INFLUENCE OF PROPORTION BETWEEN LANDSLIDE AND NONE LANDSLIDE SAMPLE TO LANDSLIDE SUSCEPTIBILITY MODELING

Van-Trung Chu[1], Shou-Hao Chiang[1,2], and Tang-Huang Lin[1]
chuvantrung@tuaf.edu.vn, gilbert@csrsr.ncu.edu.tw, thlin@csrsr.ncu.edu.tw
1. Center for Space and Remote Sensing Research, National Central University, Taoyuan 32001, Taiwan
2. Department of Civil Engineering, National Central University, Taoyuan 32001, Taiwan

KEYWORDS: landslide susceptibility, land surface disturbance index, artificial neural network, influence, sample ratio.

ABSTRACT: The quality of the statistical-based and machine-based landslide susceptibility map depends highly on the dataset's quality for model development. When investigating the training samples in the susceptibility analysis, the unbalance area ratio between landslide and non-landslide in any given study area could be an issue in the model training procedure. Therefore, determining a suitable ratio for sampling data of landslide and none landslide can be important to optimize the modeling procedure and improve the quality of the landslide susceptibility map. So, this study introduces a practical method to reduce the uncertainty of none landslide sampling and also experiments with various ratios between landslide and none-landslide samples. The synthesis of time-series land surface disturbance index (produced by Landsat products), the bivariate statistical Frequency ratio (FR) with a budget of landslide, and the experience is considered trustworthy data for reducing the uncertainty when extracting non-landslide samples. In addition, to investigate the suitable ratio of the sample subset, the range from 1:1 to 1:10 of respective landslides and none-landslide are examined. The hybrid of Frequency ratio (FR) and artificial neural network (ANN) is applied in this study to conduct the landslide susceptibility analysis in the Thu Lum watershed in Lai Chau province, Viet Nam. Comparatively, for accuracy assessment, increasing the number of absence samples leads to the problem of specificity value (true negative rate) increase, but sensitivity (true positive rate) value change downward. Overall, the Area under ROC (receiver operating characteristic) curve decreases while we increase the portion of the non-landslide sample of the training dataset. Eventually, this research shows that the unbalance sample ratio does not produce a satisfying model. For example, the unbalance ratio can be obtained when directly using the actual landslide and non-landslide area ratio. On the other hand, a balanced ratio is recommended in this study for statistical-based and machine-based landslide susceptibility analysis because it generally produces a landslide susceptibility map with better model performance.

1. INTRODUCTION

1.1 Introduction

Landslide susceptibility modeling is important to mitigation the risk of occurrence in the future. Landslide susceptibility is one of the key information which illustrates the spatial distribution of landslide occurred potentially (Guzzetti et al., 2006). The reliability of a landslide susceptibility analysis depends on (i) the usage dataset quality, which includes independence variable (factors) and dependence variable (landslide sample), and (ii) the method of conducting the modeling. Basically, in terms of usage data, besides the accessible landslide conditioning factors, the qualitative and quantitative landslide sample subset is significantly vital conducting good landslide susceptibility modeling. However, very few studies have analyzed the influence of sample quality on modeling results. The sampling location and sampling method for training and control files have also been discussed in previous studies (Lai and Tsai, 2019; Heckmann et al., 2014). For example, how the change in the ratio of the number of landslides and non-landslide points added in the sample subset affects the results of the model. In addition to landslide samples, assessing free landslide areas is still vague but should be useful when extracting non-landslide samples (Hong et al., 2019; Zhu et al., 2018).

