IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 22, 2019, accepted January 23, 2020, date of publication February 6, 2020, date of current version February 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972005

Linking the Random Forests Model and GIS to Assess Geo-Hazards Risk: A Case Study in Shifang County, China

PEI HUANG[1,2], LI PENG[1,2], AND HONGYI PAN[1]

[1]College of Geography and Resources, Sichuan Normal University, Chengdu 610101, China
[2]Institute of Mountain Hazards and Environment, Chinese Academy of Sciences, Chengdu 610041, China

Corresponding author: Li Peng (pengli@imde.ac.cn)

This work was supported in part by the Science and Technology Service Network Initiative under Grant KFJ-STS-QYZD-060, in part by the Sichuan Philosophy and Social Science 397 Planning Project under Grant SC18B095, and in part by the Cultivation Funding of Excellent Graduate Thesis by Sichuan Normal University under Grant CSY201903-28.

ABSTRACT This study proposes an objective and accurate geo-hazards risk assessment method to address the challenge of increasingly severe hazards around the world. Previous studies mostly began from the perspectives of hazard and vulnerability, ignoring the role of survey data at disaster sites during risk assessments. The random forests (RF) model was applied in this study. Combined with detailed data from hazard sites, a geo-hazards risk assessment model was constructed, with the two dimensions of disaster hazard and vulnerability, was constructed. We analyzed the spatial pattern characteristics and the internal patterns of disaster risk and discussed the risk controlling factors and their contributions. The results showed the following. (1) The RF model, when combined with hazard, vulnerability conditions, and detailed data from disaster sites, can be used to zone and verify regional geo-hazards risks, providing a method for point-to-surface disaster risk mapping. (2) The RF-based geo-hazards risk assessment results were relatively consistent with the evaluation results from the support vector machine (SVM) model, but the accuracy and stability of the RF model were higher. (3) This method can be used to avoid the subjectivity in determining the weights and threshold values for indexes and can calculate the contribution of each index to geo-hazards risks.

INDEX TERMS Geo-hazards, random forests model, risk assessment, Shifang county, support vector machine model.

I. INTRODUCTION

Geo-hazards, such as landslides, debris flows, collapse, ground fractures, are an important type of natural hazards. The occurrence of geo-hazards directly induces infrastructure damage and property loss and is sometimes life-threatening [1]. As a result of the aggravation by global climate change and the impacts of human activities, the frequency of geo-hazards has shown an increasing trend, which will also increase the harm to society [2]. In 1992, the United Nations Department for Humanitarian Affairs (UNDHA) published the definition of disaster risk: Risk is the expected losses value of people's lives, property and economic activities caused by specific natural disasters in a given region and in a given period of time, and disaster risk regarded as a function of hazard and vulnerability [3]. This definition is widely accepted by scholars and institutions. Geo-hazards risk assessments have long been the focus of scholars and government agencies [4]. By means of 3S technologies (global positioning system (GPS), geographic information system (GIS), and remote sensing (RS)), constructing an assessment index system for regional hazard risk from the perspectives of hazard and vulnerability, delimiting the range of each index threshold, determining the weight of each index, and realizing the monitoring and assessing regional geo-hazards risks, has become the most widely used assessment method [5]–[7]. Unfortunately, this assessment mode is very subjective. For example, in terms of the slope index, because of the lack of a unified standard for the slope threshold, different scholars have reached very different results when determining the ranges of a slope, resulting in uncertainty in the assessment results. There are numerous risk assessment indices for geo-hazards, including hazard-forming environments, hazard factors, and hazard-bearing bodies [8]–[10].

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

However, it remains a challenge to scientifically determine the main factors controlling hazard risk and their individual contributions. Geo-hazards risk assessments involve many scales, including cities [11], terrain areas [12], communities [13], watershed areas [14], and grids [15]. Scholars have applied many assessment models, such as the analytic hierarchy process [16], geostatistics [17], logistics regression [18], the fuzzy comprehensive evaluation method [19], the grey system theory [20], the projective strategy [21], and the attribute interval evaluation theory (AIET) [22]. In addition, the development of artificial intelligence (AI) technology offers more possibilities for scientific assessments of geo-hazards risks. Numerous machine learning models, such as decision tree (DT) [23], support vector machine (SVM) [24], artificial neural network (ANN) [25], BP-artificial neural network (BP-ANN) [26], and Bayesian network (BN) models [27], have been applied for geo-hazards risk assessments. Among them, the SVM model, which is an efficient and reliable AI algorithm, has a very strong nonlinear processing ability and is one of the significant methods in risk assessment [28]. This paper compared the assessment results between the SVM model and the RF model. However, the SVM still cannot directly estimate the contribution of each index to the total risk.

RF, which is considered an enhanced bagging technique, is an ensemble machine learning method that uses an ensemble of DT [29]. The nonlinear characteristics of RF make it applicable to multivariate prediction; thus, this approach is applied in many fields [30]–[33]. Compared with other methods, RF has the following advantages. There is no need for dimensionless processing such as normalization, as it can not only process multiple forms of data but also adapt to situations where certain attribute values are missing. RF is suitable for handling high dimensionality and complex data, and can overcome multicollinearity of data. RF is more tolerant to outliers and noise and is unlikely to suffer from overfitting issues. More importantly, the contribution of each index to the total risk can be directly obtained, avoiding the effects from subjective human assignment, and the model exhibits high stability and accuracy [34]. China is one of the countries that is most severely affected by geo-hazards, and both the intensity and frequency of hazard occurrence show increasing trends [35], especially in the southwest region. The casualties and economic losses result from geo-hazards are very serious. According to statistics, in 2015 alone, a total of 8224 geo-hazards occurred in China, resulting in a total of 229 deaths and direct economic losses of 2.49 billion RMB. To this end, the Chinese government has conducted a comprehensive investigation of geo-hazard sites and constructed a database of national hazard sites. This database provides a good verification tool for the geo-hazards risk assessments on different scales [36]. However, the limited hazard sites can reflect the hazard risk on only a point scale, and it is difficult to map risks at the regional level [37]. In addition, the assessment of geo-hazards risks involves multi-index variables and high-dimensional data processing. RF exhibits superior performance compared to many other assessment methods, especially in terms of risk prediction [31]. This study regarded Shifang county in the southwest mountainous area of China as the research area and applied the RF model to the geo-hazards risk assessments. Regional hazard and vulnerability conditions were combined with detailed data from hazard sites to explore new ways to improve the accuracy of geo-hazards risk assessments and hoping to provide a reference for geo-hazards risk management and hazard prevention planning.

