







Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models


CHEN Tao^{1,2*}  <https://orcid.org/0000-0001-6965-1256>;  e-mail: taochen@cug.edu.cn

ZHU Li¹  <https://orcid.org/0000-0002-5616-6814>; e-mail: 1085754231@qq.com

NIU Rui-qing¹  <https://orcid.org/0000-0002-0862-7890>; e-mail: rqnium@163.com

TRINDER C John³  <https://orcid.org/0000-0003-0165-5685>; e-mail: j.trinder@unsw.edu.au

PENG Ling⁴  <https://orcid.org/0000-0003-3293-3713>; e-mail: penglmail@126.com

LEI Tao⁵  <https://orcid.org/0000-0001-9639-6043>; e-mail: leitao@sust.edu.cn

* Corresponding author

¹ Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China

² Geomatics Technology and Application key Laboratory of Qinghai Province, Xining 810001, China

³ School of Civil and Environmental Engineering, The University of New South Wales, Sydney, NSW 2052, Australia

⁴ China Institute of Geo-Environment Monitoring, Beijing 100081, China

⁵ School of Electrical and Information Engineering, Shaanxi University of Science and Technology, Xi'an 710021, China

Citation: Chen T, Zhu L, Niu RQ, et al. (2020) Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models. *Journal of Mountain Science* 17(3). <https://doi.org/10.1007/s11629-019-5839-3>

© Science Press, Institute of Mountain Hazards and Environment, CAS and Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract: This work was to generate landslide susceptibility maps for the Three Gorges Reservoir (TGR) area, China by using different machine learning models. Three advanced machine learning methods, namely, gradient boosting decision tree (GBDT), random forest (RF) and information value (InV) models, were used, and the performances were assessed and compared. In total, 202 landslides were mapped by using a series of field surveys, aerial photographs, and reviews of historical and bibliographical data. Nine causative factors were then considered in landslide susceptibility map generation by using the GBDT, RF and InV models. All of the maps of the causative factors were resampled to a resolution of 28.5 m. Of the 486289 pixels in the area, 28526 pixels were landslide pixels, and 457763 pixels were non-landslide pixels. Finally, landslide

susceptibility maps were generated by using the three machine learning models, and their performances were assessed through receiver operating characteristic (ROC) curves, the *sensitivity*, *specificity*, *overall accuracy* (OA), and *kappa coefficient* (KAPPA). The results showed that the GBDT, RF and InV models in overall produced reasonable accurate landslide susceptibility maps. Among these three methods, the GBDT method outperforms the other two machine learning methods, which can provide strong technical support for producing landslide susceptibility maps in TGR.

Keywords: Mapping landslide susceptibility; Gradient boosting decision tree; Random forest; Information value model; Three Gorges Reservoir

Introduction

The socioeconomic stability of the area near

Received: 09-Oct-2019
1st Revision: 02-Dec-2019
2nd Revision: 03-Feb-2020
Accepted: 17-Feb-2020

the Three Gorges Reservoir (TGR) in China is suffering serious threats caused by many large and small active landslides in the region (Liu et al. 2009). This area is well known for its cyclical fluctuations in the reservoir water level, which causes bank slope instability: a serious but inevitable problem for residents and infrastructures (Peng et al. 2014).

Landslide susceptibility mapping (LSM) has been proven to be a complex task (Brabb 1991). Over the last several decades, numerous LSM methods have been proposed, including analytical hierarchy processes (AHPs) (Yalcin 2008), multivariate regression (Guzzetti et al. 2006), the evidential belief function (Althuwaynee et al. 2012), the weight of evidence method (Neuhauser et al. 2012), certainty factors (Devkota et al. 2013), multivariate adaptive regression splines (Felicísimo et al. 2013), logistic regression (Chen et al. 2015; Wang et al. 2017), and the information value method (Chen et al. 2016; Zêzere et al. 2017). Although these methods have achieved some good results, they still have some drawbacks in determining the relationship between the occurrence of landslides and related causative factors, such as environmental, geological and topographical factors (Sarkar et al. 2006; Yunus et al. 2019).

Recently, in addition to the models mentioned above, a variety of machine learning methods have been used for LSM, such as artificial neural networks (ANNs) (Zhou et al. 2018), support vector machines (SVMs) (Lee et al. 2018), neuro-fuzzy techniques (Lee et al. 2015), decision trees (Tsangaratos and Ilia 2016), and random forests (Youssef et al. 2016; Dou et al. 2019). Currently, some proposed integrated approaches combine statistical methods with machine learning techniques, such as rough sets-BPNNs (Wu et al. 2013), AHP-linear combination methods (Hung et al. 2016), feedback-loop-based extreme learning machine methods (Vasu and Lee 2016), statistical index-AHP techniques (Zhang et al. 2016a), ANN-MaxEnt-SVM (Chen et al. 2017a), and the fuzzy weight of evidence method (Hong et al. 2017).

With the development of ensemble learning, bagging and boosting methods are increasingly being used for classification and regression. Ensemble learning approaches have also achieved excellent results in many machine learning

competitions. There have also been studies of LSM using ensemble learning methods, such as AdaBoost (Bai et al. 2016) and bagging (Hong et al. 2018). However, as an ensemble learning method, the gradient boosting decision tree (GBDT) approach has rarely been explored for LSM even though this method achieved good performance in other fields (Zhou et al. 2018). In addition, the resulting landslide susceptibility region may be significantly impacted by even a small improvement in the prediction precision (Bui et al. 2016). For these reasons, the investigation and comparison of ensemble learning methods with traditional approaches for drawing reasonable conclusions for LSM is very important.

Therefore, the purpose of this work was to assess and compare the performance of the gradient boosting decision tree, random forest, and information value methods and to provide a detailed analysis of landslide susceptibility of the TGR, China.

1 Study area and Geological Setting

The study area is located in Yichang City, Hubei Province, China, including Zigui and Badong Counties, with latitudes between 30°50′ N and 31°05′ N and longitudes between 110°06′ E and 110°55′ E (Figure 1). The geographical range is approximately 2 to 4 km along the banks of the Yangtze River and its main tributaries, with a total length of approximately 80 km.

The Yangtze River crosses the study area broadly in a WNW-ESE direction, at an elevation ranging from 800 to 2000 m, and the geomorphology is characterized by a rugged topography. The climate is typically monsoonal and subtropical, with hot and humid summers but cold and dry winters. The average annual rainfall is approximately 1100 mm, and the rainy season is mainly concentrated in spring and summer (He et al. 2008). The maximum rainfall generally appears from June to August, while the minimum rainfall usually occurs from December to March.

The geological foundation of the study area consists of crystalline, pre-Sinian rocks with a Sinian–Jurassic sedimentary cover (Wu et al. 2001). The Huangling anticline formed a structure of approximately 73 km in northeastern Zgui

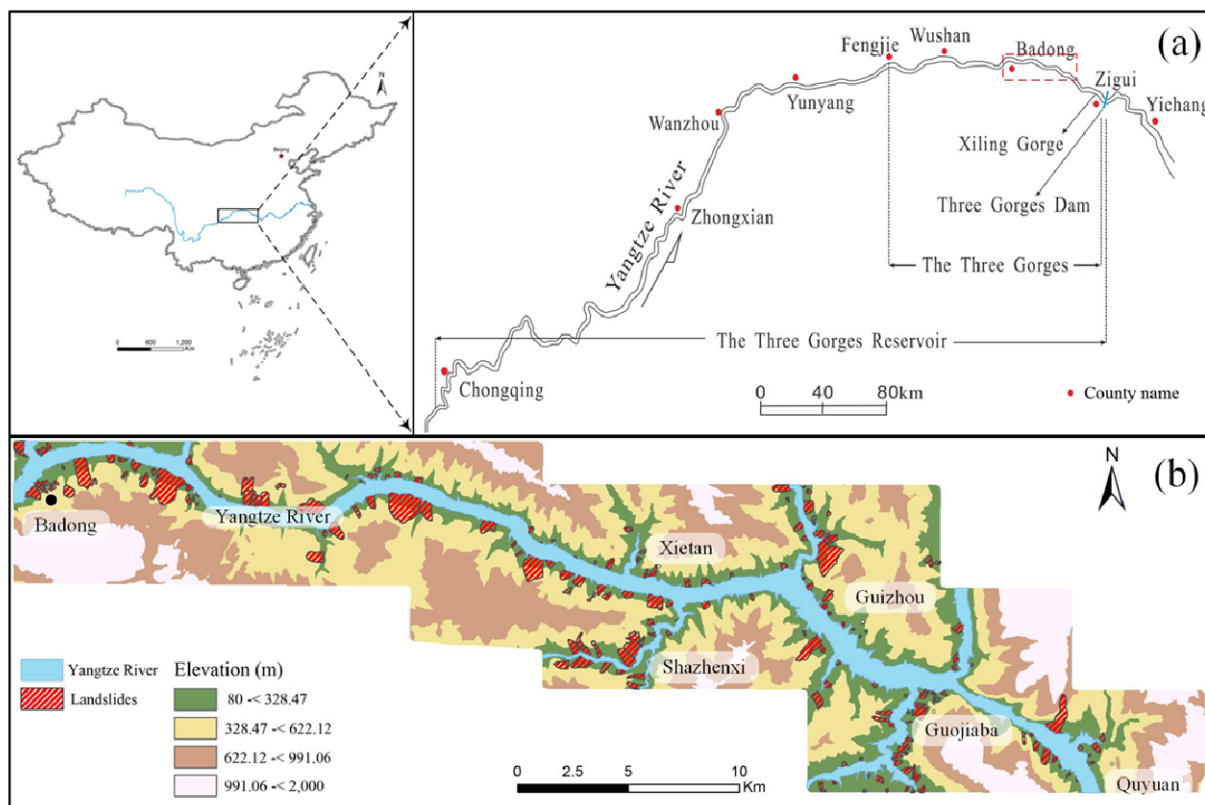


Figure 1 Location of the study area. (a) Sitemap of the Three Gorges Reservoir (TGR), China. (b) Digital elevation model (DEM) map overlaid with landslides.

