JOURNAL OF SCIENCE
SAKARYA UNIVERSITY

Sakarya University Journal of Science

ISSN 1301-4048 | e-ISSN 2147-835X | Period Bimonthly | Founded: 1997 | Publisher Sakarya University |
http://www.saujs.sakarya.edu.tr/

Title: Landslide Susceptibility Assessment Using Skyline Operator And Majority Voting

Authors: Alev Mutlu, Furkan Goz, Kubra Koksal, Arzu Erener

Recieved: 2018-11-07 13:04:26

Accepted: 2019-03-20 10:44:41

Article Type: Research Article
Volume: 23
Issue: 5
Month: October
Year: 2019
Pages: 782-787

How to cite
Alev Mutlu, Furkan Goz, Kubra Koksal, Arzu Erener; (2019), Landslide Susceptibility Assessment Using Skyline Operator And Majority Voting. Sakarya University Journal of Science, 23(5), 782-787, DOI: 10.16984/saufenbilder.479801
Access link
http://www.saujs.sakarya.edu.tr/issue/44066/479801

Sakarya University Journal of Science 23(5), 782-787, 2019

Landslide Susceptibility Assessment using Skyline Operator and Majority Voting

Alev Mutlu*[1], Furkan Goz[2], Kubra Koksal[3], Arzu Erener[4]

ABSTRACT

Landslide susceptibility assessment is the problem of determining the likelihood of a landslide to occur in a particular area based on the geological and morphological properties of the area. In this study, we propose a method wherein skyline operator is used to model landslides and majority voting is used to assess landslide susceptibility. Experiments conducted on a real life data set showed that the proposed method achieves 83.07% classification accuracy and is superior over most commonly used techniques for landslide susceptibility assessment such as logistic regression, support vector machines and artificial neural network.

Keywords: Landslide Susceptibility Assessment, Majority Voting, Skyline Operator

1. INTRODUCTION

Modeling and assessing natural hazards is a challenging and active research problem. Studies regarding natural hazards such as earthquakes [1,2], floods [3,4], and volcanic activities [5,6] have been proposed. In this study, we focus on landslides and propose a landslide modeling and landslide susceptibility assessment method based on the skyline operator and majority voting principle.

Landslide is a natural phenomenon defined as the outward and downward movement of soil making materials. According to UN Office for Disaster Risk Reduction, landslides are among the top five most frequent natural hazards and caused some 130000 deaths and US$ 50 billion economic loss [7]. Landslide susceptibility assessment is the problem of determining the likelihood of a landslide to occur in a particular area based on morphological and geological properties of the area [8]. In literature, there exist numerous studies regarding landslide susceptibility assessment. However, these studies mainly differ by means of study area rather than the underlying landslide modeling and susceptibility assessment techniques. Bivariate and multivariate statistical methods, decision trees, support vector machines, and neural networks are among the most frequently applied techniques in landslide susceptibility assessment.

In this study, we propose a hybrid method based on the skyline operator and majority voting principle for landslide susceptibility assessment. Given properties of landslide occurring zones, the skyline operator is utilized to retrieve skyline points that define the minimum values of landslide triggering factors. To assess landslide susceptibility of a zone, a two-step majority voting is implemented. In the first step, properties of a test zone are compared against each skyline point. A skyline point votes *yes* for the test instance if majority of the features of the test instance have greater values than those of the skyline point, otherwise the skyline point votes *no*. In the second step, votes for *yes* and *no* are counted and the final decision is made according to the majority of the votes. The proposed method primarily distinguishes from state-of-the-art methods by the following aspects:

[1]* Corresponding Author
[1]Kocaeli University, Faculty of Engineering, Department of Computer Engineering, alev.mutlu@kocaeli.edu.tr
[2]Kocaeli University, Faculty of Engineering, Department of Computer Engineering, furkan.goz@kocaeli.edu.tr
[3]Kocaeli University, Faculty of Engineering, Department of Computer Engineering, 150201172@kocaeli.edu.tr
[4]Kocaeli University, Faculty of Engineering, Department of Geomatic Engineering, arzu.erener@kocaeli.edu.tr

– It requires only positive data (properties of landslide occurring areas) to model landslides,
– As to our knowledge there is no other study that utilizes skyline operator in landslide susceptibility assessment,
– The method can be categorized as a hybrid rather than an ensemble.

