the factors may influence the accuracy of the susceptibility models. The Variance inflation factors (VIF) and Tolerances were applied to test the multicollinearity among the twelve factors, a Tolerance of less than 0.2 or a VIF of 5 and above indicates a multicollinearity problem (O’Brien, 2007). As shown in Table 4, the smallest tolerance in the colluvial landslide and rockfall assessment are 0.741 and 0.702, respectively, the highest VIF of them are 1.350 and 1.425, respectively. No multicollinearity was found between the causal factors.

Table 4 Multicollinearity of the causal factors

Factor	Colluvial landslide		Rockfall	
	VIF	Tolerances	VIF	Tolerances
Altitude	1.406	0.711	1.235	0.810
Slope	1.097	0.912	1.112	0.899
Aspect	1.024	0.977	1.054	0.949
Plan curvature	1.097	0.911	1.055	0.948
Profile curvature	1.152	0.868	1.180	0.847
SPI	1.315	0.761	1.276	0.783
TWI	1.452	0.702	1.350	0.741
Lithology	1.114	0.898	1.107	0.903
Bedding structure	1.122	0.891	1.094	0.914
Distance to faults	1.040	0.961	1.045	0.957
Distance to rivers	1.390	0.720	1.290	0.775
Distance to roads	1.063	0.940	1.158	0.864

5.1.2 Selection and elimination of the less important causal factors

Twelve factors were initially prepared and considered as landslide causal factors, the factors often show different contribution for susceptibility modeling. The IGR technique was used to quantitatively assess the importance of each factor. The average merit of each factor is shown in Fig. 7. The causal factors with higher average merit values are more important. The results indicate that the distance to roads is the dominant factor for rockfall with an highest average merit value of 0.109. The altitude with the average merit of 0.023 is the most important factor for colluvial landslide (Fig. 7).

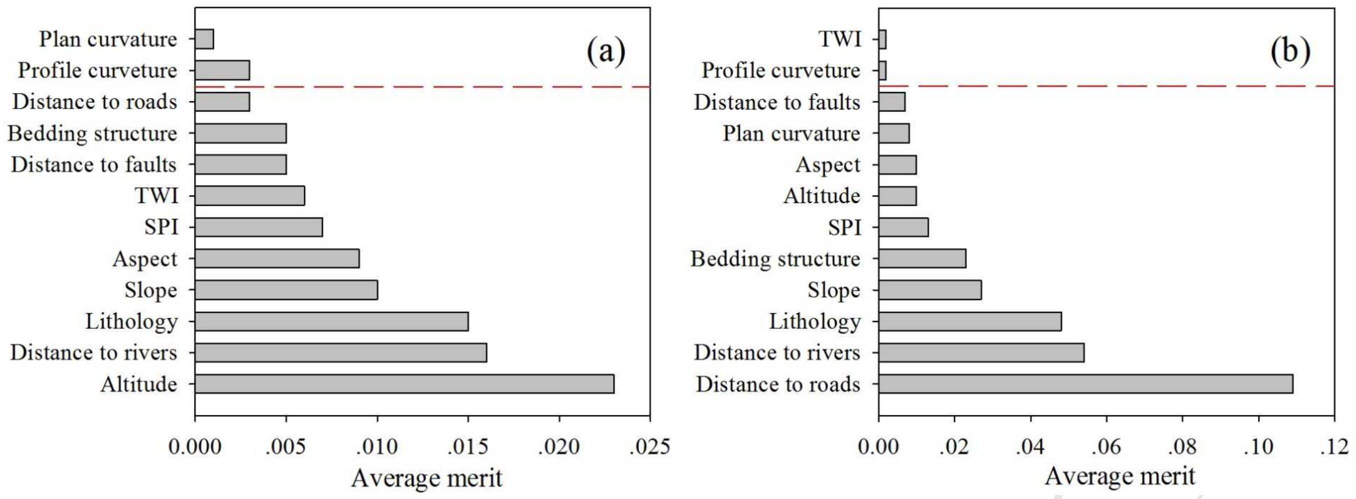

Fig. 7 The average merit of each causal factor in (a) colluvial landslide (b) rockfall

Although all the selected factors are relevant to landslides, but it is proved that the less important factors may cause noise and reduce the prediction accuracy (Pradhan and Lee, 2010b; Pham et al., 2016a). In order to find the most effective combination of the causal factors, the factors were eliminated one by one starting from the least important factor, and the SVM was used to test their prediction accuracy. As shown in Table 5, the accuracy of both the colluvial landslide and rockfall modeling increased when the less important factors were eliminated. The highest performance was achieved when the two least important factors were removed. Thus, the plan and profile curvatures were removed in colluvial landslide modeling, while the TWI and profile curvature were eliminated in rockfall modeling (Table 5).

