Landslide susceptibility analysis using decision tree method, Phidim, Eastern Nepal

Chandra Prakash Poudyal

Central Department of Geology, Tribhuvan University, Kirtipur, Kathmandu, Nepal

ABSTRACT

The decision tree is one of the new methods used for the determination of landslide susceptibility in the study area. The Phidim area is selected for the application of this method. The total surface area is 168.07 sq. km, and is located at the eastern part of Nepal. There are total of 10 different data bases used for this study which are; geological formation, elevation, slope, curvature, aspect, stream power index, topographic wetness index, distance from drainage, lineaments, and slope length, and are considered as landslide conditioning factors. Geographical information system (GIS) is used as basic tools and ARC/View is used for the processing data analysis and final map preparation. For the decision tree analysis the PASW 18 (statistical tool) is used to generate values of each factor. According to the results of decision tree, two geological formations; stream power index and slope are found as the most effective parameters on the landslide occurrence in the study area. Using the predicted values, the landslide susceptibility map of the study area is produced. To assess the performance of the produced susceptibility map, the area under curve (AUC) is drawn. The AUC value of the produced landslide susceptibility map has been obtained as 95.9%. According to the results of the AUC evaluation, the produced map has showed a good performance. As to wrap up, the produced map is able to be used for medium scaled and regional planning purposes.

INTRODUCTION

The landslide susceptibility and hazard assessments can be carried out using either direct or indirect mapping technique. In the former the degree of hazard is determined by mapping geomorphology based on experience of mapper, and imaging the characteristics of the terrain conditions (van Westen et al., 2000). The indirect technique is made either by statistical models or deterministic models, and is used to predict the landslide prone areas, based on information obtained from the interrelation between the landscape factors and the landslide distribution (van Westen et al., 2000).

Landslides are the most common component of the natural hazards in Nepalese mountains (Upreti and Dhital, 1996). Landslides are one of the most visible

*Corresponding author:
E-mail address: cppoudyal@gmail.com

and commonly perceived destructive phenomena in hilly mountainous terrain. Weak geological structures, steep and rugged land surfaces and extreme climatic conditions result in a high degree of fragility. Dynamics of hydrological processes and altitude-specific patterns of vegetation and land-use structures due to sharp changes in altitutde lead to substantial differences in environmental characteristics in mountain system.

The study area is mountainous, having highly folded, low grade metamorphic rocks, and lying at the southern part of Taplejung Window, a tectonic structure. The rocks of Lesser Himalaya are exposed within the window. Recently, numerous landslides have occurred in the study area. The field investigation, aerial photo interpretation and geological data analyses suggest that various factors have acted together on slope instability.

It is possible to produce a landslide susceptibility map utilizing various indirect mapping techniques such as combination of index maps (i.e. Pachauri and Pant, 1992; Donati and Turrini, 2002), bivariate and multivariate statistical analyses (i.e. Guzzetti et al., 1999; Lee and Min, 2001; Clerici et al., 2002; Lee, 2005; Can et al., 2005; Gokceoglu et al., 2005; Duman et al., 2006; Guzzetti et al., 2006; Gorum et al., 2008; Turer et al., 2008), neural networks (i.e. Lee et al., 2001, 2007; Gomez and Kavzoglu, 2005; Kanungo et al., 2006; Poudyal et. al, 2010) and fuzzy approach (i.e. Juang et al., 1992; Binaghi et al., 1998). Each landslide susceptibility assessment method has some advantages and drawbacks as well. GIS is used as basic analysis tool for the manipulation and management of spatial data. There are several different types of technologies that are commonly utilized in landslide analysis practice. They are probabilistic model, logistic regression model, fuzzy logic, artificial neural network models.

The decision tree analysis were applied by various researchers (Pal and Mather, 2003; Guzzetti et al., 2005; Düzgün and Özdemir, 2006; Saito et al., 2009; Wang and Niu, 2009; Wan et al., 2010) for landslide analysis, but most used other methods such as fuzzy logic, frequency ratio, logistic regression, etc. The decision tree is another method in the field of statistical analysis and is used in many field of sciences and also used for the landslide susceptibility analysis. A decision tree makes no statistical assumptions and can handle data that are represented on different measurement scales, unlike other statistical methods. The decision tree normally requires discrete class labels and they have limitations with respect to the landslide susceptibility representation with continuous value. To overcome the limitations, the discrete classes values have been used for identifying landslide susceptibility or only case analysis have been carried out. Traditional spatial techniques cannot easily discover new and unexpected patterns, trends and relationships that can be hidden deep within very large diverse geographic datasets (Miller and Han, 2001). At this point, data mining can be used to overcome the problem when processing very large datasets. When producing landslide susceptibility maps, very large datasets should be processed.

