Application of Machine Learning Algorithm in China National Geohazard Susceptibility Assessment

Bin Tong, Hao Fang*, Yuan Li, Xudong Yang, Chunrong Yin, Xueyan Qu, Yanling Zhang, Qiang Yang, Xi Chen
China Institute of Geo-environmental Monitoring, Beijing, Haidian District, 100081.

Corresponding Email Address: fhao@mail.cgs.gov.cn

Abstract. Landslide susceptibility is an important activity in landslide hazard assessment. With the advancement of artificial intelligence, the machine learning algorithm is applied in the landslide susceptibility assessment, has recently gained immense attention due to the advantages of obtaining the insights of landslide events and conditioning parameters based on data mining, which is important when tackling the challenge of mapping landslide prone areas in regional scale due to the complex nonlinear correlations among landslides and parameters and uncertainties associated during parameters reclassification. Therefore, the machine learning algorithm has become a standard approach for modeling landslide susceptibility over large regions. In this study, the random forest method is applied to produce the China national landslide susceptibility mapping based on the national landslide database containing more than 300 thousand landslide events. Thirty different categories of conditioning parameters related to the development, triggering, and potentially vulnerable elements of the landslide were collected using a scale of approximately 1:1,000,000. Through the data mining process, lithology, faults, topography, soil erosion, precipitation, and human activities were found to be the top six important contribution factors to landslide susceptibility. The mapping results show the areas of four degrees of susceptibility from high to low are 101, 191, 337, and 331 thousand square kilometers, respectively. The receiver operating curve (ROC) and area under curve (AUC) value was calculated to 0.81, indicating that results are well satisfying and could guide landslide mitigation on the national level.

1. Introduction
Landslide susceptibility mapping (LSM) is a critical activity in landslide hazard assessment that involves the relative spatial probability of landslide occurrence over appropriate spatial units. LSM is a key part of disaster management strategies by producing a map of spatial probabilities of landslide occurrence in a geographical region. The overwhelming literature (Pourghasemi and Rahmati 2020; Corominas J. et al. 2013) on landslide susceptibility is rich in works presenting applications to case studies from the local to the global scale obtained primarily using statistical techniques, artificial neural networks, and machine learning (ML) algorithms. Broadly, there are four main types of LSM approach: (1) physical-based models, (2) opinion-driven (i.e., heuristic) models, (3) statistical models, and (4) more recently ML models (Corominas et al. 2013). Each approach has been shown to have its advantages and limitations. Unlike statistical methods, ML methods can “learn” from data without
relying on rules-based functions or based on predominant assumptions, whereas statistical modeling streamlines relationships between variables in the data using mathematical equations. The above benefits allow ML methods to become more prevalent in recent studies. Compared to the majority of ML models, commonly used statistical models such as frequency ratio and weight of evidence method are less computationally intensive and can be computed at a range of different spatial scales but still require authoritative landslide inventories to effectively calculate susceptibility for landslide mapping over a large region with intermediate to small scale. Particularly, when statistical pre-assumptions are unreliable, the inappropriate assumptions may cause bias that may be unknown. Another important method in statistical analysis is the heuristic approach, which applies a weighted linear combination (WLC) of variables based on expert opinion or other sources of prior knowledge. WLC does not require landslide inventories for model calibration, but categorizing a continuous variable into discrete bins (e.g., slope from 0° to 5°, 5° to 10°, etc.) typically results in a loss of some information. WLC introduces subjectivity in each step: selecting explanatory variables, categorizing the data, setting bin weights, and establishing variable weights.