County. It is oriented mainly in the NNE-SSW direction, and the core is composed of pre-Sinian metamorphic and magmatic rocks (Figure 2).

2 Materials and Methods

2.1 Landslide inventory

A landslide inventory map for the study area was provided by the Headquarters of Geological Hazard Prevention and Treatment in the TGR. The field surveys and image interpretations were based on Google Earth ®. Finally, 202 landslides were identified, with the smallest and largest landslides having areas of 2100 m² and 1.5 km², respectively. Most of the landslides were soil landslides, accounting for approximately 52.9% of all landslides. Then, the landslides were subsequently digitized and rasterized in the Environmental Systems Research Institute (ESRI)'s ArcGIS software (version 10.3.0) at a spatial resolution of 28.5 m, which is the same resolution as the DEM data used in this work.

2.2 Landslide causative factors

The correlation between landslide occurrence and various causative factors has been difficult to predict because of the complex nature and development of the landslides, although several researchers have attempted it in the TGR (Bai et al. 2010; Bi et al. 2014; Peng et al. 2014; Chen et al. 2015; 2016; Xu et al. 2015; Zhang et al. 2016b; Wang et al. 2017; Zhang et al. 2017; Zhou et al. 2018). Based on previous studies by Bai et al. (2010), Peng et al. (2014) and Zhang et al. (2017) as well as our field survey, this research selects nine causative factors, namely, the elevation, aspect, slope, plan curvature, profile curvature, lithology, bedding structure, distance to drainage, and fractional vegetation cover (FVC), to predict the potential landslide distribution. The nine factor values are listed in Table 1. The continuous variables were reclassified into four classes by using the natural breaks method (Smith et al. 2014). The bedding structure, which shows the angular correlation between topography and strata attitude, is a continuous raster layer (Peng et al.

2014). The classification scheme for the bedding structure is shown in Table 2.

In this work, five topographic factors, elevation, slope, aspect, plan curvature, and profile

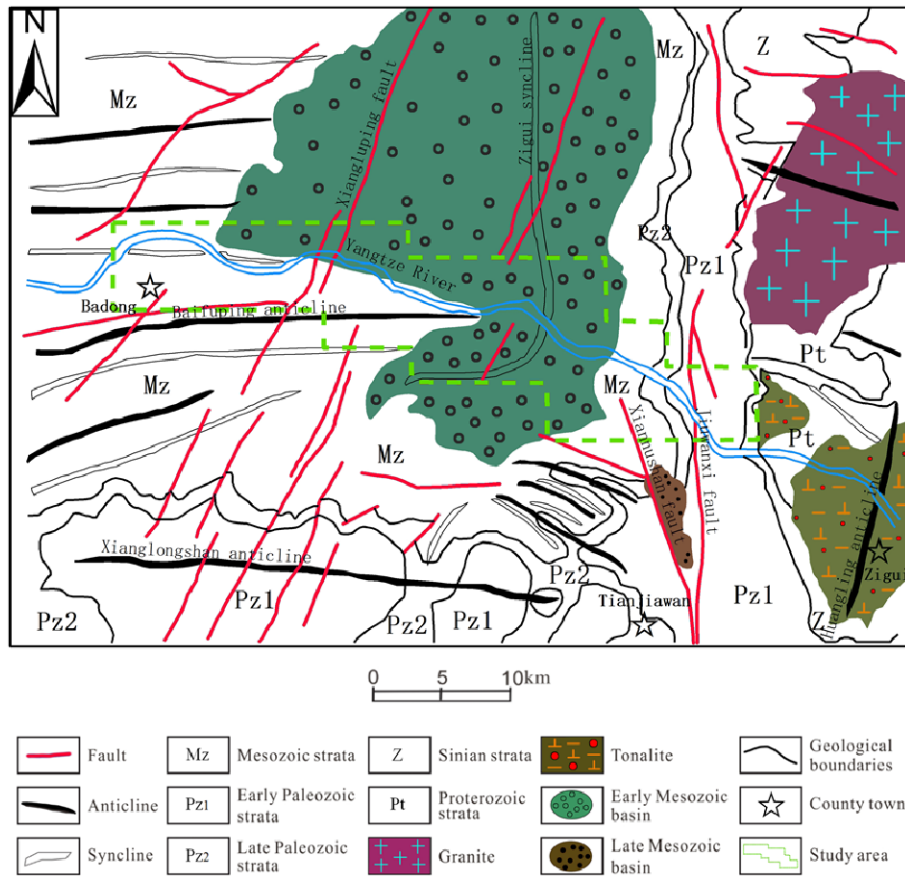


Figure 2 Regional geological and tectonic framework of the study area.

Table 1 Values of landslide causative factors

Causative Factors	Values
Elevation (m)	1) 80 -< 328.47; 2) 328.47 -< 622.12; 3) 622.12 -< 991.06; 4) 991.06 - 2000
Aspect	1) Flat; 2) North; 3) Northeast; 4) East; 5) Southeast; 6) South; 7) Southwest; 8) West; 9) Northwest
Slope (°)	1) 0 -< 11.07; 2) 11.07 -< 24.6; 3) 24.6 -< 36.9; 4) 36.9 - 78.42
Plan curvature (°/100 m)	1) -11.96 -< -1.50; 2) -1.50 -< -0.36; 3) -0.36 -< 0.44; 4) 0.44 - 16.93
Profile curvature (°/100 m)	1) -33.47 -< -1.12; 2) -1.12 -< -0.16; 3) -0.16 -< 1.7; 4) 1.7 - 32
Lithology	1) mudstone, shale and Quaternary deposits; 2) sandstones and thinly bedded limestones; 3) limestones and massive sandstones
Bedding structure	1) over-dip slope; 2) under-dip slope; 3) dip-oblique slope; 4) transverse slope; 5) anaclinal-oblique slope; 6) anaclinal slope
Fractional vegetation cover	1) 0 -< 0.16; 2) 0.16 -< 0.35; 3) 0.35 -< 0.6; 4) 0.6 - 1
Distance to drainage (m)	1) 0 -< 1230.16; 2) 1230.16 -< 1866.87; 3) 1866.87 -< 2744.49; 4) 2744.49 - 4689

Table 2 Classification for the bedding structure

Slope Types	Definition
Over-dip	$ \alpha - \beta \in [0^\circ, 30^\circ)$ or $ \alpha - \beta \in [330^\circ, 360^\circ)$, $\gamma > 10^\circ$ and $\delta > \gamma$
Under-dip	$ \alpha - \beta \in [0^\circ, 30^\circ)$ or $ \alpha - \beta \in [330^\circ, 360^\circ)$, $\gamma > 10^\circ$ and $\delta < \gamma$
Dip-oblique	$ \alpha - \beta \in [30^\circ, 60^\circ)$ or $ \alpha - \beta \in [300^\circ, 330^\circ)$
Transverse	$ \alpha - \beta \in [60^\circ, 120^\circ)$ or $ \alpha - \beta \in [240^\circ, 300^\circ)$
Anaclinal-oblique	$ \alpha - \beta \in [120^\circ, 150^\circ)$ or $ \alpha - \beta \in [210^\circ, 240^\circ)$
Anaclinal	$ \alpha - \beta \in [150^\circ, 210^\circ)$

Note: α is the slope aspect, β is the bed dip direction, γ is the bed dip angle and δ is the slope angle

curvature, were generated from a DEM, which was collected from the Headquarters of Geological Hazard Prevention and Treatment in the TGR area by using ESRI's ArcGIS software (version 10.3.0) (Figure 1b, Figure 3a-d).

