Environmental Modelling & Software 25 (2010) 747–759

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling

Biswajeet Pradhan [a,*], Saro Lee [b,1]

[a] Institute for Cartography, Faculty of Forestry, Geo and Hydro-Science, Dresden University of Technology, 01062 Dresden, Germany
[b] Geoscience Information Center, Korean Institute of Geoscience and Mineral Resources (KIGAM), 30, Kajung-Dong, Yusung-Gu, Daejon, South Korea

ARTICLE INFO

Article history:
Received 14 May 2009
Received in revised form
29 October 2009
Accepted 31 October 2009
Available online 8 December 2009

Keywords:
Landslide
Susceptibility
Artificial neural network
GIS
Klang Valley
Malaysia

ABSTRACT

Data collection for landslide susceptibility modeling is often an inhibitive activity. This is one reason why for quite some time landslides have been described and modelled on the basis of spatially distributed values of landslide-related attributes. This paper presents landslide susceptibility analysis in the Klang Valley area, Malaysia, using back-propagation artificial neural network model. A landslide inventory map with a total of 398 landslide locations was constructed using the data from various sources. Out of 398 landslide locations, 318 (80%) of the data taken before the year 2004 was used for training the neural network model and the remaining 80 (20%) locations (post-2004 events) were used for the accuracy assessment purpose. Topographical, geological data and satellite images were collected, processed, and constructed into a spatial database using GIS and image processing. Eleven landslide occurrence related factors were selected as: slope angle, slope aspect, curvature, altitude, distance to roads, distance to rivers, lithology, distance to faults, soil type, landcover and the normalized difference vegetation index value. For calculating the weight of the relative importance of each factor to the landslide occurrence, an artificial neural network method was developed. Each thematic layer's weight was determined by the back-propagation training method and landslide susceptibility indices (LSI) were calculated using the trained back-propagation weights. To assess the factor effects, the weights were calculated three times, using all 11 factors in the first case, then recalculating after removal of those 4 factors that had the smallest weights, and thirdly after removal of the remaining 3 least influential factors. The effect of weights in landslide susceptibility was verified using the landslide location data. It is revealed that all factors have relatively positive effects on the landslide susceptibility maps in the study. The validation results showed sufficient agreement between the computed susceptibility maps and the existing data on landslide areas. The distribution of landslide susceptibility zones derived from ANN shows similar trends as those obtained by applying in GIS-based susceptibility procedures by the same authors (using the frequency ratio and logistic regression method) and indicates that ANN results are better than the earlier method. Among the three cases, the best accuracy (94%) was obtained in the case of the 7 factors weight, whereas 11 factors based weight showed the worst accuracy (91%).

1. Introduction

Landslide presents a significant constraint to development in many parts of Malaysia which experience frequent landslides, the most recent ones occurring in 2000, 2001, 2004, 2007, 2008 and 2009 (Pradhan and Lee, 2009a). In Malaysia, landslides mostly occur due to heavy tropical rainfalls. Mitigation of landslide hazards can be successful only with a detailed knowledge about the expected frequency, character and magnitude of mass movements in an area. Hence, the identification of landslide-prone regions is essential for carrying out quicker and safer mitigation programs, as well as future planning of the area. In recent years, greater awareness of disasters due to landslides has brought attention to the Malaysian government. However, so far few

* Corresponding author. Tel: +49 351 463 33099; fax: +49 351 463 37028.
E-mail addresses: biswajeet@mailcity.com, biswajeet.pradhan@mailbox.tu-dresden.de (B. Pradhan).
[1] Tel: +82 42 868 3057; fax: +82 42 861 9714.

1364-8152/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2009.10.016

attempts have been made to predict the landslides or prevent the damages caused by them.