Therefore, ML models can benefit from the statistical and heuristic models by increasing the transparency, applicability, and repeatability of the calculation process. Nevertheless, ML models emphasize optimization and performance rather than inference, which is the primary concern of statistical models. Mapping and modeling have been gradually shifted from subjective qualitative analysis to objective quantitative or semi-quantitative analysis by reviewing studies on LSM based over large regions with intermediate to small scales (the continental region with the mapping scale of 1:200,000 or below) (Corominas et al. 2013). In this study, with the objective to produce a susceptibility map for rockfall, landslide, and debris flow on a national scale of 1:1,000,000, the random forest (RF) algorithm, as one of the widely applied ML approaches, was applied. The national landslide database provides input of events, and multiple parameters of different categories were involved in establishing the dataset of parameters. Finally, receiving operating curve was plotted, and the associated AUC value was obtained as the verification process. The verification result shows good applicability and accuracy of mapping result using the RF method over a large region with intermediate to small scale. The national susceptibility mapping shows the distribution of four different classes of susceptible areas in China.

2. RF method

2.1. The characters of RF method

The RF method is a supervised classification algorithm and an ensemble method that uses decision tree (DT) models to fit a data subset sampled independently using bootstrapping. RF is composed of many individual DT models, hence the use of the term “forest.” The RF method is known to provide high accuracy rates for outliers in predictors due to random selection at each split node depending on two data objects, namely, Out-Of-Bag and proximities. Overall, randomness in RF algorithms can reduce overfitting by (1) building several trees; (2) portraying observations with replacements (i.e., bootstrapped); (3) within a random subset, and they split the nodes on the best split. In RF models, some hyper-parameters must be, including (1) the number of trees to be combined, (2) the maximum depth of the trees, and (3) the maximum number of features considered at each split. The main advantage of RF is that it prevents overfitting, a common issue in generic DT models. However, the RF method is regarded as a black-box model because of the limited interoperability of decisions. Nevertheless, the RF method or its variants (e.g., EXT) are considered robust models that reduce the subjectivity in the modeling process. Current studies (Raja N. et al. 2017; Merghadi et al. 2020; Kirschbaum et al. 2015) have shown that an improved mapping accuracy could be achieved with an even smaller landslide inventory dataset, and could produce more accurate results than other ML approaches, including artificial neural networks, boosted regression tree, classification and regression trees, generalized linear model, generalized additive model, multivariate adaptive regression splines,
naïve Bayes (NB), quadratic discriminant analysis, RF, and support vector machines based on Pourghasemi and Rahmati (2020). Figure 1 presents the typical mechanism of the RF method.

Figure 1. The typical working mechanism of the RF approach.

2.2. The calculation process of RF approach in this study

The calculation algorithm applied in this study is the combined RF method and the information content (IC) model. The IC model, a common statistical method, is widely used to build the DT in the RF algorithm so that each DT becomes an individual calculation and mapping process based on a specific data sample (dataset) and a list of parameters. Therefore, numerous susceptible results can be generated based on a specific data sample and parameters. The most representative evaluation result (the one of greatest frequency of occurrence) in the DTs can be obtained via statistical probability analysis of a large number of mapping results.

The revised method has the following characteristics and advantages by integrating the IC model in the RF calculation process: the IC model is a well-established method for analyzing and evaluating geological disasters on a regional scale. RF algorithm (Figure 2) is a classic method in the field of artificial intelligence and extensive data mining. The IC model is integrated into the RF calculation process to achieve a random selection of evaluating sample data and batch construction of decision tree in a random forest; thus, the calculation is transformed from “one-time” evaluation calculation in
the previous process to “exhaustive” evaluation calculation in the present one. The calculation error of sample data anomaly and noise in the evaluation results can be effectively reduced to improve the accuracy and quality of geological disaster evaluation and to reduce the subjective influence of human factor evaluation results. Compared with the complex computing principle of other ML algorithms, the core computing model of the method is based on the information algorithm, making it easier for geological disaster professionals to understand and popularize the evaluation method.