Table 5 The prediction accuracy with elimination of the less important factors

Model	Eliminating unimportant factors	AUC
Colluvial landslide		
Model-1	Without eliminating any factor	0.893
Model-2	Plan curvature	0.901
Model-3	Plan curvature, Profile curvature	**0.912**
Model-4	Plan curvature, Profile curvature, Distance to roads	0.902
Rockfall		
Model-5	without eliminating any factor	0.902
Model-6	TWI	0.911
Model-7	TWI, Profile curvature	**0.932**
Model-8	TWI, Profile curvature, Distance to faults	0.906

5.1.3 Landslide susceptibility modeling

The machine learning models of SVM and ANN and the multivariate statistical model of LR were applied to assess the susceptibility of colluvial landslide and rockfall, respectively; the modeling processing was carried out in Clementine 12. As stated in Section 5.1.2, ten important factors, namely altitude, slope, aspect, TWI, SPI, lithology, bedding structure, distance to rivers, faults and roads were selected to establish the colluvial landslide model. Meanwhile, altitude, slope, aspect, plan curvature, SPI, lithology, bedding structure, distance to rivers, faults and roads were selected as inputs of rockfall modeling.

In this study, the parameters of SVM and ANN were obtained by error and trial method, which is shown in Table 6. Regarding ANN, the four layers ANN was adopted, and its learning rate was calculated automatically by formula (10). In the modeling of LR, the logistic regression equation of colluvial landslide index (CLI) and rockfall index (RI) are shown as follows:

$$\begin{aligned}\text{CLI} = {} & ((-4.843) + (\text{Distance to roads} * (-0.398)) + (\text{Lithology} * 1.553) \\ & + (\text{SPI} * 0.249) + (\text{Aspect} * 1.407) + (\text{Distance to faults} * 0.096) \\ & + (\text{Bedding structure} * (-0.180)) + (\text{Distance to rivers} * 1.384) \\ & + (\text{TWI} * 0.704) + (\text{Altitude} * 0.696) + (\text{Slope} * 1.056)\end{aligned} \tag{12}$$

$$\begin{aligned}\text{RI} = {} & ((-7.628) + (\text{Distance to roads} * 2.544) + (\text{Plan curvature} * 0.200) \\ & + (\text{Bedding structure} * 0.855) + (\text{Aspect} * 1.124) + (\text{SPI} * 0.642) \\ & + (\text{Distance to faults} * (-2.247)) + (\text{Distance to rivers} * 1.494) \\ & + (\text{Lithology} * 2.979) + (\text{Slope} * 1.200) + (\text{Altitude} * 0.628))\end{aligned} \tag{13}$$

Table 6 The parameters of SVM and ANN models

Models	Parameters	Notes
SVM of colluvial landslide	$c = 20, \gamma = 1.5$	c is the penalty factor, γ is the parameter of the kernel function;
SVM of rockfall	$c = 20, \gamma = 1.5$	
ANN of colluvial landslide	$n_1 = 80, n_2 = 30, \alpha = 0.9$	n_1, n_2 are the neurons number of the 1st and 2nd hidden layers respectively, α is the momentum.
ANN of rockfall	$n_1 = 70, n_2 = 30, \alpha = 0.9$	

The colluvial landslide and rockfall susceptibility index were calculated applying the SVM, ANN and LR models respectively, the results are shown in Fig. 8. Then, the final landslide susceptibility index was obtained by selecting the larger value of each pixel between the colluvial landslide and rockfall susceptibility index. At last, the landslide susceptibility index was divided into five levels: Very High (10%), High (20%), Moderate (20%), Low (20%) and Very Low (30%), which is shown in Fig. 9. Furthermore, in order to verify the significance of landslide classification, the susceptibility modeling without landslide classification was conducted using the three models as well, and the parameters of machine learning models are same to the colluvial landslide modeling (Table 6).

