Seismic landslide susceptibility assessment based on ADASYN-LDA model

Shuhao Zhang* and Peiqiao Yu
Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, Sichuan, 611756, China
*Corresponding author’s e-mail: zshgostop@126.com

Abstract. Seismic landslide susceptibility (SLS) assessment can be used to estimate the susceptibility of landslides induced by an earthquake, which has great significance for emergency measure making and land use planning to mitigate the landslide hazard and risk. At present, one of the key problems in SLS evaluation is that the sample is extremely unbalanced, that is, the number of landslide units is usually far less than that of non-landslide units, which easily leads to poor prediction performance of SLS model. In addition, when using high-resolution mapping units, the sample size greatly increases, and thus the training time cost is considerable especially for complicated nonlinear artificial intelligence such as the Support Vector Machine and the Neural Networks. In view of this, this paper proposes an SLS model (called ADASYN-LDA) combining ADASYN sample method and linear discriminant analysis (LDA), and evaluates it using 2014 Ludian earthquake event. The results show that the two ADASYN-LDA models are highly efficient in training, and have good prediction performance with the high AUC of 0.8737 and 0.8535 respectively, and with the balanced accuracy of 0.7950 and 0.6174 respectively. Compared with the traditional LDA model without ADASYN method, ADASYN-LDA model improves the results of the SLS assessment affected by the sample imbalance. The further evaluation is needed in other earthquake events, and the further improvement can be made with more landslide conditioning factors and seismic landslide data set.

1. Introduction
Earthquake-induced landslides usually have the characteristics of huge volume and extremely high velocity, and can occur suddenly and cause great damage and losses. Seismic landslide susceptibility assessment can be used to estimate the spatial susceptibility (or relative probability) of landslide occurrence induced by an earthquake, and therefore it has great significance for emergency measures making and land use planning to minimize the damages and losses caused by landslides. In recent years, a variety of studies of seismic landslide susceptibility applying data-driven method have been carried out. Su [1] developed a landslide susceptibility model using logistic regression to evaluate landslide susceptibility in Wenchuan earthquake. Kritikos [2] developed a regional seismic landslide susceptibility model based on data-driven fuzzy logic model combined with GIS. Parker [3] developed a global seismic landslide susceptibility model based on logistic regression model and nine seismic landslide inventories. Other advanced machine learning models also have been adopted for seismic landslide susceptibility, for example, the Support Vector Machine [4], the Neural Network [5] and Bayesian network [6]. A key problem that needs to be resolved in current study is the sample imbalance that the number of landslide units is far less than the number of non-landslide units. The sample...
imbalance will seriously affect the prediction performance of the susceptibility model, and current studies often adopt a random under-sampling method to reduce the effect. However, the random under-sampling method will greatly reduce the number of non-landslide units, which leads to the high uncertainty of the susceptibility model outputs. In addition, model training samples are large when using high resolution grid cells in landslide susceptibility assessment, and the training time cost is considerable especially for complicated artificial intelligence which has a lot of hyper parameters to tune. To address above drawbacks, this paper develops a seismic landslide susceptibility model (called ADASYN-LDA) combining adaptive synthetic sampling approach (ADASYN) with linear discriminant analysis (LDA). Landslides triggered by 2014 Ludian earthquake event are used as an example to evaluate this model.

2. Method

2.1. Adaptive synthetic sampling approach (ADASYN)

The ADASYN algorithm is proposed by He [7], and is briefly introduced below:

The original data set \( D_o \) contains \( N \) samples, in which \( n_s \) and \( n_l \) are the number of minority class samples (landslide units) and the number of majority class samples (non-landslide units), respectively. The number of each class in \( D_o \) may be unbalanced, and we can define the level of class imbalance as:

\[
I = n_l / n_s \tag{1}
\]

Where \( I \in (0,1] \). The ADASYN algorithm will generate some minority class samples into the original data set \( D_o \) to obtain a new data set \( D_n \) with a balance level \( \alpha \). The number of minority class samples which needs to be generated synthetically can be calculated as:

\[
C = \alpha (n_l - n_s) \tag{2}
\]

Where \( \alpha \in [0,1] \). For each minority class instance \( m_i \) (\( m_i \) is an instance with \( n \) features), determine its \( M \) nearest samples using the Euclidean distance method, and the normalized \( \bar{r}_i \) ratio[7] can then be calculated. The number \( (c_i) \) of synthetical samples which needs to be generated for \( m_i \) is calculated as:

\[
c_i = C \bar{r}_i \tag{3}
\]