Performance of the proposed method is evaluated on a real life dataset regarding Savsat region of Turkey. Savsat is a landslide intensive area located in the eastern part of Turkey and has been subject to several studies [9–12]. Experimental results based on 10-fold cross validation showed that the proposed method achieves 83.07% accuracy. When compared to the applications of support vector machines, neural networks, and logistic regression on the same dataset, the proposed method achieves higher accuracy and almost the same accuracy results when compared to decision trees.

The rest of the paper is organized as follows. In Section 2, we briefly introduce methods used for landslide susceptibility assessment and introduce the techniques used in this study. In Section 3 we introduce the proposed method. In Section 4 we first describe the study area and later discuss the experimental findings. Section 5 concludes the paper.

2. BACKGROUND

In this section, we first summarize some of the most commonly used techniques for landslide susceptibility assessment and later introduce the methods used in this study.

Bivariate statistics involve determining the relationship between two variables, generally called dependent and independent. Multivariate statistics, on the other hand, aim to figure out the relationship between one dependent variable and multiple independent variables [13]. In the case of landslide susceptibility assessment, the dependent variable indicates whether a landslide is present or absent, and independent variables are values of the landslide triggering factors. Weights of evidence [14, 15] and frequency ratio [14, 16] are two of the most commonly employed bivariate statistical methods and logistic regression [10, 17] is the most commonly used multivariate statistical method in landslide susceptibility assessment.

Support vector machines are supervised learning algorithms that map labeled instances in a space and construct hyperplanes such that they are as far as from the nearest training data of any class. To predict the class for a test instance, it is mapped into the same space and class label of the hyperplane it falls into is assigned. In case of landslide susceptibility assessment, landsliding and non-landsliding zones are plotted on an *n*-dimensional space, where *n* is the number of landslide triggering factors, and support vector machine models are built [18–20].

In literature there also exist decision tree-based methods for landslide susceptibility assessment. Such models represent the learning model as a tree like structure where each internal node is condition and leaf nodes are class labels and conjunction of test conditions from root to a leaf form a classification rule. In case of landslide susceptibility assessment, conditions are tests on values of landslide triggering factors [18, 21].

Neural networks are computational models that mimic the human brain. These models are formed of input, output, and hidden layers. Neurons at each layer are connected to those at the next layer. In case of landslide susceptibility assessment problem, neurons of the input layer indicate values of landslide triggering factors and neurons of the output layer indicate the presence or absence of a landslide [16].

In literature there are hybrid studies proposed for landslide susceptibility assessment that utilize Naïve Bayes Trees [22], Random Forest [23], neuro-fuzzy inference [24].

In this study skyline operator is utilized to model landslides. Skyline operator is concerned with retrieving objects, called *skyline points*, from a set of objects such that the retrieved objects are not dominated by any other object in the set. The skyline operator is implemented in domains such as recommendation [25, 26], scientometrics [27, 28].

Object p is said to dominate object q, donated as $p \prec q$, if p is as good as q in all dimensions and better in at least one dimension. Domination operator is formulated in (1) for objects with d dimensions where $\geq$ operator means as good as or better and operator $>$ means better. Goodness of an object is determined based on some utility function that is monotone on all attributes of the objects.

$$p \prec q \leftrightarrow \forall i\{1,2,\dots,d\} p_i \geq q_i \wedge \exists j \in \{1,2,\dots,d\} p_i > q_j \quad (1)$$

Skyline points of a set consist of objects that are not dominated by any other object in the set. The skyline operator is formulated in (2) and its SQL extension is provided in (3).

$$S_p = \{p \mid p \in P \wedge \nexists\, q \in P : q \prec p\} \quad (2)$$

$$SELECT \dots FROM \dots WHERE \dots GROUP\ BY \dots HAVING \dots$$
(3)
$$SKYLINE\ OF\ d_1[MIN|MAX], \dots, d_n[MIN|MAX]$$

