GIS-based coseismic landslide susceptibility assessment using ensemble learning approach: a case of the 2017.8.8 Jiuzhaigou earthquake event

Chuanjie Xi, Xiewen Hu, Kun He, Bo Liu
Faculty of Geoscience and Environmental Engineering, Southwest Jiaotong University, Sichuan Chengdu 610031, China

Corresponding author: xichuanjie@my.swjtu.edu.cn

Abstract. Exploring more effective landslide susceptibility assessment methods play an important role in mitigating landslide effects. This paper aims to compare the performance of different popular ensemble learning models that combined with GIS system to assess coseismic landslide susceptibility in the 2017.8.8 Jiuzhaigou earthquake area. Eight influencing factors (slope, elevation, aspect, relief altitude, lithology, peak ground acceleration, distance to river, distance to fault) were considered to construct a spatial database after the Pearson correlation analysis. The 4834 landslides from inventory data are randomly divided into 70% train samples, and 30% validate samples. We construct the random forest (RF), gradient-boosting decision tree (GBDT), and adaptive boosting (AdaBoost), which all three models use decision tree model as basic unit, and utilize the receiver operating characteristic (ROC) curve, area under curve (AUC) values, Kappa value to validate the performance of three ensemble models. The results indicate that the AdaBoost model achieved the best performance (AUC= 94.4%, Kappa=0.766), outperforming the GBDT (AUC=92.5%, Kappa=0.720), RF (AUC= 93.6%, Kappa=0.756) focused on the validation data. This study can provide an insight into evaluating coseismic landslide susceptibility with high accuracy.

1. Introduction

Coseismic landslides are common secondary disasters triggered by the earthquake in the steep mountainous terrain with extremely destructive impacts [1]. Those landslides can cause significant economic and societal damage, destruction of buildings and transportation. Recently, landslides susceptibility mapping (LSM) has been identified as an effective measure to mitigate hazard perniciousness which has absorbed major attention of worldwide scholars [2]. The purpose of LSM aims at predicting the potential spatially occurring likelihood of landslides in a specific area that given the local geology conditions. This frame rests on a basic assumption that the influencing factors associated with landslides in the past may also contribute to the same hazard development in the future [3]. Over the past decades, various methods have been proposed to carry out experiments for generating LSM. Among them, the machine learning models achieved excellent performance in landslides predicting and susceptibility mapping and became the more popular ones, such as decision tree [4], support vector machine (SVM) [5], artificial neural network (ANN) [6], logistic regression (LR) [7]. For example, Tian et al. [8] utilized the artificial neural network (ANN) model for LSM affected by the 2013 Minxian, Gansu, China, Mw5.9 earthquake. Bui et al. [9] improves the SVM model and applied it
to the spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam). Hong et al. [10] introduce a novel multi-layer perceptron network based on stochastic gradient descent optimized for LSM. However, the single machine learning model exists the limitation of adaptation with poor robustness caused by samples distribution differences and the calculations may differ by various prediction models. To settle this problem, ensemble machine learning methods have been gradually applied to improve the capability of predicted models for LSM. This method trains multiple weak classifiers for the same issue simultaneously and obtains a strong classifier by combining those weak classifiers. The advantages of ensemble learning are improving the robustness and anti-noise ability of the model. Common ensemble algorithms can be divided into a bootstrap aggregate (Bagging) based and boosting based class. The representative model of the Bagging algorithm is random forest (RF) [11] which is a forest composed of multiple decision trees, for boosting algorithm, such as famous adaptive boosting [12] (Adaboost) and gradient boost decision tree (GBDT) [13]. However, few previous studies have been conducted on the comparing the performance of diverse ensemble learning models for coseismic landslide susceptibility assessment.