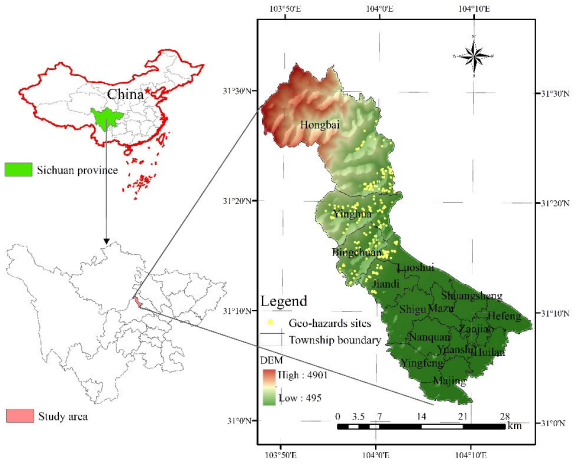


FIGURE 1. Geographic location of Shifang county.

II. MATERIALS AND METHODS

A. STUDY AREA AND DATASETS

Shifang county is affiliated with Deyang city, Sichuan Province, on the eastern margin of the Hengduan Mountain region (103°50′ 30″- 104°26′ 40″E, 31°08′- 31°28′ 24″N), covers an area of 864 km^2 and governs 14 towns and 2 districts with a total population of 430,000. The terrain gradually changes from northwest to southeast, with a sequential transition of mountains, hills, and plains. The altitude difference is large. Mountainous areas account for approximately 60% of the total area of the city. Shifang county is located in a subtropical monsoon climate zone, with heavy rainstorms in the summer. The yearly rainfall is 830 - 1200 mm. This region is located on the front edge of the middle segment of the Longmenshan fault zone, where neotectonic movement is strong, faults, folds and fissures are developed, surface weathering is strong, and rock masses are broken. Due to the influence of various human activities, the occurrence of hazards, such as collapses, landslides, debris flows, and ground subsidence is intensified. Since the Wenchuan earthquake, geo-hazards in Shifang county have increased. According to the detailed results of geo-hazards, Shifang county has a total of 310 hazard sites, which mainly experienced small geo-hazards. The geographic location of Shifang county is shown in Fig. 1.

The data in this study mainly include natural factor data and socioeconomic data. Among the natural factor data, slope,

the distance to river, and altitude difference are extracted from digital elevation model (DEM) data. The data set is provided by the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC), with a spatial resolution of 30 m × 30 m. The precipitation station data are from the Earth System Research Laboratory Physical Sciences Division (ESRL PSD), and the precipitation data are interpolated by using ANUSPLIN software. Geological data mainly include seismic intensity, lithology, and fault data. Lithology and fault data are mainly from the National Geological Archives of China. Seismic intensity data are from the China Earthquake Administration. Normalized difference vegetation index (NDVI) data are downloaded from the Level-1 and Atmosphere Archive & Distribution System Distributed Active Archive Center (LAADS DAAC). Per capita GDP and population density in the study area are obtained as the socioeconomic data from the spatial distribution of the discretized GDP and population data. Both GDP and population data are from the Statistical Yearbook. Building density, cultivated land density, and roading density are extracted from the 2015 land-use data provided by the Ministry of Land and Resources.

B. METHODOLOGY

In previous studies, index threshold values delimiting and weight assignment were subjective, and detailed data from geo-hazard sites were ignored during risk assessments. In view of this, we mainly used the following process for risk assessment. First, from the perspectives of hazard and vulnerability, we selected the codrivers that affect the various subcategories of geo-hazards in the study area. However, we did not assign threshold values and weights to each index, thus avoiding the subjectivity of the assessment results. At the same time, detailed data from hazard sites were introduced to calculate the influence coefficient for each hazard site. Next, based on the assessment index system and the hazard influence coefficient, the geo-hazards risk samples in the study area were evenly selected using the grid as the unit, and the risk level was assigned. In the total sample set, the grids with and without hazard sites were all considered. Third, based on the five-fold cross-validation method, the total risk sample set was input into the RF model. The training samples were used for modeling, validation samples were used for validating, and then the accuracy of the model was obtained. To highlight the advantages of RF in risk assessments of geo-hazards, the SVM algorithm was selected for comparative analysis. Finally, all the data to be tested in the study area were input into the above two models, and the risk results were analyzed and compared to clarify the spatial distribution pattern of hazard risk. A flowchart of this process is shown in Fig. 2.

1) ESTABLISHMENT OF THE EVALUATION INDEX SYSTEM

There are various types of geo-hazards in the study area, including landslides, ground collapses, debris flows, and unstable slopes. It is well known that different hazard types

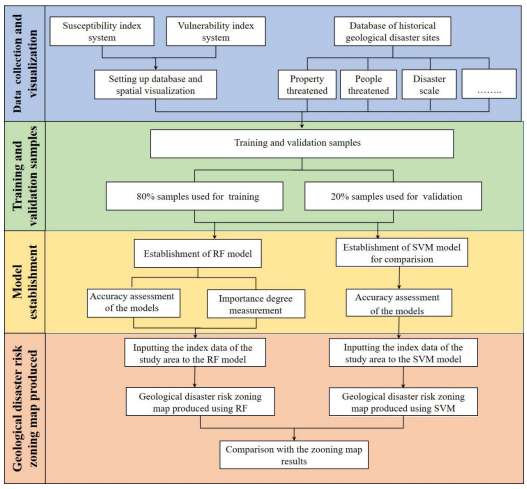


FIGURE 2. Flowchart.

have different driving mechanisms. Traditional geo-hazards risk assessments mainly define thresholds and assign weights. However, the thresholds of the same impact factor for different hazards are certainly different, and it is unscientific to artificially delimit unified threshold ranges for multiple hazards. This study attempted to evaluate the comprehensive risk of all subcategories of geo-hazards. Unfortunately, at present, there is still a lack of a unified standard system for geo-hazards risk assessment. We referred to the connotation of disaster risk published by the UNDHA to construct an evaluation index system from the two dimensions of hazard susceptibility and vulnerability. Therefore, we comprehensively selected the codriving factors for each subcategory of disasters. Using RF, we effectively avoided threshold delimiting and weight assignment for various indices. From the relevant previous studies [38]–[41] and with the help of GIS technology, 13 indices were selected (Table 1). Among them, the hazard index system has 8 indices, including slope, altitude difference, precipitation, seismic intensity, etc; the vulnerability index system has 5 indices, including per capita GDP, population density, building density, etc. The spatial distribution of each index is shown in Fig. 3.

Based on the vector range of the study area, a grid dataset for Shifang county was created. The size of each grid cell was 300 m × 300 m, and there were a total of 9576 grid cells. The indices were spatially distributed, and the index data for each grid unit were extracted to construct the database for the risk assessment index of Shifang county.