Several different approaches to the numerical evaluation of landslide hazard can be found in current literature, including direct and indirect heuristic approaches, and deterministic, probabilistic and statistical approaches. Guzzetti et al. (1999) summarized many landslide hazard evaluation studies based on geomorphological relationships between landslide types, pattern, and the morphological, lithological and structural settings. Recently, studies on landslide susceptibility assessment made use of GIS, and many applied probabilistic models such as frequency ratio and logistic regression methods (Akgün and Bulut, 2007; Dahal et al., 2008; Lee et al., 2002a,b; Lee and Talib, 2005; Lee and Dan, 2005; Akgün et al., 2008; Tunusluoglu et al., 2008; Gorum et al., 2008; Lamelas et al., 2008; Süzen and Doyuran, 2004; Lee and Sambath, 2006; Gokceoglu et al., 2000; Zerger, 2002; Pradhan et al., 2006, 2008; Pradhan and Youssef, 2009). Other conventional methods for landslide susceptibility analysis also exist in the literature. For example, Iovine et al. (2003a,b) applied cellular-automata based deterministic modeling approaches for assessing debris-flow susceptibility in Italy. Among the recent approaches to landslide susceptibility evaluation, some studies adopted fuzzy logic and artificial neural network models (Ercanoglu and Gokceoglu, 2002; Neaupane and Achet, 2004; Arora et al., 2004; Lee, 2007; Lee et al., 2003a,b, 2004; Gomez and Kavzoglu, 2005; Pradhan and Lee, 2007, 2009a, 2009b, 2009c; Pradhan et al., 2009; Youssef et al., 2009). In the neural network method, Nefeslioglu et al. (2008) showed that ANNs give a more optimistic evaluation of landslide susceptibility than logistic regression analysis, whereas Melchiorre et al. (2008) improved the predictive capability and robustness of ANNs by introducing a cluster analysis. Moreover, Kanungo et al. (2006) showed that landslide susceptibility map derived from combined neural and fuzzy weighting procedure is the best amongst the other weighting techniques. Ermini et al. (2005) compared two neural architectures (Probabilistic Neural Network and Multi-Layered Perceptor), obtaining a better prediction with this latter. Also, in Catani et al. (2005), the temporal hazard was estimated with ANNs via the translation of state of activity in recurrence time and hence probability of occurrence. Ermini et al. (2005) and Catani et al. (2005) used unique conditions units for the terrain units definition in ANNs analysis. Pradhan and Lee (2009a) assessed landslide susceptibility analysis at Pennag, Malaysia using frequency ratio, logistic regression and ANN model in which they concluded that logistic regression model showing slightly better prediction accuracy than the latter. Finally, Lui et al. (2006) assessed the landslide hazard using ANNs for a specific landslide typology (debris flow), considering among the triggering factors frequency of flooding, covariance of monthly precipitation, and days with rainfall higher than a critical threshold.

Up to date, few studies have been carried out on landslide susceptibility and risk analysis in Malaysia. Lee and Pradhan (2006) performed landslide susceptibility and risk analyses for Penang Island using a frequency ratio and logistic regression model. Recently, Pradhan and Lee (2009b) used ANN model with different training sites for landslide hazard and risk analysis at Penang Island in Malaysia. A GIS-based landslide hazard analysis has been carried out by the authors in the same area, by using frequency ratio and bivariate logistic regression model (Lee and Pradhan, 2007). In order to compare the bias effect in the earlier models, the ANN concept was applied in the present study. The main difference between this study and the references is that the weights of the causative factors were here calculated three times using ANN model and validated with landslide test location data. Furthermore, a comparison of the susceptibility maps obtained in this study with those earlier generated by frequency ratio and logistic regression model is also commented.

2. Case study: Klang valley area of Selangor

2.1. Study area characteristics

This study was conducted in a landslide-prone area in the Klang Valley in Selangor, using a variety of spatial datasets. The study area is located approximately between 2° 40′ E and 3° 50′E and 101° 30′N and 102° 0′N (Fig. 1). The landuse of the study area consists mainly of peat-swamp forest, plantation forest, inland forest, scrub, grassland and ex-mining area. Since this is a follow-up paper to Lee and Pradhan (2007), the study area and the datasets used for neural network processing remain almost the same. In this paper, areas where landslides records are not available were excluded from the analysis. Additionally, two parameters (elevation and distance to roads) were added to the existing spatial database. A detailed description of the study area can be found in Lee and Pradhan (2007). The regional geology map of the study area is shown in Fig. 2.

2.2. Landslides

In the study area, the main triggering factor for landsliding is the high amount of precipitation. The study area in Klang Valley, Selangor, experienced much landslide damage mainly from soil slides occurring where the maximum daily rainfall was around 200 mm. The intensity of the rain is another factor that affects the hill slopes causing severe sheet, rill and gully erosion. During such times, many of the natural and man-made slopes would be marginally unstable. However, the neural network analysis does not include precipitation because rainfall can be assumed relatively uniform throughout the study area.

Recent studies have shown that, in this area, most slopes show a tendency to landslides of various types and dimensions (Lee and Pradhan, 2007). Many of them are active and periodically cause serious and widespread damage to buildings and people. In order to delineate a complete landslide inventory map of the area under study, many data sources were utilized. Remote sensing methods were used to obtain historical records of the landslides. Archived 1:5,000–1:50,000 aerial photographs, SPOT 5 panchromatic satellite image and landslide reports over the past 23 years were used for the visual detection of landslide occurrences in the study area. These aerial photographs and satellite images were taken during 1981–2004 and were acquired by the Malaysian Remote Sensing Agency. In addition, all historical landslides reports, newspaper records and archived data were assembled for the period under examination. The source material varies in quality with respect to precise location of each landslide event. Based on site description, archived database and aerial photo interpretation, the locations of the individual landslides were located on 1:25,000 maps and the location plotted as closely as possible. In the aerial photographs and high-resolution satellite images, historical landslides could be observed as breaks in the forest canopy, bare soil, or geomorphological features like head- and side scarps, flow tracks, and soil- and debris deposits below a scar. These landslides were then classified and sorted out based on their modes of occurrence. The landslide inventory map was very helpful in understanding that different triggering factors control different slope movement types. Most of the landslides are shallow rotational and a few translational in type. A few landslides which occurred in almost flat areas were not considered in the analysis.

