Susceptibility Analysis of Earthquake-Induced Landslide Using Random Forest Method

T.Y. Miao
State Key Laboratory of Earth Surface Processes and Resources Ecology
Beijing Normal University
Beijing, China
Academy of Disaster Reduction and Emergency Management
Ministry of Civil Affairs & Ministry of Education,
Beijing, China

M. Wang
State Key Laboratory of Earth Surface Processes and Resources Ecology
Beijing Normal University
Beijing, China
Academy of Disaster Reduction and Emergency Management
Ministry of Civil Affairs & Ministry of Education,
Beijing, China

***Abstract*--This paper used random forest method to produce susceptibility analysis for earthquake-induced landslides. A landslide susceptibility model was built based on random forest method and data of environment variables and Newmark displacement. Landslide susceptibility mapping and attributes evaluation was also preformed. This model was tested in Pingwu County of Sichuan Province. The results show that the random forest method provides better accuracy on landslide prediction, and also shows smoother characteristic on the distribution of the landslide susceptibility when compared to wildly used regression models. The results confirm that the random forest method can be performed as an effective approach for landslide susceptibility evaluation and mapping.**

Keywords-random forest; earthquake-induced landslide; susceptibility analysis; mapping; Wenchuan earthquake

I INTRODUCTION

Asia has the highest landslides risk around the globe [1], and landslides are responsible for a considerably large property damages and human casualties [2]. Landslide can be induced by various environmental factors which tend to have complex and compound relationships. One of the most prominent features of earthquake is triggering landslides, especially in the mountainous areas. Earthquake-induced landslides usually have the characteristics of large-scale, clustering in spatial and continuity in temporal. How to effectively evaluate the landslide susceptibility is one of the major focus in disaster prevention and reduction, and also a challenge in post-earthquake recovery and reconstruction.

Landslide susceptibility analysis is a combined qualitative and quantitative analysis of the spatial distribution of existing or potential landslides [3]. There are three major landslide susceptibility models including heuristic, statistical, and physically based models [4]. The majority of landslide susceptibility analysis and mapping research relies on statistical models [5-7]. These researches mainly focused on the application of different statistic inference to reveal the relationship between spatial distribution of landslides and environmental factors. For example, Beaze used multiple linear regressions to analyze shallow sliding, and pointed out that the dominant factors are slope and vegetation coverage [8]. Logistic regression is a popular and convenient approach in landslide susceptibility mapping [9-11], and behaved relatively well in exploring major affecting factors [12, 13]. Duman et al [14] also used Logistic regression in landslide susceptibility zoning, and indicated that geological conditions have important impact on the occurrence of landslides.

Lacking complete landslide inventory, statistics methods remain to be the important approach to evaluate landslide hazard and generate susceptibility mapping. The landslides induced by major earthquakes (i.e., Wenchuan earthquake) often show contingency and complexity characters and relationships between landslides and environment factors may not be captured by traditional regression approaches. Different from existing literature, we used random forest algorithm to analyze the susceptibility of earthquake-induced landslides and investigate the dominant affecting factors. The susceptibility map generated by using the random forest method was also compared with that using the logistic regression.

II STUDY AREA AND DATA

The study area is located in Nanba-Shuiguan in Pingwu County, Mianyang, Sichuan. This area lies approximately 180 kilometers northeast to the 2008 Wenchuan earthquake epicenter, and it is one of the most damaged areas during the earthquake. This mountainous area has elevation ranging from 671 to 2667 meters. The Fujiang river and Hongxi creek flow through the area. Figure 1 shows the basic geographic information of the study area. The orange polygons mark the location of the earthquake-induced landslides trigged by Wenchuan earthquake. The landslides information is obtained and visually interpreted by comparing two SPOT-5 satellite images taken before the Wenchuan earthquake (on September 5, 2006) and after the earthquake (on June 4, 2008). The landslides information is validated and calibrated by multiple field trips conducted during 2012 to 2014.

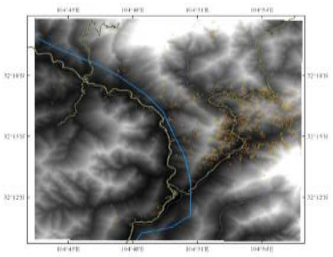

FIGURE 1: THE STUDY AREA AND LANDSLIDES.