2) SAMPLING AND RISK LEVEL DETERMINATION

The detailed data on geo-hazards that were obtained by the relevant government agencies through field surveys included various attributes, such as the location and scale of hazard sites, the number of people threatened, and the amount of property threatened. Detailed data on hazard sites provide support for geo-hazards risk assessments, which can make the assessment results more consistent with the actual hazard situations. In addition, during the construction of the RF model,

TABLE 1. List of the various indices.

Target layer	Criteria layer	Index	Unit	Meaning
Geo-hazards Risk	Hazard	Slope (SL)	°	Within a certain threshold range, with increasing slope, the probability of occurrence of geo-hazards increases.
		Altitude difference (AD)	m	On a regional scale, the greater the AD value is, the more likely a hazard will occur.
		Precipitation (PR)	mm	There is a positive correlation between the amount of summer precipitation (May-September) and the number of hazards.
		Seismic intensity (SI)	g	The higher the degree of seismic intensity is, the stronger the destructiveness and the greater the impact of an earthquake.
		Lithology (LI)	/	LI is closely related to hazards, the looser the rock mass is, the more the area is prone to hazards.
		Distance to fault (DF)	km	The closer to the fault is, the more frequent geo-hazards are.
		Distance to river (DR)	m	The development of geological hazard sites is closely related to DR, but the way of impact is uncertain.
		Normalized difference vegetation index (NDVI)	/	NDVI is closely related to the development of geological hazard sites, but its impact is uncertain.
	Vulnerability	Per capita GDP (PCGDP)	yuan	In the similar hazard-forming environments, the higher the per capita GDP is, the greater the threat and the more serious the losses.
		Population density (PD)	/	In the similar hazard-forming environments, the higher the PD is, the greater the damage caused by hazards.
		Building density (BD)	/	In the similar hazard-forming environments, the greater the BD is, the greater the damage caused by hazards.
		Cultivated land density (CLD)	/	In the similar hazard-forming environments is, the greater the CLD is, the greater the damage caused by hazards.
		Road density (RD)	/	In the same hazard-forming environments is, the greater the RD is, the greater the losses caused by hazards.

sample selection is the key step. The core of constructing an RF evaluation model based on an index variable is to evaluate the corresponding risk level for each sample. The attributes of hazard sites also provide the basis for the determination and verification of sample risk levels. Therefore, based on the risk assessment indices in the study area, the attributes of the hazard sites, and government survey data, geo-hazards risk samples were evenly selected from the study area and the corresponding risk levels were assigned to the samples: low risk, medium risk, high risk and highest risk. During the process of selecting sample data, we chose areas with and without the occurrence of geo-hazards. The specific process is shown in Fig. 4.

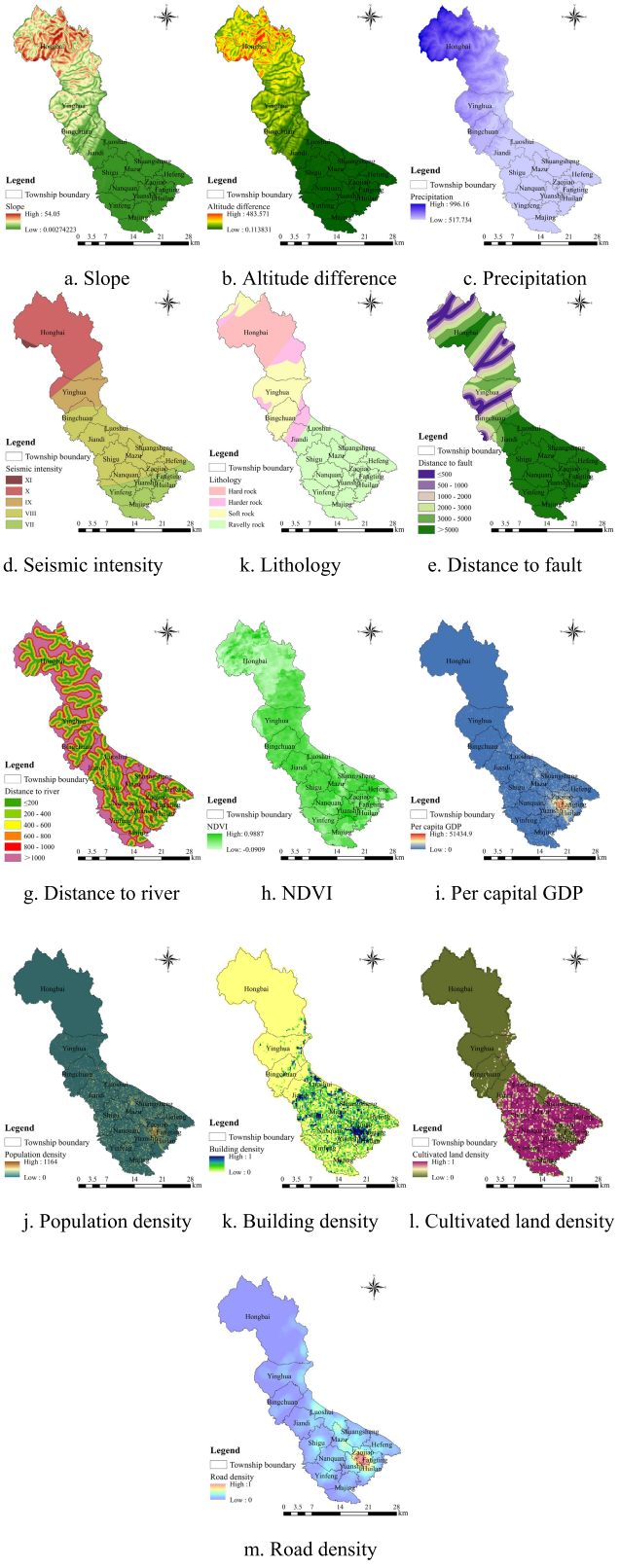


FIGURE 3. Evaluation index.

The hazard risk levels of the samples were determined based on detailed survey data of geo-hazards sites, various evaluation indicators and relevant government data

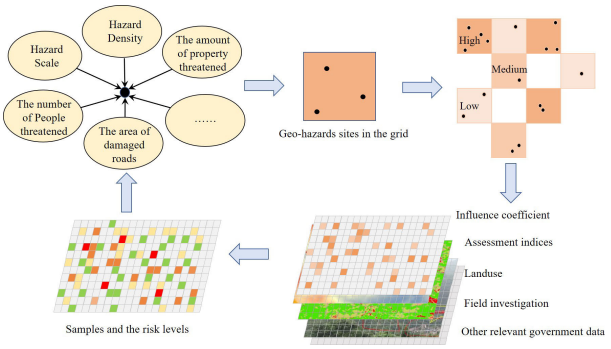


FIGURE 4. Schematic of sample selection.