The scar locations and the vectorised landslide outlines are shown in Fig. 3. Multiple field investigations were carried out (12–18 June 2006; 19–25 September 2006; 3–9 January 2007; 18–22 October 2007; and 6–10 October 2008) and fresh landslide locations were collected using GPS survey.

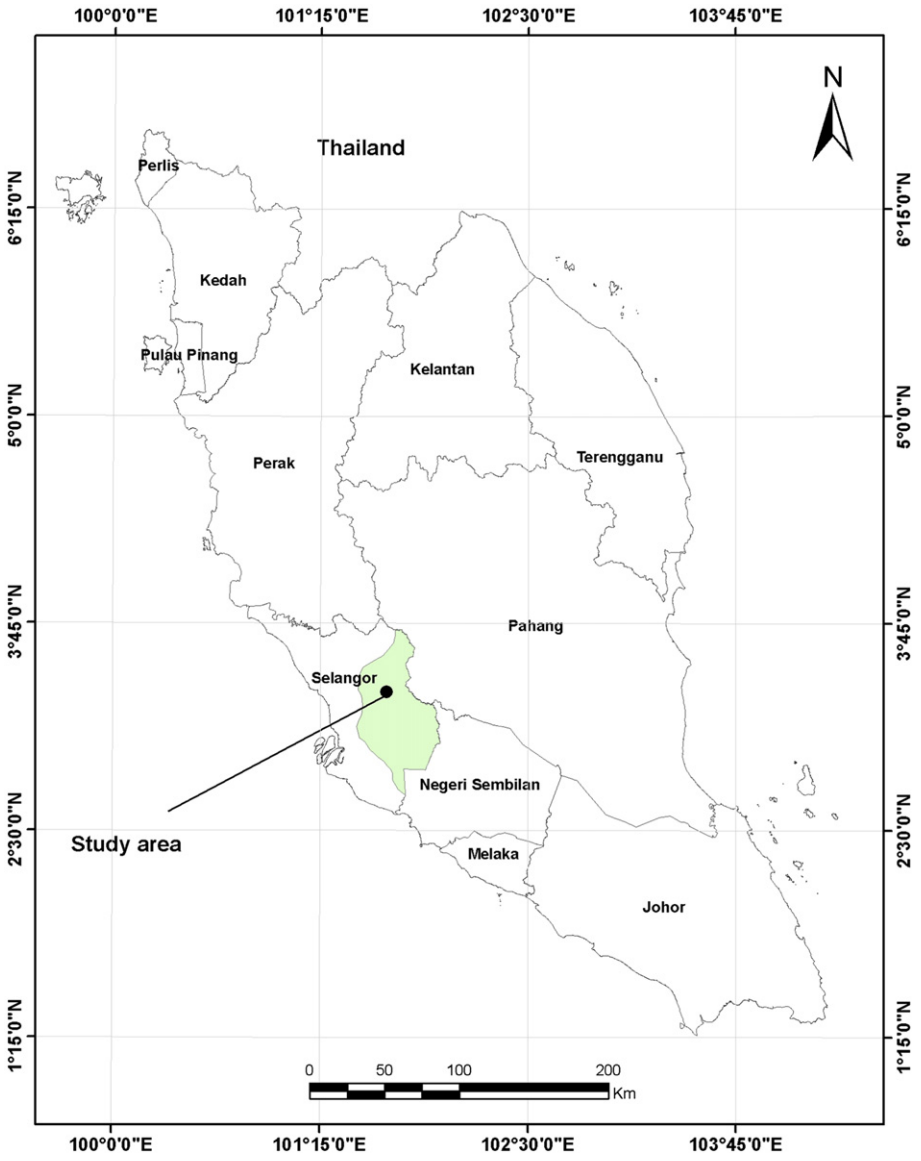

Fig. 1. Study area.

A total number of 390 landslides, covering 2,124,300 grid cells (10 m × 10 m), were mapped in the study area. These landslide-affected grid cells were given a value of 1 while a 0 value was given to 3,289,894,400 grid cells not affected by landslides.