According to previous works, both landslide and none-landslide sample strategy selection have existing problems. In terms of landslide sampling: (i) the training sample and the test sample should not be in the same landslide block because it will lead to the problem of lack of reliability when testing (San, 2014), (ii) different perspectives of scientists about the type of sample should be single point or pixel-based has been summarized in previous research (Chu et al., 2020a), and (3) the best landslide polygon sample should be the main scarp of the landslide. For non-landslide sampling, Park and Kim (2019) employed The normalized frequency ratio to extract the non-landslide dataset for the absence of a dataset. Another attempted an empirical threshold of landslide density located in the slope unit to separate stable and unstable areas (Rossi et al., 2010). However, the real absence data may be hard to obtain directly like what we can see (Hong et al., 2019).

This study investigates the influence of integration between the much commonly recommended landslide sample point selection of the single event-based and increasing non-landslide point ratio. Long tern satellite data were

applied to investigate the free landslide regions to assist the non-landslide sampling procedure. In this study, the Frequency Ratio (FR) was adopted to quantify all applied conditioning factors in an Artificial Neural Network (ANN). The study area, Nam Ma basin – Lai Chau –Viet Nam is selected to perform the analysis, considering thirteen conditioning factors and different ratios between landslide and none landslide points of training subsets.

1.2 Study area and data usage

The selected study area is Thu Lum as a small basin of the Da river system located in Muong Te district - Lai Chau province, Vietnam. The watershed is about 179.64 km^2 in the high mountainous terrain and low-density population (Fig.1.). This area was recorded as a noticeable and serious rainfall-induced landslide occurrence in 2018. The single event-based landslide inventory was investigated based on the integration of satellite images and field surveys. Then a total of 702 single points of the landslide was selected (black dot in Figure 1). According to the accessible data, thirteen landslide conditioning factors were considered, created from various data sources (Table 1).

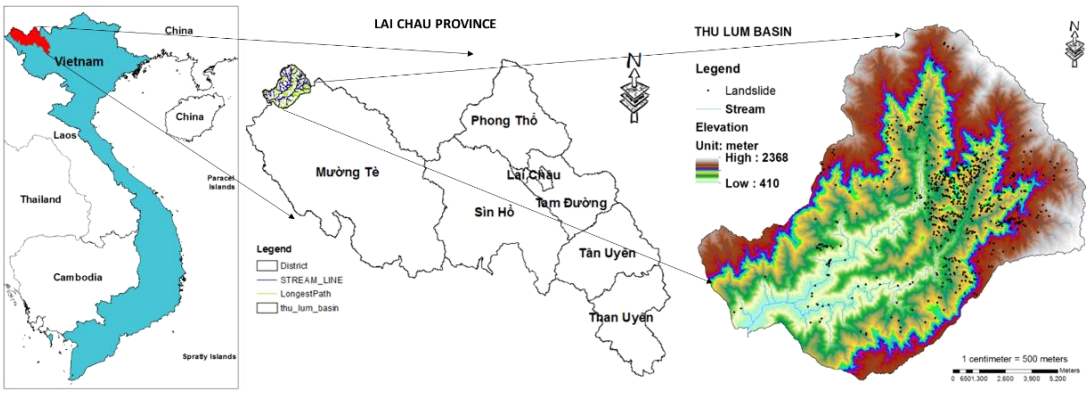

Figure 1. Location of the study area

Table 1. Data collection and sources

Data	Type	Source	Annotation	Propose dataset
Cadastral map	Vector	DNRE*	Detailed surveying scale 1:1000	road network
Topographical map	DEM (raster)	DNRE*	8 pieces DEM 10 m resolution	slope, elevation, aspect, TWI, curvature, stream network
Satellite image (Sentinel 2)	Images	USGS	Post-event (2018.03.11) Pre-event (2018.02.21)	land cover, NDVI
Geology map	Vector	Institute of Geological Sciences VAST	Original scale: 1:200.000	lithology, fault line
Soil map	Vector	DNRE*	Original scale: 1:100.000	soil types, depth of soil

* Department of Natural Resources and Environment

The detailed landslide conditioning factors are described in our previous work (Chu et al., 2020b)