In this study we selected the following variables to examine the relationship between landslide and environmental factors: elevation (Digital Elevation Model, DEM), lithology, Normalized Difference Vegetation Index (NDVI), land-use type, vegetation type, roughness, relief degree of land surface (RDLS), profile curvature, and plan curvature. The 30 by 30 meters DEM data were used to compute roughness, RDLS, profile curvature, and plan curvature in the study area. NDVI was obtained from the SPOT-5 satellite image of September 5th, 2006. Lithology data was normalized from 1:2,500,000 geology map of Sichuan province. Land-use information was obtained from the 1:500,000 land-use map of China. Vegetation type was obtained from the 1:1,000,000 vegetation map of China. Additionally, this study used Newmark displacement caused by the earthquake. The calculation and detailed information of Newmark displacement of the study area can be found in the previous study of Liu [15] and Wang [16]. All the data is integrated into 30 by 30 meters grids for the susceptibility analysis.

III RANDOM FOREST

The random forest algorithm is a combination of decision tree predictors presented by Leo Breiman [17] in 2001. The algorithm generates multiple classification and regression trees (CART), and the final classification result is voted among all the trees in the "forest".

The algorithm of random forest (for classification) can be stated as follow [17, 18]:

1) Multiple bootstrap sub-sample sets ware random drawn, the process is called bagging. For each bootstrap procedure, the unselected samples were called OOB (out-of-bag) data, which are used as test sample to calculate the test statistic of the random forest model.

2) A CART is grown from each bootstrap subset.

The bootstrap-CART procedure is repeated, for instance, 500 times, and 500 CARTs are grown (the so-called "forest").

3) When a test sample or a new sample enters the forest, the algorithm distribute the data to each tree for classification. The final classification outcome is voted among the result of the trees.

4) OOB sample is used as test set, and its mis-classification ratio is calculated to evaluate the effectiveness of the random forest model.

5) The randomized bagging process is designed to ensure higher stability. The two randomized sample process during sample drawn (in bootstrap) and node selection (in CART) ensured a higher accuracy in the classification prediction [19] without over-fitting the data [17]. Additionally, because a large number of trees are grown in the forest, the system error in generalization can be limited [17].

Shortly after the random forest algorithm was developed in 2001, the approach have been adopted in many different fields. Random forest approach shows outstanding performance in processing satellite and radar remote sensing images [20-22]. However, there are very few literatures involving landslide problems. Che-Wei Shen et al show that the random forest model can work well to identify the importance of factors related to mud flow triggered by Typhoon, while the principal component analysis cannot provide satisfactory results [23]. Catani et al [24] used the random forest algorithm to investigate the difference among rankings of factor importance for landslide occurrence at various spatial scales, which is one of the few studies on landslides using random forest.

IV SUSCEPTIBILITY MODEL AND RESULTS

In the study area, the data sample is generated from 420,336 grids, and only 2.4% of the total grids have landslides. Therefore, this study uses all landslide grids and randomly drawn partial non-landslide grids as the impute data for random forest model. The selection of non-landslide data uses ARCMAP automatic spatial random sampling tool to ensure that the data is collected randomly in space. The random forest model is computed using CoreLearn package in R program.

A. Data Imbalance

In general treatment of analysis imbalanced data samples, the ratio between landslide data and non-landslide data is fixed at 1:1. But for disaster research, the number of disaster sample is extremely limited. Thus, before determine the sample ratio between landslide and non-landslide data, this study will briefly discuss how different sample size ratio might affect the model final results.

The sample size of landslide data is fixed, and non-landslide data set is drawn separately with the sample size ratio at 1:1, 1:1.5, 1:2 and 1:5. Random forest models are executed for each data set to compare model accuracy.

In the Table 1, accuracy is defined as the percentage of correct classification over all sample. Specificity (also called true positive rate) is defined as the percentage of correctly classified landslide sample. Sensitivity (also called true negative rate) is defined as the percentage of correctly classified non-landslide sample. The result shows that, the ratio of non-landslide sample is definitely affecting the model prediction accuracy. When the ratio of landslide sample decreases, although the model overall accuracy increases, the classification accuracy on landslide is getting worse. The increasing overall accuracy is at the cost of decreasing classification accuracy of the landslide sample, because the model is overwhelmed by the increasing ratio of non-landslide sample. Since this study focused on a better

identification of the susceptibility of landslides, rather than only compare model overall accuracy, model specificity (which shows the accuracy on identify landslide samples) must be taken into consideration as an importance measurement.