3. Data used

Landslides have multiple and often interconnected causes which are usually of complex parameterization. For large-scale analysis of landslide activity, data acquisition regarding principal casual factors is necessary, followed by the construction of base maps representing the factors controlling the distribution of landslides in the area under study (Ermini et al., 2005; Catani et al., 2005; Ercanoglu, 2005). The parameters utilized in this analysis are variables of both categorical and continuous data. The generated landslide-related parameters are shown in Table 1. Both nominal and numeric variables were subdivided into appropriate classes, defined on the basis of the influence that they exert on landslide mechanics, and expressed in the interval 0–1.

In this study, eleven factors were selected in the first stage: slope angle, slope aspect, curvature, elevation, distance to rivers, distance to roads, distance to faults, surface geology, normalized difference vegetation index (NDVI), landcover and soil types. These parameters were transformed into a spatial vector database using GIS. Later, the thematic layers were converted into a raster grid with 10 m × 10 m cells for application in the neural network modeling. In this analysis, the weights of the factors were calculated three times, using all 11 factors in the first case, then recalculating after removal of those 4 factors that had the smallest weights, and thirdly after removal of the remaining 3 least influential factors. These three cases were applied to the study area and the results mapped using GIS. The study area is characterized by 12,131 rows and 21,258 columns grid cells (i.e., total number is 257,880,798). Both, ArcView 3.2 and ArcGIS 9.0 were used as the basic GIS analysis tools for spatial management and data manipulation.

3.1. Geomorphological parameters

The slope angle is easily relatable to the slope movement, because it is strongly linked to the forces involved. The slope angle map was derived from a DEM with a resolution of 10 × 10 m.

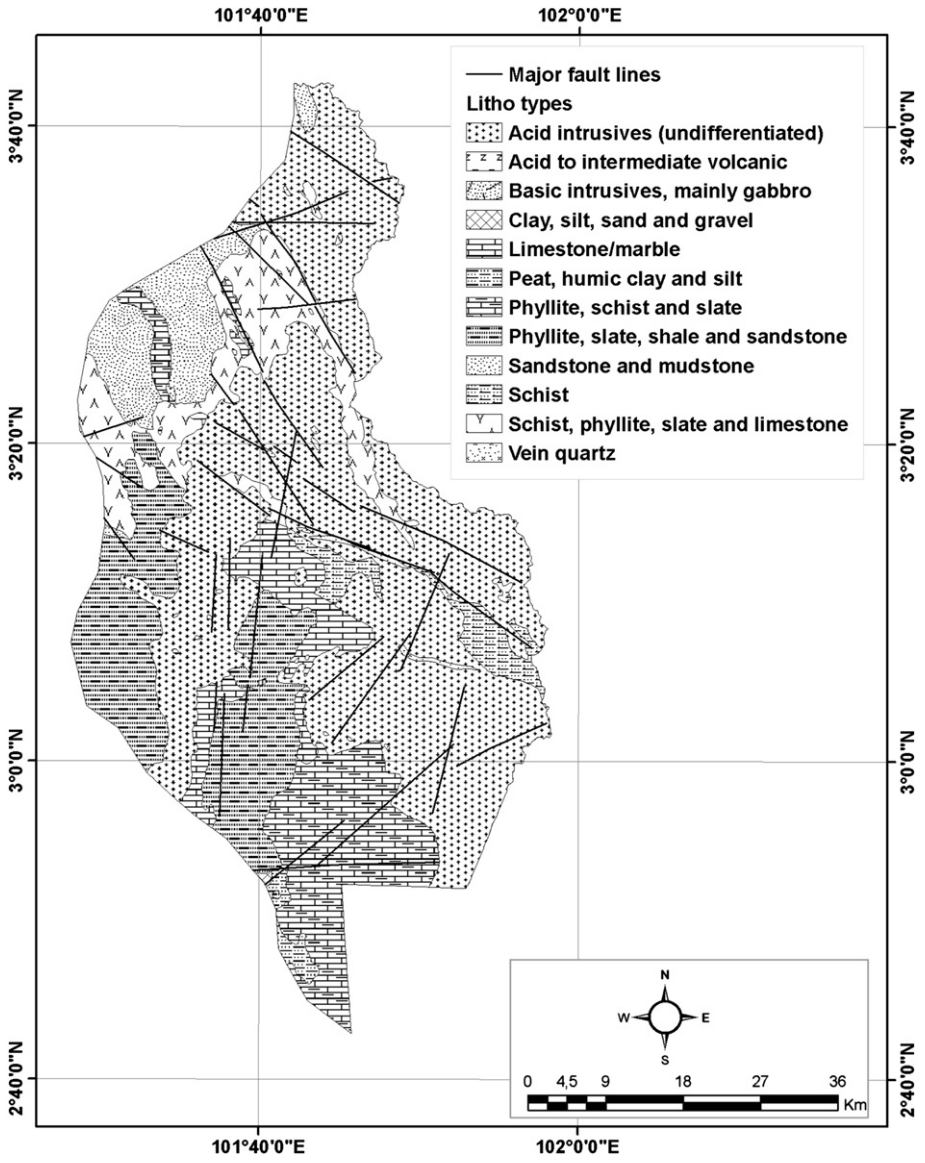

Fig. 2. Regional geology and structural map of the study area (Source: Geological Survey of Malaysia).