In this study, three ensemble learning models (RF, Adaboost, GBDT) were constructed for a case of 2017 8.8 Mw6.5 Jiuzhaigou earthquake event. Eight factors (slope, elevation, aspect, relief amplitude, lithology, peak ground acceleration, distance to river, distance to fault) are considered to construct a spatial database after the correlation test. The receiver operating characteristic (ROC) curve, area under curve (AUC) values, sensitivity, and specificity were utilized to validate the performance of model. This study clarified the differences of three ensemble learning models and thus improved the accuracy of the coseismic landslides susceptibility assessment.

Figure 1. Study area and landslides inventory map.

2. Description of the study area and materials

2.1 Study area
The study area is located in Jiuzhaigou scenic, Sichuan province, China, belonging to transition zone of Qinghai-Tibet Plateau and Sichuan Basin, which is a famous natural scenery tourist region around nationwide (Figure 1). It is situated in the mountain of Baishuihe River basin, dominated by canyons and cold alpine mountain area and well-developed karst landform. The average altitude in the area is
over 3000m, and the relative height difference is up to 2000m. The main geological formation units including Triassic (T), Carboniferous (C), Permian (P), and the lithology of exposed rock are mainly composed of limestone and sandstone. The average annual precipitation is about 555.6mm with cold semiarid monsoon and cold sub-humid climate.

After a magnitude Mw 6.5 earthquake hit the Jiuzhaigou area on August 8, 2017, the scenic area was hit by four heavy rains from June 2018 to August 2020, triggered numerous landslides that caused seriously destruction of natural landscape and transportation. Therefore, carrying out the landslide susceptibility assessment has a significance mean for tourist safety and disaster prevention.

2.2 Materials sources
In this study, the landslide inventory map has been gathered including 4834 coseismic landslides recorded via high-resolution remote sensing interpretation from Xu et al. [14]. Those landslides are mainly shallow rock slides and rock falls.

The 30m digital elevation map data (DEM) is obtained from the Geospatial Data Cloud Platform, Chinese Academy of Sciences (http://www.gscloud.cn). This data was utilized to extract elevation, slope angle, slope aspect, and river distribution information of the study area.

The 1:200000 scale geological map was adopted for extracting stratigraphic lithology and major faults information of the study region.

The ground motion caused by seismic shocks is the significant inducement of coseismic landslides, therefore, the dynamic parameter of earthquake is necessary data. This study selects the peak ground acceleration (PGA) as landslides influencing factors, downloaded from the USGS website (http://earthquake.usgsShakemap.gov).

2.3 Landslides influencing factors
This study singled eight factors out to construct a spatial database, including slope angle, elevation, slope aspect, relief amplitude, lithology, peak ground acceleration, distance to river, distance to fault, and applied the Pearson correlation coefficient [15] (PCC) to conducted factors correlation test. PCC values can reflect the degree of linear correlation between two influencing factors with a range from −1 to 1.

Slope angle has been recognized as one of the significant topographic factors that directly related to stability of nature slope. Slope aspect indirectly affects soil mass stability by influencing the rainfall precipitated. Elevation and relief amplitude (the difference between the highest and lowest points in a terrain unit) was frequently considered in previous research because it can reflect the gravitational potential energy of terrain. Lithology and faults govern the development and distribution of landslides and are associated with the nature of activity. The distance to river is used to reflect the erosion of slope materials caused by water, which reduces the stability of slope. PGA is a dynamic parameter that measures the ground motion impact of the earthquake; therefore, it has been considered as an important landslide conditioning factor. All factors utilized by this study as shown in Table 1.