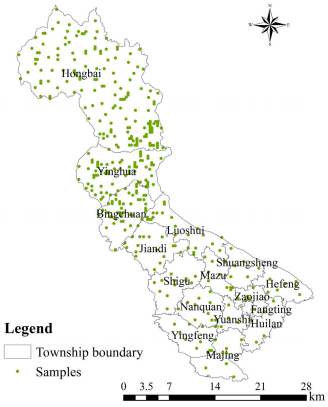


FIGURE 5. Spatial distribution of the geo-hazard samples at different levels.

(reports, planning documents, etc.). The detailed survey data of geo-hazards sites were established by the geological department based on the geo-hazards sites that occurred in various regions. The accuracy of the geo-hazards sites is relatively high and the attributes are comprehensive. First, assigning the risk level is assigned to the selected samples according to the influence coefficient, which was calculated according to the hazard scale, hazard density, number of people threatened, amount of property threatened, number of damaged houses, area of damaged roads, and other attributes of the detailed survey data of geo-hazards sites in the study area. A total of 400 samples were selected in this study (Fig. 5), with 100 risk samples at each level. Second, the risk levels that were classified were revised through various indicators, relevant government data and previous research results [42]. Then, various indicators and risk levels of the selected samples were input into the model to form the disaster risk classification rules. Finally, according to the above rules, all the data to be measured were inputted into RF model again to predict the level of geo-hazards risk in the study area. Compared with traditional algorithms, the RF model does not need to set index weights or classification criteria in advance. The weight and classification criteria were implicit in the inherent rules of the data.

3) RANDOM FORESTS MODEL

RF is a very effective classifier that is composed of a set of tree-structured classifiers {h (X, Θ k), k = 1, ...},

where {Θk} represents independent, identically distributed random vectors. At the input of independent variable X, each DT will cast a unit vote for the most popular class.

First, the bootstrap sampling method was used to extract k samples from the original training set D. The feature number (m) in each sample was the same as the original training set D. Multiple samples are drawn using the resampling bootstrap method, which improves the randomness of the training set.

Second, k DTs were generated for k samples, and k classification results were obtained {($h1$ X), $h2$ X, …, hn X}. Features (n ≤ m) are randomly chosen from every sample to create the split feature set, from which the optimal features were selected to grow the nodes. When n < m, there were differences between each DT. The minimum Gini value is the split standard of the node, and the corresponding variable is known as the optimal variable. The minimum Gini value of an internal tree node was calculated as follows:

$$Gini(t) = 1 - \sum_{q=1}^{u} [p(q|t)]^2 \tag{1}$$

where p (q | t) represents the probability of the risk class q at node t, and u is the number of classes.

Third, each tree was grown to the largest extent possible and no pruning was conducted.

Fourth, the above steps were repeated to form random forests. The final result was determined in accordance with the majority rule voting mechanism as follows:

$$f(x1) = m\_vote\{hi(x)\} \tag{2}$$

where m_vote represents the result of the vote.

Last, the generalization error and variable importance (called the "index contribution degree" in this study) were calculated using the bagging algorithm to integrate the training set. The probability that a sample is not extracted from the total training set D with a sample size N is $(1\text{-}\ 1/N)^N$ When N is sufficiently large, $(1-1/N)^N \rightarrow \frac{1}{e} = 0.368$. This indicates that more than 1/3 of the samples in set D are left out of the bootstrap sample; these samples are called out-of-bag (OOB) data. After classification tree generation, the OOB data are used to calculate the error classification rate, known as the OOB error. The model generalization error is the average OOB error of all trees in the random forests. Generally, there are two methods to calculate the importance degree of each index. In this study, the decreases in the Gini index at the node split are used to calculate the importance of each index to the result of the risk classification. The formula is as follows:

$$P_r = \frac{\sum_{i=1}^{k}\sum_{j=1}^{t} D_{Grij}}{\sum_{r=1}^{m}\sum_{i=1}^{k}\sum_{j=1}^{t} D_{Grij}} \times 100\% \tag{3}$$

where m represents the total number of indices, k is the number of texturing trees, t is the number of nodes in each tree, D_{Grij} is the Gini decrease value at the jth node in the ith tree that belongs to the rth index, and P_r is the degree of contribution from the rth index from all available indices.

The principle of RF is shown in Fig. 6.

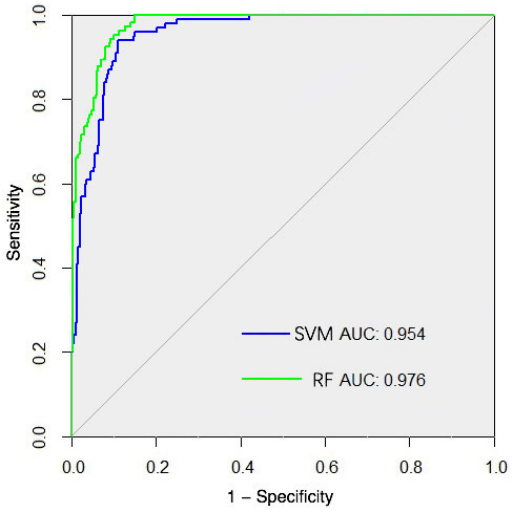


FIGURE 6. Average ROC curves.

TABLE 2. Binary confusion matrix.

	Classified positive	Classified negative
Positive	TP	FN
Negative	FP	TN

4) RROC CURVE AND THE AUC VALUE

The receiver operating characteristics (ROC) curve and the area under the ROC (AUC) are often used to evaluate the performance of the model. The binary confusion matrix (Table 2) can be used to reflect these two indicators.

The positive and the negative represent positive and negative samples, respectively, and the classified positive and classified negative samples represent correctly classified and misclassified samples, respectively. We can thus statistically analyze the TP (true positive), FP (false positive), FN (false negative), and TN (true negative). Furthermore, the TPR (true positive rate) and FPR (false positive rate) can be calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

When the threshold is adjusted, each threshold corresponds to one TPR and FPR set, and the ROC curve is the curve with the TPR as the vertical axis and FPR as the horizontal axis. The area under the curve is the AUC, with a value between 0 and 1, generally greater than 0.5. The closer to 1, the better the performance of the classifier and the higher the accuracy of the model.