3. Data preparation

3.1. Landslide inventory
China has established a comprehensive geological hazard database that includes landslide, rockfall and debris flow, and other types. The total number of landslide, rockfall, and debris flow reaches more than 280,000. This paper aims to produce the susceptibility mapping for the abrupt geological hazard that mainly indicates landslide, rockfall, and debris flow. Considering the different characters of landslide, rockfall, and debris flow distributed across the national territory in different geological environments, treating the entire inventory data set as one uniform data set in the mapping process is inappropriate. The triggering conditions and controlling parameters for landslides, rockfall, and debris flows can be significantly different across the national territory in different geological environments. Conversely, treating each individual event as an equally uniform event (point) may cause unacceptable bias in the mapping result.

![Figure 3. A series of landslide, rockfall, and debris flow datasets subdivided based on category and scale](image-url)
Therefore, the entire inventory dataset was subdivided into 12 different sub-dataset based on categories (landslide, rockfall, and debris flow) and scale (vast, large, intermediate, and small), with each inventory in sub-dataset have the same category and scale. The susceptibility mapping process was used to determine the susceptibility potential for each sub-dataset. Eventually, the mapping result of each sub-dataset was summarized so that every sub-dataset with the same category and scale of geological hazards was considered in the mapping process and result. The primary advantage of creating a series of inventory sub-dataset was that it could eliminate the overwhelming influence of one sub-dataset that contains a significantly larger portion of events in the final mapping result, as well as improve the representation and accuracy of the final result.

3.2. Conditioning parameters applied in this study

Considering the data availability, systematicness, independence, timeliness, reliability, and moderate accuracy of evaluation factors, a comprehensive system of conditioning parameters included 4 categories, 15 sub-classes, and 29 three-level classifications (Segoni et al. 2020). The four categories included basic geography, geological environment conditions, inducing factors, and disaster-bearing bodies. The digital and raster processing and resampling are completed based on 250 m by 250 m standard raster data. Based on the evaluation of the main control factors (with the scale from 1:500,000 to 1:1,000,000), the index system of this evaluation is preliminary determined as follows: landform division, geological structure distribution, seismic division, stratum lithology division, engineering rock group division, soil erosion intensity, rainfall distribution, human activity intensity, etc. (Figure 4). Due to the limit of the paper length, only six important conditioning parameters (topographic and geomorphic, lithological unit, active faults distribution, intensity of subsoil erosion map, categories of land use, and distribution of plantation) were listed in this paper. In view of special requirements of wide-area small-scale evaluation on the selection and preparation of evaluation indexes, the impact of evaluation elements on geological disasters is reasonably reflected in the evaluation element system as follows:

1. In order to consider the impact of human activities on geological disasters, the national GDP economic data (since 2000) were collected to calculate the annual average growth rate of GDP, and the results can be used to express the spatial distribution of human development intensity;

2. In order to consider the influence of active faults on the development and distribution of geological disasters, the influence of active faults are expressed by a representative factor, that is, the influence intensity index of active faults is set, which is a comprehensive expression of the density and activity intensity of active faults, which can reflect at any spatial position;

3. In order to consider the influence of loose degree (soft and hard degree) of topsoil on the development and formation of landslide geological disasters, soil erosion data were collected and prepared in this evaluation to express the influence of wind erosion, hydraulic erosion, and freeze-thaw erosion factors on topsoil to reflect this factor.
4. Calculation

Based on the susceptibility process described in Figures 1 and 2, the susceptibility result for each inventory dataset was produced (Figure 5). The final susceptibility result based on each dataset, as expressed by numerical index, was re-calculated through normalization (from 0 to 1) so that the index can be added numerically with equal weight for every sub-susceptibility result. As is well known, the higher the number of a certain raster value (closer to 1) indicates that the susceptibility of the raster is relatively higher. Based on this principle, the final susceptibility results, also known as the summation of the index numerical results, can be obtained and then reclassified using the Natural Break method (a built-in classification modeling in ArcGIS), resulting in the susceptibility mapping result (Figure 5). Applying the Natural Break method for the mapping process reduces the influence of humans in the calculation process for improving transparency and repeatability. For verification of the results, THE ROC was plotted based on the derived susceptibility result, and the AUC value is 0.81, indicating very good mapping accuracy.
5. Discussion
According to the results, the following comments can be made:

(1) The area that is extremely susceptible to rockfall, landslide, and debris flow is about 1 million square kilometers, accounting for 10.5% of the national territorial area; more than 90 thousand geological hazard potentials were found in the area, accounting for 30% of the total number. The overall density of geological hazard potential is nearly 900 per 10 thousand square kilometers.