The decision tree investigation for landslide susceptibility assessment constitutes the main purpose of this study. This study also includes assessment of the landslide conditioning factors, some theoretical information on the decision tree technique and application of the decision trees to landslide susceptibility assessment.

OUTLINES OF THE STUDY AREA

The study area lies in the Panchthar District of Eastern Nepal and is bounded by the latitudes 27°07'30" and 27°15'00" N and the longitudes 87°45'00" and 87°52'30"E (Fig. 1). The study area is about 168.07 sq. km (12.25 km x 13.72 km). The topography of the area is highly dissected and rugged, the NE part has high relief with dense forests. Most of the ridges run approximately E to W, and rivers flow from E to W. The altitude values in the application site are between 440 and 2500 m while the dominant altitude ranges are 1500–2200 m. The application site has a dendritic drainage pattern, because of presence of soft lithological units and high slope angles.

The general physiographic trend of the area is more or less E-W. The range of the slope gradient values is 0 to 87 degrees. The area is more restrained in the center with V-shaped valleys and gently sloped ridges. The N-facing slopes are steeper than the S-facing slopes with more cultivated area. The slopes are modified by terracing to make small plots for cultivation. The gentle slopes support colluvium and variable extents of residual soil. The colluvium is scattered in its distribution among the foothills and occurs as ribbon deposits infilling drainage courses. Small areas of alluvial deposits are found along the river valleys.

Fig. 1 Map of Nepal showing study area.

The bedrock geology of the study area consists predominantly of highly weathered low to medium-grade metamorphic rocks. The rocks are meta-sandstone, phyllite, quartzite, and schist. The succession of the area (Poudyal et al. 2010) comprises Bharata Phyllite; Phyme Khola member; Phidim Member; Jorpati Augen gneiss; in the Lesser Himalayan sequence. The Higher Himalayan sequence consists of banded gneiss, from lower to upper sequences. The Bharapa Phyllite consists of fine-grained, crenulated phyllite intercalations with medium-grained, medium-bedded quartzite. The Phyme Khola member consists of white, coarse-grained quartzite with a well-developed slaty cleavage. The Phidim member consists of fine-grained, crenulated phyllite with medium to thick bedded grey quartzite. The Jorpati Augen gneiss consists of well developed augen structures at lower part and upper part contains both hornblende and garnet-ferrous schist. The strength of the banded gneiss is higher than other rocks found in the study area. The geological map of the area is shown in Fig. 2.

The study area contains a large scale folded structure within the window and have high prone to landslide in comparison with the Higher Himalayan rocks. There are 111 large landslides observed in this area. The characteristics of landslide movements vary as rockslides along a plane parallel to foliation or bedding or rapid earth flows. The depth of these landslides varies from shallow to deep in south facing slopes, whereas north facing slopes have rock sliding, toppling or wedge failure. The landslide occurs during the monsoon season and the events are generally destructive to property, roads, buildings, and agricultural land. Most slides originated from gulley erosion and failure occurs due to the removal of lateral support and load exerts on the slope. The slides also occurred due to road construction where the natural hill slope is modified by the removal of lateral support to make a roadway.