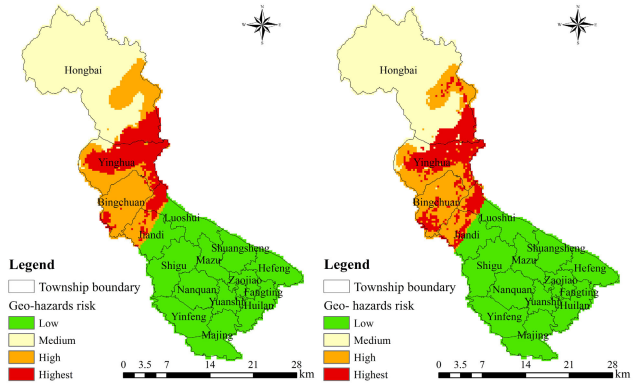


FIGURE 7. Zoning map of the geo-hazard risks. a) Assessment results from the RF model; b) Assessment results from the SVM model.

III. RESULTS

A. ACCURACY ASSESSMENT OF THE MODELS

In an RF model, it is not necessary to use cross-validation to establish an unbiased estimate of the errors because the RF estimates the errors during modeling using the OOB error estimate [43]. However, to construct the same comparative environment as the SVM model, which utilizes the five-fold cross-validation method, the total sample set was randomly divided into 4 training data sets and 1 test data set with equal volume, and the division was repeated 5 times. The sample data were substituted into the RF model and the SVM model for training, and the training accuracy was compared. After many cycles of debugging, optimal parameters were obtained. The number of classification trees in the RF model was set to 1000; the number of splits at the nodes was set to 4; the kernel function in the SVM model was the radial basis function (RBF), with the coefficient $\gamma = 0.1$; and the penalty coefficient $C = 1$.

From the ROC, the data (Fig. 6) show that the average AUCs for both the RF model and SVM model in the 5-fold cross-validation were greater than 0.9, indicating that both models had relatively high evaluation accuracy. However, the average AUC of the RF model was0.976, which was 2.2%greater than that of the SVM model. This result further verified that the RF model has excellent robustness and generalization ability and that its overall accuracy is superior to that of the SVM model.

B. CLASSIFICATION RESULT ANALYSIS

By combining the detailed data from the Shifang county hazard sites, all the data to be tested in the study area were input into the RF and SVM models to obtain two zoning maps of the geo-hazard risks in Shifang county (Fig. 7).

The data in Figs. 7(a)-7(b) show that the spatial distributions of the geo-hazards risk zones at various levels were relatively consistent. The main distribution difference between the two was reflected in the local high risk and highest risk zones. Meanwhile, Table 3 shows that the differences in the proportions of the risk zones at each level were also small.

TABLE 3. Risk area ratios in the RF and SVM models (%).

	Low risk	Medium risk	High risk	Highest risk
RF	44.6	27.97	16.52	10.91
SVM	44.31	29.47	14.43	11.79
Difference rate	0.29	1.5	2.08	0.88

The accuracy (Fig. 6) of the evaluation results of the two models showed that the RF model had a significantly high accuracy than the SVM model. However, to further verify the accuracy of the classification results of the two models, we conducted field verification in the local zones of the study area. For example, the scattered highest risk zones in the north-central region in Figure 7 (b) have high seismic intensity, dense faults, large altitude differences, steep slopes, and many secondary geo-hazards. However, the data in the RS images show that this region was sparsely populated. The population was sporadically distributed along roads, where the risk was lower than that in the highest risk zones to the south. The results of the RF model were more reasonable than the results of the SVM model. The risk assessment of geo-hazards based on RF showed high accuracy and stability.

The risk of geo-hazards in Shifang county exhibited large spatial differentiation, while the spatial agglomeration of each hazard risk level was prominent. The highest risk zones were located in the mid-mountain areas and the local low-mountain or hilly areas in the central region of Shifang county; the area of the highest risk zones was the smallest among all risk levels, i.e., 94.24 km^2, accounting for 10.91% of the total study area. High risk zones were mainly located in the high-mountain area in the north and the medium-mountain area in the central region, with an area of approximately 142.7 km^2, accounting for 16.52% of the total area of the study area. The medium risk zones were concentrated in the mountainous area in the north, with an area of approximately 241.7 km^2, accounting for 27.97% of the total area studied. The low risk zones were concentrated in the southern plains and had the largest area of all the risk zones (385.35 km^2), accounting for 44.6% of the total area studied. From a spatial point of view, the low risk zones and the high and highest risk zones were separated by the dividing line between the plain area and the hilly areas. This result is consistent with the distribution of the geo-hazards sites in the study area.

As a result of the joint action of multiple factors, the central region represents the high and highest risk zone for geo-hazards. This area is a key area for hazard prevention and management. This area is not suitable for major construction projects, and human activities should be minimized. It is also important to strengthen the monitoring of hazard sites, take corresponding engineering and biological measures at the major geo-hazards sites, and adopt resettlement measures to reduce hazard risk if necessary. The population and buildings in the southern plain area are dense with a high intensity of human activities. However, because of the flat terrain, the conditions are insufficient for the formation of geo-hazards. Therefore, the southern plain area becomes a low risk zone.

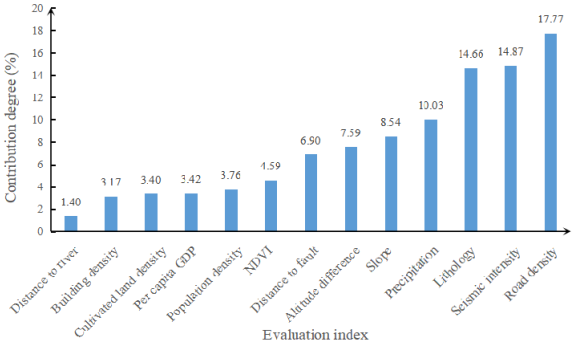


FIGURE 8. Contributions of various indices.

The mountainous area in the northern part of the study area is steep and located in the Longmenshan fault zone, which has a high vulnerability to hazards. However, because of the relatively low precipitation compared to that in the southern and central regions and the sparse population and economic backwardness, there are fewer objects threatened by hazards, making this region a medium risk zone for geo-hazards. In this region, the protection of mountainous vegetation and water and soil loss control should be emphasized and the local water and soil conservation and the water conservation capacities should be enhanced.

C. INDEX CONTRIBUTION DEGREE ANALYSIS

In this paper, according to the mean decreases of Gini index of all nodes in the RF model, the contribution of each index to the geological hazard risk was calculated (Fig. 8). Among the indices, RD, SI, LI, and PR showed the highest contributions to the risk of geo-hazards, all exceeding 10%, indicating that these four indices had the most significant impact on hazard risk in the study area. DR, PCGDP, BD, PD, and CLD had the lowest contributions to the hazard risk; their contribution degrees were less than 5%, indicating that these five indices had weaker impacts on the hazard risk of Shifang county. The cumulative contribution of the first seven indices accounted for 80.44% of the total, indicating that the seven indices played a decisive role in the hazard risk in the study area.