Table 1. Feature parameters of landslide factors in study area.

| Influencing factors | Classification | Area /km² | Landslide ratio /% | Landslide area /km² | Landslides density A/km² | Frequency ratio |
|---------------------|----------------|-----------|--------------------|---------------------|-------------------------|----------------|
| Distance to river   | >1             | 216.8829  | 57.22              | 5.36                | 12.75                   | 3.86           |
| /Km                 | 1~2            | 205.6401  | 30.76              | 2.96                | 7.23                    | 2.25           |
|                     | 2~3            | 196.1244  | 9.25               | 0.99                | 2.28                    | 0.79           |
|                     | 3~4            | 188.8902  | 2.23               | 0.20                | 0.57                    | 0.17           |
|                     | >4             | 696.0663  | 0.54               | 0.12                | 0.04                    | 0.03           |
|                     | 0.12           | 31.0149   | 0.00               | 0.00                | 0.00                    | 0.00           |
|                     | 0.16           | 252.5211  | 7.07               | 0.63                | 1.35                    | 0.39           |
| PGA                 | 0.2            | 490.7988  | 42.66              | 3.83                | 4.20                    | 1.22           |
|                     | 0.24           | 673.3656  | 32.46              | 3.28                | 2.33                    | 0.76           |
|                     | 0.26           | 54.5526   | 17.87              | 1.89                | 15.84                   | 5.41           |
| Influencing factors | Classification | Area /km² | Landslide ratio /% | Landslide area /km² | Landslides density A/km² | Frequency ratio |
|---------------------|----------------|-----------|--------------------|----------------------|--------------------------|----------------|
| Elevation /m        |                |           |                    |                      |                          |                |
| <2500               | 117.1026       | 3.85      | 0.24               | 1.59                 | 0.32                     |                |
| 2500~3000           | 256.5657       | 36.49     | 3.34               | 6.88                 | 2.03                     |                |
| 3000~3500           | 407.1996       | 44.60     | 4.51               | 5.29                 | 1.73                     |                |
| 3500~4000           | 470.2248       | 13.18     | 1.32               | 1.35                 | 0.44                     |                |
| >4000               | 250.2981       | 1.88      | 0.24               | 0.36                 | 0.15                     |                |
| <15                 | 164.2257       | 3.58      | 0.08               | 1.05                 | 0.08                     |                |
| 15~25               | 328.3407       | 9.35      | 0.55               | 1.38                 | 0.26                     |                |
| 25~35               | 445.7997       | 24.22     | 1.77               | 2.63                 | 0.62                     |                |
| 35~42               | 313.7328       | 35.97     | 3.51               | 5.54                 | 1.75                     |                |
| 42~75               | 249.2919       | 26.87     | 3.71               | 5.21                 | 2.32                     |                |
| <68                 | 322.1415       | 24.93     | 2.33               | 3.74                 | 1.13                     |                |
| 68~138              | 271.1628       | 23.15     | 2.06               | 4.13                 | 1.19                     |                |
| 138~213             | 308.6199       | 17.46     | 1.67               | 2.73                 | 0.84                     |                |
| 213~286             | 300.2976       | 16.49     | 1.76               | 2.65                 | 0.91                     |                |
| 286~360             | 299.169        | 17.98     | 1.81               | 2.90                 | 0.94                     |                |
| <217                | 239.0706       | 5.21      | 0.16               | 1.05                 | 0.10                     |                |
| 217~298             | 411.8256       | 16.98     | 0.93               | 1.99                 | 0.35                     |                |
| 298~375             | 433.1421       | 31.11     | 2.55               | 3.47                 | 0.92                     |                |
| 375~467             | 317.5839       | 35.00     | 4.01               | 5.33                 | 1.97                     |                |
| 467~950             | 101.1708       | 11.69     | 1.98               | 5.58                 | 3.05                     |                |
| Distance to faults /km |                |           |                    |                      |                          |                |
| <0.5                | 197.9721       | 19.57     | 1.52               | 4.78                 | 1.20                     |                |
| 0.5~1               | 171.0279       | 14.92     | 1.24               | 4.22                 | 1.13                     |                |
| 1~1.5               | 137.664        | 14.44     | 1.42               | 5.07                 | 1.61                     |                |
| 1.5~2               | 113.2551       | 7.30      | 0.77               | 3.12                 | 1.06                     |                |
| >2                  | 881.433        | 43.77     | 4.68               | 2.40                 | 0.83                     |                |
| P2                  | 254.2635       | 8.87      | 0.66               | 1.69                 | 0.41                     |                |
| P1                  | 64.426         | 2.21      | 0.15               | 1.66                 | 0.36                     |                |
| T3                  | 64.9386        | 3.95      | 0.22               | 2.94                 | 0.53                     |                |
| T2                  | 361.5687       | 3.66      | 0.29               | 0.49                 | 0.13                     |                |
| T1                  | 76.3542        | 0.00      | 0.00               | 0.00                 | 0.00                     |                |
| N                   | 3.1815         | 0.12      | 0.005              | 1.89                 | 0.25                     |                |