TABLE I. RANDOM FOREST MODEL ACCURACY UNDER DIFFERENT SAMPLE SIZE RATIO.(IN PERCENTAGE).

Sample size ratio	1:1	1:1.5	1:2	1:5
Accuracy	95.0	95.4	95.5	97.4
Specificity	98.0	97.2	96.0	90.8
Sensitivity	92.0	94.2	95.3	98.6

B. *Cost Matrix*

To deal with the effect of imbalanced data on machine learning, Nguyan [25] introduced the concept of cost matrix. It is designed to reduce the accuracy "bubble" caused by imbalanced data. Based on the landslide problem in this study, the mis-classification of a landslide sample is given a higher cost than a non-landslide sample. The CARTs is optimized upon the cost matrix, so that the model total "cost" will be minimized. The question of how to design the cost ratio is still debatable among researchers, Margineantu [26] suggested that the cost ratio should be reciprocal of the sample size ratio, but Hand [27] suggests that the reciprocal ratio is only a convenience approach which does not on the ground describe the severity of misclassification.

Thus, this study applies 2 cost matrix on each sample ratio, and compare the model prediction accuracy (Table 2).

TABLE II. RANDOM FOREST MODEL ACCURACY UNDER DIFFERENT SAMPLE SIZE RATIO AND DIFFERENT COST MATRIX.(IN PERCENTAGE).

Sample and cost matrix ratio	sample 1:1 cost2:1	sample 1:1 cost5:1	sample1: 1.5 cost2:1	sample1: 1.5 cost5:1	sample 1:2 cost2:1	sample 1:2 cost5: 1	sample1 :5 cost2:1	sample 1:5 cost5: 1
Accuracy	94.89	94.98	95.36	95.44	95.51	95.52	97.34	97.37
Specificity	97.93	97.96	97.24	97.04	95.89	96.05	90.89	91.07
Sensitivity	91.85	92.00	94.11	94.04	95.32	95.25	98.64	98.63

Unfortunately, the cost matrix only has a miner improvement on specificity for various sample ratios. The reason can be that the model without cost matrix already have a considerably high prediction accuracy, the misclassification ratio of landslide sample have always under 10% for all models. Therefore, the effect of cost matrix do not have a distinctly difference on the prediction results.

C. *Model and Results*

The final model chooses the landslide and non-landslide sample set with 1:2 ratio, and no cost matrix is used. The random forest model is built using R program and CoreLearn package. The model prediction accuracy result is shown in Table 3. The overall model accuracy is 92.79%, and area under ROC curve reaches 0.9935.

TABLE III. RANDOM FOREST MODEL OOB ERROR AND PREDICTION MATRIX.

OOB error：7.21%			
prediction matrix			
	No	Yes	Error rate
No	19324	1408	0.0679
Yes	833	9533	0.0804

All samples are calculated through the random forest model, and in the total 420,336 samples, only 32,460 samples are mis-classified, and the model overall accuracy reaches 92.4%. Among all the mis-classified data, only 417 of them are landslide samples, that means only 4% of landslide samples are mis-classified. The earthquake-induced landslide susceptibility mapping is composed by ARCGIS, Figure. 2. Figure 2 shows the values of the landslide susceptibility calculated by the random forest model. The red color denotes high susceptibility of landslides, and these locations are most sensitive toward external perturbation.

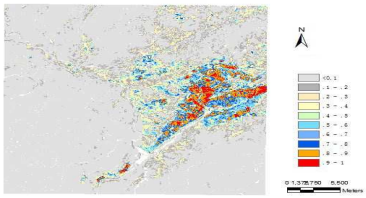

FIGURE II. LANDSLIDE SUSCEPTIBILITY (RANDOM FOREST MODEL).

For all the sample size and cost matrix combinations studied above, the factor importance ranking stays the same through all models. Lithology is the factor with the foremost importance, followed by distance to fault zone, vegetation type, and land-use type, also elevation, RDSL and Newmark displacement , Figure. 3.

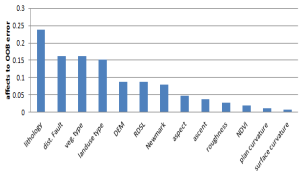

FIGURE III. IMPORTANCE RANKING OF ENVIRONMENTAL FACTORS.