According to USGS (1993), the positional accuracy needed for 1:25,000 scale maps must be 12.5 m. For this reason, a pixel size of 10 m was selected for the DEM.

Using the above DEM, slope aspect and curvature, was extracted using the Digital Elevation Model Analysis Tool (DEMAT) of ArcView 3.2. Slope aspects in the study area generally tend towards all 8 directions. In the curvature map, negative curvatures represent concave, zero curvatures represent flat and positive curvatures represents convex.

Altitude is useful to classify the local relief and locate points of maximum and minimum heights within terrains. Altitude refers to the height of the study area and it varies between 0 and 1760 m.

3.2. Distance to rivers

The distance to rivers is represented by the proximity of the rivers and drainages in the area. The drainage map at the scale of 1:25,000 was obtained from the Department of Agriculture. The most significant drainage type observed in the study area is dendritic pattern. The drainage buffers were calculated based on the Euclidean distance method in ArcView 3.2.

3.3. Distance to roads

Road-cuts are usually sites of anthropologically instability. A given road segment may act as a barrier, a net source, a net sink or a corridor for water flow, and depending on its location in the area, it usually serves as a source of landslides. The complete transportation network was included in this GIS-based analysis. However, roads on flatlands and on top of ridges were excluded in this study.

3.4. Geological parameters

It is widely recognized that geology greatly influences the occurrence of landslides, because lithological and structural variations often lead to a difference in strength and permeability of rocks and soils. The Klang Valley, geological map, produced for the study area and its surroundings by Mineral and Geosciences Department of Malaysia consist of 12 rock units. The study area is mainly covered by acid intrusive (49%).

Fault lines and lineaments were extracted from a mosaic of SPOT 5 images using edge enhancement and filtering techniques as well