2. METHODOLOGY

2.1. Landslide training subset selection

In this study, we adopted the single point at the highest position within the landslide polygon. Hence, 702 landslide polygons from landslide inventory were used to define the landslide subsets. Additionally, none landslide samples also need to be carefully investigated to avoid uncertainties. For non-landslide sampling, the long-term information associated with land surface disturbance can be assessed using the LandTrendr algorithm (Kennedy et al., 2010). It has been roofed and pointed out that time series land surface disturbance data has a positive contribution to avoid uncertainties for none-landslide sample investigation (Chu et al., 2020a). Subsequently, we randomly select the none landslide sample points with a ratio of 1:1 to 1:10 to produce ten different sample subsets. The robustness and consistency of the sample subset have been roofed in our recent previous work, so in this experiment, we just simply randomly separate the training set with 70% and remain part is testing (this ratio is much commonly used).

2.2. Model conduction

2.2.1. Frequency ratio

The relationships between spatial landslide distribution and conducted factors are assumed to be significant, and availability for predicting landslides occurred in the future (Nsengiyumva et al., 2019). The area landslide ratio occurred, and the ratio of an existing class of factors matched the related landslide ratio. The meaningful of this step is that simplifying the data into the homogenous form of arithmetic data (Eq.1: W_{ij} is the frequency ratio of class i of parameter j; FL_{ij} is the rate of landslides points in class i of parameter j; FN_{ij} is the rate of point of class i of parameter j; n is the number of parameters). The detail of FR calculation was introduced in our previous work (Chu et al., 2021)

$$W_{ij} = \sum_{j}^{n} \frac{FL_{ij}}{FN_{ij_{ij}}} \qquad (1)$$

2.2.2. Artificial neural network

Artificial Neural Network (ANN) is an advanced and computational information processing model. The multilayer perceptron is the most popular type where the number of the input, hidden, and output layers are defined. The most advantage is that it can generate a lot of input data simultaneously for complicated computing, then showing reliable results (Polykretis and Chalkias, 2018). The three-layered feed-forward network type included one input layer with thirteen factors, hidden layers, and one output layer. The number of hidden layers and the number of nodes in a hidden layer can be defined as 2**Ni*+1 (*Ni* is the number of factors) (Hecht-Nielsen, 1987). The completed network is used as a feed-forward structure to simulate the entire study area. All of the training and simulation steps were performed in MATLAB software. As the ANN model requires, all the range of input factors should be normalized to the range 0.1 to 0.9 following by equation (2) (Park et al., 2013). The quantified conditioning factors using FR.

$$\text{Normalized} = \frac{\text{Pixel(i)} - \text{Min}}{\text{Max} - \text{Min}}(0.8) + 0.1 \quad (2)$$

where Normalized is the new value, Pixel(i) is the original value of the pixel i^{th}; Min is the minimum, and Max is the maximum value of the original range.

2.3. Model performance and validation

The reliability of the landslide models could be measured using statistical criteria for evaluation as Sensitivity, Specificity, False Positive Rate, False Negative Rate, Positive Predictive power, Negative Predictive Power Overall accuracy, and Kappa index. Additionally, the AUC (area under the ROC curve) was also used for quantitative model assessment. The value close to zero means the model is non-informative, while the value close to 1 indicates a perfect model, while values in the range of (0.5–0.6), (0.6–0.7), (0.7–0.8), (0.8–0.9), and (0.9–1) can be categorized as poor, average, good, very good, and excellent, respectively (Yesilnacar and Topal, 2005).

3. RESULTS

3.1. Landslide susceptibility map

Different training subsets represent the landslide susceptibility maps in Figure 2: the number of no landslide increases from equal to ten times of landslide sample points. All the training with mean square error set smaller than 0.03. Subsequently, the accuracy assessment was calculated for detailed comparison both with training and independent testing subset. The color stands for the magnitude of susceptibility, mean close to green (value = 0) very low susceptibility, and close to the red area (value = 1) very high susceptibility.