3. Methodology

3.1 Random Forest

Random forest is one of the mushroomed machine learning algorithms at present, pertain to the Bagging ensemble algorithm system, and was originated proposed by Breiman in 2001. The character of RF model is that it was constructed by numerous decision trees as a benchmark decision unit, and combines those single decision units by voting or averaging method to achieve a strong classifier. It was noted that those decision trees are uncorrelated with each other. The samples data of each tree in the ensemble is built from drawing of source data repeatedly randomly and with replacement. This sampling strategy is referred to as Bootstrap. Therefore, for each benchmark tree, the sample data was taken advantaged of training is different. Actually, there is third of data not involved in the training process, which is called as out of bag (OOB) data and used to carry out an unbiased estimate of the error. The probability calculation formula of RF is as follows:
\[ p_c = \max \left\{ p_l = \frac{\sum_{i=1}^{m} p_{lj}}{m} \mid i \in I \right\} \] (1)

Where \( p_c \) is the probability of landslides occurrence, \( p_{lj} \) is the predict probability of the \( j \) decision tree, \( m \) is the sum of the tree.

### 3.2 adaptive boosting

The popular boosting algorithm AdaBoost belong to the ensemble learning frame, induced by Freund and Schapire\[^{[12]}\] in 1997. Similar to other ensemble learning models, AdaBoost combines all of the weak learners that were fitted on repeatedly modified versions of the data to build a strong classifier. It can update the weight of the training samples to improve the performance of those weak learners. If there are \( N \) groups of training samples, and the weight of each group are \( w_1, w_2 ... w_n \), respectively. Initially, those weights are all set as of \( 1/N \), and individually modified in each successive iteration latter. At a given step, those training examples that were incorrectly predicted by the boosted model induced at the previous step have their weights increased, whereas the weights are decreased for those that were predicted correctly. Finally, the AdaBoost model combined those weak learners via a weighted majority vote (or sum) to produce a prediction.

### 3.3 gradient boost decision tree

The GBDT model pertain to boosting algorithm also, which is a generalization of boosting to arbitrary differentiable loss functions. All weak learners utilized in the GBDT model are decision trees, which are constructed by a way of continuous. GBDT consisted by the multiple decision tree and each tree training based on the negative gradient of the loss function of previous one. The finally predicted output obtained by summing up the results of all the trees. The advantage of GBDT algorithm is that it has excellent robustness to outliers in the output space. It was recognized as a strong generalization algorithm similar to SVM.

### 4. Result

#### 4.1 Correlation and importance analysis

The Pearson correlation coefficient calculated results of the eight influencing factors as shown in Figure 2. Figure 3 shows the importance analysis consequences of each influencing factors based on the random forest approach. As a result, the PGA (0.365) has the significant importance for landslide susceptibility assessment among those factors, followed by the elevation (0.181), lithology units (0.159), slope aspect (0.079), slope angle (0.064), relief amplitude (0.063), distance to faults (0.046), distance to river (0.042).

The PCC value exceeds 0.8 means that there is high correlation between two factors. According to analysis results, all of the factors were not strongly correlated with each other, therefore, a total of eight factors were utilized for construing landslides predict model.

