LANDSLIDE SUSCEPTIBILITY MAPPING USING RANDOM FOREST MODEL IN LAO CAI PROVINCE, VIETNAM

Xuan Quang Truong[1], Xuan Luan Truong[2], Thi Khanh Linh Dang[1]

Thuy Dung Nguyen[1], Duc An Nguyen[1]

[1]Faculty of Information Technology, Hanoi University of Natural Resources and Environment
41A, Phu Dien Road, Bac Tu Liem, Hanoi, Vietnam
Email: txquang@hunre.edu.vn
[2]Faculty of Information Technology, Hanoi University of Ming and Geology
18 Pho Vien, Duc Thang Ward, Tu Liem District, Hanoi , Vietnam
Email: truongxuanluang@gmail.com

ABSTRACT

The main objective of this study is to investigate potential of random forest model in landslide susceptibility mapping in the Lao Cai province of Vietnam. Landslide locations were randomly selected in 70% for training and 30% for validation of the models. Nine factors related with landslide in Lao Cai province were used and built in a spatial database as the thematic maps. The relationships between the location and shape of landslide occurred and these factors were identified by using the random forest model. The model were used to create a landslide susceptibility maps. The final step is validation of the random forest model. For the Random Forest model, the validation accuracy in regression algorithms showed ~ 87%

1. INTRODUCTION

Landslide is the 7th in the table ranking major natural hazards by number of deaths report, From 1903-2004 approximately 16,000 people were killed by landslide in Europe (Nadim et al., 2006). In Vietnam the area affected by landslide spreads widely in the northern mountain area and is distributed narrowly in the middle of country near the border with Lao PDR. Lao Cai province is located in the north of the country and it is well known that landslides occur frequently in Lao Cai. Social and economic losses due to landslides can be reduced by means of effective planning and management. In recent years, many regions around the world have effects of climate changes such as extreme rainfall events (CRV 2015). The rainfall-triggered landslide is especially exacerbated in countries that are located in storm centers of the world, such as Vietnam (Truong et al, 2018).

Landslides susceptibility assessment has to be conducted to identify prone areas and guide risk management. Recently, some new approaches for landslide susceptibility assessment using soft computing techniques such as statistical models (Yilmaz I. 2010); logistic regression (Chen W 2016); artificial neural networks (Pradhan& Lee 2010); support vector machines (Peng et al. 2014); Naive Bayes (Tsangaratos & Ilia 2016) decision trees models (Pradhan 2013); and random forest (Pourghasemi & Kerle 2016).

The main purpose of this study is to use random forest algorithm for selection of useful features and generate landslides susceptibility map based on landslide dataset is composed 279 landslide polygons (data taken from northwest Vietnam geological Mapping Division)

2. STUDY AREA

The study area is a part of Lao Cai province, it is located on the upper part of Red River Basin (21o59'33.35" to 22o46'49.9") and (103o34'50.75" to 104o33'4.136") (Figure 1). The elevation is from 48 to 2392 m above the sea level. 52% of study area is occupied by slope angles, those are higher than 15 degree, areas with slope less than 8^o are 15%. The average temperature ranges between 22 and 24^{o}C and rainfall varies between 1400 mm and 3000 mm. Average annual humidity is over 80%. There are two main seasons: a rainy season spanning from April to October, and a dry season starting from October to March. The study area is located in an active tectonic region with the relatively fast movement of the Red River fault zone that results in continuously landslide occurrences over the years (Truong et al, 2018).