VDiscussions

A. *Result of Susceptibility Analysis of Earthquake-Induced Landslide*

The result of the 92.4% overall accuracy of all sample illustrates that the random forest method can provide a very accurate classification. The model gives even more accurate result on landslide samples, and the mis-classification sample is concentrated in the non-landslide samples. Considering the random forest algorithm is design to resist model over-fitting, it is reasonable to imply that the mis-classification of non-landslide sample is not entirely contributed by system error. The mis-classification suggests that numerous non-landslide samples have the similar characteristic as the landslide samples. Those non-landslide samples have high susceptibility indicate that the corresponding grids have relatively high potential of landslide occurrence. Although landslides did not occur during the Wenchuan earthquake, those high susceptibility areas should be monitored for a longer term in the future study.

B. *Choice over Cost Matrix*

Although cost matrix does not show a significant improvement on the results in this study, the method should still be taken serious consideration in the future study. Weiss [28] compared over-sample, under-sample and cost matrix under different data sets, and the result shows inconclusive on the optimal approach. The tests in this study suggest that the effect of cost matrix may be significant when relatively a large positive samples (i.e., landslide samples) can be obtained.

C. *Comparison with Logistic Regression Model*

Wang and Liu [15, 16] used Logistic regression model to analyze the landslide susceptibility in the same study area. The logistic regression model has a 81.2% overall prediction accuracy. The corresponding susceptibility map is shown in Figure. 4.

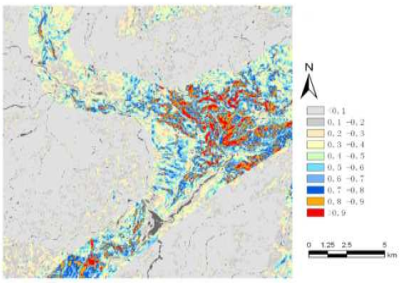

FIGURE IV. LANDSLIDE SUSCEPTIBILITY -- LOGISTIC REGRESSION.

Comparing the two susceptibility maps (Figure. 2 and Figure. 4), the earthquake-induced landslide susceptibility shows a similar spatial distribution trend. However, the accuracy is improved significantly by the random forest model. Furthermore, the Logistic regression model shows highly sensitive over the lithology pattern. The belt shape pattern of susceptibility in Figure. 4 is the consequence of lithology variation. Although lithology also has the highest ranking of importance in the random forest model, the susceptibility map shows a smoother pattern and does not severely influenced by the distribution of lithology especially at the class boundaries.

On the other hand, the factor importance ranking is slightly different between two approaches, but maintains certain consistency. Lithology has the highest importance in both models. Land-use type, Newmark displacement and DEM are factors with high importance in both models.

VICONCLUSIONS

Landslide susceptibility analysis and mapping can significantly contribute to regional disaster evaluation and disaster prevention. This study discussed the usage of random forest approach in landslide analysis, and its applicability on susceptibility mapping. The result is consistent with traditional regression approach, but more accurate in model prediction and smoother in pattern mapping. This study confirms that the random forest method is an effective approach in landslide susceptibility analysis and mapping.

To better understand the disaster mechanism for secondary disasters after Wenchuan earthquake, the future studies can focus on the non-landslide areas but with high susceptibility values. More detailed classification and continuous observation will be conducted to establishing a quantitative relationship between susceptibility and newly evolving landslide or landslide expansion in the future research.

ACKNOWLEDGEMENTS

This research is supported by National Natural Science Foundation of China under grand 41101505 and State Key Laboratory of Earth Surface Processes and Resource Ecology under grand 2014-ZY-05.

REFERENCES

[1] CRED, CRED newsletter, 2006, http://www.emdat.be/publications

[2] Schuster, R. L., Socioeconomic significance of landslides, *Landslides: investigation and mitigation, Special Report*. pp. 247:12–35, 1996.

[3] Glade, T., Anderson, M. & Crozier, M. J., *Landslide hazard and risk*. Wiley Online Library, 2005.

[4] van Westen C. J., Saibal, G., Pankaj, J., *et al*, From Landslide Inventories to Landslide Risk Assessment: An Attempt to Support Methodological Development in India. *Proc. of Landslide Science and Practice*. Springer, pp. 3–20, 2013.