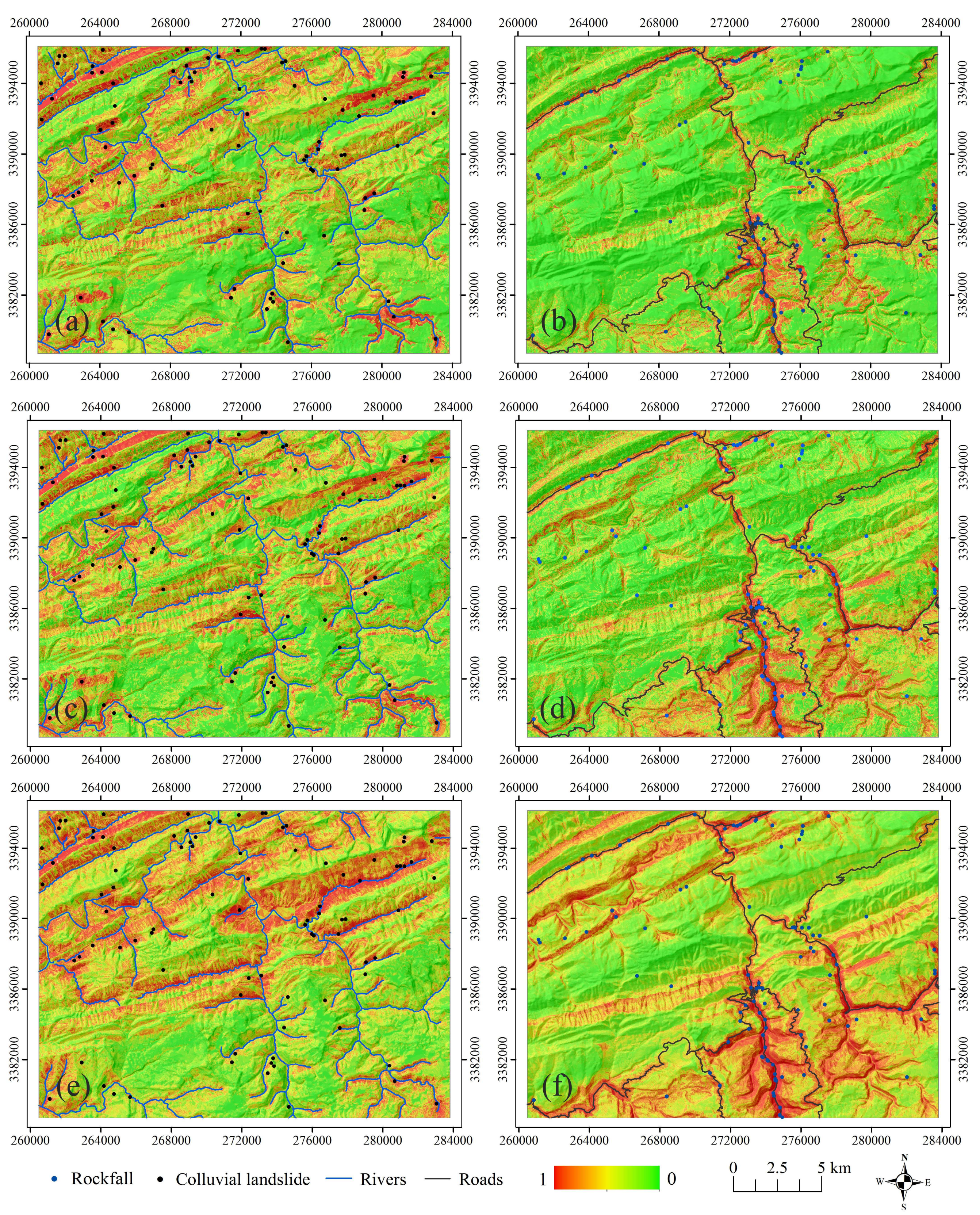

Fig. 8 Susceptibility index of **(a)** colluvial landslide using SVM, **(b)** rockfall using SVM, **(c)** colluvial landslide using ANN, **(d)** rockfall using ANN, **(e)** colluvial landslide using LR and **(f)** rockfall using LR