The ADASYN algorithm generates each of the \( c_i \) synthetic samples as:

\[
s_i = m_i + (m_{zi} - m_i) \beta \tag{4}
\]

Where \( \beta \in [0,1] \), is a random number. Through the above process, the new data set \( D_n \) with \( N + C \) samples is obtained and can be used to train a Linear discriminant analysis (LDA). The model (ADASYN-LDA) combines the ADASYN algorithm with LDA will be applied in seismic landslide susceptibility assessment in this paper.

2.2. Linear discriminant analysis (LDA)

LDA is a classical linear classification model and is widely adopted in landslide susceptibility assessment [8, 9]. LDA calculates the probability of landslide occurrence based on Bayesian theory, and regards this probability as a posteriori probability:

\[
P(Y = k \mid X) = \frac{P(Y = k)P(X \mid Y = k)}{P(X)} \tag{5}
\]

To calculate the posteriori probability \( P(Y = k \mid X) \), \( P(X \mid Y = k) \) is assume to follow multivariate Gaussian distribution, and we can calculate \( P(X \mid Y = k) \) as:

\[
P(X \mid Y = k) = \frac{1}{(2\pi)^{n/2} |\Sigma_k|^{1/2}} \exp \left(-\frac{1}{2}(X - \mu_k)^T \Sigma_k^{-1} (X - \mu_k) \right) \tag{6}
\]

Where \( Y \) is the target variable of an instance, landslide presence or landslide absence, \( Y \in \{0,1\} \), and \( k \) is the classification value. \( X \) is a \( n \)-dimensional vector denoting the landslide conditioning factors with
The covariance matrix of the training samples is given by 

\[ \Sigma_k \] for classification 

and the mean vector of the training samples is given by 

\[ \mu_k \] for classification.

Figure 1. The location of the study area and the distribution of landslides[10] induced by 2014 Ludian earthquake

In landslide susceptibility assessment the landslide mapping unit should be determined first. The grid cell is selected as the landslide mapping unit in this study, because this unit can better match the unit of earthquake parameters. The seismic landslide susceptibility model is created based on ADASYN-LDA model, and uses 12 seismic landslide conditioning factors as independent variable and landslide presence (or absence) as target variable. The spatial partition method [11] is used for model training and validation, and then the results from the susceptibility assessment can be evaluated.

3. Case study

The study area in Ludian County, Yunnan Province, lies in the belt between Yunnan-Guizhou Plateau and Sichuan Basin (figure 1). The elevation of the study area ranges from 866m to 2807m, with an average value of 1771m. Niulan river flows through this study area, forming a magnificent landscape of high mountains and deep valleys. In recent years 6 moderate and strong earthquakes with magnitude Ms 5.0 or above have occurred in this study area [12]. On August 3, 2014, an earthquake with the magnitude of Mw 6.2 hit Ludian County, with the epicenter of 27.10°N, 103.34°E and focal depth of 12 km, and induced a great death toll with 617 deaths and 3143 injuries, and triggered extensive landslides[13]. In this study, the landslides triggered in 2014 Ludian earthquake is used to conduct a seismic landslide susceptibility assessment.

The digital elevation model used in this study is Aster GDEM V2 with a resolution of 20 meters and 30 meters in vertical and horizontal respectively, and is derived from Geospatial Data Cloud (http://www.gscloud.cn/). The ground shake parameters of 2014 Ludian earthquake including PGV, PGA and epicenter location, are derived from ShakeMap of the U.S. Geological Survey. Landslide
inventory with 1024 landslide observations in this earthquake, is derived from a database of world earthquake-triggered ground failure inventories[14], and made by Ying-ying [10]. Figure 1 shows the epicenter of 2014 Ludian earthquake and the distribution of the co-seismic landslides. We divide the study area into zone A and B for model training and validation for later seismic landslide susceptibility assessment. In this paper terrain (slope gradient, curvature, profile curvature and plane curvature), location (elevation and topographic position index), direction (slope aspect and epicentral direction), and epicenter and earthquake parameters (PGV, PGA and epicentral distance) are used to build the seismic landslide susceptibility model.

