SEPARATION OF LANDSLIDE SOURCE AND RUN-OUT AREAS WITH MACHINE LEARNING FOR LANDSLIDE INVENTORY REFINEMENT

Jhe-Syuan Lai

Feng Chia University, No. 100, Wenhwa Rd., Taichung, 40724, Taiwan
Email: jslai@fcu.edu.tw

KEY WORDS: Digital elevation model, landslide inventory, machine learning, random forests.

ABSTRACT: There are three common features typical of natural terrain landslides, i.e., source, trail and deposition. The term run-out, generally used to describe the downslope displacement of failed geo-materials by landslides, is used in this study to represent the combination of the landslide trail and deposition. In general, the area of a landslide in the landslide inventory, detected from remotely sensed images, might contain the run-out area (called landslide affected area in this study), unless manually removed by the geologist or expert using the auxiliary data. However, the run-out area should be excluded in a strict definition of real landslides because it is caused by different mechanisms. This might produce biases and reduce the reliability of landslide inventory, i.e., landslide samples including run-outs. To this end, this study integrates topographic variables derived from the digital elevation model with Random Forests, one of machine learning algorithms, for separating landslide source and run-out areas from the landslide affected area. Preliminary results indicate that the accuracies of developed models can reach 80% in most cases.

1. INTRODUCTION

Landslide inventory (database) is an important material for landslide analysis (Guzzetti et al., 2012), such as landslide susceptibility and hazard assessments. From a geotechnical or geological point of view, there are three common features typical of natural terrain landslides (Dai and Lee, 2002), i.e., source, trail and deposition. The term run-out, generally used to describe the downslope displacement of failed geo-materials by landslides (Mondini et al., 2011), is used in this study to represent the combination of the landslide trail and deposition. In general, the area of a landslide in the landslide inventory, detected by means of automatic or semi-automatic algorithms from remotely sensed images, might contain the run-out area (called landslide affected area in this study), unless manually removed by the geologist or expert using aerial stereo-photos or other auxiliary data. The run-out area should be excluded in a strict definition of real landslides because it is caused by different mechanisms. This might produce biases and reduce the reliability of landslide inventory, i.e., landslide samples including run-outs. To address this issue, this study integrates topographic variables, such as aspect, curvature, elevation and slope, derived from the Digital Elevation Model (DEM) with Random Forests, one of machine learning algorithms, for separating landslide source and run-out areas from the landslide affected area as well as refining the present landslide inventory.

2. DATA AND METHODS

An area of 117 km^2 of the Kaoping watershed in southern Taiwan is selected as the study site (Figure 1). The used 10-m DEM was produced by Chiang et al. (2012). A landslide inventory generated after Typhoon Morakot was further interpreted manually to separate the source and run-out classes according to stereo aerial photos and auxiliary data. This study selects top 3 polygons which are the largest area size for exploring the separability between landslide source and run-out areas based on the Random Forests algorithm. The inventory polygons are converted into the pixel format (10 by 10 meters) to extract the corresponding topographic factors for analysis.