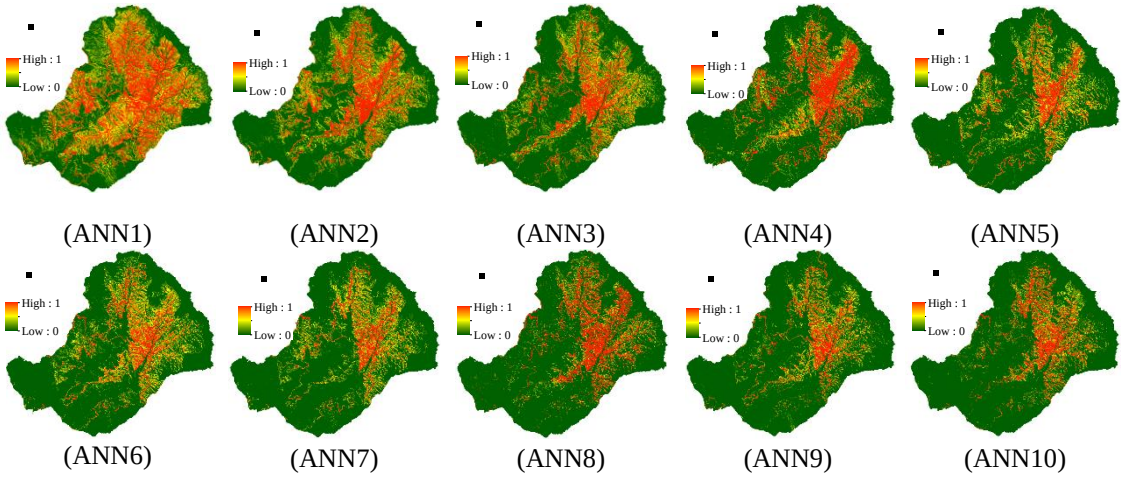

Figure 2. Landslide susceptibility maps based on different training subsets

3.2. Model accuracy comparison

Table 2. Accuracy assessment of different training cases

	Model success (training)									
	ANN1	**ANN2**	**ANN3**	**ANN4**	**ANN5**	**ANN6**	**ANN7**	**ANN8**	**ANN9**	**ANN10**
Sensitivity	0.95	0.84	0.81	0.81	0.67	0.73	0.63	0.8	0.67	0.72
Specificity	0.85	0.93	0.97	0.94	0.96	0.96	0.98	0.97	0.98	0.98
False Positive Rate	0.15	0.07	0.03	0.06	0.04	0.04	0.02	0.03	0.02	0.02
False Negative Rate	0.05	0.16	0.19	0.19	0.33	0.27	0.37	0.2	0.33	0.28
Positive Predictive power	0.86	0.93	0.96	0.93	0.95	0.95	0.97	0.96	0.98	0.97
Negative Predictive power	0.95	0.86	0.84	0.83	0.75	0.78	0.73	0.83	0.75	0.78
Overall accuracy	0.9	0.89	0.89	0.87	0.82	0.85	0.81	0.88	0.83	0.85
Kapa	0.8	0.77	0.78	0.75	0.64	0.69	0.61	0.77	0.65	0.7
AUC	0.954	0.941	0.942	0.929	0.925	0.922	0.913	0.933	0.933	0.921
S.E.	0.007	0.008	0.008	0.009	0.009	0.009	0.01	0.009	0.009	0.01
	Model prediction (testing)									
	ANN1	**ANN2**	**ANN3**	**ANN4**	**ANN5**	**ANN6**	**ANN7**	**ANN8**	**ANN9**	**ANN10**
Sensitivity	0.87	0.78	0.77	0.75	0.64	0.69	0.64	0.66	0.62	0.65
Specificity	0.91	0.94	0.97	0.93	0.96	0.96	0.981	0.94	0.972	0.991
False Positive Rate	0.09	0.06	0.03	0.07	0.04	0.04	0.019	0.06	0.028	0.009
False Negative Rate	0.13	0.22	0.23	0.25	0.36	0.31	0.36	0.34	0.38	0.35
Positive Predictive	0.87	0.93	0.96	0.92	0.94	0.95	0.97	0.92	0.96	0.99
Negative Predictive power	0.91	0.81	0.81	0.79	0.73	0.75	0.73	0.74	0.72	0.74
Overall accuracy	0.9	0.86	0.87	0.84	0.8	0.82	0.81	0.8	0.8	0.82
Kapa	0.78	0.72	0.74	0.69	0.6	0.65	0.62	0.6	0.59	0.64
AUC	0.944	0.924	0.919	0.906	0.91	0.913	0.912	0.888	0.914	0.912
S.E.	0.011	0.013	0.014	0.015	0.015	0.015	0.015	0.016	0.014	0.015