ASSESSMENT OF LANDSLIDE CONTRIBUTING FACTORS

In this section, the contribution factors for landslide in the study area are explained. In the present study, the digital elevation model (DEM) was prepared from the data that collected from the department of survey, Nepal. From the DEM various factors for the landslide
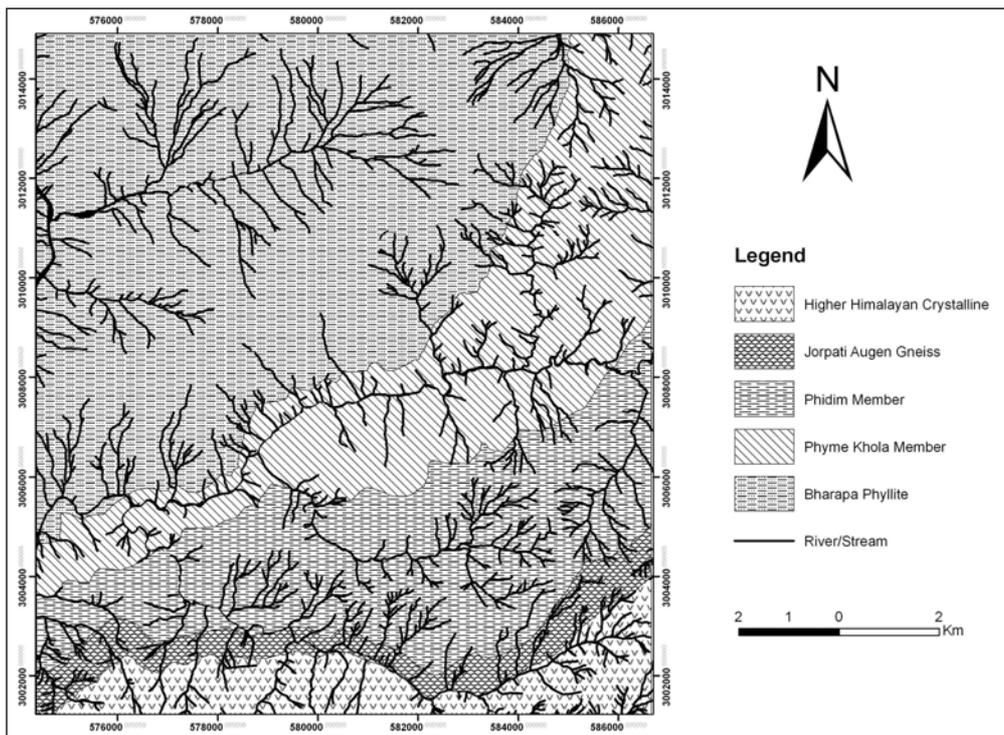
Fig. 2 Geology and drainage network of the Phidim area.

occurrence contribution are extracted such as slope, elevation, curvature, aspect, land use, stream power index (SPI), topographic wetness index (TWI), length slope (LS). Also the distance from lineaments, distance from drainage is used for this analysis. All map produced are in the raster format with the pixel size 10 x 10 sq. m. A landslide inventory map was prepared using the remote sensing and verified with the field visit before the analysis done.

One of the most important stages of landslide susceptibility mapping is to describe the factors governing the landslides identified in the area. A landslide susceptibility mapping procedure for the application site has been performed previously by Poudyal et al. (2010) considering the logistic regression and artificial neural networks technique. In this study, only the decision tree analysis is applied using 10 different factors that contribute for the occurrence of landslides. The characters of landslides identified in the region are mainly shallow and active.

DECISION TREE METHOD

A decision tree is a technique for finding and describing structural patterns in data as tree structures; a decision tree does not require advance knowledge of the relationship between all of the input variables and an objective variable. A decision tree can also handle data measured on different scales, without any assumptions concerning the frequency distributions of the data, as it is non-linear (Pal and Mather, 2003).

Among many classification algorithms, the decision tree is an efficient algorithm for classifying large datasets. A decision tree is a top-down classification strategy based on automatically selected rules that partition a set of given entities into smaller classes (Breiman et al. 1984, Quinlan 1986). Moreover, decision-trees can identify a target (dependent) variable. This is not possible with unsupervised machine learning methods like cluster analysis, factor analysis (principal component analysis) and statistical measures that treat all variables equally without predicting the value of a variable (Saadia et al., 2006).

The decision-tree model has a number of advantages compared to numerically-oriented techniques such as linear and nonlinear regression (function fitting), logistic regression, artificial neural networks ANNs, and genetic algorithms. Decision-trees are easy to build and interpret, and they can automatically handle interactions between both continuous (ordinal, interval) and categorical (nominal) variables. Decision-trees can identify the most decisive variables, which are those that are used to create the splits near the top of the tree. In addition, they do not require specification of the form of a function to be fitted to the data, as is necessary for other competing procedures.

Each class, denoted as a leaf of the decision tree, consists of a subset of all records belonging to one or several categories according to the values of a specific variable, termed the target variable. Each rule is hierarchically represented by a path from the root node to a leaf via intermediate nodes and branches. The nodes of the path represent the supporting variables" that maximize the distinction among the classes and minimize the diversity within each class. Each branch represents the values of the supporting variables used as criteria to classify the dataset. The hierarchical process is equivalent to a minimization of the entropy produced by the classification at each successively lower level of the hierarchy.