RD had the highest contribution among all indices, reaching 17.77%, indicating that road construction for human activities was the most important factor affecting the risk of geo-hazards in the study area. Road construction and other activities disturbed the soil, increased the degree of surface rock fragmentation, and formed numerous unstable slopes, which provided the necessary conditions for hazard occurrence. According to the statistics, 82 hazard sites existed along the highway in the study area, accounting for 26.45% of the total hazard sites in the study area. Meanwhile, the central and northern parts of the study area were severely affected by the Wenchuan earthquake (earthquake intensity of VII -XI). The contribution degree of the earthquake to hazard risk was 14.87%. After the Wenchuan earthquake, there were a large number of secondary hazards in the earthquake-stricken area, which provided conditions for the outbreak of geo-hazards. Therefore, the level of hazard risk is relatively high, which is consistent with the objective law. The lithology in the study area is dominated by clastic rocks (accounting for 79.84%), which are easily weathered; an increase in clastic matter further exacerbates the probability of hazard occurrence. Therefore, the contribution degree of lithology to hazard risk was relatively high, reaching 14.66%. Precipitation is an important triggering factor for geo-hazards with a contribution rate of 10.03%. Precipitation in the study area (May-September) was greater than 517.73 mm, and the abundant rainfall in the mountainous area provides a powerful hydrodynamic condition for the occurrence of hazards. Meanwhile, fault development in the central and northern parts of the study area also promotes the occurrence of hazards. The differences in altitude and slope were also important factors for the risk of geo-hazards in the study region, with contributions of 7.59% and 8.54%, respectively. Over the terrain with large altitude differences and steep slopes, the geo-hazards have greater potential energy than those in other areas.

IV. CONCLUSION AND DISCUSSION

Based on regional hazard system theory and detailed data on hazard sites, 13 indices were selected from the dimensions of hazard and vulnerability and the RF model was used to perform the risk assessment for Shifang county. The main conclusions are as follows:

1) Based on the RF model, combined with hazard and vulnerability conditions of geo-hazards and detailed data on hazard sites, a method was proposed for point-to-surface mapping of hazard risk.

This method overcomes the shortcoming that it is difficulty to map geo-hazard risks at the regional level based on hazard sites data. This method also avoids the need for threshold segmentation and weight assignment of the risk indices. The contribution of indicators (which can be regarded as weights) is directly calculated by the training function of the RF model, resulting in more objective evaluation results. Although RF models have been widely used in disaster risk assessment, previous studies mainly carried out model accuracy verification and disaster risk assessment based on the sample attributes of whether disasters occurred or not [44], [45]. These studies ignored the risk attribute of the geo-hazard sites themselves and it was difficult to realize point-to-surface disaster risk mapping. However, the detailed survey data on geological disaster points can provide a good verification tool for geo-hazards risk assessments at various scales. The only studies that are similar to the methods in this paper mostly considered the impact ranges of historical disaster points [31], and it was difficult to reflect the heterogeneity of risks within the same impact range. Therefore, the evaluation results may be ambiguous. In contrast, the evaluation method in this paper can be used for objective classification and verification of geo-hazards risks. Due to the lack of a unified standard system for geo-hazards risk assessments, we referred to the connotation of disaster risk that was put forward by UNDHA to constructed an evaluation index system from the two dimensions of hazard and vulnerability. This system needs

further improvement, especially in terms of vulnerability. The geo-hazards risk assessments in our study area was based on the grid scale (300 m×300 m), and it was difficult to spatially distributed the above indicators over each evaluation unit. In the future, we will seek new methods to incorporate geological engineering, refuge sites, age structure, disaster education, and other indicators into the evaluation system.

2) Compared to the SVM model, the RF model exhibited a better effect in terms of the geo-hazards risk assessment results. The RF model could better reflect the spatial differentiation characteristics of hazard risk in the study area. The averaged AUC value for the RF model reached 0.976, which was 2.2% greater than that of the SVM model. The RF model was accurate and stable. The high and highest risk zones of the geo-hazards in Shifang county were mainly located in the central mountainous areas and low-mountain or hilly areas in the central part of the study area, which are adjacent to each other and account for 10.72% and 16.43% of the total area, respectively. The medium risk zones were located in the mountainous area in the north of the study area, accounting for 27.93% of the total area. The area of low risk zones was the largest, and these zones were concentrated in the southern plain area.

The RF model exhibited significantly higher accuracy than the SVM model, which is consistent with the conclusions of other studies [46], [47]. This result likely because the RF model includes an unbiased self-verification function, the OOB error (also known as the generalization error), therefore, the accuracy is high. The OOB error in the RF model used in this study was 3.6%, which implies that the RF model has strong generalization ability and is highly applicable to geo-hazards risk assessment.

3) The RF model can directly calculate the contribution of each index to the geo-hazards risk. The risk of geo-hazards in Shifang county is affected by many factors. RD, SI, LI and PR are the main controlling factors of hazard risk in the study area. These 4 factors all contribute more than 10% to the total hazard risk and have a dominant influence on the hazard risk.

According to the ranking results of the contribution degrees of various indices, the four indices of road density, seismic intensity, lithology and precipitation were found to have the greatest impact on the geological disaster risk of the study area, with a total contribution of 57.33%. This result shows that during disaster risk prevention and management in the future, the above four indicators should be the focus of attention. The number of unstable slopes in the study area accounted for 23.55% of the total hazards, and the road engineering construction increased the geo-hazards risk. Therefore, in the process of disaster prevention and management, attention should be paid to the impact of road construction on disaster risk. The establishment of disaster prevention projects and biological measures is particularly important. Second, the monitoring and warning mechanisms for earthquakes and hazard sites should be improved to reduce the risk of geo-hazards in the study area.

The RF model also has some limitations. The working process for the RF model is a black box operation, and there is no way to control the internal operation of the model. The best parameters are determined by trying different parameters and random seeds. The RF model and the SVM model are important machine learning models, and each has its own strengths. Geo-hazards risk assessments should combine the advantages of various algorithms to explore a more scientific and reasonable method to more accurately cope with geo-hazard risks.

REFERENCES

[1] M. Li, J. Lv, X. Chen, and N. Jiang, "Provincial evaluation of vulnerability to geological disaster in China and its influencing factors: A three-stage DEA-based analysis," *Natural Hazards*, vol. 79, no. 3, pp. 1649–1662, Dec. 2015.

[2] N. Komendantova, R. Mrzyglocki, A. Mignan, B. Khazai, F. Wenzel, A. Patt, and K. Fleming, "Multi-hazard and multi-risk decision-support tools as a part of participatory risk governance: Feedback from civil protection stakeholders," *Int. J. Disaster Risk Reduction*, vol. 8, pp. 50–67, Jun. 2014.