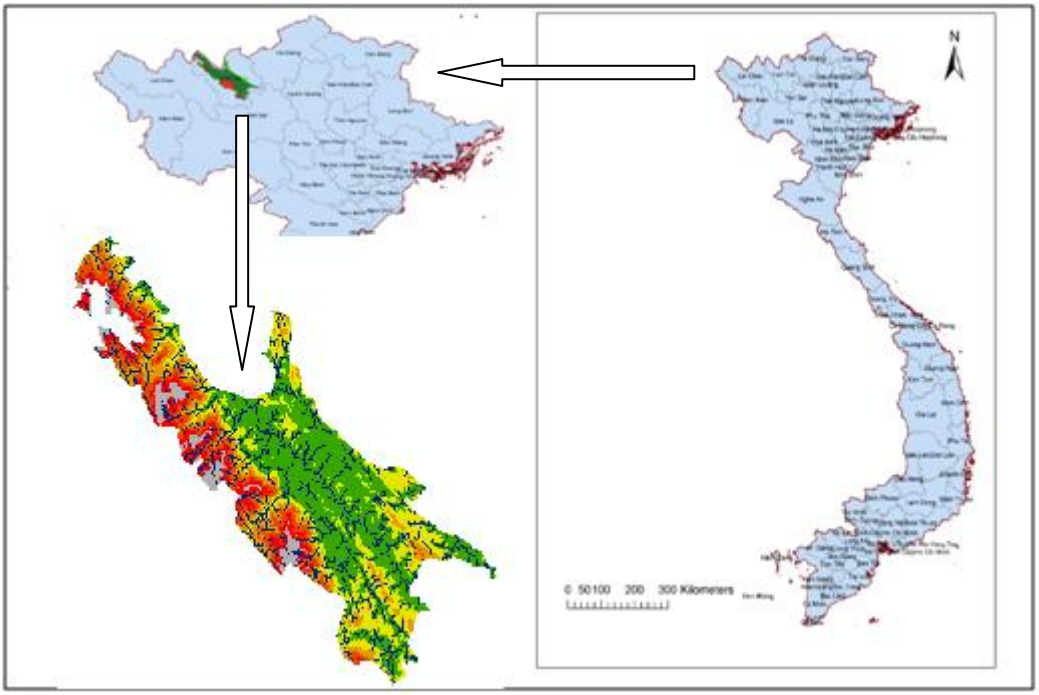

Figure 1: Study area

3. LANDSLIDE CONDITIONING FACTORS

Because the rainfall-trigged landslides in this study area occurred due to interactions of various geo-environmental factors, including topography, land cover, lithology, soil type, and river network (Bui et al, 2017). In this paper, 10 factors were chosen according to the data available and general geo-environment of the study area as shown in Table 1. The condition factors are slope, aspect, altitude, landcover, lithology, soil type, distance to fault, distance to river, rainfall. A digital elevation model (DEM) from ASTER GDEM data with resolution 25x25 m was used to process slope, aspect, elevation of the study area using QGIS 2.16.3. Landcover was extracted from LandsatTM with 30 m resolution in 2013 using ENVIN software. The lithology, geology and fault maps at scale 1:200,000 provided by the Ministry of Natural Resource. The river and rainfall maps were used from the Vietnam national atlas with scale (1:200,000). In this research, distance to fault map with six classes and distance to river with five classes for the study area were constructed by buffering the fault and river lines extracted from fault map and river map above.