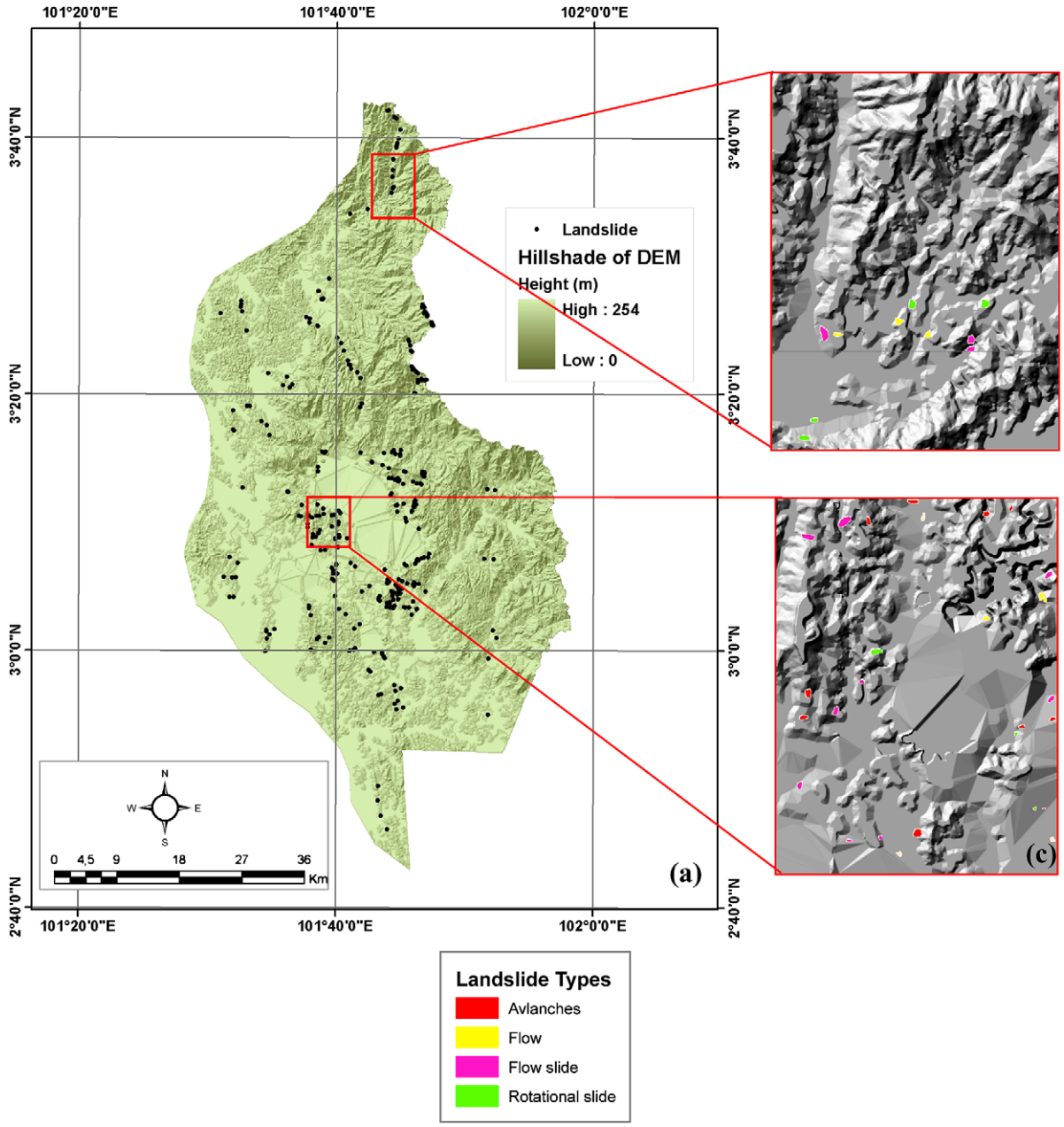

Fig. 3. Landslide inventory map of the Klang Valley area. (a) Landslide locations overlaid on hill shaded DEM; (b) enlarged views of the north eastern part of the area; (c) enlarged views of the central part of the area.

as subsequent field verifications. Lineaments include tectonic structures and geomorphic signatures such as topographic breaks. A total of 39 lineaments were extracted in this study aided by abrupt relief changes, valley and cliffs with lithological variation, erosional features and changes in drainage patterns. The lineaments were enclosed by 100 m buffer zones. This diameter is an average threshold set based on a comprehensive assessment of how far slope failures extend from linear features (Ayalew and Yamagishi, 2005).

3.5. Landcover

In this study, landcover information was extracted from Landsat ETM image using object-based classification method. The segmentation function and classification method performed using eCognition software was found to be a better way to produce the landcover information with different scale parameter. The object-oriented segmentation not only segregates the objects or pixel groups using their spectral characteristics but can also distinguish various classes in the images based on their shape and texture. The landsat ETM scene of 31 May 2001 is classified into different landcover classes. Nine landcover classes were pre-identified for reclassification are dense forest, sparse forest, settlements, rubber, oil palm, mangrove, lake, mining area and urban. A segmentation technique is used to build up a hierarchical network of image objects (see Fig. 4 for illustration). Hence, each image object knows its neighbouring objects and it is possible to define relations between these objects. This is done within a class hierarchy, which is the frame for formulating the knowledge based for the classification process. The eCognition software allows the advanced (semi)

Table 1
Various attribute data layers used in the analysis.

Classification		GIS data type		Scale or resolution	
Landslide	Landslide	ARC/INFO polygon coverage	ARC/INFO GRID	1:25,000	10 m × 10 m
Topographic map	Slope angle Slope aspect Curvature Altitude	ARC/INFO line, point and polygon coverage		1:25,000	
Drainage map	Distance to rivers	ARC/INFO line coverage		1:25,000	
Road network	Distance to roads	ARC/INFO line coverage		1:25,000	
Soil map	Types	ARC/INFO polygon coverage		1:100,000	
Geology map	Litho types Distance to faults	ARC/INFO polygon, line coverage		1:63,300	
Land cover	Land cover	ARC/INFO GRID		30 m × 30 m	
NDVI	NDVI	ARC/INFO GRID		10 m × 10 m	

automatic image analysis relying on image objects by using tools for object-oriented fuzzy-rule classification including contextual and shape information. The basic difference, especially when compared to pixel-based procedures, is that eCognition software does not classify single pixels, but rather image objects, which are extracted in a previous image segmentation step. It is meaningful to analyze groups of contiguous pixels as objects instead of using the conventional pixel-based classification unit. This will reduce the local spectral variation caused by crown textures, gaps and shade.

3.6. Soil types

The different soil types present in the area under study were grouped into a number of types that are homogenous in terms of chemical composition. The digital soil layer was obtained from the hard copy of 1:100,000 soil map from the Department of Soils of Malaysia (which is the only existing soil map of the study area). A set of 21 different soil types were used in this analysis.