[3] *Internationally Agreed Glossary of Basic Terms Related to Disaster Management*. United Nations Dept. Humanitarian Affairs, Geneva, Switzerland, 1992.

[4] J. Zhu, H. Zhang, X. Yang, L. Yin, Y. Li, Y. Hu, and X. Zhang, "A collaborative virtual geographic environment for emergency dam-break simulation and risk analysis," *J. Spatial Sci.*, vol. 61, no. 1, pp. 133–155, Jan. 2016.

[5] D. S. Hadmoko, F. Lavigne, J. Sartohadi, P. Hadi, and Winaryo, "Landslide hazard and risk assessment and their application in risk management and landuse planning in eastern flank of Menoreh Mountains, Yogyakarta Province, Indonesia," *Natural Hazards*, vol. 54, no. 3, pp. 623–642, Sep. 2010.

[6] R. Hadji, A. E. Boumazbeur, Y. Limani, M. Baghem, A. E. M. Chouabi, and A. Demdoum, "Geologic, topographic and climatic controls in landslide hazard assessment using GIS modeling: A case study of Souk Ahras region, NE Algeria," *Quaternary Int.*, vol. 302, pp. 224–237, Jul. 2013.

[7] C. Niu, Q. Wang, J. Chen, W. Zhang, L. Xu, and K. Wang, "Hazard assessment of debris flows in the reservoir region of wudongde hydropower station in China," *Sustainability*, vol. 7, no. 11, pp. 15099–15118, Nov. 2015.

[8] A. Akgun, C. Kıncal, and B. Pradhan, "Application of remote sensing data and GIS for landslide risk assessment as an environmental threat to Izmir city (west Turkey)," *Environ. Monit. Assessment*, vol. 184, no. 9, pp. 5453–5470, Sep. 2012.

[9] A. Si, J. Zhang, S. Tong, Q. Lai, R. Wang, N. Li, and Y. Bao, "Regional landslide identification based on susceptibility analysis and change detection," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 10, p. 394, Oct. 2018.

[10] D. Tiranti, S. Crema, M. Cavalli, and C. Deangeli, "An integrated study to evaluate debris flow hazard in alpine environment," *Frontiers Earth Sci.*, vol. 6, pp. 60–73, May 2018.

[11] A. Kaitantzian, C. Loupasakis, and D. Rozos, "Assessment of geo-hazards triggered by both natural events and human activities in rapidly urbanized areas," in *Engineering Geology for Society and Territory*, vol. 5. Cham, Switzerland: Springer, 2015, pp. 675–679.

[12] M. Parise, "Karst geo-hazards: Causal factors and management issues," *Acta Carsol.*, vol. 44, no. 3, pp. 401–414, Jan. 2015.

[13] W. Jiang, Y. Deng, Z. Tang, R. Cao, Z. Chen, and K. Jia, "Adaptive capacity of mountainous rural communities under restructuring to geo-hazards: The case of Yunnan Province," *J. Rural Stud.*, vol. 47, pp. 622–629, Oct. 2016.

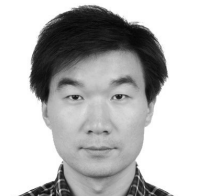
[14] I.-J. Chiou, C.-H. Chen, W.-L. Liu, S.-M. Huang, and Y.-M. Chang, "Methodology of disaster risk assessment for debris flows in a river basin," *Stochastic Environ. Res. Risk Assessment*, vol. 29, no. 3, pp. 775–792, Mar. 2015.

[15] T. K. Raghuvanshi, L. Negassa, and P. M. Kala, "GIS based grid overlay method versus modeling approach—A comparative study for landslide hazard zonation (LHZ) in Meta Robi District of West Showa Zone in Ethiopia," *Egyptian J. Remote Sens. Space Sci.*, vol. 18, no. 2, pp. 235–250, Dec. 2015.

[16] S. S. Moustafa, "Application of the analytic hierarchy process for evaluating geo-hazards in the Greater Cairo area, Egypt," *Electron. J. Geotech. Eng.*, vol. 20, no. 6, pp. 1921–1938, 2015.