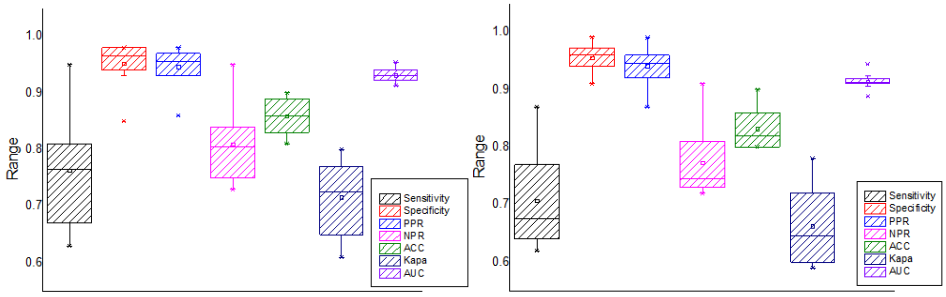

Figure 3. Box plot of accuracy assessment ten different proportions of the sample subsets (left – training result, right – testing result)

According to results of landslide susceptibility maps (figure 2) based on different sample ratios between landslide and none-landslide ANN1 to ANN10 (the number stands for the ratio that model used for conducting), the threshold 0.5 was applied for classified landslide (greater than 0.5) and none-landslide (remain area). The confusion matrix method was used to calculate accuracy results for comparison (see Table 2). The model train success is the result of using training data to estimate, and the model prediction was used as an independent testing subset for calculating. Figure 3 summarizes and shows visually the results of the variability accuracy with the box and whisker plot based on the information in Table 2. All of that information can conclude the influence of the ratio change between landslide and no landslide in the sample set.

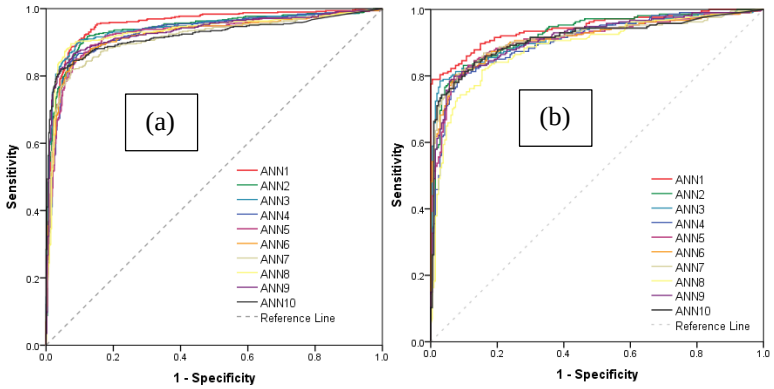

Figure 4. Receiver operating curves (ROC) of all training cases (a), and testing cases (b)

The results of successive and predictive ability also were presented using the ROC diagram (Figure 4). The different color line stands for different ROC; the curve was created by calculated sensitivity and 1-specificity value when the threshold increased. Since the area under the ROC curve (AUC) was calculated for compression.