To assess landslide susceptibility in the proposed method, majority voting principle is implemented. Trained on the same data set, different classifiers may assign different class labels for a particular test instance. In majority voting, votes of individual classifiers for a test instance is counted and prediction with the highest vote is assigned to the test instance. Majority voting principle can be formulated as in (4) where $h_1, h_2, \dots, h_N$ are individual classifiers, $w_1, w_2, \dots, w_N$ are weights that sum to 1 and I(·) is indicator function.

$$c(x) = argmax_i + \sum_{j=1}^{N} w_j(h_j\,(x) = 1) \tag{4}$$

3. THE PROPOSED METHOD

The proposed method consists of three main steps: data preprocessing, model building, and landslide susceptibility assessment. In the following subsections we present these steps.

3.1 Data Preprocessing

In literature there is no consensus on the exact set of landslide triggering factors. In this study, similar to [10], aspect, slope, soil-map, altitude, erosion, land-use, distance to fault, depth of soil, distance to drainage, distance to road, and lithology are considered as landslide triggering factors. Data related to landslide triggering factors are usually obtained from different organizations in raw format; hence the data needs to be preprocessed. The data-preprocessing step in this study involves discretization, weight assignment, and data cleaning.

3.1.1 Discretization

In this step features that come from continuous domain are discretized using equal width binning. In equal width binning method, data to be discretized is firstly sorted and then partitioned into intervals of almost equal width. The equal width binning method is formulated in (5) where *w* is width of an interval; *k* is the number intervals determined a priori, *min* and *max* are, respectively, the smallest and largest value in the data set. Intervals boundaries are defined as [-∞, *min* + *w*], (*min* + *w*, *min* + *2w*], ... (*min* + (*k-1)*w, ∞].

$$\boldsymbol{k} = \frac{\mathbf{max} - \boldsymbol{min}}{\boldsymbol{w}}$$
(5)

In the proposed method aspect, slope, altitude, erosion, distance to fault, depth of soil, distance to drainage, distance to road are the features that have continuous values.

3.1.2 Weight Assignment

After discretization, each discrete value is assigned with a weight that indicates its effect on landslide occurrence. In this study, weights are assigned using Frequency ratio (FR) model. FR model is formulated in (6) where the numerator indicates the fraction of the zones with landslides and discrete value i (N_{Li}) over the total number of landsliding zones (N_i). The denominator indicates the fraction of the number of landsliding zones with any value of the feature that *i* belongs to over total number of zones of the study area. FR = 1 is assumed to be average, FR values less than 1 indicate low correlation between landslide occurrence and the value under consideration, and vice versa. The weight of *i* is the natural logarithm of its frequency ratio (7).

$$\boldsymbol{FR_i} = \frac{N_{Li}/N_{Ni}}{\sum N_{Li}/\sum N_{Ni}} \tag{6}$$

$$\boldsymbol{w_i} = \ln(\boldsymbol{FR_i})$$
(7)

Weight of a feature value indicates its effect on landslide occurrence. The higher weight indicates more effect.

3.1.3 Data Cleaning

After the discretization and class weight calculation steps, feature vectors with exactly the same values in every feature position but class label may be generated. We consider such representations as inconsistent and remove them. Discretization and weight assignment may also generate repetitive feature vectors, such repeating vectors are also removed from the training data in data cleaning step.

3.2 Model Building

In this study, skyline operator is used to model landslides. For this purpose, representations of only landslide occurring zones are considered and skyline operator is used to find dominating feature vectors with the lowest features values. Hence (3) is implemented with *MIN* keyword. In this study we assume that each skyline point is a model that defines the minimum values a feature vector should have in order to represent a landslide-zone.

In Table 1 we provide 10 vectors each indicating properties of a landslide zone. The first 11 attributes represent different characteristics of a zone and the last attribute is the class label.