Several decision tree learning algorithms have been proposed, such as Classification and Regression Trees (CART) (Breiman et al., 1984), Iterative Dichotomizer version 3 (ID3) (Quinlan, 1986), and C4.5 (an industrial version of ID3) (Quinlan, 1993). These algorithms share a common hypothesis, but differ in the way that they quantify the distinction and diversity criteria, that is, entities of the input dataset might be independent. In this study, we used a PASW 18 statistic data editor for landslide susceptibility mapping. PASW uses the decision tree analysis model QUEST (Quick, Unbiased, Efficient, and Statistical Tree) algorithm. It is a tree-structured classification algorithm that yields a binary decision tree like C & RT.

Decision trees are often compared with Artificial Neural Networks (ANNs) because both methods can model data that have nonlinear relationships between variables, and both can handle interactions between variables. However, neural networks have a number of drawbacks. They do not present an easily understandable model that allows researchers to get a full explanation for the underlying nature of the data being analyzed. In

Fig. 3 Decision tree model of analysis.

addition, neural networks handle only binary categorical input data, not those with multiple classes. It is also difficult to incorporate a neural network model into a computer system; this requires a good interpreter. In contrast, once a decision tree model has been built, it can be converted to statements that are implemented easily in most computer languages without requiring a separate interpreter. Fig. 3 represents the decision tree analysis model.

LANDSLIDE SUSCEPTIBILITY MAPPING AND RESULTS

In order to perform the research reported in this paper, PASW Analysis software was chosen as the analysing platform as it supports decision trees with continuous variables (called as regression trees). It supports for nested table, automatic feature selection, automatic cardinality reduction features and other reasons for choosing this data mining platform. The decision tree is used for the purpose of predicting landslide susceptible areas to develop a model, a well known classification technique of analysis. The PASW analysis employs its own decision tree algorithm and this algorithm can handle both discrete and continuous variables or attributes and shows useful parameters for configuring tree induction step such as tree splitting

options, and it can build dependency network graphs which show the effects of variables between them and impact degrees of independent variable(s) on predictable variable(s). Analysis via PASW statistics 18 software requires that all of the factors be derived from ArcView GIS and introduced into the PASW. Total landslides contain 14,189 pixels in the study area, and this analysis consists of ten input attributes and one predictable attribute (grid cells with landslide, 1, and grid cells without landslide, 0). After data preparation, the PASW algorithm is run on 70% training and 30% test cases of the whole dataset.

The model summary is shown in Table 1. By using the trained model, the landslide susceptibility values of each pixel on the map are determined via the prediction tool of the software, and the landslide susceptibility map is produced with the help of predicted numerical values of each pixel on the map. The final map was prepared using these output values imported into ArcView. The final susceptibility map produced by the PASW decision tree algorithm is shown in Fig. 4.

The result of this analysis shows that only a few places have high and very high susceptible areas. The proportion of very low and low susceptible areas is high. From viewing map, it can be concluded that the role of drainage and slope angle is most important in determining the susceptible area.

Table:1 Model summary used for the decision tree analysis

Model Summary		
Specifications	Growing Method	EXHAUSTIVE CHAID
	Dependent Variable	landslide
	Independent Variables	twi, spi, ls, lineament, drainage, aspect, curvature, geology, landuse
	Validation	Split Sample
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	ls, curvature, drainage, twi, spi, landuse, aspect, geology, lineament
	Number of Nodes	51
	Number of Terminal Nodes	27
	Depth	12

* Decision Tree.
TREE landslide [n] BY twi [s] spi [s] ls [s] lineament [s] drainage [s] aspect [n] curvature [n] geology [n] landuse [n]

Poudyal et al. (2010) produced previously the landslide susceptibility map of the same area by frequency ratio analysis and artificial neural networks (ANNs). The model produced by Poudel et al. (2010) is used to generate a landslide susceptibility map that correctly classified 82.2% and 78.2% of the landslide-prone areas. In the present study, to check the performance of the produced map, the AUC is drawn (Fig. 5) and obtained as 95.9%.