4. DISUSSION AND CONCLUSION

4.1. Discussion

Landslides usually occur in remote and high terrain areas, the inventory activity conduct by the human resource is time-consuming and difficult. So the contribution of remote sensing technology is a tremendous contribution for produce reliable data quickly and accurately. That issue is once again useful for this work in terms of producing samples and data creation. Guzzetti et. al, mentioned that the landslide maps cover no more than one percent of the steepness area, and also diverse systematic information needs to investigate. So, the progress of data collection is kind of time-consuming and resources (Guzzetti et al., 2012). The application of remote sensing to aid the work is an absolute and tremendous contribution. Thus, in our case study, the landslide event was recorded right after the massive rain period in 2018. Fine resolution 10 meters of free assessed Sentinel-2 data pre and post-event were conducted to estimate the occurred areas (Chu et al., 2020b). Additionally, our free-landslide estimation scheme was proposed and applied in our previous works with excellent assistance for landslide susceptibility modeling (Chu et al., 2020a; Chu et al., 2020b; Chu et al., 2021). Few researchers have focused on the effect of the data quality of none-landslide samples (Hong et al., 2019; Zhu et al., 2018). Also, because of limited effort on this problem, so it still needs more experiments to make a more confident conclusion.

The main issue in the present work is proposed evidence that the problem is when the number of none landslide increases. The results displayed above obviously point out that the limited budget of landslides is the only way to

increase the number of samples. However, it entirely negatively impacts the ability of the sample in the calculation of probability value. As evidenced by the data in Table 3, the increase in the number of non-landslide samples makes the bias gradually shift towards that side. Along with that, the error of predicting the landslide decreases (Table 2). The same problem has also been pointed out in a few previous studies (Lai and Tsai, 2019; Heckmann et al., 2014).

Last but not least, each model has pros and cons, so the term novel hybrid model, model integration, or model combination are becoming more popular recently (Chu et al., 2020b; Chang and Chiang, 2009; Felicisimo et al., 2013; Wang and Hu, 2015; Yan et al., 2019). It was our concern that we also decided to combine the two models, and of course, the results have improved.

4.2. Conclusions

Our work mainly focused on figuring out the influence of changing proportion between the number of landslide points and none landslide points contained in the sample subset for modeling landslide susceptibility. The thirteen conditioning factors were collected and processed carefully, especially the novel method of investigating the sample subset with the significant contribution of remote sensing data. To overcome the limitation of each method, we introduce the idea to integrate a bivariate statistic Frequency ratio and Artificial Neural Network to conduct the landslide susceptibility model. Based on those issues, the study worked so smoothly and conducted a clear conclusion: (1) 702 landslide sample points were selected nearby the initial location of every single landslide based on the differencing method using pre and post Sentinel-2 products. (2) With The assistance of remote sensing technique and standard other exclusion criteria, our method is effective for long-term stable areas where will be placed for none-landslide sample selection. (3) According to the results shown above, it can be concluded that the balance number of landslide and none landslide samples should be conducted rather than imbalanced to avoid the bias issue.