Table 1. Example training dataset

L1	[0, 2, 1, 5, 1, 2, 3, 2, 7, 0, 3, 1]
L2	[0, 3, 3, 4, 3, 1, 4, 5, 8, 0, 6, 1]
L3	[0, 3, 2, 6, 2, 4, 4, 3, 7, 5, 9, 1]
L4	[0, 3, 3, 3, 1, 3, 2, 3, 8, 4, 4, 1]
L5	[0, 3, 3, 6, 3, 2, 5, 4, 9, 1, 4, 1]
L6	[0, 4, 6, 6, 3, 2, 4, 7, 8, 0, 3, 1]
L7	[0, 6, 8, 9, 6, 5, 7, 4, 8, 4, 7, 1]
L8	[0, 3, 3, 3, 2, 1, 2, 4, 4, 0, 2, 1]
L9	[0, 5, 4, 7, 3, 2, 6, 3, 9, 4, 4, 1]
L10	[0, 7, 6, 4, 2, 4, 7, 4, 6, 2, 4, 1]

The skyline operator will return L1, L4, and L8 as skyline points.

3.3 Landslide Susceptibility Assessment

To assess landslide susceptibility of a test zone a two-step majority voting mechanism is implemented. In the first step, feature vector representing the test zone is compared against each skyline point. A skyline point votes *yes* if majority of feature values representing the test zone are greater than the values of the skyline otherwise it votes *no*. In the second step, votes for *yes* and *no* are counted, and the final decision is made according to the majority of the votes.

In Table 2, we list the properties of the skyline points discovered in the previous section and two test instances.

Table 2. Example test dataset

L1	[0, 2, 1, 5, 1, 2, 3, 2, 7, 0, 3, 1]
L4	[0, 3, 3, 3, 1, 3, 2, 3, 8, 4, 4, 1]
L8	[0, 3, 3, 3, 2, 1, 2, 4, 4, 0, 2, 1]
T1	[0, 1, 2, 4, 0, 2, 1, 1, 4, 0, 2, 0]

For T1, L1 will vote *no* (7 features have smaller values), L4 will vote *no* (10 features have smaller values), and L8 will vote *yes* (5 features have smaller values). The final decision will be *no*.

4. EXPERIMENTS

To evaluate the performance of the proposed method, a real life dataset describing highly landslide intensive area namely Savsat is used. Savsat is town in north east of Turkey located between 42°24'N and 42°50'N latitudes and 41°12'E and 41°37'E longitudes. The area has been subject to several landslide studies including [9, 10, 21]. Figure 1 shows the study area. Landslides indicated with yellow are dormant and landslide indicated with blue are active.

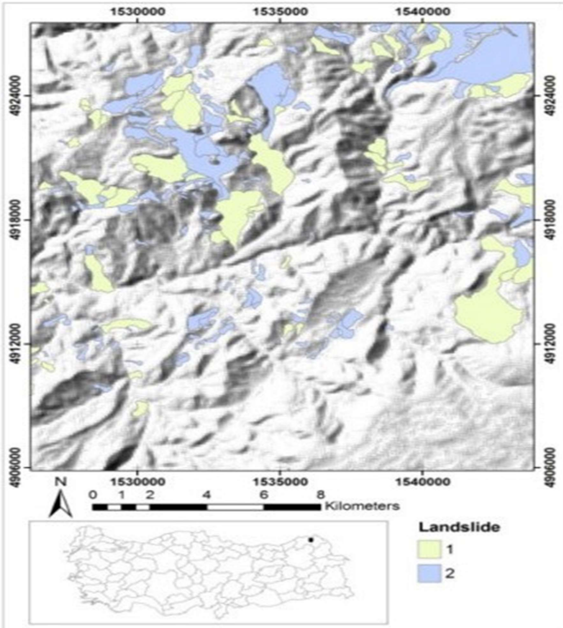

Figure 1. Map of the study area, yellow areas represent dormant landslides, blue areas represent active landslides

The dataset describing the area are obtained from the General Directorate of Mineral Research and Exploration (MRE) of Turkey. The dataset describing the area consists of properties of 13645 zones. Each zone is described using eleven features and a class label. Eight of the features are from continuous domain and the remaining three are from categorical domain.

Table 3 lists the attributes discretized, the number of bins, and bin width. Soil depth and erosion are the remaining two features that are discretized. Soil depth is discretized into four bins as follows: very deep (depth > 90 cm), deep (90 >= depth > 50 cm), shallow (50 >=depth > 20 cm) and shallow (20 >= depth >= 0 cm). Erosion is also discretized into 4 bins, namely none or low, medium, high, and very high. The bin widths are consistent with those recommended in literature [10, 14, 30].