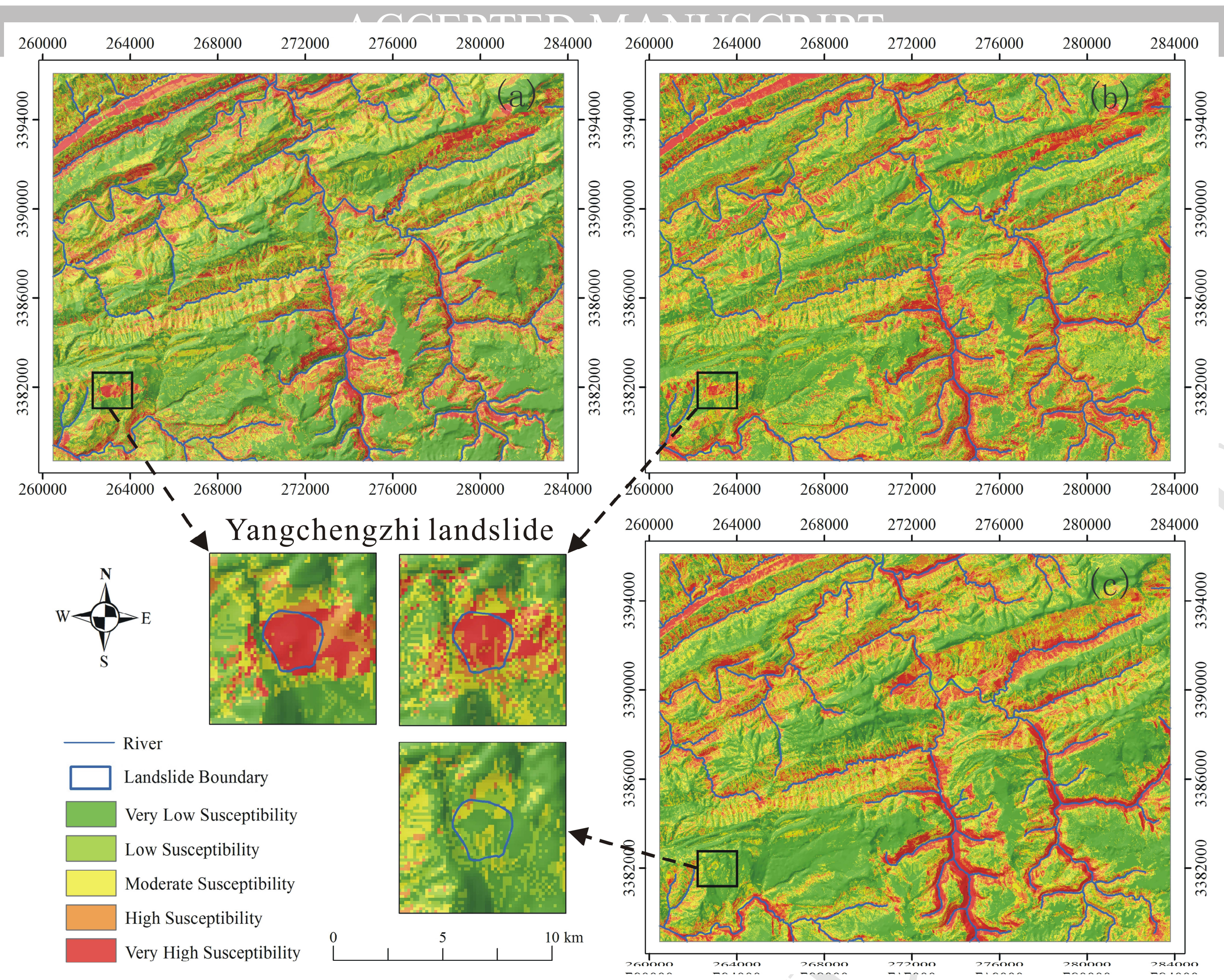

Fig. 9 (a) Landslide susceptibility mapping using SVM, **(b)** Landslide susceptibility mapping using ANN and **(c)** Landslide susceptibility mapping using LR

5.2 Validation and comparison

5.2.1 Using accuracy statistic

Validation is an essential component in landslide susceptibility modeling to attest the effectiveness and scientific significance of the used method (Frattini et al., 2010). The landslide distribution in different susceptibility levels was statistically analyzed. The results are shown in Table 7 and Fig. 10:

(**a**) In case of the SVM model, 87.79% of the landslides were distributed in the High and Very High susceptibility groups, while the results of the ANN and LR models were 78.51% and 57.07%, respectively.

(**b**) The area of the Very High level of SVM model accounted for 9.284% of the total domain with a proportion of landslide in the total landslide of 65.13%. The frequency ratio of SVM model in the Very High level was the largest (7.015), while the ANN and LR models were much smaller i.e. 5.241 and 2.233, respectively.

Table 7 Accuracy statistics of the SVM, ANN, and LR models

Susceptibility Level	Probability of landslide	Pixels in landslide(**A**)	Pixels in domain(**B**)	Proportion of landslide in domain(**A/B**)	Proportion of landslide in total landslide(**C**)	Proportion of domain in total domain(**D**)	Frequency ratios (**C/D**)
SVM							
Very Low	0.000 ~ 0.023	31	197837	0.016%	0.550%	30.28%	0.018
Low	0.023 ~ 0.082	191	144568	0.132%	3.390%	22.13%	0.153
Moderate	0.082 ~ 0.240	466	129432	0.360%	8.269%	19.81%	0.417
High	0.240 ~ 0.824	1277	120875	1.056%	22.66%	18.50%	1.225
Very High	0.824 ~ 1.000	3670	60657	6.050%	65.13%	9.283%	7.015
ANN							
Very Low	0.000 ~ 0.169	228	199042	0.115%	4.046%	30.60%	0.132
Low	0.169 ~ 0.275	364	133837	0.272%	6.459%	20.58%	0.314
Moderate	0.275 ~ 0.400	619	129085	0.480%	10.98%	19.84%	0.553
High	0.400 ~ 0.620	1671	127778	1.308%	29.65%	19.65%	1.509
Very High	0.620 ~ 1.000	2753	60627	4.541%	48.85%	9.322%	5.241
LR							
Very Low	0.000 ~ 0.151	462	198051	0.233%	8.198%	30.45%	0.269
Low	0.151 ~ 0.234	702	131161	0.535%	12.45%	20.17%	0.618
Moderate	0.234 ~ 0.343	1255	129462	0.969%	22.27%	19.91%	1.119
High	0.343 ~ 0.532	1988	128235	1.550%	35.28%	19.72%	1.789
Very High	0.532 ~ 1.000	1228	63460	1.935%	21.79%	9.758%	2.233

(**c**) As for the level of Very Low, the area of SVM model accounted for 30.28% of the total domain, while its landslide only accounted for 0.550% of the total landslide. The frequency ratio of SVM model in the Very Low level was the lowest of 0.018; and the ANN and LR models were 0.132 and 0.269, respectively, which were much larger than the SVM model.