The lithology as well as the bedding structure maps were derived from the geological map at a

scale of 1:50000 from the Hubei Province Geological Survey (HPGS) and digitized in ESRI's ArcGIS software (version 10.3.0) (Figure 3e-f). The FVC map was prepared using a backpropagation neural network based on a scene from the China-Brazil Earth Resources Satellite data with a path/row of 04/65, which was acquired in April

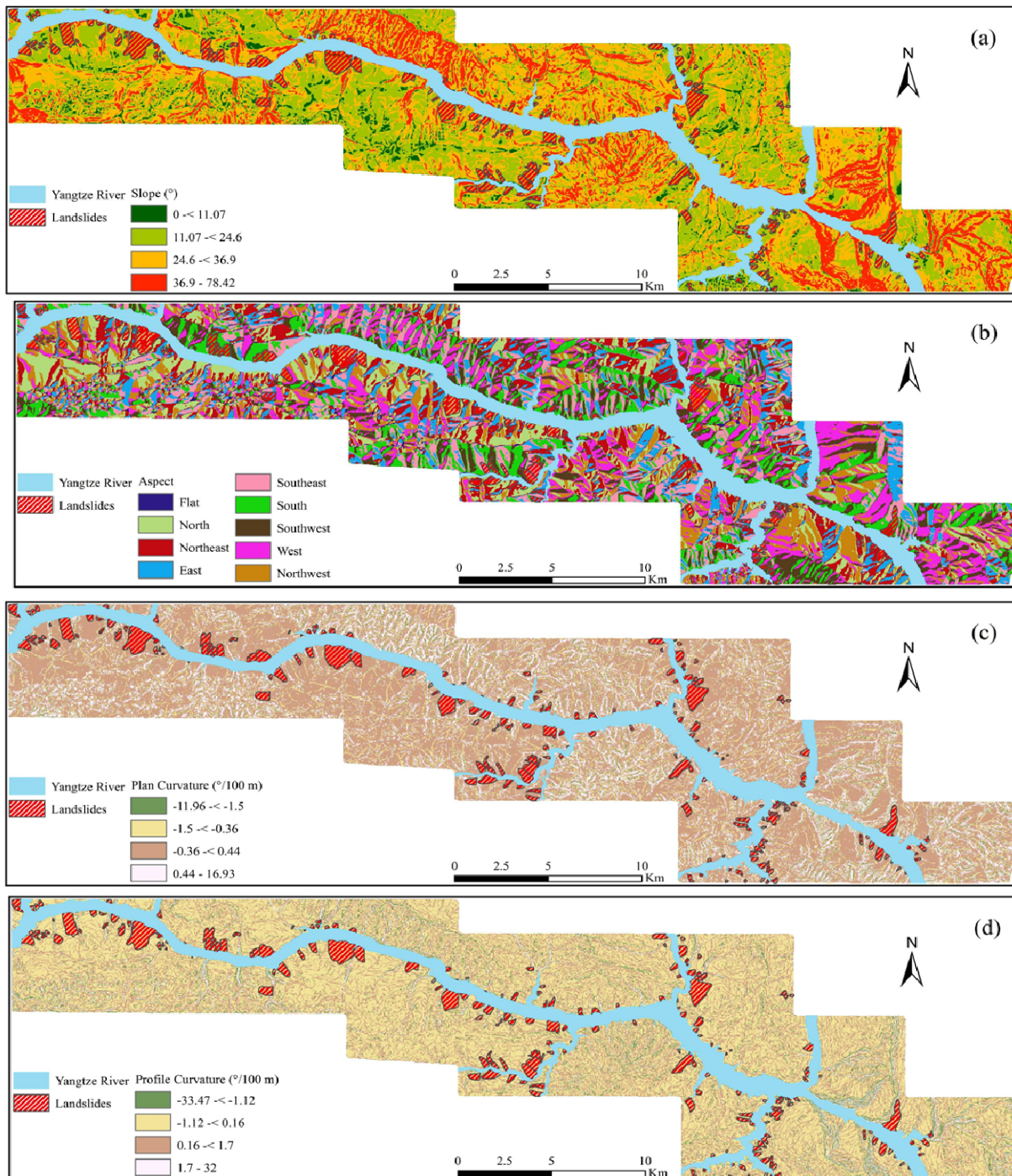
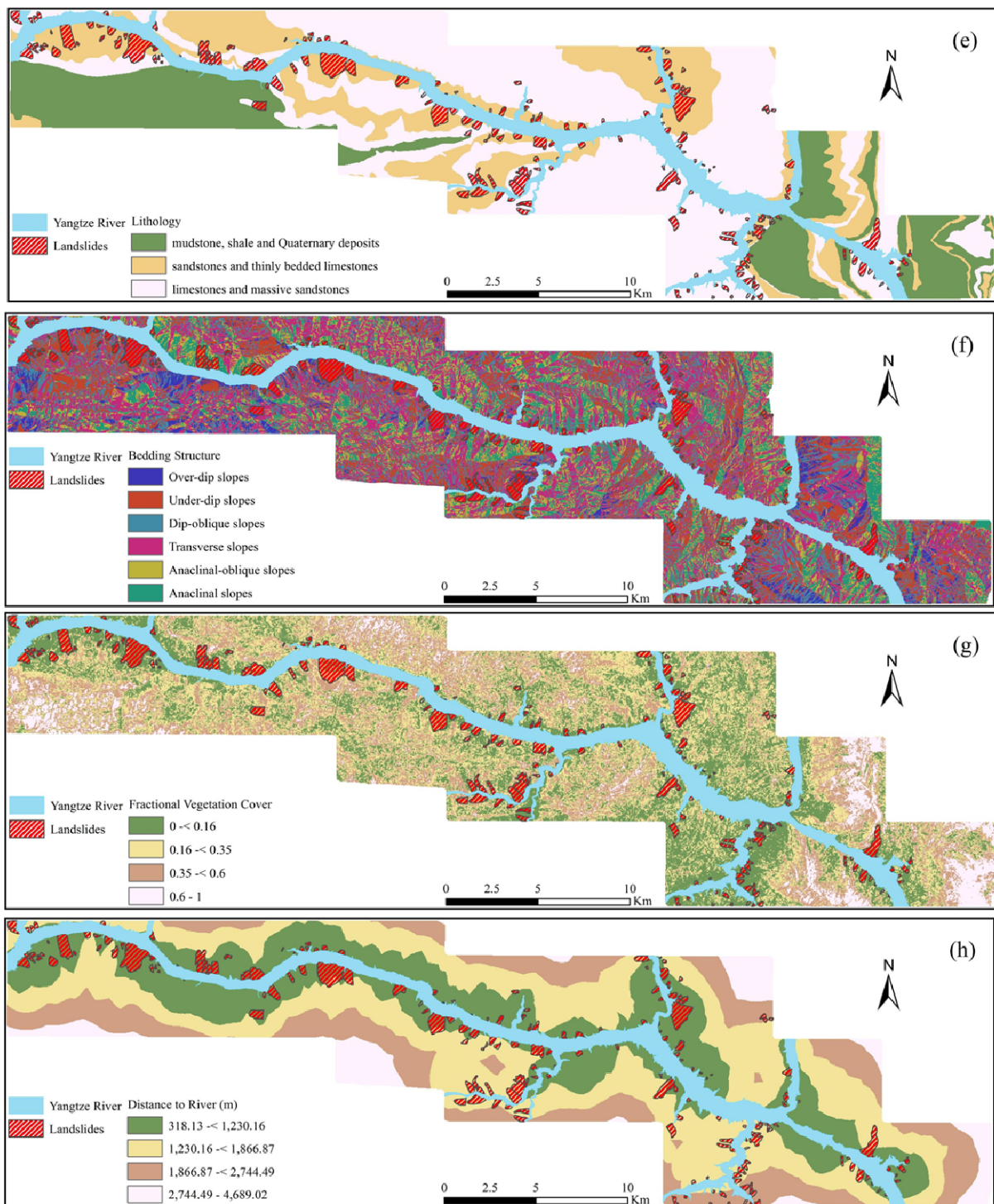


Figure 3 Landslide causative factors maps: (a) Slope; (b) Aspect; (c) Plan Curvature; (d) Profile Curvature; (e) Lithology; (f) Bedding Structure; (g) Fraction Vegetation Cover; (h) Distance to River. (-To be continued-)



(-Continued-) **Figure 3** Landslide causative factors maps: (a) Slope; (b) Aspect; (c) Plan Curvature; (d) Profile Curvature; (e) Lithology; (f) Bedding Structure; (g) Fractional Vegetation Cover; (h) Distance to River.

2004 (Chen et al. 2015, 2016) (Figure 3g). The distance to the drainage map was derived from the topographical map by buffering the river lines (Figure 3h). Finally, all the landslide causative factor layers were transformed into raster layers

with a spatial resolution of 28.5 m.

2.3 Training and validation datasets

Pixels were classified into two subsets by

following previous studies (Chen et al. 2018; Wang et al. 2019): 80% of the landslides and the same number of non-landslide pixels were randomly selected to form the training subsets; the remaining 20% of the landslides were assigned as the validation data.

2.4 Methodology

2.4.1 Gradient boosting decision tree (GBDT)

The gradient boosting algorithm is an ensemble learning approximation to the gradient descent method (Fridman 2002). The gradient boosting decision tree (GBDT) method is a combination of decision tree and ensemble learning techniques and was designed to improve the performance of a single predictive model by combining many models (Schapire 2003). The GBDT algorithm uses two processes: boosting and classification. It has the ability to correct the training results and reduce the degree of overfitting by a regularization function (Elith et al. 2008). Decision trees are convenient for researchers to understand and realize. The trees can handle missing features well because each node of the decision tree relies on only one feature (Breiman et al. 1984). Boosting is a method for improving the performance of the GBDT by constructing a predictive function sequence group, which transforms several weak learning classifiers into a strong learning classifier with high precision (Schapire 2003). The GBDT algorithm was developed in Python language by using the GBDT class library of scikit-learn.