[17] V. Chiessi, M. D'Orefice, G. S. Mugnozza, V. Vitale, and C. Cannese, "Geological, geomechanical and geostatistical assessment of rockfall hazard in San Quirico Village (Abruzzo, Italy)," *Geomorphology*, vol. 119, nos. 3–4, pp. 147–161, Jul. 2010.
[18] W. Chen, H. R. Pourghasemi, and Z. Zhao, "A GIS-based comparative study of Dempster–Shafer, logistic regression and artificial neural network models for landslide susceptibility mapping," *Geocarto Int.*, vol. 32, no. 4, pp. 367–385, Apr. 2017.
[19] H. M. Liao, X. G. Yang, F. G. Xu, H. Xu, and J. W. Zhou, "A fuzzy comprehensive method for the risk assessment of a landslide-dammed lake," *Environ. Earth Sci.*, vol. 77, no. 22, 750-763, Nov. 2018.
[20] J. Zhu, Z. Lizhong, Z. Xiaoyuan, L. Guoling, W. Qian, C. Zizhao, and W. Wei, "Application of entropy-based grey model in geological hazard assessment: A case study of Qingchuan County, Sichuan Province," *J. Catastrophol.*, vol. 27, no. 1, pp. 78–82, Jan. 2012.
[21] X. J. Xie, F. Q. Wei, and J. Zhang, "Application of projection pursuit model to landslide risk classification assessment," *Earth Sci.-J. China Univ. Geosci.*, vol. 40, no. 9, pp. 1598–1606, Dec. 2015.
[22] S. C. Li, Z. Q. Zhou, L. P. Li, P. Lin, Z. H. Xu, and S. S. Shi, "A new quantitative method for risk assessment of geo-hazards in underground engineering: Attribute interval evaluation theory (AIET)," *Tunnelling Underground Space Technol.*, vol. 53, pp. 128–139, Mar. 2016.
[23] P. Tsangaratos and I. Ilia, "Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece," *Landslides*, vol. 13, no. 2, pp. 305–320, Apr. 2016.
[24] D. T. Bui, T. A. Tuan, N.-D. Hoang, N. Q. Thanh, D. B. Nguyen, N. Van Liem, and B. Pradhan, "Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization," *Landslides*, vol. 14, no. 2, pp. 447–458, Apr. 2017.
[25] J.-W. Lin and J.-S. Chiou, "Active probability backpropagation neural network model for monthly prediction of probabilistic seismic hazard analysis in Taiwan," *IEEE Access*, vol. 7, pp. 108990–109014, 2019.
[26] M. Liu, Y. He, J. Wang, H. P. Lee, and Y. Liang, "Hybrid intelligent algorithm and its application in geological hazard risk assessment," *Neurocomputing*, vol. 149, pp. 847–853, Feb. 2015.
[27] W.-J. Liang, D.-F. Zhuang, D. Jiang, J.-J. Pan, and H.-Y. Ren, "Assessment of debris flow hazards using a Bayesian Network," *Geomorphology*, vols. 171–172, pp. 94–100, Oct. 2012.
[28] D. Tien Bui, T. A. Tuan, H. Klempe, B. Pradhan, and I. Revhaug, "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*, vol. 13, no. 2, pp. 361–378, Apr. 2016.
[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[30] M. I. Sameen, B. Pradhan, and S. Lee, "Self-learning random forests model for mapping groundwater yield in data-scarce areas," *Natural Resour. Res.*, vol. 28, no. 3, pp. 757–775, Jul. 2019.
[31] Z. Wang, C. Lai, X. Chen, B. Yang, S. Zhao, and X. Bai, "Flood hazard risk assessment model based on random forest," *J. Hydrol.*, vol. 527, pp. 1130–1141, Aug. 2015.
[32] L. Tang, F. Cai, and Y. Ouyang, "Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China," *Technol. Forecasting Social Change*, vol. 144, pp. 563–572, Jul. 2019.
[33] R. Hänsch and O. Hellwich, "Classification of PolSAR images by stacked random forests," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 2, pp. 74–89, Feb. 2018.
[34] B. T. Pham, B. Pradhan, D. T. Bui, I. Prakash, and M. Dholakia, "A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India)," *Environ. Model. Softw.*, vol. 84, pp. 240–250, Oct. 2016.
[35] J. Hou, J. Lv, X. Chen, and S. Yu, "China's regional social vulnerability to geological disasters: Evaluation and spatial characteristics analysis," *Natural Hazards*, vol. 84, no. S1, pp. 97–111, Nov. 2016.
[36] G. P. Chen, J. S. Zhao, L. Yuan, Z. J. Ke, M. Gu, and T. Wang, "Implementation of a geological dis-aster monitoring and early warning system based on multi-source spatial data: A case study of Deqin Country, Yunnan Province," *Hazards Earth Syst. Sci.*, vol. 15, pp. 1–15, Jul. 2017.
[37] O. F. Althuwaynee and B. Pradhan, "Semi-quantitative landslide risk assessment using GIS-based exposure analysis in Kuala Lumpur City," *Geomatics, Natural Hazards Risk*, vol. 8, no. 2, pp. 706–732, Dec. 2017.
[38] S. Mandal and R. Maiti, "Application of analytical hierarchy process (AHP) and frequency ratio (FR) model in assessing landslide susceptibility and risk," in *Semi-quantitative Approaches for Landslide Assessment and Prediction*. Singapore: Springer, 2015, pp. 191–226.
[39] A. Strehmel, S. Schönbrodt-Stitt, G. Buzzo, C. Dumperth, F. Stumpf, K. Zimmermann, K. Bieger, T. Behrens, K. Schmidt, R. Bi, J. Rohn, J. Hill, T. Udelhoven, W. Xiang, X. Shi, Q. Cai, T. Jiang, N. Fohrer, and T. Scholten, "Assessment of geo-hazards in a rapidly changing landscape: The three Gorges Reservoir Region in China," *Environ. Earth Sci.*, vol. 74, no. 6, pp. 4939–4960, Sep. 2015.
[40] J. Hou, J. Lv, X. Chen, and S. Yu, "China's regional social vulnerability to geological disasters: Evaluation and spatial characteristics analysis," *Natural Hazards*, vol. 84, no. 1, pp. 97–111, Aug. 2016.
[41] A. Erener and H. B. S. Düzgün, "A regional scale quantitative risk assessment for landslides: Case of Kumluca watershed in Bartin, Turkey," *Landslides*, vol. 10, no. 1, pp. 55–73, Feb. 2013.
[42] Y. M. Li, "Risk assessment of geological hazards in shifang," M.S. thesis, Dept. Environ Sci., CDUT Univ., Chengdu, China, 2010.
[43] D. Zhao, Q. Wu, F. Cui, H. Xu, Y. Zeng, Y. Cao, and Y. Du, "Using random forest for the risk assessment of coal-floor water inrush in Panjiayao coal mine, northern China," *Hydrogeol. J.*, vol. 26, no. 7, pp. 2327–2340, Nov. 2018.
[44] W. Chen, X. Xie, J. Wang, B. Pradhan, H. Hong, D. T. Bui, Z. Duan, and J. Ma, "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility," *Catena*, vol. 151, pp. 147–160, Apr. 2017.
[45] H. R. Pourghasemi and N. Kerle, "Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran," *Environ. Earth Sci.*, vol. 75, no. 3, pp. 185–201, Jan. 2016.
[46] J. Goetz, A. Brenning, H. Petschko, and P. Leopold, "Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling," *Comput. Geosci.*, vol. 81, pp. 1–11, Aug. 2015.
[47] N. Micheletti, L. Foresti, S. Robert, M. Leuenberger, A. Pedrazzini, M. Jaboyedoff, and M. Kanevski, "Machine learning feature selection methods for landslide susceptibility mapping," *Math. Geosci.*, vol. 46, no. 1, pp. 33–57, Jan. 2014.

PEI HUANG received the B.S. degree in human geography from Sichuan Normal University, Chengdu, China, in 2017. He is currently pursuing the Ph.D. with the Institute of Mountain Hazards and Environment, Chinese Academy of Sciences. He is good at machine learning algorithms, such as random forests, support vector machines, and artificial neural networks. His research interests include disaster risk assessment and urban resilience construction.



LI PENG received the M.S. and Ph.D. degrees in physical geography from the University of Chinese Academy of Sciences, Beijing, China, in 2008 and 2013, respectively. He is currently a Professor with the College of Geography and Resources, Sichuan Normal University. He is the author who published more than 20 SCI/SSCI articles. His research interests include hazard risk assessment and land use planning. He is a reviewer of SCI/SSCI journal, such as IEEE Access, *Natural Hazards*, the *International Journal of Disaster Risk Science*, and *Land Degradation & Development.*

HONGYI PAN received the B.S. degree in agricultural resources and environment, and the M.S. degree in land resources management from the Agricultural University of Hebei Province, China, in 2004 and 2007, respectively, and the Ph.D. degree in physical geography from the University of Chinese Academy of Sciences, Beijing, China, in 2010. Since 2012, he has been an Associate Professor with the College of Geography and Resources, Sichuan Normal University. He has rich scientific research experience, and research ability and advantages. His main research areas include land use and ecological environment change, and resources and environment management. As the first author or corresponding author, he has published more than 30 related articles.