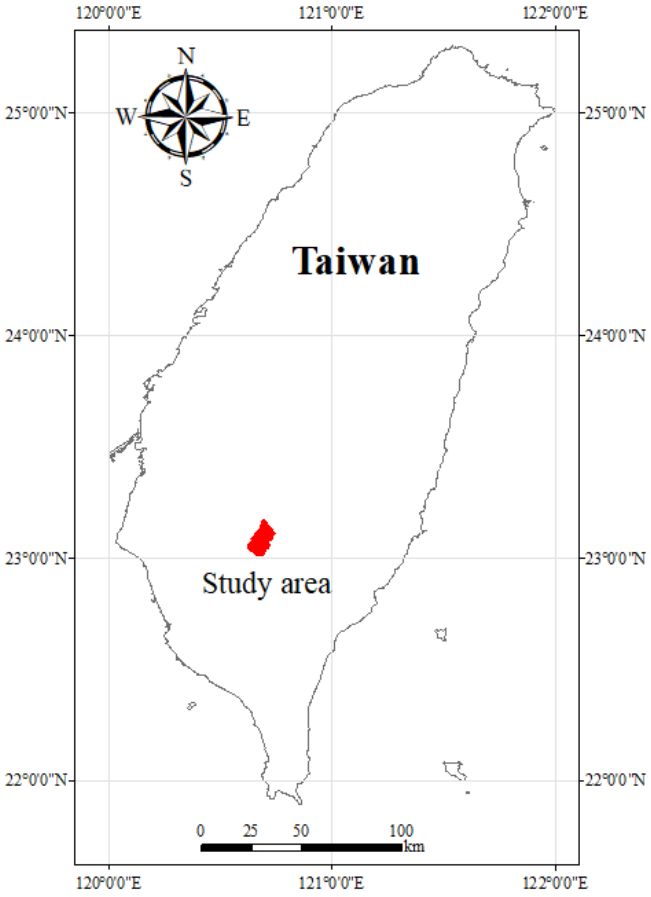

Figure 1. Study site

The Random Forests (RF) classifier (Breiman, 2001) is employed for constructing the landslide susceptibility model in this study. This is an extension of Decision Tree (DT) algorithm which is a classical and popular approach in the machine learning domain. The concept of both RF and DT classifiers is similar and both adopt the Information Gain (IG) measure to evaluate the degree of impurity of causative factors. The larger IG indicates that the corresponding causative factor should be selected in a higher priority to construct a conditional node and ignore this factor in next computation. After several iterations, a tree model, which comprises a sequence of "If-Then" rules, is extracted to classify other instances. The difference between the random forests and decision tree algorithm is that the former randomly separates training data into many subsets to build many trees (so called the forest) and optimize them.

For modeling process, 2/3 of landslide source and run-out samples (training data) are randomly selected polygon by polygon to construct the RF models. The developed models are used to predict other polygon samples (check data). This study applies three commonly used quantitative indices for verification, including Overall Accuracy (OA), User's Accuracy (UA), and Producer's Accuracy (PA).

3. RESULTS AND DISCUSSION

According to the algorithm and procedure mentioned in the previous section, the constructed models are verified by the OA, UA and PA indexes derived from confusion matrices. The quantitative evaluations for verification are shown in Table 1. The table shows that the preliminary results can reach 80% in most cases. However, some disagreements can be observed, such as the models constructed from No. 1 and No. 2 polygons for prediction of No. 3 polygon. More precisely, there are higher omission (mission) errors for landslide source predictions. To address the unbalanced prediction results, the constructed models with an extremely false alarm or missing

error (called over-fitting effect), the impact of using cost-sensitive analysis to adjust the decision boundary (e.g., Lai, 2018; Tsai et al., 2016) for the improvement of the RF algorithm will be explored in the future.

Table 1. Quantitative evaluations for separating landslide source and run-out areas

Training Data	Check Data	OA (%)	Run-out		Source	
			UA	PA	UA	PA
No. 1	No. 2	98.75	1	0.98	0.97	1
	No. 3	**61.72**	0.61	1	1	**0.06**
No. 2	No. 1	98.38	0.97	1	1	0.96
	No. 3	**60.69**	0.6	1	1	**0.03**
No. 3	No. 1	89.51	1	0.83	0.79	1
	No. 2	85.81	1	0.77	0.74	1

4. CONCLUSION

In this study a procedure is developed for the integration of topographic data and machine learning for separation of landslide source and run-out areas in order to refine landslide inventory. The preliminary results show that the accuracies can reach 80% in most cases. Future works might usefully extend the present use of the constructed models to adjust the decision boundary for improving the prediction capability, and to predict all inventory polygon samples.

ACKNOWLEDGEMENT

This study was supported, in part, by the Ministry of Science and Technology of Taiwan under the project 108-2119-M-035-003.

REFERENCES

Breiman, L., 2001. Random forests. Machine Learning, 45, pp. 5-32.

Chiang, S.-H., Chang, K.-T., Mondini, A.C., Tsai, B.-W., Chen, C.-Y., 2012. Simulation of event-based landslides and debris flows at watershed level. Geomorphology, 138, pp. 306-318.

Dai, F.C., Lee, C. F., 2002. Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. Geomorphology, 42, pp. 213-228.

Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K.T., 2012. Landslide inventory maps: new tools for an old problem. Earth-Science Reviews, 112, pp. 42-66.

Lai, J.-S., 2018. Improving regional landslide susceptibility assessments by integrating geo-spatial data and data mining algorithms. PhD Dissertation, National Central University, Taiwan, 232 pages.

Mondini, A.C., Chang, K.-T., Yin, H.-Y., 2011. Combing multiple change detection indices for mapping landslides triggered by typhoons. Geomorphology, 134, pp. 440-451.

Tsai, F., Lai, J.-S., Lu, Y.-H., 2016. Land-cover classification of full-waveform LiDAR point cloud with volumetric texture measures. Terrestrial, Atmospheric and Oceanic Sciences, 27(4), pp. 549-563.