2.4.2 Random forest (RF)

The random forest (RF) algorithm is a classification technique that uses a CART decision tree as the base classifier. The RF algorithm also belongs to the category of ensemble learning algorithms based on bagging theory (Breiman 2001). Each decision tree that makes up the random forest is generated in parallel, and these trees can be classification trees or regression trees. In each decision tree, each node is split using the best features that can generate the optimal solution among all the features (Kausar and Majid 2016). The RF algorithm has been a popular technique for extracting useful but hidden information within

large amounts of data (Chen et al. 2018). In a random forest, the training sets are generated first by using the bootstrap method (Breiman 2001), and then a decision tree is constructed for each training set. The different classifiers of the same type were trained by using each training subset. Then, the individual classifiers are combined by simple majority voting. The RF algorithm was developed in MATLAB by using Random Forest for MATLAB, an open source toolbox developed by Abhishek Jaiantilal from the University of Colorado, Boulder (<https://code.google.com/p/randomforest-MATLAB/>).

2.4.3 Information value (InV)

The information value (InV) method is a statistical approach for the prediction of spatial events on the basis of associated parameters and landslide relationships (Sarkar et al. 2006). It is constructed with the influencing factors of known landslide areas and calculates the sensitivity of each influencing factor. Then, the evaluation prediction model is established based on the sensitivity and can be expanded into adjacent areas according to the principle of analogy (Yin and Yan 1988). The information value I_i of each causative factor X_i can be expressed as:

$$I_i = \log \frac{S_i/N_i}{S/N} \quad (1)$$

where S_i is the landslide pixel number in the presence of a causative factor X_i , N_i is the number of pixels associated with causative factor X_i , S is the total number of landslide pixels, and N is the total number of pixels in the study area. Then, the overall information value I can be calculated by

$$I = \sum_{i=1}^n \log \frac{S_i/N_i}{S/N} \quad (2)$$

Negative and positive values of I_i represent the irrelevant and relevant correlation between the presence of a certain causative factor and landslide event, respectively. The stronger the correlation is, the higher the value of I_i (Yan 1988).

2.4.4 Assessment methods

The results of the three landslide susceptibility mapping methods were verified by comparing them with the locations of existing landslides based on a success-rate curve and prediction-rate curve (Pradhan and Lee 2010). The success-rate curve is based on the training datasets used in the construction of the landslide susceptibility model,

while the prediction-rate curve is based on the validation datasets. The success-rate curve can help in understanding how well the models used to derive landslide susceptibility maps have classified the area of existing landslides, while the prediction-rate curve can explain how well the models can predict landslides (Chung and Fabbri 2003).

The accuracy of these three susceptibility mapping models was assessed by using receiver operating characteristic (ROC) curves, the *sensitivity*, *specificity*, *overall accuracy (OA)*, and *kappa coefficient (KAPPA)*. The X axis in the ROC curve shows the sensitivity of the model (the false-positive rate), while the Y axis indicates one minus the specificity of the model (the true-positive rate). The area under the ROC curve (AUC) is used to represent the quality of a probabilistic model used to predict the occurrence or nonoccurrence of predefined “events”. The AUC values can be classified into five categories: poor (0.5-0.6), moderate (0.6–0.7), good (0.7–0.8), very good (0.8–0.9), and excellent (0.9-1) (Chen et al., 2017b). The higher the AUC value is, the better the performance of the model (Youssef et al. 2015).

The *sensitivity*, *specificity*, and *OA* can be expressed by the following equations:

$$\text{Sensitivity} = TP/(FN+TP) \quad (3)$$

$$\text{Specificity} = TN/(TN+FP) \quad (4)$$

$$OA = (TP+TN)/(TP+FN+ FP+TN) \quad (5)$$

where FP (false positive) and FN (false negative) are the numbers of misclassified pixels and TP (true positive) and TN (true negative) are the numbers of correctly classified pixels.

3 Results

3.1 Application of the gradient boosting decision trees model

To improve the generalization ability of the model, a number of parameters were tested in the import algorithm for the landslide training data, and the optimized model was finally obtained. The network search in Python's scikit-learn library was used to adjust these parameters, that is, to allow the machine to traverse and cross-validate automatically according to the given model parameters and to determine the selection of

parameters by tracking the scoring results. In the GBDT model, seven important parameters need to be adjusted:

(1) the maximum number of weak learners (*n_estimators*)

(2) the weight reduction factor for each weak learner (*learning_rate*)

(3) the maximum depth of decision trees (*max_depth*)

(4) the minimum number of samples for the leaf nodes (*min_samples_leaf*)

(5) the minimum number of samples required for the internal nodes to be divided (*min_samples_split*)

(6) the maximum number of features when dividing (*max_features*)

(7) the sampling rate of the training sets (*subsample*).

After several trial-and-error attempts to find the appropriate values of the parameters to obtain an acceptable accuracy of landslide susceptibility mapping, these seven parameters are set to 60, 0.05, 11, 70, 1200, 9, and 0.7, respectively.

3.2 Application of random forest model

The random forest model was developed in MATLAB and the Random Forest for MATLAB open source toolbox. There were two important parameters in Random Forest for MATLAB open source toolbox: *ntree* and *mtry*. The *ntree* parameter is the number of decision trees, while the *mtry* parameter means the number of variables used in the node for binary trees. In this study, after several trials, the *ntree* parameter was set to 700 to achieve the optimal parameters. The number of variables used in the node for binary trees should be the square root of the number of causative factors, so the *mtry* parameter was set to 3 in this work.

3.3 Application of the information value model

The information value model was constructed in ESRI's ArcGIS software (version 10.3.0). First, all the continuous causative factors were reclassified into four classes. Then, the nine causative factor layers were overlapped with the landslide inventory map by using the overlay

Table 3 Typical Information values calculated for each categories of causative factors, based on information value method

Causative factors	Categories	S_i	N_i	S_i/N_i	Information value
Lithology	Mudstone, shale, and Quaternary deposits	939	110422	0.008	-0.839
	Sandstones and thinly bedded limestones	14737	129752	0.113	0.287
	Limestones and massive sandstones	12741	244861	0.052	-0.052
Bedding structure	Over-dip slopes	626	24549	0.026	-0.362
	Under-dip slopes	6476	87420	0.074	0.101
	Dip-oblique slopes	5627	84423	0.067	0.055
Slope (°)	11.07 -< 24.6	16365	175409	0.093	0.202
	36.9 - 78.42	1061	87185	0.012	-0.683
Aspect	Flat	30	1878	0.016	-0.565
	North	6468	67591	0.096	0.213
	Northeast	4932	65409	0.075	0.109
Elevation (m)	80 -< 328.47	20159	107174	0.188	0.506
	622.12 -< 991.06	358	135479	0.003	-1.346
Profile curvature (°/100 m)	-33.47 -< -1.12	402	25885	0.016	-0.577
	-1.12 -< 0.16	15256	240761	0.063	0.034
	0.16 -< 1.7	12542	201881	0.062	0.025
Plan curvature (°/100 m)	-0.36 -< 0.44	23722	325839	0.073	0.094
	0.44 - 16.93	2104	91720	0.023	-0.408
Fractional vegetation cover	0 -< 0.16	17048	166692	0.102	0.241
	0.6 - 1	136	38936	0.003	-1.225
Distance to river (m)	318.13 -< 1230.16	18424	146552	0.126	0.331
	1230.16 -< 1866.87	8935	205896	0.043	-0.13
	1866.87 -< 2744.49	1167	102332	0.011	-0.711

Notes: S_i is the landslide pixel number in the presence of a causative factor X_i , N_i is the number of pixels associated with causative factor X_i

analysis function in ArcGIS. Finally, the information value of each evaluation factor layer of various types was calculated by using Eq. (2) (Table 3).

3.4 Landslide susceptibility maps

The landslide susceptibility maps were generated by using the three abovementioned models, and the susceptibilities were reclassified into five classes (very low, low, moderate, high, and very high) by using the natural breaks method (Figure 4).

A good map of landslide susceptibility should demonstrate the predictability of the occurrence of new or reactivated landslides. In the field validation step, many existing landslides lie in the very high susceptibility areas. Figure 5 shows the Qianjiangping and Baishuihe landslides in Zigui County. These two landslides lie in the very high landslide susceptibility area. But we can see some differences among these three landslide susceptibility maps. For the GBDT based landslide susceptibility map, the area of Qianjiangping and Baishuihe landslides are main in the very high landslide susceptibility level, while the area around

Table 4 Percentages of different landslide susceptibility classes (%).

Landslide Susceptibility	GBDT	RF	InV
Very Low	77.54	65.32	10.23
Low	11.33	13.25	21.26
Moderate	6.03	9.19	24.69
High	3.61	7.10	26.54
Very High	1.49	5.14	17.28

Notes: Gradient boosting decision tree (GBDT), random forest (RF) and information value (InV) models.

these two landslides are in moderate or low landslide susceptibility level. The situation in the other two machine learning based, RF and InV, landslide susceptibility maps, the area around these two landslides are also in very high, or high landslide susceptibility level, especially in InV based landslide susceptibility map.