CONCLUSIONS

The performance of the produced map in this study is higher than that of the map produced earlier. As a result of the application of decision tree method, the drainage network has prominent role on the decisive role for the occurrences of the landslide in the application site. If the field conditions are taken into account, this result is physically conceivable. The main difference of the decision trees employed in the present study from the other methods exhibits the order of the conditioning parameters. This situation provides a comparison between analysis results and field observations for the expert. The decision trees analysis shows a high prediction capacity, and therefore is a useful tool when producing reliable landslide susceptibility maps.

REFERENCES

Binaghi, E., Luzi, L., Madella, P., Pergalani, F., and Rampini, A., 1998. Slope instability zonation: a comparison between certainty factor and fuzzy Dempster – Shafer approaches, Natural Hazards, 17, pp. 77–97. http://dx.doi.org/10.1023/A:1008001724538

Breiman, L., 2001. Random forests. Machine Learning, 45(1), pp. 5–32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J., 1984. Classification and Regression Trees, Wadsworth International Group.

Clerici, A., Perego, S., Tellini, C., Vescovi, P., 2002. A procedure for landslide susceptibility zonation by the conditional analysis method. Geomorphology, 48, pp. 349-364. http://dx.doi.org/10.1016/S0169-555X(02)00079-X

Donati, L. and Turrini, M. C., 2002. An objective method to rank the importance of the factors predisposing to landslides with the GIS methodology: application to an area of the Apennines (Valnerina; Perugia, Italy), Engineering Geology, 63, pp. 277–289. http://dx.doi.org/10.1016/S0013-7952(01)00087-4

Duman, T.Y., Can, T., Gokceoglu, C., Nefeslioglu, H.A., Sonmez, H., 2006. Application of logistic regression for landslide susceptibility zoning of Cekmece Area, Istanbul, Turkey. Environmental Geology, 51, pp. 241-256. http://dx.doi.org/10.1007/s00254-006-0322-1

Düzgün HSB, Özdemir A., 2006. Landslide Risk Assessment and Management by Decision Analytical
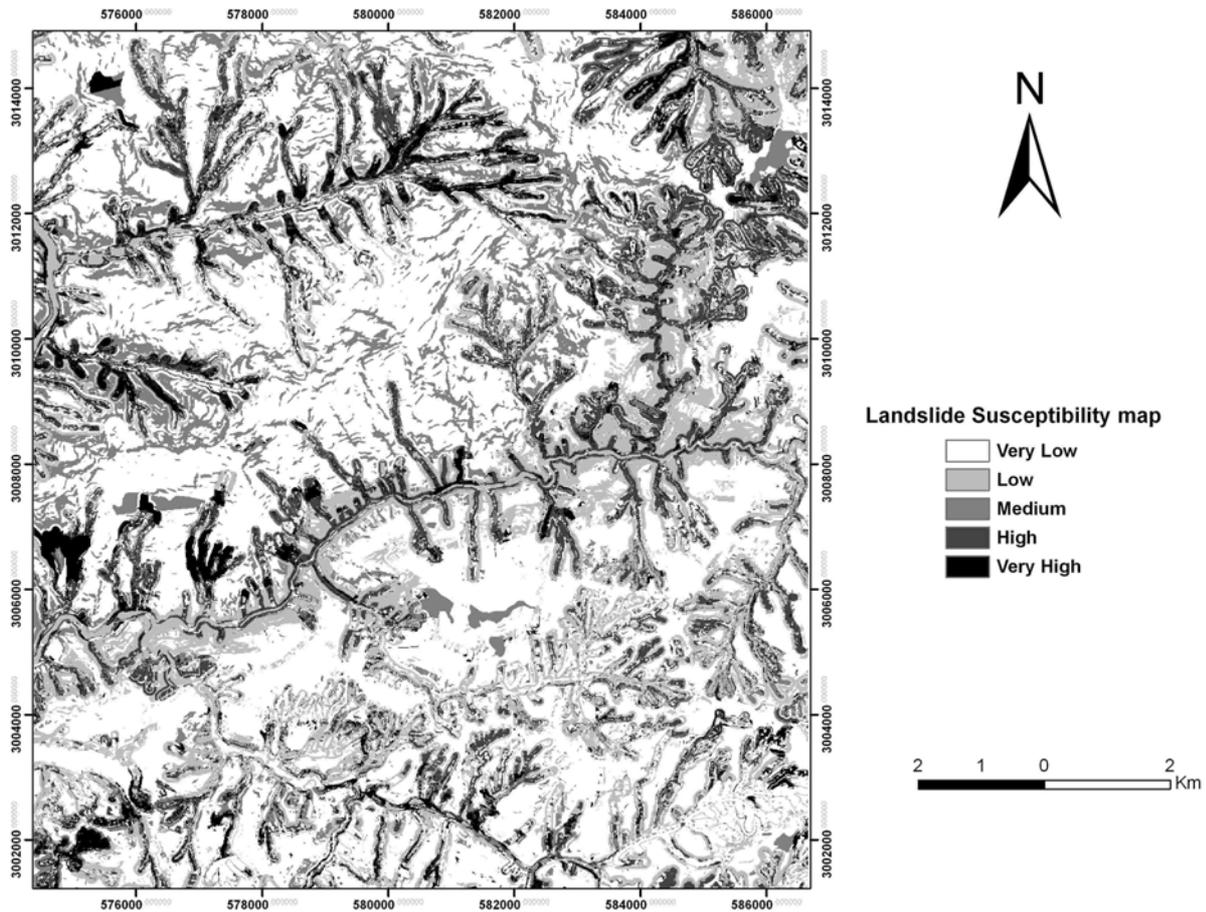
Fig. 4 Landslide susceptibility map produced by using the decision tree method.
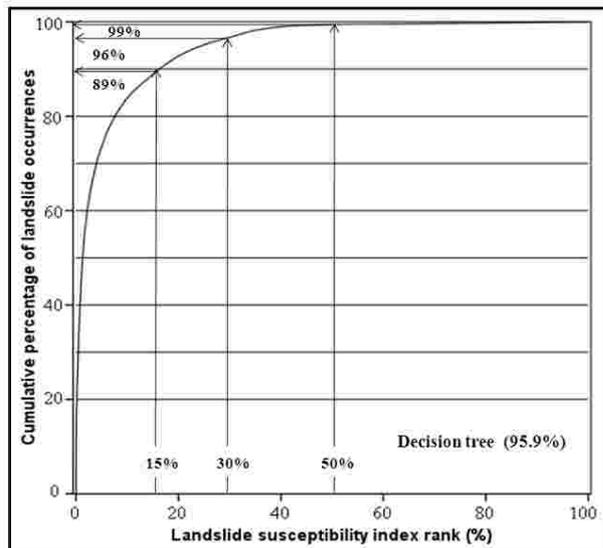
Fig. 5 Area under curve (AUC) evaluation of the decision tree analysis.