(2) The area that is highly susceptible to rockfall, landslide, and debris flow is about 1.91 million square kilometers, accounting for about 20% of the national territorial area; more than 123 thousand geological hazard potentials were found in the area, considering 41% of the total number. The average density of geological hazard potential is 646 per 10 thousand square kilometers.

(3) The area that is intermediately susceptible to rockfall, landslide, and debris flow is about 3.58 million square kilometers, accounting for 35.1% of the national territorial area; more than 83 thousand geological hazard potentials were found in the area, accounting for 27.5% of the total number. The average density of geological hazard potentials is 231 per 10 thousand square kilometers.

(4) The area that is less susceptible to landslides and debris flows is about 3.31 million square kilometers, accounting for 34.5% of the national territorial area; more than five thousand geological hazard potentials were found in the area, accounting for 0.2% of the total number. The average density of geological hazard potential is nearly 15 per 10 thousand square kilometers.

(5) The national susceptibility result shows that the highly susceptible region of rockfall, landslide, and debris flow lies in the southwest of China, the eastern margin of the Qinghai-Tibet Plateau, the Yunnan-Guizhou Plateau, and the Qinba mountain, the Loess Plateau. The distribution of various susceptible classes can be used as guidance to support the strategic planning of geological hazard mitigation and prevention on the national level.

6. Conclusion
In this study, a revised RF method is applied to produce the China national LSM based on more than 288 thousand of landslide, rockfall, and debris flow events. More than 30 different categories of
conditioning parameters related to the development, triggering, and potentially vulnerable elements of
the landslide were collected using a scale of approximately 1:1,000,000. The top six crucial
contribution factors to landslide susceptibility were found through data mining are lithology, faults,
topography, soil erosion, precipitation, and human activities. The mapping results show that the areas
of four degrees of susceptibility from high to low are 101, 191, 337, and 331 thousand square
kilometers, respectively. The ROC and AUC value was calculated to 0.81, satisfying the results.

Acknowledgment
Funding: This study was funded by the China Geological Survey Projects (Grant No. 1212011140016
and DD20160273).

References
[1] Pourghasemi and Rahmati (2020). Prediction of the landslide susceptibility: Which algorithm,
which precision? Catena. 162 (2018) 177–192.
[2] Corominas J. et al (2013) Recommendations for the quantitative analysis of landslide risk. Bull
Eng Geol Environ. DOI 10.1007/s10064-013-0538-8.
[3] Raja N. et al (2017) Landslide susceptibility mapping of the Sera River Basin using logistic
regression model. Nat Hazards (2017) 85:1321-1346. DOI 10.1007/s11069-016-2591-7.
[4] De Graff et al (2012) Producing landslide-susceptibility maps for regional planning in data-scarce
regions. Nat Hazards (2012) 64:729-749. DOI 10.1007/s11069-012-0267-5.
[5] Merghadi et al (2020) Machine learning methods for landslide susceptibility studies: A
comparative overview of algorithm performance. Earth-Science Reviews.
https://doi.org/10.1016/j.earscirev.2020.103225.
[6] Segoni et al (2020). Landslide susceptibility assessment in complex geological settings: sensitivity
to geological information and insights on its parameterization. Landslides. 17:2443–2453.
[7] Kirschbaum et al (2015). Spatial and temporal analysis of global landslide catalog.
Geomorphology. 245 (2015): 4-15.