Furthermore, the distribution of each landslide susceptibility class is presented in Table 4. The distributions of the susceptibility classes in each model are significantly different. The predictions from the GBDT and RF models tend to be similar, where the values of percentages decrease from very low to very high. However, the predictions from the InV model are different, since the very low susceptibility class covers 10.23% of

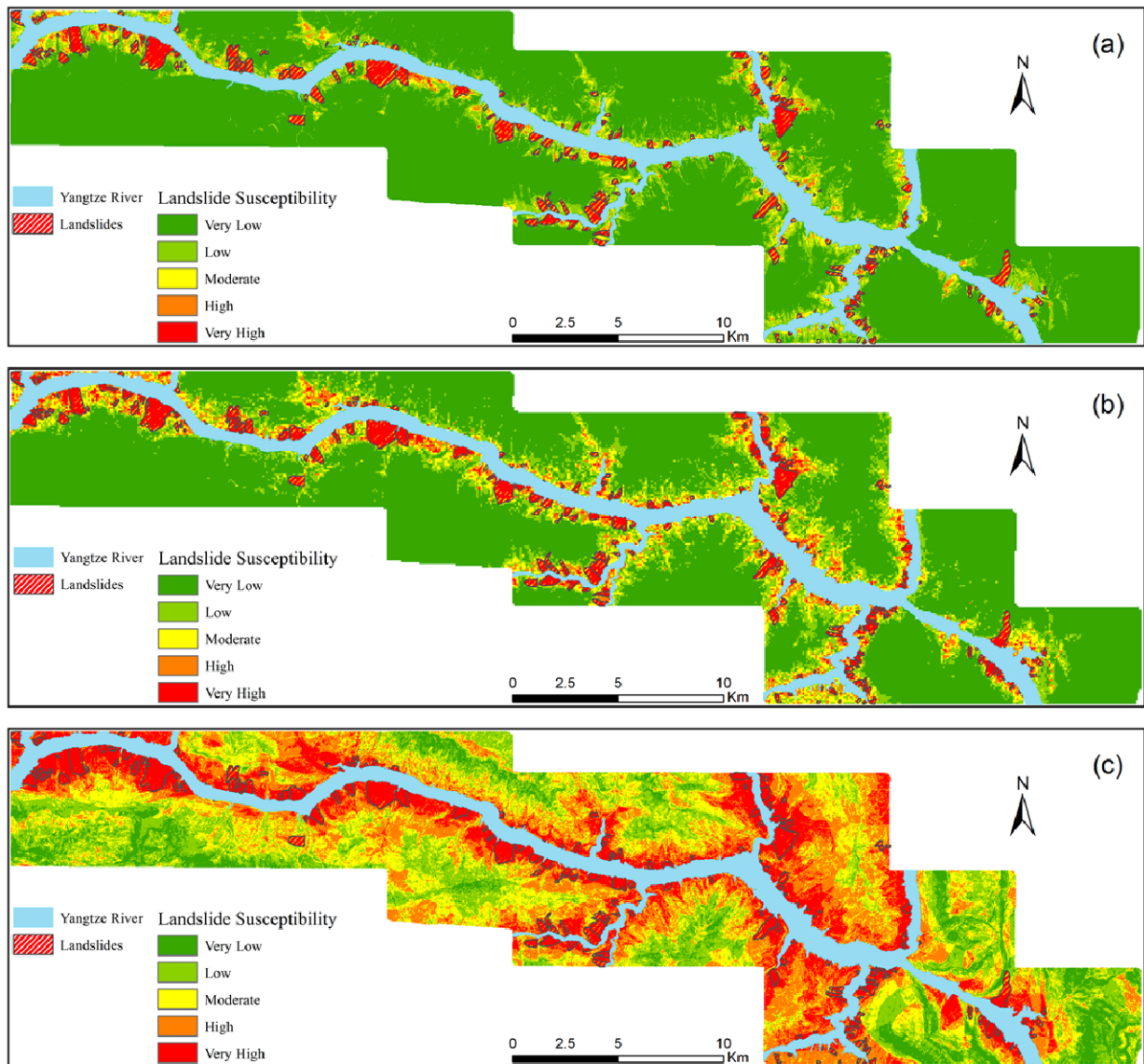


Figure 4 Landslide Susceptibility Maps derived from: (a) gradient boosting decision tree (GBDT); (b) random forest (RF); (c) information value (InV).

the study area, while the low, moderate, high, and very high susceptibility classes cover 21.26%, 24.69%, 26.54%, and 17.28% of the study area, respectively.

3.5 Validation of the landslide susceptibility maps

The success-rate curve and prediction-rate curve of the three models are shown in Figure 6a-b, and the accuracy statistics are presented in Table 5. The AUCs of the success-rate curves for the GBDT, RF, and InV models are 0.962, 0.937, and 0.887, respectively, while the AUCs of the prediction-rate curve for those three models are 0.963, 0.937, and 0.883, respectively. It can be

concluded that all three models can give reasonable LSM results because the AUCs of the success-rate curves and prediction-rate curves for these models are above 0.8. Among these three models, the GBDT model gives the highest AUC values both in the success-rate curve and the prediction-rate curve, followed by the RF and InV models.

The *sensitivity*, *specificity*, *OA*, and *KAPPA* for the training and validation datasets are shown in Table 6.

From the point of view of the *sensitivity*, the InV model achieved the highest performance (96.8% for both the training and validation datasets), followed by the GBDT (86.6% for the training

datasets and 86.7% for the validation datasets) and RF models (80.7% for the training datasets and 80.8% for the validation datasets). From Table 6, it can be seen that the InV model achieves the largest number of TPs and FNs, which indicates that the InV model mapped a larger number of landslide susceptibility areas. This phenomenon may cause a high commission error, although it did achieve the highest sensitivity.

The GBDT achieved the best specificity, OA, and KAPPA than the other two models for both the training and validation datasets, followed by the RF and InV models.

Based on the above findings, it is clear that the GBDT model achieves the overall best performance among these three methods, since it resulted in the four highest accuracies out of the five assessment indices. Then, the statistical significance of the three results was evaluated by using McNemar’s

tests with a 5% significance level for each pair of LSM results (Table 7). Based on Table 7, the null hypothesis that the distributions of the different values across each pair are equal is rejected. This means that the GBDT method is beneficial for mapping landslide susceptibility.

4 Discussion

In this work, three advanced machine learning approaches were applied to map the landslide susceptibility at the TGR area, China, and the results of these three models were compared by using the AUC, *sensitivity*, *specificity*, *OA*, and *KAPPA* values. The results showed that there is a difference between these three methods for the training/validation datasets. The GBDT achieved

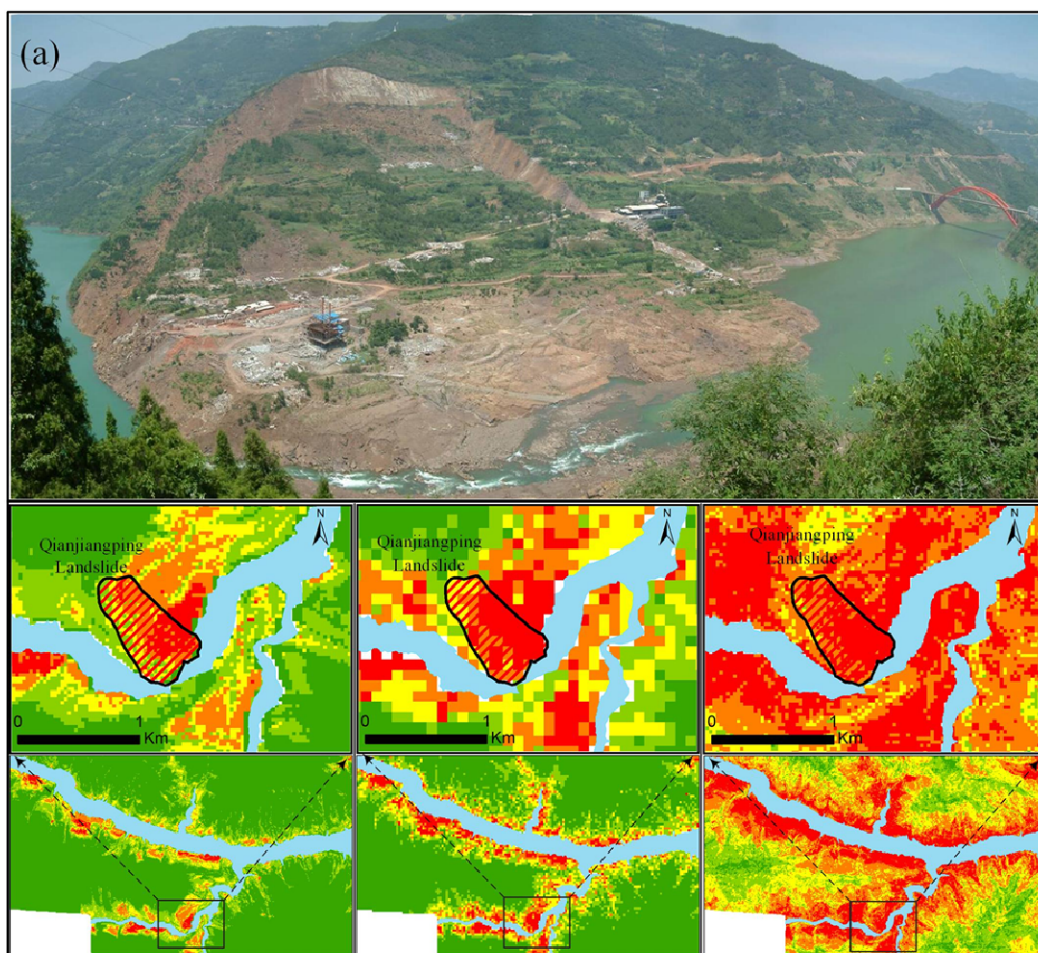


Figure 5 Typical landslides in the study area in Zigui County. (a) Qianjiangping landslides and (b) Baishuihe landslides. The results at the bottom of (a) and (b) are gradient boosting decision tree (GBDT), random forest (RF) and information value (InV) models from left to right. Original pictures were from the Headquarters of Geological Hazard Prevention and Treatment in the TGR. (-To be continued-)