#### 4.2 Evaluation of the predict models

To compare the performance of three ensemble models, all three ensemble models use the decision tree algorithm as basic decision unit. The receiver operating characteristic (ROC)\[^{[16]}\] curve, area under the curve (AUC), kappa value and confusion matrix were employed in this study, and the results are indicated as Table. 2. The FP (false positive) is the sum of landslide samples that has been misclassified. FN (false negative) is the number of samples of non-landslide occurrence that has been misclassified. The TP and FP are the number of landslides that were correctly classified as positive and negative units, respectively. The ROC curve and AUC value are common indexes for verifying the performance of landslide predict model. Figure 4 shows the ROC analysis results of three ensemble models.

According to Figure4 and Table 2, all three models have obtained excellent capability that the AUC values exceed 0.9 for landslide prediction. Among them, the AdaBoost model achieved the best
performance (AUC= 94.4%, Kappa=0.766), outperforming the GBDT (AUC=92.5%, Kappa=0.720), RF (AUC= 93.6%, Kappa=0.756) focused on validation data. It can be obvious that the gap in the performance between the training and validation dataset of three models has not existed a great difference. It indicated that the three models were not overfitting. Therefore, the results have shown that the RF model achieved the most desired predictive performance for landslide susceptibility assessment.

Table 2. A slightly more complex table with a narrow caption.

| model  | TP   | FP   | TN   | FN   | AUC   | Kappa |
|--------|------|------|------|------|-------|-------|
| RF     | 1388 | 265  | 1160 | 88   | 93.6% | 0.756 |
| AdaBoost | 1318 | 236  | 1244 | 103  | 94.4% | 0.766 |
| GBDT   | 1291 | 257  | 1204 | 149  | 92.5% | 0.720 |

4.3 landslide susceptibility map
The probability of coseismic landslide occurrence has been transformed into a raster data using the Jenks natural break approach based on GIS system. Those raster data have been reclassified into five landslide susceptibility groups and the results of susceptibility map as shown in Figure 5.

It can be obvious that the three models produced similar susceptibility maps. Based on evaluation results, in the RF model, 83.3% landslides located in 10.09% very high susceptibility area and the value of frequency ratio is 8.25, higher than GBDT (7.32) and AdaBoost (7.57) model. Therefore, the
RF model produced a higher quality of very high susceptibility map than GBDT and AdaBoost ensemble algorithm.

| model   | Susceptibility groups | Total area ratio (%) | Landslides Ratio (%) | Frequency ratio |
|---------|-----------------------|-----------------------|-----------------------|-----------------|
| RF      | Very high             | 10.09                 | 83.3                  | 8.25            |
|         | high                  | 8.73                  | 13.3                  | 1.52            |
| GBDT    | Moderate              | 9.41                  | 2.75                  | 0.29            |
|         | Low                   | 15.01                 | 0.6                   | 0.04            |
|         | Very low              | 56.76                 | 0.1                   | 0.00            |
|         | Very high             | 10.36                 | 75.8                  | 7.32            |
| AdaBoost| Moderate              | 9.33                  | 5.4                   | 0.58            |
|         | Low                   | 16.76                 | 3.2                   | 0.19            |
|         | Very low              | 54.90                 | 0.6                   | 0.01            |
|         | Very high             | 12.73                 | 96.28                 | 7.57            |
|         | high                  | 4.01                  | 1.46                  | 0.36            |
|         | Low                   | 6.17                  | 0.91                  | 0.15            |
|         | Very low              | 73.23                 | 0.74                  | 0.01            |

**Figure 5.** Landslide susceptibility map result, a RF model, b GBDT model, c AdaBoost model.

5. Conclusions

This study compared the three popular ensemble models for coseismic landslide susceptibility mapping with the case of the Jiuzhaigou earthquake event. Eight influencing factors were selected to construct landslide predict model, including slope angle, elevation, slope aspect, relief amplitude, lithology, peak ground acceleration, distance to river, and distance to fault. The Kappa, ROC curve, and AUC values are performed to assess and validate the landslide models.