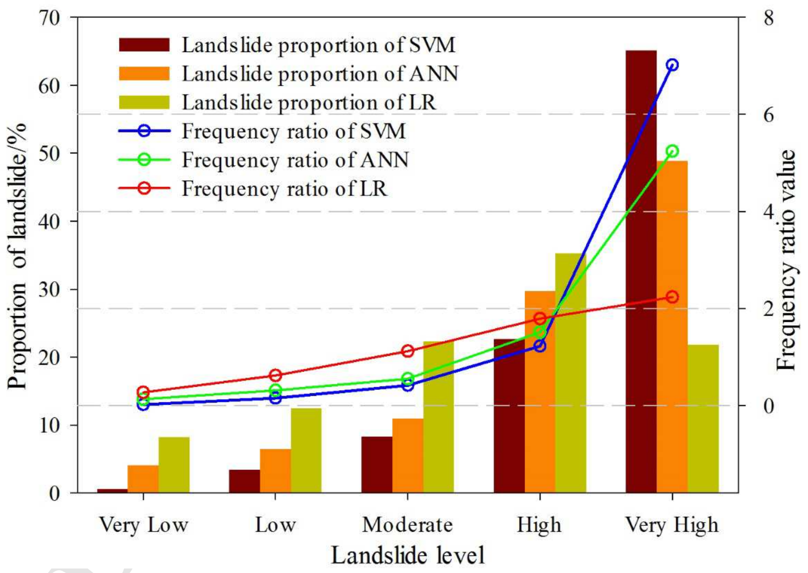

Fig. 10 Landslide proportion and frequency ratio of each susceptibility level

5.2.2 Using ROC curve

The statistical method is effective to evaluate the model performance. However, it is a cutoff-dependent approach that requires reclassification of landslide susceptibility index. The evaluation results may vary with the breakpoints of reclassification. The receiver operating characteristics (ROC) curve (Hanley and McNeil, 1983) is cutoff-independent. The area under the ROC curve (AUC) can be used to assess the performance of models, and the model with a larger AUC is considered better.

The ROC curves in Fig. 11 and 12 show the training and verifying performance of the used models in the colluvial landslide and rockfall modeling, respectively. The machine learning models of SVM and ANN achieved excellent performance in both of the colluvial landslide and rockfall assessment. The SVM model outperformed ANN, its AUC of training and verifying are 0.937 and 0.912, respectively in colluvial landslide assessment, 0.967 and 0.932, respectively in rockfall assessment. The SVM model achieved higher prediction accuracy, because it is based on the principle of structural risk minimization, instead of traditional experience risk minimization, and its solution is globally optimal. On the contrary, the ANN is based on the principle of experience risk minimization which often leads to locally optimal solution.

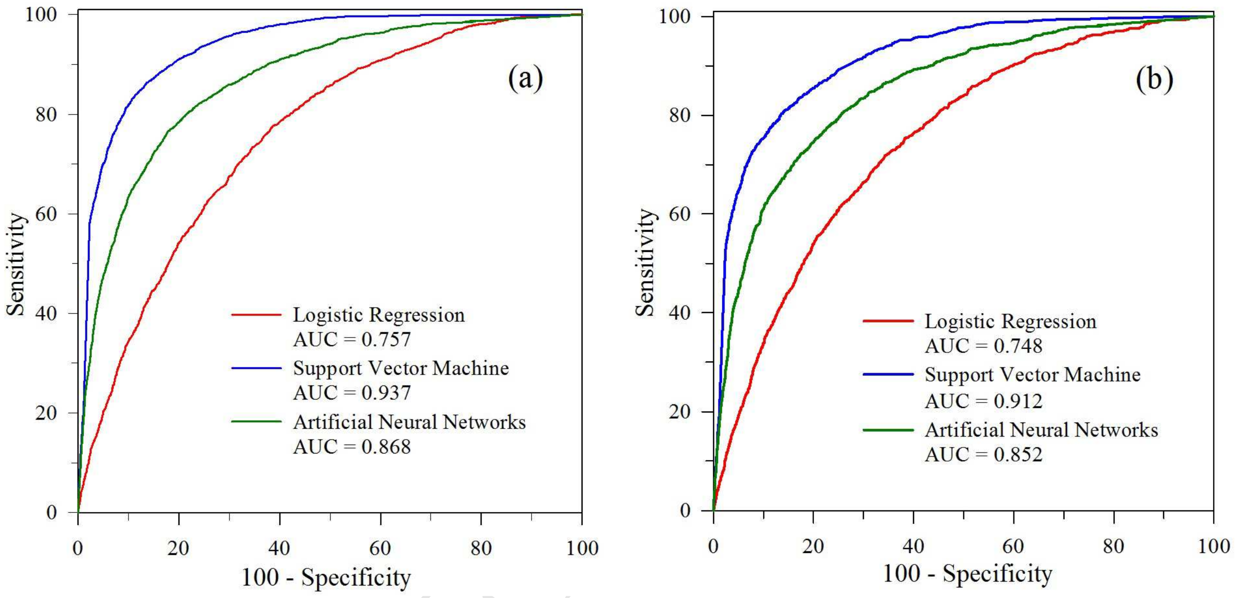

Fig. 11 The ROC curves of the SVM, ANN and LR models in colluvial landslide susceptibility assessment: **(a)** training and **(b)** verifying

The LR model shows the worst performance in both cases with the verifying AUC value of 0.748 and 0.884, respectively. The colluvial landslide development is strongly and jointly affected by many factors, which is a more complex nonlinear problem than the rockfall (Fig. 7). The LR model uses linear combinations of variables, which is not adept at modeling grossly complex nonlinear problem. This is the reason why LR model showed worse performance in colluvial landslide modeling but better in rockfall. As shown in Fig. 9, the Yangchengzhi landslide was predicted accurately by the two machine learning models, but not predicted accurately by LR model. Overall, the machine learning models of SVM and ANN achieved better performance than the multivariable statistics model of LR, and the SVM performed the best.