(-Continued-)

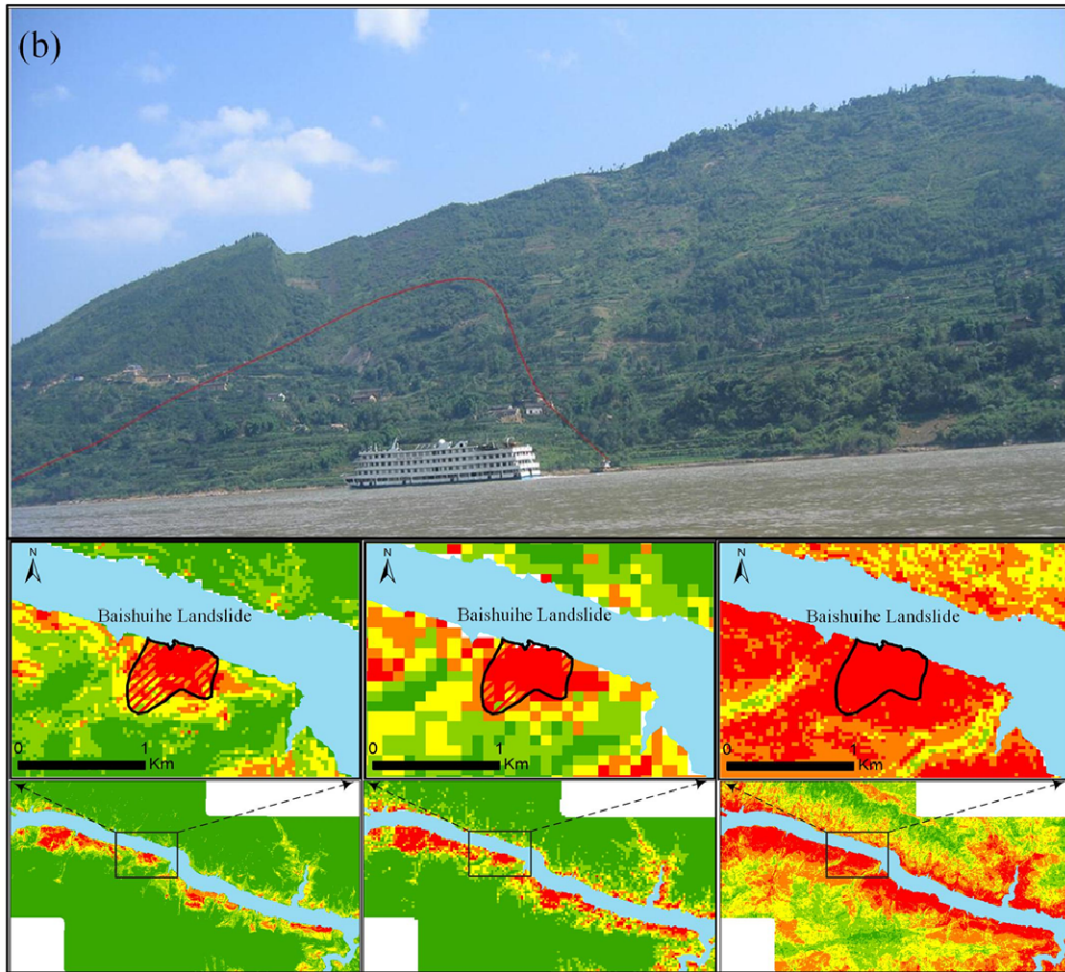


Figure 5 Typical landslides in the study area in Zigui County. (a) Qianjiangping landslides and (b) Baishuihe landslides. The results at the bottom of (a) and (b) are gradient boosting decision tree (GBDT), random forest (RF) and information value (InV) models from left to right. Original pictures were from the Headquarters of Geological Hazard Prevention and Treatment in the TGR.

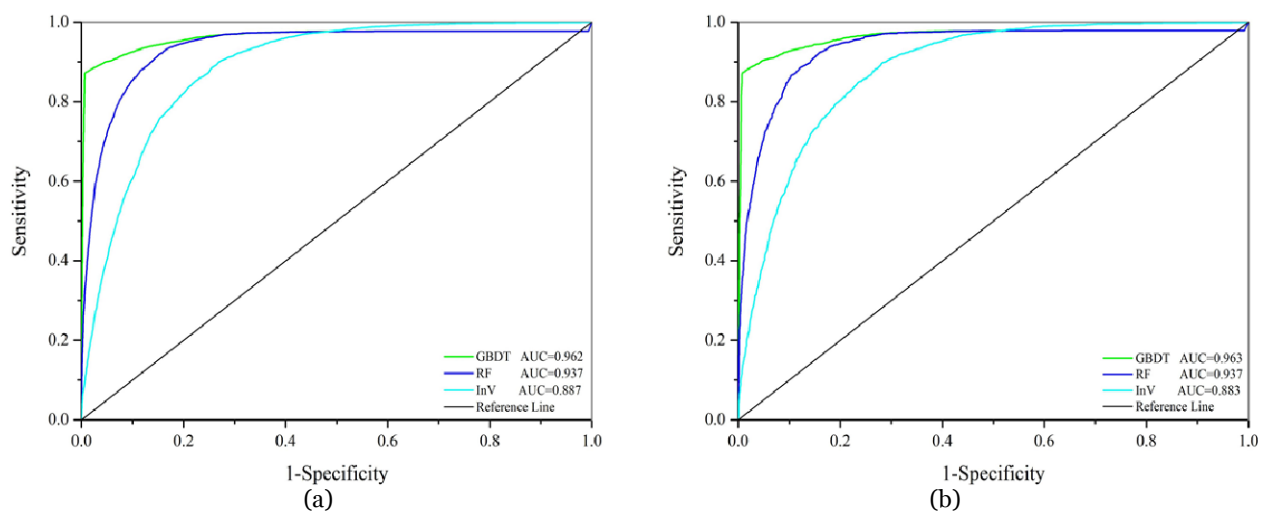


Figure 6 Receiver operating characteristic (ROC) and area under the ROC curve (AUC) for the results of landslide susceptibility mapping based on gradient boosting decision tree (GBDT), random forest (RF) and information value (InV) models. (a) Success-rate curve; (b) Prediction-rate curve.

Table 5 Accuracy statics of GBDT (gradient boosting decision tree), RF (random forest), and InV (information value) models based on the training and validation datasets. The bolded and underlined number means the highest accuracy in each row.

Parameters	Training datasets			Validation datasets		
	Models			Models		
	GBDT	RF	InV	GBDT	RF	InV
Area	<u>0.962</u>	0.937	0.887	<u>0.963</u>	0.937	0.883
SE	<u>0.001</u>	<u>0.001</u>	0.002	<u>0.002</u>	<u>0.002</u>	0.003
95% Lower CL	<u>0.960</u>	0.934	0.884	<u>0.959</u>	0.932	0.876
95% Upper CL	<u>0.964</u>	0.939	0.890	<u>0.967</u>	0.942	0.889

Notes: SE=Standard error; CL=Confidence Level. The bolded and underlined number means the highest accuracy in each row.

Table 6 Model performance based on the training and validation datasets.

Parameters	Training datasets			Validation datasets		
	Models			Models		
	GBDT	RF	InV	GBDT	RF	InV
True positive (TP)	19764	18424	22102	4945	4611	5523
True negative (TN)	22678	21063	12968	5664	5225	3159
False positive (FP)	143	1758	9853	41	480	2546
False negative (FN)	3057	4397	719	760	1094	182
<i>Sensitivity (%)</i>	86.6	80.7	<u>96.8</u>	86.7	80.8	<u>96.8</u>
<i>Specificity (%)</i>	<u>99.4</u>	92.3	56.8	<u>99.3</u>	91.6	55.4
<i>Overall accuracy (%)</i>	<u>93.0</u>	86.5	76.8	<u>93.0</u>	86.2	76.1

Note: The bolded and underlined number means the highest accuracy in each row.

the highest *specificity* (0.994/0.993), *OA* (0.93/0.93), and *KAPPA* (0.86/0.86) values with an AUC of 0.962/0.963, which means that it has the best results among these three methods. The InV method achieved the lowest values compared with the GBDT and RF methods for the two datasets, but the values of the *sensitivity* from the InV model for the two datasets are the highest among these three methods, indicating that it can map the largest number of landslide susceptibility areas, while some very low or low landslide susceptibility areas are mapped as landslides with high or very high susceptibility. This finding explains why the values of the other four evaluation indices for the InV method are the lowest among the three methods.

Table 7 Hypothesis test summary. Three models are GBDT (Gradient boosting decision tree), RF (random forest), InV (information value models) respectively.

	Null hypothesis	Test	Sig.	Decision
1	Distribution of different values across GBDT and InV are equally likely	Related-samples McNemar test	0.000	Reject the null hypothesis
2	Distribution of different values across GBDT and RF are equally likely	Related-samples McNemar test	0.000	Reject the null hypothesis
3	Distribution of different values across RF and InV are equally likely	Related-samples McNemar test	0.000	Reject the null hypothesis

Note: Asymptotic significances are displayed. The significance level is 0.05.