Table 1: Classes of landslide conditioning factors

Type	Classes	Number of class pixels	Landslide pixels	Bel
distance to faults (m)	0-100	189237	90	0.171
	100-200	185406	74	0.143
	200-350	602887	124	0.258
	250-350	434748	149	0.123
	>350	1433830	1209	0.304
Geology	Pebble gravel, pebble, gravel	203363	179	0.253
	Chang Pung Formation	241111	192	0.229
	Granodiorite, granite biotite	391716	147	0.108
	Pegmatite, granit split	10765	2	0.053
	Basalt and komatiite basalt	9152	0	0.000
	Quartz and quartz sericite	100143	124	0.356
	Mia Lé Formation	203363	179	0.253
landcover	River	12785	5	0.068
	Residential land	959	0	0.000
	Agricultural land	253360	104	0.071
	Bare land	2742	0	0.000
	Shrubs land	50180	14	0.048
	Agricultural land	176853	281	0.276
	Dense forest land	1524283	768	0.088
	Medium forest land	33200	60	0.314
	Forest land	535206	413	0.134
rainfall (mm)	2400 - 2600	12207	25	0.204
	2200 - 2400	1246475	160	0.013
	2000 - 2200	638461	293	0.046
	1800 - 2000	400772	413	0.102
	1600 - 1800	184817	326	0.175
	1200-1600	106719	429	0.460
River	0-50	181011	119	0.161
	50-100	174999	141	0.197
	100-150	165706	157	0.232
	150-200	160187	180	0.275
	>200	1907582	1049	0.135
Soil	Goup1	624	21	0.013
	Goup2	119	33	0.767
	Goup3	321	72	0.214
	Goup4	10610	662	0.001
	Goup5	7356	858	0.004
Slope	0-8.3^{o}	401275	196	0.065
	8.3-14.7^{o}	482772	239	0.055
	14.7-20.6^{o}	510900	295	0.060
	20.6-26.4^{o}	490887	341	0.075
	26.4-33.2^{o}	392803	292	0.101
	33.2-41.8^{o}	234582	241	0.233
	41.8-78.5^{o}	76343	45	0.411
Aspect	-1	352	0	0.000
	35-72 and 320-360	352597	121	0.058
	72-113	443296	201	0.061
	113-156	350151	206	0.101
	156-198	284492	176	0.131
	198-238	285811	237	0.174
	238-279	313892	415	0.253
	279-320	282413	181	0.136
DEM	8 classes from 48-2400 m	Bel values: 0.04 - 0.3		