5. REFERENCES

1. Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M. and Galli, M. 2006. Estimating the quality of landslide susceptibility models. Geomorphology, 81 (1-2), 166-184.
2. Lai, J.-S. and Tsai, F. 2019. Improving GIS-based landslide susceptibility assessments with multi-temporal remote sensing and machine learning. Sensors-Basel, 19 (17), 3717.
3. Heckmann, T., Gegg, K., Gegg, A. and Becht, M. 2014. Sample size matters: investigating the effect of sample size on a logistic regression susceptibility model for debris flows. Natural Hazards Earth System Sciences, 14 (2), 259.
4. Hong, H., Miao, Y., Liu, J. and Zhu, A.-X. 2019. Exploring the effects of the design and quantity of absence data on the performance of random forest-based landslide susceptibility mapping. Catena, 176, 45-64.
5. Zhu, A.-X., Miao, Y., Yang, L., Bai, S., Liu, J. and Hong, H. 2018. Comparison of the presence-only method and presence-absence method in landslide susceptibility mapping. Catena, 171, 222-233.
6. San, B.T. 2014. An evaluation of SVM using polygon-based random sampling in landslide susceptibility mapping: the Candir catchment area (western Antalya, Turkey). International journal of applied earth observation geoinformation, 26, 399-412.
7. Chu, V.T., Chiang, S.-H. and Lin, T.-H. Sample Position Affect Landslide Susceptibility Models in Hotspot Area of Nam Ma Basin, Lai Chau, Viet Nam. In EGU2020: Sharing Geoscience Online, p. 18278.
8. Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A.C. and Peruccacci, S. 2010. Optimal landslide susceptibility zonation based on multiple forecasts. Geomorphology, 114 (3), 129-142.
9. Chu, V.T., Chiang, S.H. and Lin, T.H. Landslide Susceptibility Modeling with Frequency Ratio, Logistic Regression, Artificial Neural Network, and the Model Combination Method In Progress of Remote Sensing Technology for Smart Future pp. 2911-2921.
10. Kennedy, R.E., Yang, Z. and Cohen, W.B. 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr—Temporal segmentation algorithms. Remote Sens Environ, 114 (12), 2897-2910.
11. Nsengiyumva, J.B., Luo, G.P., Amanambu, A.C., Mind'je, R., Habiyaremye, G., Karamage, F. et al. 2019. Comparing probabilistic and statistical methods in landslide susceptibility modeling in Rwanda/Centre-Eastern Africa. Sci Total Environ, 659, 1457-1472.
12. Chu, V.T., Chiang, S.-H. and Lin, T.-H. A Quantitative Approach for Evaluating the Contribution of Conditioning Factors to Landslide Susceptibility Moedeling. In International Symposium on Remote Sensing 2021, pp. 338-341.
13. Polykretis, C. and Chalkias, C. 2018. Comparison and evaluation of landslide susceptibility maps obtained from weight of evidence, logistic regression, and artificial neural network models. Nat Hazards, 93 (1), 249-274.
14. Hecht-Nielsen, R. Kolmogorov's mapping neural network existence theorem. IEEE Press New York, pp. 11-14.

15. Park, S., Choi, C., Kim, B. and Kim, J. 2013. Landslide susceptibility mapping using frequency ratio, analytic hierarchy process, logistic regression, and artificial neural network methods at the Inje area, Korea. Environ Earth Sci, 68 (5), 1443-1464.
16. Yesilnacar, E. and Topal, T. 2005. Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Eng Geol, 79 (3-4), 251-266.
17. Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M. and Chang, K.-T. 2012. Landslide inventory maps: New tools for an old problem. Earth-Sci Rev, 112 (1-2), 42-66.
18. Chang, K.T. and Chiang, S.H. 2009. An integrated model for predicting rainfall-induced landslides. Geomorphology, 105 (3-4), 366-373.
19. Felicisimo, A., Cuartero, A., Remondo, J. and Quiros, E. 2013. Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. Landslides, 10 (2), 175-189.
20. Wang, J. and Hu, J. 2015. A robust combination approach for short-term wind speed forecasting and analysis–Combination of the ARIMA (Autoregressive Integrated Moving Average), ELM (Extreme Learning Machine), SVM (Support Vector Machine) and LSSVM (Least Square SVM) forecasts using a GPR (Gaussian Process Regression) model. Energy, 93, 41-56.
21. Yan, F., Zhang, Q., Ye, S. and Ren, B. 2019. A novel hybrid approach for landslide susceptibility mapping integrating analytical hierarchy process and normalized frequency ratio methods with the cloud model. Geomorphology, 327, 170-187.