The overall outcomes of this study indicate that the results of the GBDT, RF and InV methods are satisfactory for LSM. The distribution of the landslide-prone areas was in line with previous studies in this area (Peng et al. 2014; Chen et al. 2015; Chen et al. 2016; Wang et al. 2017), which confirms that the landslide-prone areas are distributed along the Yangtze River, and the areas farther away from the Yangtze River are subject to a lower landslide risk. It can be concluded that the comparison of these three methods can provide a promising way to generate landslide susceptibility maps in landslide hazard prone zones of the TGR, China. More attention should be paid by policy makers to the landslide-prone areas determined for this area.

However, some uncertainties remain. First, the selection of causative factors may introduce some uncertainties in the LSM because the importance of each causative factor

with the three methods is not the same (Figure 7). The causative factor of elevation achieves the highest importance index, followed by the distance to the river, fractional vegetation cover, slope, lithology, bedding structure, plan curvature, and aspect. The profile curvature achieves the lowest importance index. This finding means that in the TGR area, the causative factor of elevation plays an important role in landslide susceptibility mapping because of the wide range of the elevation. In the TGR area, landslides often occurred along the river mostly because of cyclical fluctuations in the reservoir water level, so we can see from Figure 7 that the causative factor of distance to the river ranked 2nd among the nine causative factors. Although their importance can be calculated, and

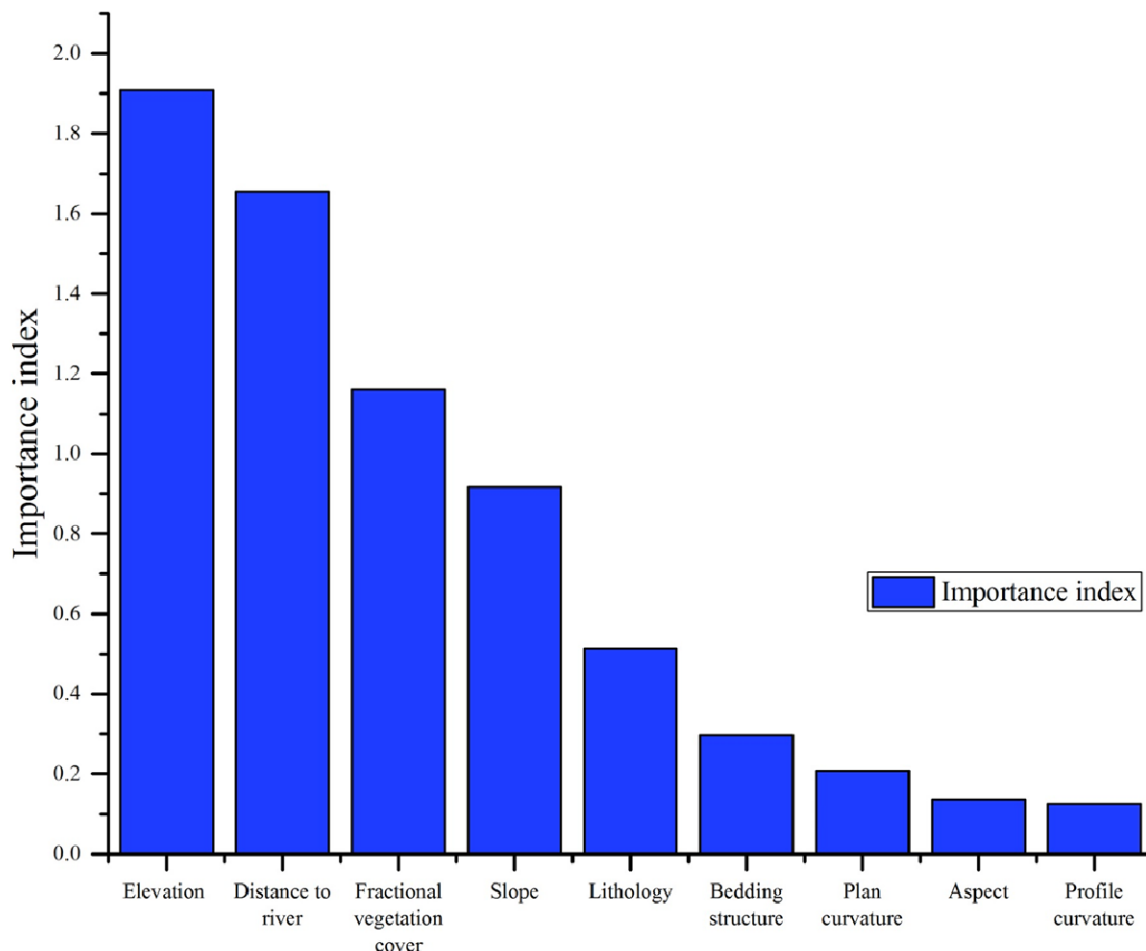


Figure 7 Importance of the causative factors in landslide susceptibility mapping.

some researchers have used many methods to select factors, errors are inevitable, which will cause a decrease in the LSM accuracy (Bui et al. 2016; Dou et al. 2015; Pham et al. 2019). Second, in this work, different types of landslides were not classified when we applied the three LSM methods, which may cause uncertainty because previous studies have proven that landslide susceptibility can be evaluated individually for each landslide type based on their different spatial incidence (Martha et al. 2010; Ahmed and Dewan 2017). Third, the complete and coherent landslide inventory data in this area were deficient. This deficiency may also result in uncertainties, while some new landslides in this study area may not have been identified in the landslide inventory datasets.

5 Conclusions

The objective of this study was to apply three

machine learning methods on LSM at the TGR area, China, and evaluate their performance. The three methods have never before been compared in the entire literature related to landslide susceptibility studies for the study area. In total, 202 landslide areas and 9 landslide causative factors were applied to train the methods. Then, the ROC curves, *sensitivity*, *specificity*, *OA*, and *KAPPA* values were used to evaluate the performance of the methods. In the current study, the GBDT model obtained the highest AUC of the success-rate curve (0.962), AUC of the prediction-rate curve (0.963), *specificity* (0.994), *OA* (0.93) and *KAPPA* (0.86) and moderate *sensitivity* (0.866), while the InV method achieved the lowest AUC value for the success-rate curve (0.887) and prediction-rate curve (0.883), *specificity* (0.568), *OA* (0.768), and *KAPPA* (0.537), and the highest *sensitivity* (0.968). However, all three methods can provide reasonable LSM results in the study area. The outcomes of this work could be helpful to city planners and

engineers in future development and land-use

planning for the TGR area in China.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61601418, 41602362, 61871259), in part by the Opening Foundation of Hunan Engineering and Research Center of Natural Resource Investigation and Monitoring (2020-5), in part by the Qilian

Mountain National Park Research Center (Qinghai) (grant number: GKQ2019-01), and in part by the Geomatics Technology and Application Key Laboratory of Qinghai Province, Grant No. QHDX-2019-01.

References

- Ahmed B, Dewan A (2017) Application of bivariate and multivariate statistical techniques in landslide susceptibility modeling in Chittagong City Corporation, Bangladesh. *Remote Sensing* 9(4): 304. <https://doi.org/10.3390/rs9040304>
- Althuwaynee F, Pradhan B, Lee S (2012) Application of an evidential belief function model in landslide susceptibility mapping. *Computers & Geosciences* 44(44): 120–135. <https://doi.org/10.1016/j.cageo.2012.03.003>
- Bai SB, Wang J, Lu GN, et al. (2010) GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges Area, China. *Geomorphology* 115: 23–31. <https://doi.org/10.1016/j.geomorph.2009.09.025>
- Bi RN, Schleier M, Rohn J, et al. (2014) Landslide susceptibility analysis based on ArcGIS and Artificial Neural Network for a large catchment in Three Gorges region, China. *Environmental Earth Sciences* 72(6): 1925–1938. <https://doi.org/10.1007/s12665-014-3100-5>
- Brabb E (1991) The world landslide problem. *Episodes* 14:52–61
- Breiman L, Friedman H, Olshen A (1984) Classification and regression trees. Chapman & Hall, New York.
- Bui D, Tuan T, Klempe H (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. <https://doi.org/10.1007/s10346-015-0557-6>
- Chen T, Niu R, Du B (2015) Landslide spatial susceptibility mapping by using GIS and Remote Sensing techniques: a case study in Zigui County, the Three Georges reservoir, China. *Environmental Earth Sciences* 73(9): 5571–5583. <https://doi.org/10.1007/s12665-014-3811-7>
- Chen T, Niu R, Jia X (2016) A comparison of information value and logistic regression models in landslide susceptibility mapping by using GIS. *Environmental Earth Sciences* 75(10): 1–16. <https://doi.org/10.1007/s12665-016-5317-y>
- Chen W, Peng J, Hong H (2018) 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, Pourghasemi H, Kornejady A (2017a) Landslide spatial modeling: introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. *Geoderma*, 305: 314–327. <https://doi.org/10.1016/j.geoderma.2017.06.020>
- Chen W, Pourghasemi H, Naghibi S (2017b) A comparative study of landslide susceptibility maps produced using support vector machine with different kernel functions and entropy data mining models in China. *Bulletin of Engineering Geology and the Environment* 77(2):647–664. <https://doi.org/10.1007/s10064-017-1010-y>
- Chung C, Fabbri A (2003) Validation of spatial prediction models for landslide hazard mapping. *Natural Hazards* 30(3): 451–472. <https://doi.org/10.1023/B:NHAZ.0000007172.62651.2b>
- Devkota K, Regmi A, Pourghasemi H (2013) Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling–Narayanghat road section in Nepal Himalaya. *Natural Hazards* 5(1): 135–165. <https://doi.org/10.1007/s11069-012-0347-6>
- Dou J, Yamagishi H, Pourghasemi HR, et al. (2015) An integrated artificial neural network model for the landslide susceptibility assessment of Osado Island, Japan. *Natural Hazards* 78(3): 1749–1776. <https://doi.org/10.1007/s11069-015-1799-2>
- Dou J, Yunus AP, Bui DT, et al. (2019) 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>
- Elith J, Leathwick J, Hastie T (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* 77: 802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- Felicísimo A, Cuartero A, Remondo J (2013) Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. *Landslides* 10(2): 175–189. <https://doi.org/10.1007/s10346-012-0320-1>
- Friedman J (2002) Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38: 367–378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
- Guzzetti F, Reichenbach P, Ardizzone F (2006) Estimating the quality of landslide susceptibility models. *Geomorphology* 81(1): 166–184. <https://doi.org/10.1016/j.geomorph.2006.04.007>
- He K, Li X, Yan X (2008) The landslides in the Three Gorges Reservoir Region, China and the effects of water storage and rain on their stability. *Environmental Geology* 55: 55–63. <https://doi.org/10.1007/s00254-007-0964-7>
- Hong H, Ilia I, Tsangaratos P (2017) A hybrid fuzzy weight of evidence method in landslide susceptibility analysis on the Wuyuan area, China. *Geomorphology* 290: 1–16. <https://doi.org/10.1016/j.geomorph.2017.04.002>
- Hong H, Liu J, Bui D (2018) Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *Catena* 163: 399–413. <https://doi.org/10.1016/j.catena.2018.01.005>
- Hung L, Van N, Duc D (2016) Landslide susceptibility mapping by combining the analytical hierarchy process and weighted linear combination methods: a case study in the upper Lo River catchment (Vietnam). *Landslides* 13(5): 1285–1301. <https://doi.org/10.1007/s10346-015-0657-3>
- Kausar N, Majid A (2016) Randomforest-based scheme using