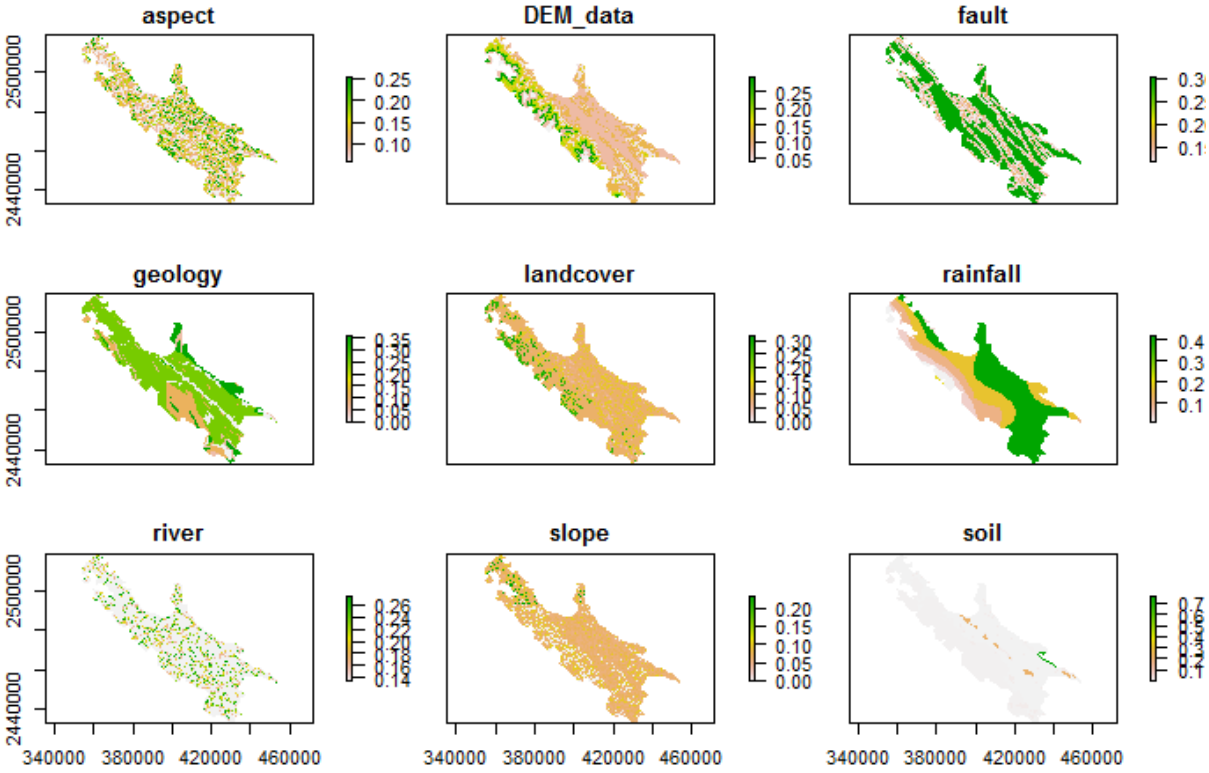

Figure 2: Evidential belief function of factors based on Table 1

Suppose that we have a set of landslide conditioning factors $C=(C_i, i=\overline{1..n})$ Probability of $P(C) \rightarrow [0,1]$, if *A* is subset of C The *m(A)* measures the degree to which the evidence support *A*; it is denoted belief function (Bui et al 2012). The evidential belief functions (EBF) model has been widely used for presenting the distribution of landslide occurrences. Equation (1) shows the EBF function (Shrestha et al 2017):

$$Bel(C_{ij}) = \frac{W(C_{ij}D)}{\sum_{j=1}^{m} W(C_{ij}D)} \quad \text{Where} \quad W(C_{ij}D) = \frac{\frac{N(C_{ij} \cap D)}{N(C_{ij})}}{\frac{N(D) - N(C_{ij} \cap D)}{N(T) - N(C_{ij})}} \qquad (1)$$

Where, the study area T consists of a total number of pixels N(T), landslide D occurs in a number of pixels *N(D)*. The classes of C_i in T are given by *Cij (j = 1,..,m)*, then by overlaying a binary map (i.e., 0 = absence, 1 = presence). The numerator to calculate parameter $W(C_{ij}D)$ is the conditional probability of the existence of the landslide. Nine landslide factors above were used to prepare the landslide susceptibility maps. The results of calculation shown in Figure 2 and Table 1.

4. RANDOM FOREST ALGORITHM

Radom Forest (RF) classification is a machine learning algorithm, the idea of random forest is to expand multiple decision trees for a random subset of variables associated with

training data. RF combined with randomized node optimization and bagging, RF uses few hundred to several thousand trees, depending on the size and nature of the training set. An optimal number of trees can be found using cross-validation, or by observing the out-of-bag (OOB) error. RFs give various outputs for analyzing results that consist of out-of-bag (OOB) accuracy and evaluate the contribution of Factors (Shrestha et al 2017). In order to investigate potential of random forest model in study area, in this study Random Forest algorithm was used CRAN package, and implemented in the R statistical programming environment.

5. RESULT AND DISCUSSION

The spatial distributions of landslides in each class of factors were analyzed using the spatial analysis tool in ArcGIS 10.2. All 279 landslide polygons were considered in order to calculate the spatial relationship between class factors and landslide distribution. Only 70% of the total landslides (training data) were considered in the RF model. Number of landslide in pixels in training dataset was converted from landslide polygons is 512 pixels. After calculate *Bels* in Equation 1, All values were assigned to a corresponding landslide factors and their classes for each factor. To produce the ensemble, the RF algorithm with 501 trees was applied to prepare landslide susceptibility map. The map obtained from the RF method showed that approximately 37% of study area had none of very low landslide susceptibility, 29%, 22% and 12% were flagged as having low, moderate, high, and very high susceptibilities.