After the data cleaning step, the size of the data set is reduced to 9762 instances, 6329 of which describe landslide safe zones and the remaining 3433 zones describe landslide-occurring zones.

Table 3. Properties of discretized attributes

Feature name	Value range	Num. of bins	Bin width
Aspect	[-1, 360]	8	45
Slope	[0, 52]	9	5
Altitude	[697, 3000]	5	500
Distance to fault	[0, ∞]	9	500
Distance to drainage	[0, ∞]	8	100
Distance to road	[0, ∞]	8	100

In Table 4 we report the predictive accuracy achieved by the proposed method, logistic regression, support vector machines, neural network, and decision trees. The reference models are built using Weka tool [31]. The reported results are obtained via 10-fold cross validation. On the average, the skyline operator retrieved 160 skyline points at each fold.

Table 4. Accuracy results

Method Name	Accuracy
Logistic Regression	77.13%
Support Vector Machine	76.35%
Neural Network	80.48%
Decision Tree	84.22%
The Proposed Method	83.07%

As the results show, the proposed method is superior over the reference studies other than decision support tree based model.

5. CONCLUSION

In this study we introduced a method based on the skyline operator and majority voting principle for landslide susceptibility assessment. In the proposed method the skyline operator is utilized to discover skylines that describe land- slide models, and majority voting is used to assess landslide susceptibility of a test instance. Experiments on a real life data set describing properties of a highly landslide intensive area namely Savsat show that the proposed method is superior over most commonly used methods in landslide susceptibility assessment such as logistic regression, support vector machines and neural networks. The proposed method performs slightly worse than decision tree based model, 83.07% vs. 84.22%.

Landslides usually occur in large areas. Hence a factor that has a high effect in a particular part of a landslide may have low effect in another part of the landslide. As a feature work we plan to extend the proposed method to handle such situations.

REFERENCES

[1] A. Morales-Esteban, F. Martínez-Álvarez, A. Troncoso, J. Justo, C. Rubio-Escudero, “Pattern recognition to forecast seismic time series”, *Expert Systems with Applications*, vol. 37, no. 12, pp. 8333-8342, 2010.

[2] K. Asim, F. Martínez-Álvarez, A. Basit, T. Iqbal, “Earthquake magnitude prediction in Hindukush region using machine learning techniques”, *Natural Hazards*, vol. 85, no. 1, pp. 471-486, 2016.

[3] H. Cloke, F. Pappenberger, “Ensemble flood forecasting: A review”, *Journal of Hydrology*, vol. 375, no. 3-4, pp. 613-626, 2009.

[4] B. Bhattacharya, D. Solomatine, "Neural networks and M5 model trees in modelling water level–discharge relationship", *Neurocomputing*, vol. 63, pp. 381-396, 2005.

[5] H. Langer, S. Falsaperla, A. Messina and S. Spampinato, "Perfomance of a new multistation alarm system for volcanic activity based on neural network techniques", *in Second European Conference on Earthquake Engineering and Seismology*, 2014.

[6] J. Parra, O. Fuentes, E. Anthony, V. Kreinovich, “Use of Machine Learning to Analyze and – Hopefully – Predict Volcano Activity”, *Acta Polytechnica Hungarica*, vol. 14, no. 3, 2017.

[7] UNISDR: Landslide Hazard and Risk Assessment", Unisdr.org. [Online]. Available: https://www.unisdr.org/files/52828_03landslidehazardandriskassessment.pdf. [Accessed: 31- Aug- 2018].

[8] F. Dai, C. Lee, Y. Ngai, “Landslide risk assessment and management: an overview”, *Engineering Geology*, vol. 64, no. 1, pp. 65-87, 2002.

[9] E. Topsakal, T. Topal, “Slope stability assessment of a re-activated landslide on the Artvin-Savsat junction of a provincial road in Meydancik, Turkey”, *Arabian Journal of Geosciences*, vol. 8, no. 3, pp. 1769-1786, 2014.

[10] A. Erener, A. Mutlu, H. Sebnem Düzgün, “A comparative study for landslide susceptibility mapping using GIS-based multi-criteria decision analysis (MCDA), logistic regression (LR) and association rule mining (ARM)”, *Engineering Geology*, vol. 203, pp. 45-55, 2016.