- feature and decision levels information for multi-focus image fusion. *Pattern Analysis and Applications* 19(1): 221–236.
<https://doi.org/10.1007/s10044-015-0448-4>
- Lee J, Sameen M, Pradhan B (2018) Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. *Geomorphology* 303: 284–298.
<https://doi.org/10.1016/j.geomorph.2017.12.007>
- Lee M, Park I, Lee S (2015) Forecasting and validation of landslide susceptibility using an integration of frequency ratio and neuro-fuzzy models: a case study of Seorak mountain area in Korea. *Environmental Earth Sciences* 74(1): 413–429.
<https://doi.org/10.1007/s12665-015-4048-9>
- Liu C, Liu Y, Wen M (2009) Geo-Hazard Initiation and Assessment in the Three Gorges Reservoir. In *Landslide Disaster Mitigation in Three Gorges Reservoir, China*; Wang, F.W., Li, T.L., Eds.; Springer: Berlin/Heidelberg, Germany, 3–40
- Martha T, Kerle N, Jetten V (2010) Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods. *Geomorphology* 116(1–2): 24–36.
<https://doi.org/10.1016/j.geomorph.2009.10.004>
- Neuhauser B, Damm B, Terhorst B (2012) GIS-based assessment of landslide susceptibility on the base of the Weights-of-Evidence model. *Landslides* 9(4): 511–528.
<https://doi.org/10.1007/s10346-011-0305-5>
- Pawluszek K, Borkowski A (2017) Impact of DEM-derived factors and analytical hierarchy process on landslide susceptibility mapping in the region of Rożnów Lake, Poland. *Natural Hazards* 86(2): 919–952.
<https://doi.org/10.1007/s11069-016-2725-y>
- Peng L, Niu R, Huang B (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.
<https://doi.org/10.1016/j.geomorph.2013.08.013>
- Pham BT, Prakash I, Dou J, et al. (2019) A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. *Geocarto International* 1–25.
<https://doi.org/10.1080/10106049.2018.1559885>
- Pradhan B, Lee S (2010) Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models. *Environmental Earth Sciences* 60(5): 1037–1054.
<https://doi.org/10.1007/s12665-009-0245-8>
- Sarkar S, Kanungo D, Patra A (2006) GIS Based Landslide Susceptibility Mapping - A Case Study in Indian Himalaya in Disaster Mitigation of Debris Flows, Slope Failures and Landslides, Universal Academic Press, Tokyo, 617–624
- Schapire R (2003) The boosting approach to machine learning: an overview. *Nonlinear Estimation and Classification* 171: 149–171.
https://doi.org/10.1007/978-0-387-21579-2_9
- Smith M, Goodchild M, Longley P (2014) *Geospatial analysis—the comprehensive guide to principles, techniques and software tools*. Webversion (Oct 2007). *Trans GIS*. 12(5): 645–647.
- Tsangaratos P, Ilia I (2016) Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. *Landslides* 13(2): 305–320.
<https://doi.org/10.1007/s10346-015-0565-6>
- Vasu N, Lee S (2016) A hybrid feature selection algorithm integrating an extreme learning machine for landslide susceptibility modeling of Mt. Woomyeon, South Korea. *Geomorphology* 263: 50–70.
<https://doi.org/10.1016/j.geomorph.2016.03.023>
- Wang Q, Wang Y, Niu R (2017) Integration of Information Theory, K-Means Cluster Analysis and the Logistic Regression Model for Landslide Susceptibility Mapping in the Three Gorges Area, China. *Remote Sensing* 9(9): 938.
<https://doi.org/10.3390/rs9090938>
- Wang Y, Fang Z, Hong H (2019) Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. *Science of The Total Environment* 666: 975–993.
<https://doi.org/10.1016/j.scitotenv.2019.02.263>
- Wu S, Shi L, Wang R (2001) Zonation of the landslide hazards in the fore reservoir region of the Three Gorges Project on the Yangtze River. *Engineering Geology* 59: 51–58.
[https://doi.org/10.1016/S0013-7952\(00\)00061-2](https://doi.org/10.1016/S0013-7952(00)00061-2)
- Wu X, Niu R, Ren F (2013) Landslide susceptibility mapping using rough sets and back-propagation neural networks in the Three Gorges, China. *Environmental Earth Sciences* 70(3): 1307–1318. <https://doi.org/10.1007/s12665-013-2217-2>
- Xu K, Guo Q, Li Z (2015) Landslide susceptibility evaluation based on BPNN and GIS: a case of Guojiaba in the Three Gorges Reservoir Area. *International Journal of Geographical Information Science* 29(7): 1111–1124.
<https://doi.org/10.1080/13658816.2014.992436>
- Yalcin A (2008) GIS-based landslide susceptibility mapping using analytical hierarchy process and bivariate statistics in Ardesen (Turkey): Comparisons of results and confirmations. *Catena* 72(1): 1–12.
<https://doi.org/10.1016/j.catena.2007.01.003>
- Yan T (1988) Recent advances of quantitative prognoses of landslide in China. In: *Proceedings of the fifth international symposium on landslides, Lausanne, Switzerland*, 2: 1263–1268.
- Yin K, Yan T (1988) Statistical prediction models for slope instability of metamorphosed rocks. In: *Proceedings of the fifth international symposium on landslides, Lausanne, Switzerland*. 2: 1269–1272
- Youssef A, Al-Kathery M, Pradhan B (2015) Landslide susceptibility mapping at Al- Hasher area, Jizan (Saudi Arabia) using GIS-based frequency ratio and index of entropy models. *Geosciences Journal* 19 (1): 113–134.
<https://doi.org/10.1007/s12303-014-0032-8>
- Youssef A, Pourghasemi H, Pourtaghi Z (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.
<https://doi.org/10.1007/s10346-015-0614-1>
- Yunus AP, Fan X, Tang X, et al. (2020) Decadal vegetation succession from MODIS reveals the spatio-temporal evolution of post-seismic landsliding after the 2008 Wenchuan earthquake. *Remote Sensing of Environment* 236: 111476.
<https://doi.org/10.1016/j.rse.2019.111476>
- Zêzere J, Pereira S, Melo R (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 G, Cai Y, Zheng Z (2016a) Integration of the Statistical Index Method and the Analytic Hierarchy Process technique for the assessment of landslide susceptibility in Huizhou, China. *Catena* 142: 233–244.
<https://doi.org/10.1016/j.catena.2016.03.028>
- Zhang KX, Wu XL, Niu RQ, et al. (2017) The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China. *Environmental Earth Sciences* 76(11): 1–20.
<https://doi.org/10.1007/s12665-017-6731-5>
- Zhang M, Cao X, Peng L (2016b) Landslide susceptibility mapping based on global and local logistic regression models in Three Gorges Reservoir area, China. *Environmental Earth Sciences* 75(11): 1–11.
<https://doi.org/10.1007/s12665-016-5764-5>
- Zhou C, Yin K, Cao Y (2018) Landslide susceptibility modeling applying machine learning methods: a case study from Longju in the Three Gorges Reservoir area, China. *Computers & Geosciences* 112: 23–37.
<https://doi.org/10.1016/j.cageo.2017.11.019>