Figure 2. Seismic landslide susceptibility maps for zone A and B in Ludian earthquake event. (a) seismic landslide susceptibility maps of ADASYN-LDA_BA, and (b) seismic landslide susceptibility maps of ADASYN-LDA_AB.

4. Seismic landslide susceptibility assessment

4.1. Process of the seismic landslide susceptibility assessment
ADASYN-LDA proposed in the method section is used in this seismic landslide susceptibility assessment. After using the training data set to build the landslide susceptibility model, a validation data set independent of the training data set is needed to objectively evaluate the prediction results of the susceptibility model. In this paper the validation method of spatial partition [11] is used. The study area is partitioned into two zones A and B (Figure 1). We use the actual landslide samples in zone A as the training set to build the seismic landslide susceptibility model, and use the actual landslide samples in zone B as the validation set to evaluate the model. The model using above building and evaluation sequence is called ADASYN-LDA_AB. The second landslide susceptibility assessment is performed in an opposite way, with zone B as the training set, and zone A as the validation set, and the model is called ADASYN-LDA_BA. It is usually necessary to take much time to train the complicated nonlinear machine learning model (e.g. the Support Vector Machine (SVM) or Neural Networks (NN)). By comparison, only 5 seconds are spent on training the two ADASYN-LDA models above using a computer with a CPU of i74930mx and a memory of 16g.
4.2. Results and discussion

Fig. 2a, b are the seismic landslide susceptibility maps obtained by ADASYN-LDA_AB and ADASYN-LDA_BA models, respectively. In both two maps, the high susceptibility zones are distributed in the middle or low elevation parts of the bank slopes of the main river valleys. The areas far away from the main river valleys, even if close to the ridges where the seismic amplification effect may be significant, are low susceptibility zones. In addition, most of the actual landslides are distributed in the higher susceptibility area, while only a few landslides are distributed in the lower susceptibility area. The above analyses demonstrate that the predicted susceptibility values of the two models are in a good agreement with the actual landslide distribution.

| Model       | TP    | TN    | FP    | FN    | TPR   | TNR   | AUC    | FPR   | BA    |
|-------------|-------|-------|-------|-------|-------|-------|--------|-------|-------|
| ADASYN-LDA_AB | 3400  | 157081| 39507 | 899   | 0.791 | 0.799 | 0.8737 | 0.201 | 0.7950|
| LDA_AB       | 0     | 196588| 0     | 4299  | 0     | 1     | 0.8272 | 0     | 0.5000|
| ADASYN-LDA_BA| 399   | 154589| 5097  | 1097  | 0.267 | 0.968 | 0.8575 | 0.032 | 0.6174|
| LDA_BA       | 0     | 159686| 0     | 1496  | 0     | 1     | 0.8226 | 0     | 0.5000|

We further evaluate the model prediction performance by quantitative method. First, we use the confusion matrix, and calculate True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) [15] from the model outputs, as shown in Table 1. TP Rate (TPR), FP Rate (FPR) and TN Rate (TNR) are often used in landslide susceptibility assessment, and can be calculated as: TPR = TP/(TP+FN), FPR = FP/(FP+TN) and TNR = TN/(TN+FP), respectively. Considering that the landslide units are very rare compared with non-landslide units in seismic landslide susceptibility assessment. This paper uses some measures less susceptible to the sample imbalance. The balanced accuracy (BA) is a measure suitable for imbalanced datasets and can be calculated as: BA = (TPR + TNR) / 2. For evaluating the improvement of ADASYN algorithm on the sample imbalance in landslide susceptibility assessment, we also calculated the results from the traditional LDA (linear discriminant analysis) model, which are LDA_AB and LDA_BA, respectively. Table 1 shows that the BA of the two LDA is only 0.5. By comparison, two LDA models with ADASYN method have good prediction performance with a balanced accuracy of 0.7950 and 0.6174, respectively.

![Figure 3. ROC curve of the 4 models](image)

ROC curve is also a measure less susceptible to sample imbalance. In ROC curve, the abscissa is FPR and the ordinate is TPR. AUC is the area under ROC curve, and the larger AUC indicates the better model prediction performance [15]. Figure 3 shows that the AUC of ADASYN-LDA_AB and ADASYN-LDA_BA are 0.8737 and 0.8575, respectively. The AUC value greater than 0.8 indicates that the prediction performances of above two models are good enough. LDA_AB and LDA_BA models without ADASYN method are 0.8272 and 0.8226, respectively. According to the above analysis, the
LDA model using ADASYN method improves the prediction performance in seismic landslide susceptibility assessment affected by sample imbalance that the number of landslide units is far less than the number of the non-landslide units.