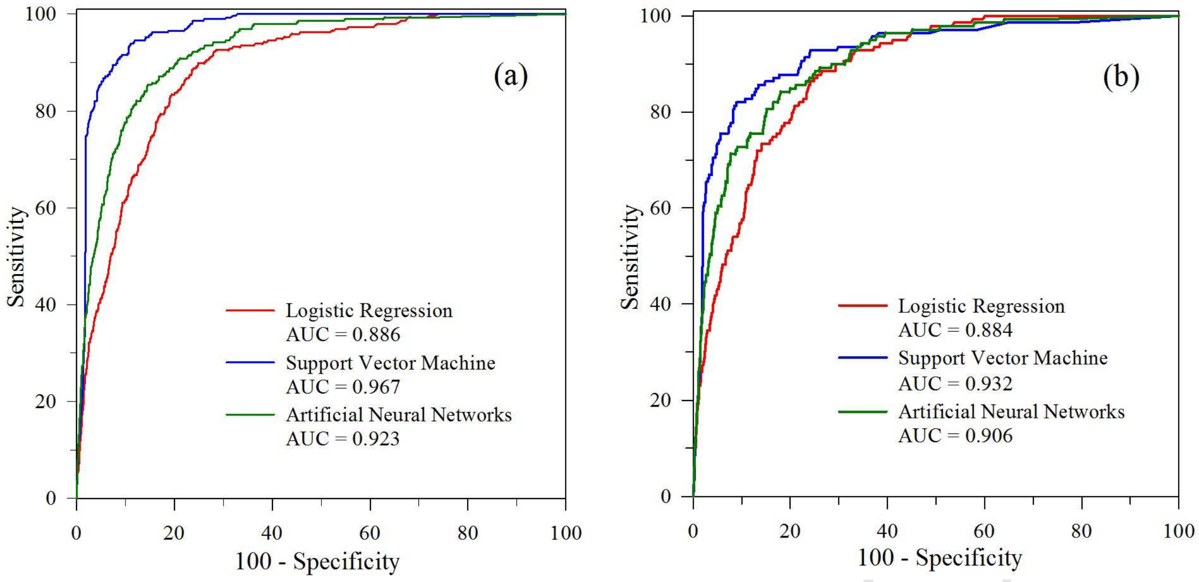

Fig. 12 The ROC curves of the SVM, ANN and LR models in rockfall susceptibility assessment: **(a)** training and **(b)** verifying

As shown in Table 8, the prediction AUC of SVM, ANN, and LR in susceptibility assessment without landslide classification are 0.881, 0.836 and 0.697, respectively. All of them are less than the prediction AUC of the separate colluvial landslide and rockfall assessment. Due to the separation of landslide type, the prediction accuracy of the three models were improved 0.041, 0.043 and 0.119, respectively (Table 8). The susceptibility assessment with landslide classification can achieve more accurate prediction than the susceptibility assessment without landslide classification, especially for the model without strong classification capacity.

Table 8 The prediction performance comparison

Models	AUC of prediction			Accuracy improvement
	Landslide without classification	Colluvial landslide	Rockfall	
SVM	0.881	0.912	0.932	0.041
ANN	0.836	0.852	0.906	0.043
LR	0.697	0.748	0.884	0.119

Note: Accuracy improvement = (Colluvial landslide + Rockfall) / 2 - Landslide without classification

5.2.3 Statistical significance test

It is important to check if the performance among the used models has a statistically significant difference when comparing their performance (Pham et al., 2016b; Tien Bui et al., 2016). The Friedman test (Friedman, 1937) is an effective non-parametric method and widely used in statistical significance test. In this study, the Friedman test at 95% significant level was carried out to check if there are statistically significant differences between the three susceptibility models. All the p-values of the colluvial landslide and rockfall modeling were extremely low (< 0.000) and less than 0.005. It indicates the null hypothesis (i.e., no differences between the performance of the test models) is rejected. Consequently, the performance of the three models is significantly different and comparable.

6. Discussion

In this study area, the landslides were divided into two types: colluvial landslide and rockfall. The causal factors had different significances on landslide occurrences. For example, the distance to roads was the dominant factor for rockfall, while the colluvial landslide was strongly affected by the distance to rivers, altitude, and lithology (Fig. 7). Various levels of causal factor influenced landslides types inversely. As shown in Table 1, the slope of 6~18° had the most positive effect on colluvial landslide, while the slope > 39° showed the most positive effect on rockfall. Taking all these issues into account, the development law of each landslide type was found different, it was suggested that each LSM should be conducted separately when the study area is threatened by more than one landslide type, then a more accurate prediction can be achieved (Table 8).

The LR performed well in rockfall with the predicted AUC of 0.884, but not good in the colluvial landslide with the predicted AUC of 0.748. It indicates that the LR model is not suitable for complex nonlinear problem. The machine learning models (SVM and ANN) are excellent in both the colluvial landslide and rockfall susceptibility assessment. It demonstrates that the machine learning models are also applicative in complex nonlinear problem, and the SVM model has a better performance because of its globally optimal solution. Moreover, one issue should be noticed that the model performance is data dependent, it may vary with different cases. For instance, the Random Forest model performed well in Arno River basin (central Italy) (Catani et al., 2013), but not well in Lianhua County (China) (Hong et al., 2016). In this study, the SVM showed stable prediction performance in both the colluvial landslide and rockfall assessment. Furthermore, as reported in previous literatures, the SVM model achieved accurate prediction in almost all the cases (Pradhan, 2013; Peng et al., 2014; Pham et al., 2016a; Lee et al., 2017), which stated that the performance of SVM has a strong robustness. Hence, the SVM model can be recommended before reaching a consensus on the model of landslide susceptibility assessment.