[5] Saro, L. & Touch, S., Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. *Environmental Geology*, 50(6), pp. 847–855, 2006.

[6] Akgun. A., A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at Izmir, Turkey. *Landslides*, 9(1), pp. 93–106, 2012.

[7] Reza, P. H., Biswajeet, P. & Candan, G., Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. *Natural Hazards*, 63(2), pp. 965-996, 2012.

[8] Baeza, C. & Corominas, J., Assessment of shallow landslide susceptibility by means of multivariate statistical techniques. *Earth surface processes and landforms*, 26(12), pp. 1251–1263, 2001.

[9] Bai, S., Wang, J., L¨u, G., *et al*, GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. *Geomorphology*, 115(1), pp. 23–31, 2010.

[10] Shivani, C., Mukta, S. & Manoj, A. K., Landslide susceptibility zonation of the Chamoli region, Garhwal Himalayas, using logistic regression model. *Landsilde*, 7, pp. 411-423, 2010.

[11] Bui, D. T., Lofman, O., Revhaug, I., *et al*, Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. *Natural Hazards*,59, pp. 1413-1444, 2011.

[12] Van Den E. M., Vanwalleghem, T., Poesen, J. *et al*, Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium). *Geomorphology*, 76, pp. 392–410, 2006.

[13] Bai, S., Lu, G., Wang, J. *et al*, GIS-based rare events logistic regression for landslide-susceptibility mapping of Lianyungang, China. *Environmental Earth Sciences*, 62, pp. 139-149, 2011.

[14] Duman, T. Y., Can, T., Gokceoglu, C., *et al*, Application of logistic regression for landslide susceptibility zoning of Cekmece Area, Istanbul, Turkey. *Environmental Geology*, 51(2), pp. 241–256, 2006.

[15] Liu, M., An integrated approach to incorporate triggering and environmental factors in the analysis of regional earthquake-induced landslide hazards . Beijing Normal University, Beijing, China. 2013. (in Chinese)

[16] Wang, M., Liu, M., Yang, S. *et al*, Incorporating Triggering and Environmental Factors in the Analysis of Earthquake-Induced Landslide Hazards. *International Journal of Disaster Risk Science*, 5, pp. 125-135, 2014.

[17] Breiman Leo, Random Forest. *Machine Learning*,45, pp. 5-32, 2001.

[18] Breiman Leo, Bagging Predictors. *Machine Learning*, 24, pp. 123-140, 1996.

[19] Svetnik, V., Liaw, A., Tong, C., *et al*, Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 43(6), pp. 1947–1958, 2003.

[20] Chan, J. C. & Paelinckx, D., Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112(6), pp. 2999–3011, 2008.

[21] Martinuzzi, S., Vierling, L. A., Gould, W. A. *et al*, Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. *Remote Sensing of Environment*, 113(12), pp. 2533–2546, 2009.

[22] Fang, K., Wu, J., Zhu, J. *et al*, A Review of Technologies on Random Forests. *Statistics & Information Forum*, 26, pp. 32-38, 2011 (in Chinese).

[23] Shen, C., Hsiao, C., Ku, B., *et al*, Relationship Analysis of Debris Flow-inducing Factors in Typhoon Morakot Affected Area By Using Random Forest Algorithm. *Proc. of 2012 Taiwan disaster management seminar*, Taiwan, 2012. (in Chinese)

[24] Catani, F., Lagomarsino, D., Segoni, S., *et al*, Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Natural Hazards & Earth System Sciences*, 13(11) , pp. 2815-2831, 2013.

[25] Nguyen, T.N., Gantner, Z. & Schmidt Thieme, L., Cost-sensitive learning methods for imbalanced data. *Proc. of IEEE International Joint Conference on Neural Networks (IJCNN 2010)*, Barcelona, Spain, pp. 1–8, 2010.

[26] Margineantu, D. D., When does imbalanced data require more than cost-sensitive learning? *Proc. of Learning from Imbalanced Data Sets (Technical Report WS-00-05)*, Austin, TX, USA, pp. 47–50,2000.

[27] Hand D. J., Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), pp.103-123, 2009.

[28] Weiss, G. M., McCarthy, K. & Zabar, B., Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? *Proc. of the 2007 International Conference on Data Mining*, CSREA Press, pp. 35-41, 2007.