Prodecure for Dereköy, Konya, Turkey. Natural Hazards, 39, pp. 245–263. http://dx.doi.org/10.1007/s11069-006-0026-6

Gokceoglu, C., Sonmez, H., Nefeslioglu, H.A., Duman, T.Y., Can, T., 2005. The 17 March 2005 Kuzulu landslide (Sivas, Turkey) and landslide–susceptibility map of its near vicinity. Engineering Geology, 81, pp. 65–83. http://dx.doi.org/10.1016/j.enggeo.2005.07.011

Gomez, H., Kavzoglu, T., 2005. Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. Engineering Geology, 78, pp. 11–27. http://dx.doi.org/10.1016/j.enggeo.2004.10.004

Gorum, T., Gonencgil, B., Gokceoglu, C., Nefeslioglu, H. A., 2008. Implementation of reconstructed geomorphologic units in landslide susceptibility mapping: the Melen Gorge (NW Turkey), Natural Hazards, 46, pp. 323-351. http://dx.doi.org/10.1007/s11069-007-9190-6

Guzzetti, F., Carrara, A., Cardinali, M., Reichenbach, P.,

75

1999. Landslide hazard evaluation: a review of current techniques and their application in a multi–scale study, Central Italy, Geomorphology, 31, pp. 181–216. http://dx.doi.org/10.1016/S0169-555X(99)00078-1

Guzzetti, F., Reichenbach, P., Cardinali, M., Galli, M., Ardizzone, F., 2005. Landslide hazard assessment in the Staffora basin, Northern Italian Apennines. Geomorphology, 72, pp. 272-299. http://dx.doi.org/10.1016/j.geomorph.2005.06.002

Guzzetti, F., Galli, M., Reichenbach, P., Ardizzone, F., Cardinali M., 2006. Landslide hazard assessment in the Collazzone area, Umbria, Central Italy, Natural Hazards and Earth System Sciences, 6, pp. 115–131. http://dx.doi.org/10.5194/nhess-6-115-2006