The error of landslide susceptibility assessment is composed of the false positive part and false negative part. Statistically, the two parts have the same influence on model performance, but their cost of misclassification is totally different. To a certain extent, the false positive result may restrict the use of land, leading to economic losses slightly. But if the landslide or landslide-prone areas are erroneously identified as stable slopes, such as the Yangchenzhi landslide in the susceptibility map of LR (Fig. 9), and proceed with land planning and utilization without any risk control measures, it may lead to catastrophic consequences. In the future study of susceptibility assessment, we should pay more attention to reduce the false negative error.

7. Conclusions

Landslide susceptibility assessment is crucial for strict land-use planning and disaster risk reduction in landslide-prone areas. In this study, Longju in the TGRA was taken as a case study where two types of landslide were observed, the colluvial landslide and rockfall, and their mechanisms were different. Altitude (450m-950m), distance to rivers (<200m), and lithology (Interbeds of mudstone and sandstone) were dominant in colluvial landslide, while the crucial factors of rockfall were identified as distance to roads (<50m), distance to rivers (<200m) and lithology (Lithic sandstone with mudstone). Due to the separation of landslide type, the prediction AUC values of SVM, ANN and LR models were improved 0.041, 0.043 and 0.119, respectively. It indicates that the LSM with landslide classification can achieve more excellent performance. It is recommended to separately analyze and assess each landslide type, and combine separate susceptibility map to obtain better results.

The causal factors have different influences on landslide occurrences, which lead to different contributions of each factor to the modeling and assessment of landslide susceptibility. Information gain ratio is an effective method, which can quantify the importance of causal factors. The performance comparison of the eight models with different eliminated factors indicates that the less important factors may have a negative effect in LSM and those noise-producing factors should be eliminated to achieve greater prediction precision.

Two machine learning models (i.e. SVM and ANN) and a multivariate statistical model (namely LR) were applied to carry out the colluvial landslide and rockfall susceptibility assessment. The performance was evaluated by the ROC curves and Friedman test. The comparison results demonstrate that the machine learning models outperform the multivariate statistical method. The SVM model showed the best performance with AUC value for training and verifying of 0.937 and 0.912 respectively in colluvial landslide assessment. The training and verifying AUC value was found 0.967 and 0.932, respectively in rockfall assessment. SVM model are highly recommended to conduct landslide susceptibility assessment in the TGRA and other similar context.

Acknowledgements

This paper was prepared as part of the projects "The study of mechanism and forecast criterion of the gentle-dip landslides in The Three Gorges Reservoir Region, China" (No. 41572292) and "Study on the hydraulic properties and the rainfall infiltration law of the ground surface deformation fissure of colluvial landslides" (No. 41702330) funded by the National Natural Science Foundation of China. The comments from the three anonymous reviewers and the editors have significantly improved the quality of this article. The first author would like to thank the China Scholarship Council for funding his research at the University of Florence, Italy.

References

AGU, 2017. The human cost of landslide in 2016, The Landslide Blog, American Geophysical Union (AGU). (http://blogs.agu.org/landslideblog/).

Bai, S. B., Wang, J., Lu, G. N., Zhou, P. G., Hou, S. S., Xu, S. N., 2010. GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. Geomorphology 115(1), 23-31.

Budimir, M. E. A., Atkinson, P. M., Lewis, H. G., 2015. A systematic review of landslide probability mapping using logistic regression. Landslides 12(3), 419-436.

Catani, F., Lagomarsino, D., Segoni, S., Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat. Hazard. Earth. Sys. 13(11), 2815-2831.

Corominas, J., Van Westen, C., Frattini, P., Cascini, L., Malet, J. P., Fotopoulou, S., Catani, F., Van Den Eeckhaut, M., Mavrouli, O., Agliardi, F., Pitilakis, K., Winter, M.G., Pastor, M., Ferlisi, S., Tofani, V., Hervás, J., Smith, J. T., Pitilakis, K., 2014. Recommendations for the quantitative analysis of landslide risk. Bull. Eng. Geol. Environ. 73(2), 209-263.

Cox, D. R., 1958. The Regression Analysis of Binary Sequences. J. Roy. Stat. Soc. B. 20(2), 215-242.

Felicísimo, Á.M., Cuartero, A., Remondo, J., Quirós, E., 2013. Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. Landslides 10, 175–189

Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., Savage, W. Z., 2008. Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. Eng. Geol. 102(3), 99-111.

Frattini, P., Crosta, G., Carrara, A., 2010. Techniques for evaluating the performance of landslide susceptibility models. Eng. Geol. 111(1), 62-72.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32: 675–701.

Goetz, J. N., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput. Geosci. 81, 1-11.