[11] C. Ozgen, "An Invectigation of landslide at km:12+ 200 od Artvin-Savsat junction-Meydancik Provincial road", PhD thesis, Middle East Technical University (2012).

[12] P. Temel, "Evaluation of potential run-of river hydropower plant using multicriteria decision making in terms of environmental and social aspect", PhD thesis, Middle East Technical University (2015).

[13] J. Isotalo, "Basics of Statistics", [Online] Available: https://www.mv.helsinki.fi/home/jmisotal/BoS.pdf [Accessed: 2019-01-17]

[14] Q. Ding, W. Chen, H. Hong, "Application of frequency ratio, weights of evidence and evidential belief function models in landslide susceptibility mapping", *Geocarto International*, pp. 1-21, 2016.

[15] C. Xu, X. Xu, F. Dai, J. Xiao, X. Tan, R. Yuan, "Landslide hazard mapping using GIS and weight of evidence model in Qingshui River watershed of 2008 Wenchuan earthquake struck region", *Journal of Earth Science*, vol. 23, no. 1, pp. 97-120, 2012.

[16] B. Pradhan, S. Lee, "Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modeling", *Environmental Modelling & Software*, vol. 25, no. 6, pp. 747-759, 2010.

[17] L. Wang, M. Guo, K. Sawada, J. Lin, J. Zhang, "Landslide susceptibility mapping in Mizunami City, Japan: A comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models", *CATENA*, vol. 135, pp. 271-282, 2015.

[18] B. Pradhan, "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS", *Computers & Geosciences*, vol. 51, pp. 350-365, 2013.

[19] B. Feizizadeh, M. Roodposhti, T. Blaschke , J. Aryal, "Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping", *Arabian Journal of Geosciences*, vol. 10, no. 5, 2017.

[20] D. Kumar, M. Thakur, C. Dubey and D. Shukla, "Landslide susceptibility mapping & prediction using Support Vector Machine for Mandakini River Basin, Garhwal Himalaya, India", *Geomorphology*, vol. 295, pp. 115-125, 2017.

[21] H. Saito, D. Nakayama, H. Matsuyama, "Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan", *Geomorphology*, vol. 109, no. 3-4, pp. 108-121, 2009.

[22] A. Shirzadi, D. T. Bui, B. T. Pham, K. Solaimani, K. Chapi, A. Kavian, H. Shahabi, I. Revhaug, "Shallow landslide susceptibility assessment using a novel hybrid intelligence approach", *Environmental Earth Sciences*, vol. 76, no. 2, 2017.

[23] H. R. Pourghasemi, K. Norman, "Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran", *Environmental Earth Sciences*, vol. 75, no. 3, 2016.

[24] I. N. Aghdam, P. Biswajeet, M. Panahi, "Landslide susceptibility assessment using a novel hybrid model of statistical bivariate methods (FR and WOE) and adaptive neuro-fuzzy inference system (ANFIS) at southern Zagros Mountains in Iran", *Environmental Earth Sciences*, vol. 76 no.6, 2017.

[25] K. Kodama, Y. Iijima, X. Guo and Y. Ishikawa, "Skyline queries based on user locations and preferences for making location-based recommendations", *International Workshop on Location Based Social Networks*, 2009

[26] J. Yang et al, "Finding superior skyline points for multidimensional recommendation applications", *World Wide Web*, vol. 15, no. 1, pp. 33-60, 2012

[27] G. Stoupas et al, "Rainbow ranking: an adaptable, multidimensional ranking method for publication sets", Scientometrics, vol. 118, no. 1, pp. 147-160, 2018

[28] A. Sidiropoulos, "Gazing at the skyline for star scientists", *Journal of Informatics*, vol. 10, no. 3, 2016.

[30] E. Topsakal, "An Invectigation of landslide at km:12+ 200 od Artvin-Savsat junction-Meydancik Provincial road", PhD thesis, Middle East Technical University, (2012).

[31] Weka 3: Data Mining Software in Java, [Online] Available: https://www.cs.waikato.ac.nz/ml/weka/ [Accessed: 2018-04-10].