3.7. Normalized difference vegetation index (NDVI)

The NDVI map was obtained from Landsat TM satellite image acquired on 15 September 2005. The NDVI value was calculated using the common formula $NDVI = (IR - R)/(IR + R)$. The NDVI value denotes areas of vegetation in an image. The presence of dense green vegetation implies high NDVI values, due to high concentration of chlorophyll resulting in a low reflectance in the red band as well as due to the high stacking of leaves. Sparse vegetation, on the other hand, implies low NDVI values due to less or even no chlorophyll and leaves.

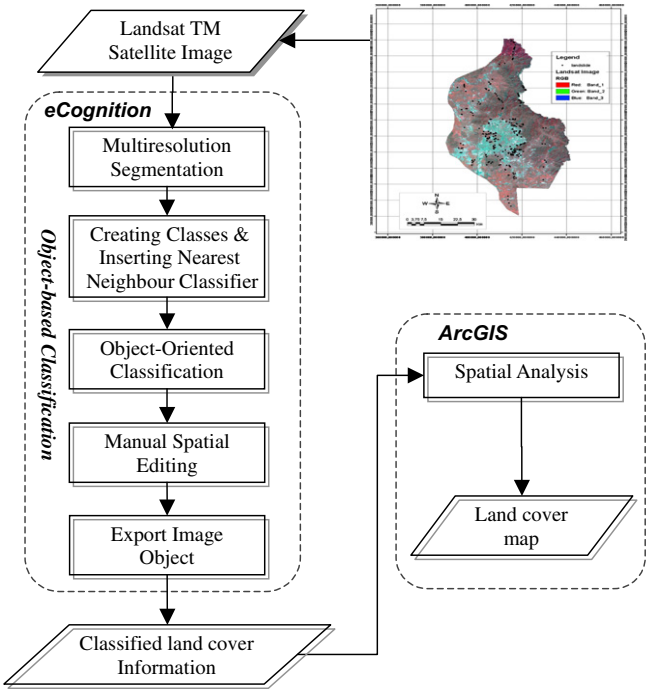

Fig. 4. Land cover information extraction process flow chart.

4. Artificial neural networks

4.1. Preview

The artificial neural network approach has many advantages compared to other statistical methods. First, it is independent on the statistical distribution of the data, and there is no need for specific statistical variables. Neural networks allow the target

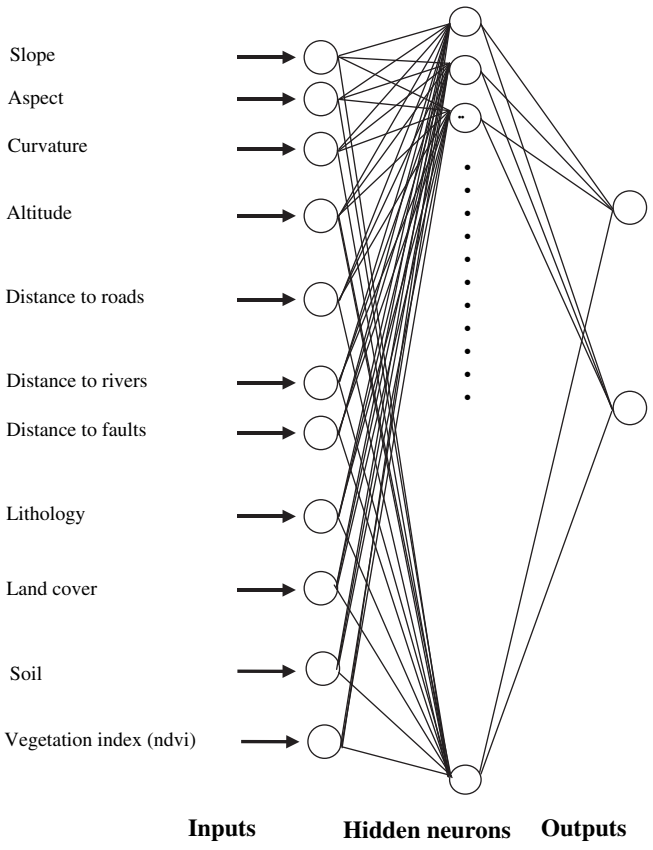

Fig. 5. Architecture of neural network model.