Gorsevski, P. V., Brown, M. K., Panter, K., Onasch, C. M., Simic, A., Snyder, J., 2016. Landslide detection and susceptibility mapping using LiDAR and an artificial neural network approach: a case study in the Cuyahoga Valley National Park, Ohio. Landslides 13(3), 467-484.

Guo, C., Montgomery, D. R., Zhang, Y., Wang, K., Yang, Z., 2015. Quantitative assessment of landslide susceptibility along the Xianshuihe fault zone, Tibetan Plateau, China. Geomorphology 248, 93-110.

Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148 (3), 839–843.

Hecht-Nielsen, R., 1988.Theory of the back propagation neural network. Neural. Networks.1(S-1), 445-448.

Hong, H., Pourghasemi, H. R., Pourtaghi, Z. S., 2016. 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.

Hungr, O., Leroueil, S., Picarelli, L., 2014. The Varnes classification of landslide types, an update. Landslides, 11(2), 167-194.

Hussin, H. Y., Zumpano, V., Reichenbach, P., Sterlacchini, S., Micu, M., van Westen, C., Bălteanu, D., 2016. Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model. Geomorphology 253, 508-523.

Lee, S., Hong, S. M., Jung, H. S., 2017.A Support Vector Machine for Landslide Susceptibility Mapping in Gangwon Province, Korea. Sustainability 9(1), 48.

Moore, I. D., Grayson, R. B., Ladson, A. R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol. Processes 5(1), 3-30.

O'Brien, R. M., 2007. A caution regarding rules of thumb for variance inflation factors. Qual. Quant. 41, 673– 690.

Peng, L., Niu, R., Huang, B., Wu, X., Zhao, Y., Ye, R., 2014. Landslide susceptibility mapping based on rough set theory and support vector machines: A case of the Three Gorges area, China. Geomorphology 204, 287-301.

Pham, B. T., Bui, D. T., Dholakia, M. B., Prakash, I., Pham, H. V., Mehmood, K., Le, H. Q., 2016. A novel ensemble classifier of rotation forest and Naïve Bayer for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS. Geomatics, Natural Hazards and Risk, 1-23.

Pham, B. T., Pradhan, B., Bui, D. T., Prakash, I., Dholakia, M. B., 2016. A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). Environ. Modell. Softw. 84, 240-250.

Pradhan, B., and Lee, S., 2010. Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. Landslide 7, 13-30.

Pradhan, B., and Lee, S., 2010. Landslide susceptibility assessment and factor effect analysis: back-propagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environ. Modell. Softw: 25(6), 747-759.

Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350-365.

Quinlan, J. R., 1996. Improved use of continuous attributes in C4. 5. Journal of artificial intelligence research, 4, 77-90.

Tien Bui, D., Tuan, T. A., Klempe, H., Pradhan, B., Revhaug, I., 2016. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides, 13(2), 361-378.

Van Westen, C. J. (1993). Application of geographic information systems to landslide hazard zonation (Doctoral dissertation, TU Delft, Delft University of Technology).

Varnes, D J., 1978. Slope movement types and processes. In: Schuster RL, Krizek RJ (eds) Landslides, analysis and control, special report 176: Transportation research board, National Academy of Sciences, Washington, DC., pp. 11–33.

Vapnik, V. N., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York.

Wu, X., Niu, R., Ren, F., Peng, L., 2013.Landslide susceptibility mapping using rough sets and back-propagation neural networks in the Three Gorges, China. Environ. Earth Sci. 70(3), 1307-1318.

Yao, X., Tham, L. G., Dai, F., 2008. Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. Geomorphology 101(4), 572-582.

Yin, Y., Huang, B., Wang, W., Wei, Y., Ma, X., Ma, F., Zhao, C., 2016.Reservoir-induced landslides and risk control in Three Gorges Project on Yangtze River, China. Journal of Rock Mechanics and Geotechnical Engineering 8(5), 577-595.

Yin, K., Yan, T., 1988. Statistical prediction model for slope instability of metamorphosed rocks. In Proceedings of the 5th International Symposium on Landslides (Vol. 2, pp. 1269-1272).

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., Al-Katheeri, M. M., 2016. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides 13(5), 839-856.

Zhu, A., Wang, R., Qiao, J., Qin, C., Chen, Y., Liu, J., Du, F., Lin, Y., Zhu, T., 2014.An expert knowledge-based approach to landslide susceptibility mapping using GIS and fuzzy logic. Geomorphology 214, 128-138.

Zhou, C., Yin, K., Cao, Y., Ahmed, B., 2016. Application of time series analysis and PSO–SVM model in predicting the Bazimen landslide in the Three Gorges Reservoir, China. Eng. Geol. 204, 108-120.

Zhou C., Yin K., Xiang Z., Yang B., 2015. Quantitative evaluation of the landslide susceptibility in Chun'an county based on GIS. Safety and Environmental Engineering 22(1), 45-50. (in Chinese)

Highlights:

1. The development laws of the colluvial landslide and rockfall were analyzed.

2. The model performance was improved by eliminating less important factors.

3. The separated modeling of each landslide type has significantly increased the prediction accuracy.

4. The performance of three models was compared and the SVM model performed the best.