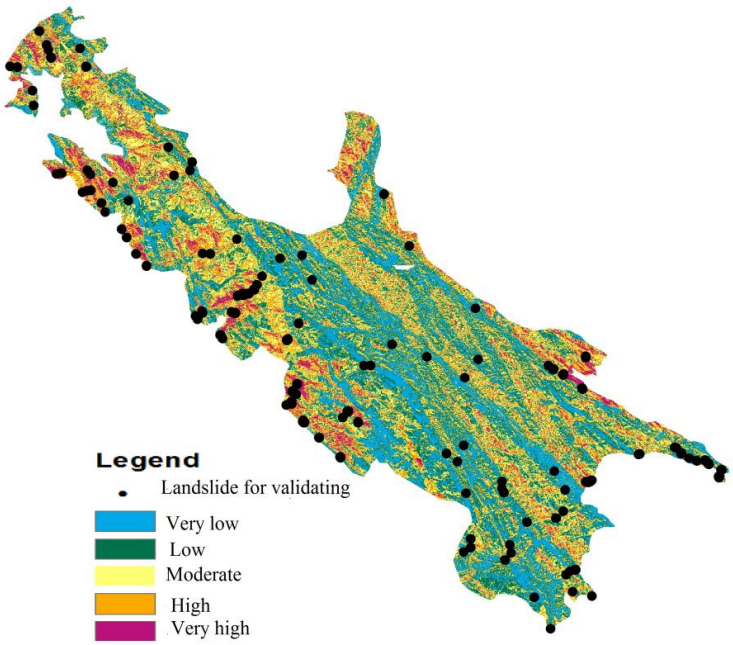

Figure 3: Landslide susceptibility mapping using Random Forest Algorithm

Validation of ensemble landslide susceptibility map, the Cohen's Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. Accuracy of study indicates that accuracy of RF is 87.7%, OOB error of 501 trees in RF is 21.1% and Kappa index is 0.63.

6. REFERENCES

Nadim, F.; Kjekstad, O.; Peduzzi, P.; Herold, C.; Jaedicke, C. (2006) Global landslide and avalanche hotspots. *Landslides* 3(2): 159-173.

Country Report Vietnam(CRV), March 2015. Natural Disaster Risk Assessment and Area Business Continuity Plan Formulation for Industrial Agglomerated Areas in the ASEAN Region. AHA Centre, Japan International Cooperation Agency

Yilmaz I. 2010. The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks. Environ Earth Sci 60:505-519.

Chen W, Pourghasemi HR, Zhao Z. 2016. A GIS-based comparative study of Dempster-Shafer, logistic regression and artificial neural network models for landslide susceptibility mapping. *Geocarto* Int. DOI: 10.1080/10106049.2016.1140824

Pradhan B, Lee S. 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environ Model Softw*. 25:747–759.

Pradhan B. 2013. A comparative study on the predictive ability of the decision tree, support vector machine andneuro-fuzzy models in landslide susceptibility mapping using GIS. Compu Geosci. 51:350–365.

Pham, B.; Tien Bui, D.; Pourghasemi, H.; Indra, P.; Dholakia, M.B. Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: A comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. Theor. *Appl. Climatol*. 2015, 128, 255–273.

XL Truong; M Mitamura; Y Kono; V Raghava; XQ Truong; TH Do; TD Bui; S Lee. Enhancing prediction performance of landslide susceptibility model using hybrid machine learning approach of bagging ensemble and logistic model tree, *Appl. Sci*. 2018, 8(7), 1046; https://doi.org/10.3390/app8071046.

Tsangaratos P, Ilia I. (2016). Comparison of a logistic regression and Naıve Bayes classifier in landslide susceptibility assessments: the influence of models complexity and training dataset size. *CATENA*. 145:164–179.

Peng L, Niu R, Huang B, Wu X, Zhao Y, Ye R. (2014). Landslide susceptibility mapping based on rough set theory and support vector machines: A case of the Three Gorges area, China. Geomorphology. 204:287–301.

Pourghasemi HR, Kerle N. (2016). Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. Environ Earth Sci. 75:1–17.

Tien Bui, D.; Anh Tuan, T.; Hoang, N.-D.; Quoc Thanh, N.; Nguyen, B.D.; Van Liem, N.; Pradhan, B. (2017). Spatial Prediction of Rainfall-induced Landslides for the Lao Cai area (Vietnam) Using a Novel hybrid Intelligent Approach of Least Squares Support Vector Machines Inference Model and Artificial Bee Colony Optimization. Landslides , 14, 447–458.

S Shrestha, TS Kang, MK Suwal. (2017). emble Model for Co-Seismic Landslide Susceptibility Using GIS and Random Forest Method. *International Journal of Geo-Information* 6 (365).

Tien Bui, D., Pradhan, B., Lofman, O., Revhaug, I., Dick, O.B., (2012). Spatial prediction of landslide hazards in Hoa Binh province (Vietnam): a comparative assessment of the efficacy of evidential belief functions and fuzzy logic models. *CATENA* 96, 25-40.