Juang, C.H., Lee, D.H., Sheu, C., 1992. Mapping slope failure potential using fuzzy sets. Journal of Geotechnical Engineering Division, ASCE, 118, pp. 475-493. http://dx.doi.org/10.1061/(ASCE)0733-9410(1992)118:3(475)

Kanungo, D.P., Arora, M.K., Sarkar, S., Gupta, R. P., 2006. A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. Engineering Geology, 85, pp. 347–366. http://dx.doi.org/10.1016/j.enggeo.2006.03.004

Lee, S., Min, K., 2001. Statistical analysis of landslide susceptibility at Yongin, Korea. Environmental Geology, 40, pp. 1095–1113. http://dx.doi.org/10.1007/s002540100310

Lee, S., 2005. Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data, International Journal of Remote Sensing, 26, (7-10), pp. 1477–1491. http://dx.doi.org/10.1080/01431160412331331012

Lee, C. F., Ye, H., Yeung, M. R., Shan, X., Chen, G., 2001. GIS based methodology for natural terrain landslide susceptibility mapping in Hong Kong, Episodes, 24, pp. 150–159.

Lee, S., Ryu, J.H., Kim, I. S., 2007. Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: case study of Youngin, Korea. Landslides, DOI 10.1007/s10346-007-0088

Miller, H.J., Han, J., 2001. Geographic Data Mining and Knowledge Discovery, CRC Press. http://dx.doi.org/10.4324/9780203468029

Pachauri, A.K. and Pant, M., 1992. Landslide hazard mapping based on geological attributes. Engineering Geology, 32, pp. 81-100. http://dx.doi.org/10.1016/0013-7952(92)90020-Y

Pal M and Mather PM., 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86, pp. 554–56. http://dx.doi.org/10.1016/S0034-4257(03)00132-9

Poudyal C. P., Oh H. J., Chang C., and Lee S., 2010. Landslide susceptibility maps comparing frequency ratio and artificial neural networks: a case study from the Nepal Himalaya. Environmental Earth Science, http://dx.doi.org/10.1007/s12665-009-0426-5

Quinlan, J. R., 1986. Induction of decision trees, Machine learning, 1: pp. 81–106. http://dx.doi.org/10.1007/BF00116251

Quinlan, J. R., 1993. C4.5: Programs for Machine Learning. Los Altos, California: Morgan Kaufmann Publishers, Inc. http://dx.doi.org/10.1023/A:1022643204877

Saito, H., Nakayama, D., and Matsuyama, H., 2009. Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan, Geomorphology, 109(3-4), pp. 108-121. http://dx.doi.org/10.1016/j.geomorph.2009.02.026

Saadia, R.T., Munir, H.S., Shaheen, N., Khalique, A., Manzoor, S., Jaffar, M., 2006. Multivariate analysis of trace metal levels in tannery ef?uents in relation to soil and water: a case study from Peshawar, Pakistan, Journal of Environmental Management, 79, pp. 20–29. http://dx.doi.org/10.1016/j.jenvman.2005.05.009

Turer, D., Nefeslioglu, H.A., Zorlu, K., Gokceoglu, C., 2008. Assessment of geo-environmental problems of the Zonguldak province (NW Turkey). Environmental Geology, 55, pp. 1001-1014. http://dx.doi.org/10.1007/s00254-007-1049-3

Turrini, M. C. and Visintainer, P., 1998. Proposal of a method to define areas of landslide hazard and application to an area of the Dolomites, Italy, Engineering Geology, 50, pp. 255–265. http://dx.doi.org/10.1016/S0013-7952(98)00022-2

Upreti BN, Dhital MR., 1996. Landslide studies and management in Nepal. ICIMOD, Nepal, 87p.

van Westen CJ, Soeters R, Sijmons K., 2000. Digital geomorphological landslide hazard mapping of the Alpago area, Italy, Int Appl Earth Observ. Geoinformatics 2(1), pp. 51–59.

Wan Sa, Lei TC. and Chou TY., 2010. A novel data mining technique of analysis and classification for landslide problems. Natural Hazards, 52, pp. 211–230. http://dx.doi.org/10.1007/s11069-009-9366-3

Wang X., and Niu R., 2009. Spatial forecast of landslides in three gorges based on spatial data mining, Sensors, 9(3), pp. 2035–2061. http://dx.doi.org/10.3390/s90302035