According to PCC test result, all factors do not have significant correlation with each other. According to importance analysis based on RF model, the PGA (0.365) has the significant importance, outperforming the elevation (0.181), lithology unis (0.159), slope aspect (0.079), slope angle (0.064), relief amplitude (0.063), distance to faults (0.046), distance to river (0.042).

Experimental outcomes showed that the AdaBoost model achieved the best performance (AUC=94.4%, Kappa=0.766), followed by RF (AUC=93.6%, Kappa=0.756), and the GBDT (AUC=92.5%, Kappa=0.720). However, the RF model achieved the higher quality of very high susceptibility map, which 83.3% landslides located in 10.09% very high susceptibility area with the value of frequency ratio is 8.25, higher than GBDT (7.32) and AdaBoost (7.57).
References
[1] Tang H M, Hu X L, Xu C, Li C D, Yong R, Wang L Q 2014 A novel approach for determining landslide pushing force based on landslide-pile interactions Eng. Geol.182 (19) 15–24
[2] Chawla A, Chawla S, Pasupuleti S, Rao A C S, Sarkar K, Dwivedi R 2018 Landslide susceptibility mapping in Darjeeling Himalayas, India Adv. Civil Eng. 2018 17
[3] Nhu V, Hoang N, Nguyen H, Ngo P T T, Thanh Bui T, Hoa P V, Samui P, Tien Bui D 2020b Effectiveness assessment of Keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area Catena. 188 104458
[4] Dou J, et al. 2019a Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan Sci. Total Environ. 662 332–346
[5] Yi Y, Zhang Z, Zhang W, Xu C 2019a Comparison of Different Machine Learning Models For Landslide Susceptibility Mapping. In: IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium pp 9318–9321
[6] Catani F, Casagli N, Ermini L, Righini G, Menduni G, 2005 Landslide hazard and risk mapping at catchment scale in the Arno River basin Landslides 2 329–342
[7] Chauhan S, Sharma M, Arora M K, 2010 Landslide susceptibility zonation of the Chamoli region, Garhwal Himalayas, using logistic regression model Landslides 7 411–42
[8] Tian Y, Xu C, Hong H, et al. 2018 Mapping earthquake-triggered landslide susceptibility by use of artificial neural network (ANN) models: an example of the 2013 Minxian (China) Mw 5.9 event Geomatics Natural Hazards & Risk. 10(1):1-25
[9] Tien Bui D, Tuan T A, Hoang N D, et al. 2017 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 14(2):1-12
[10] Hong H, Tsangaratos P, Ilia I, et al. 2020 Introducing a novel multi-layer perceptron network based on stochastic gradient descent optimized by a meta-heuristic algorithm for landslide susceptibility mapping Science of The Total Environment. 742:140549
[11] Freund Y, Schapire RE,1997 A Decision Theoretic Generalization of Online Learning and an Application to Boosting Journal of Computer and System Sciences. 55 (1):119-139
[12] Friedman J H 2001 Greedy Function Approximation: A Gradient Boosting Machine Annals of Statistic 29(5):1189-1232
[13] Tian Y, Xu C, Ma S, Xu X, W S, Z H 2019 Inventory and Spatial Distribution of Landslides Triggered by the 8th August 2017 MW 6.5 Jiuzhaigou Earthquake, China J. Earth Sci. 30 206–217
[14] Merghadi A, Abderrahmane B, Tien Bui D 2018 Landslide susceptibility assessment at Mila Basin (Algeria): a comparative assessment of prediction capability of advanced machine learning methods ISPRS Int. J. Geo-Information 7
[15] van Erkel A R, Pattynama PM T 1998 Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology Eur. J. Radiol. 27 88–94