5. Conclusions
We propose a seismic landslide susceptibility (SLS) model called ADASYN-LDA which combining the ADASYN sample method and linear discriminant analysis (LDA), and then evaluates it using 2014 Ludian earthquake event in China. 12 factors including the topography, direction, location factors and earthquake parameters are used as seismic landslide conditioning factors. The spatial partition method are used for model training and validation. The seismic landslide susceptibility maps given by the two ADASYN-LDA models are in a good agreement with the actual landslide distribution, and thus objectively represent the spatial characteristic of the seismic landslides in this earthquake. In SLS assessment, the two ADASYN-LDA models spend only a few seconds in training, which greatly reduces the time cost compared with the complicated nonlinear artificial intelligence such as the SVM and the NN. Compared with the traditional LDA model without ADASYN sample method, the two ADASYN-LDA models have better prediction performance and thus mitigate the effect caused by sample imbalance in SLS assessment. The model proposed in this paper can estimate the spatial probability of seismic landslides, and thus may provide guidelines for emergency responses making and land use planning to reduce the seismic landslide hazard and risk.

References
[1] Su, F., Cui, P., Zhang, J., et al. (2010) Susceptibility assessment of landslides caused by the wenchuan earthquake using a logistic regression model. Journal of Mountain Science, 7: 234-245.
[2] Kritikos, T., Robinson, T.R., Davies, T.R.H. (2015) Regional coseismic landslide hazard assessment without historical landslide inventories: A new approach. Journal of Geophysical Research: Earth Surface, 120: 711-729.
[3] Parker, R.N., Rosser, N.J., Hales, T.C. (2017) Spatial prediction of earthquake-induced landslide probability. Nat Hazards Earth Syst Sci Discuss, 2017: 1-29.
[4] Xu, C., Dai, F., Xu, X., et al. (2012) GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China. Geomorphology, 145-146: 70-80.
[5] Xu, C., Xu, X., Dai, F., et al. (2012) Comparison of different models for susceptibility mapping of earthquake triggered landslides related with the 2008 Wenchuan earthquake in China. Computers & Geosciences, 46: 317-329.
[6] Song, Y., Gong, J., Gao, S., et al. (2012) Susceptibility assessment of earthquake-induced landslides using Bayesian network: A case study in Beichuan, China. Computers & Geosciences, 42: 189-199.
[7] He, H., Bai, Y., Garcia, E.A., et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning; proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), F, 2008 [C]. IEEE.
[8] Carrara, A., Crosta, G., Frattini, P. (2008) Comparing models of debris-flow susceptibility in the alpine environment. Geomorphology, 94: 353-378.
[9] Rossi, M., Guzzetti, F., Reichenbach, P., et al. (2010) Optimal landslide susceptibility zonation based on multiple forecasts. Geomorphology, 114: 129-142.
[10] Ying-ying, T., Chong, X., Xi-wei, X., et al. (2015) Spatial distribution analysis of coseismic and pre-earthquake landslides triggered by the 2014 Ludian MS6. 5 earthquake. Seismology and Geology, 37: 291-306.
[11] Chung, C.J.F., Fabbri, A.G. (2003) Validation of Spatial Prediction Models for Landslide Hazard Mapping. Natural Hazards, 30: 451-472.
[12] Xi-Wei, X., Guo-Yan, J., Gui-Hua, Y., et al. (2014) Discussion on seismogenic fault of the Ludian M(s) 6.5 earthquake and its tectonic attribution. CHINESE JOURNAL OF GEOPHYSICS-CHINESE EDITION, 57: 3060-3068.
[13] Lin, X., Zhang, H., Chen, H., et al. (2015) Field investigation on severely damaged aseismic buildings in 2014 Ludian earthquake. Earthquake Engineering and Engineering Vibration, 14: 169-176.
[14] Schmitt, R.G., Tanyas, H., Nowicki Jessee, M.A., et al., 1064 [R]. Reston, VA, 2017.
[15] Fawcett, T. (2006) An introduction to ROC analysis. Pattern Recognition Letters, 27: 861-874.