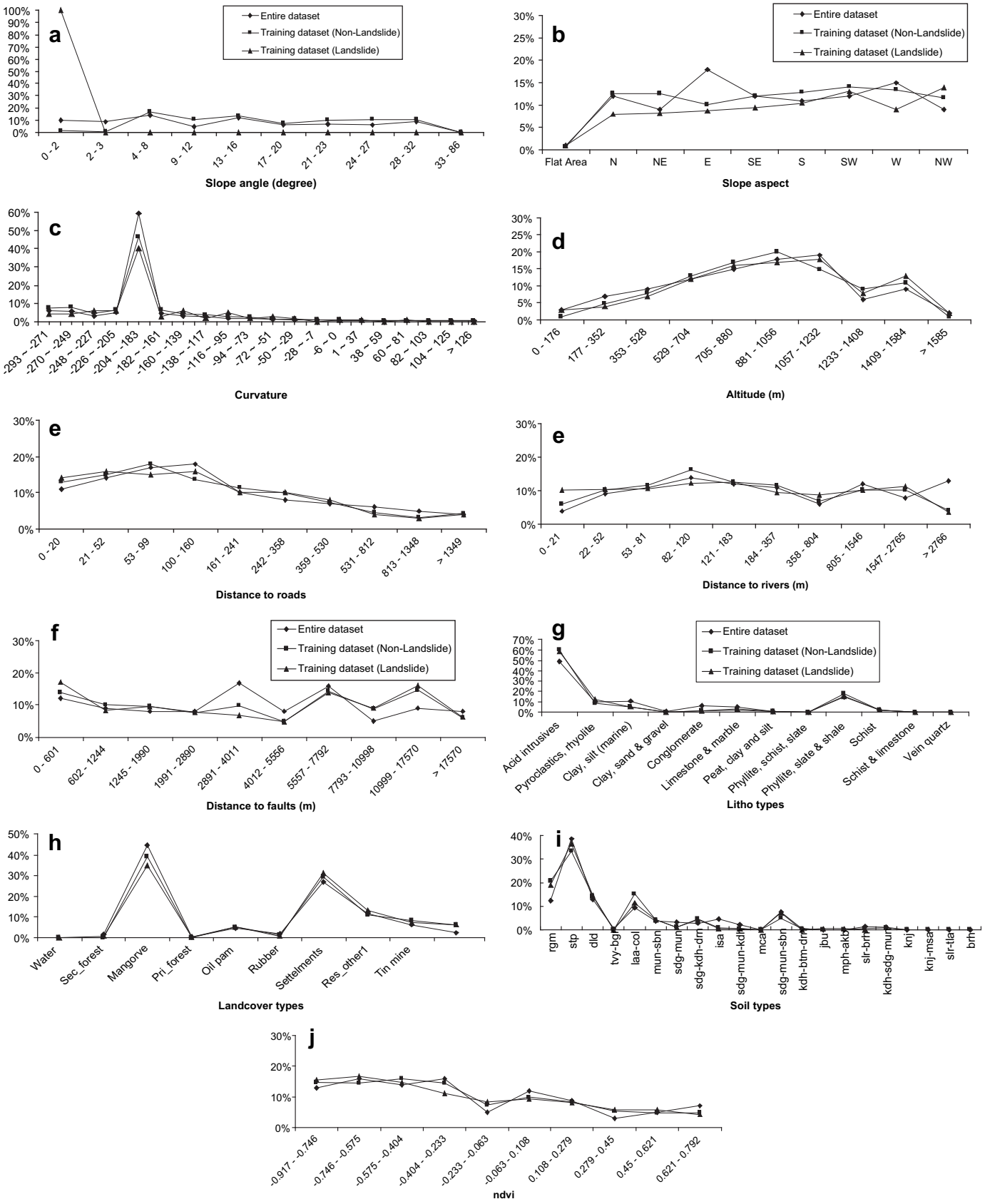

Fig. 6. Distribution of entire and training data set (in percentage).

Table 2
Weights of each factor in case 1.

Factor	No.												
	1	2	3	4	5	6	7	8	9	10	Average	SD	Weight
Slope angle	0.1611	0.1624	0.1683	0.1587	0.1573	0.1564	0.1666	0.1698	0.1718	0.1716	0.1644	0.0059	2.905
Slope aspect	0.0893	0.0867	0.0921	0.094	0.0979	0.0957	0.0843	0.0801	0.0882	0.0928	0.0901	0.0055	1.592
Curvature	0.0947	0.0934	0.0987	0.0992	0.0979	0.0984	0.0979	0.0934	0.0972	0.0998	0.0971	0.0024	1.715
Altitude	0.0905	0.0912	0.0986	0.0963	0.0974	0.0932	0.0985	0.0884	0.0983	0.0756	0.0928	0.0071	1.640
Distance to roads	0.0972	0.0979	0.0981	0.0732	0.0984	0.0789	0.0961	0.0945	0.0932	0.0929	0.0920	0.0088	1.626
Distance to rivers	0.0638	0.0823	0.0719	0.0991	0.0898	0.0887	0.0773	0.0679	0.0782	0.0803	0.0799	0.0107	1.412
Distance to faults	0.056	0.0569	0.0583	0.0667	0.0612	0.0679	0.0674	0.0649	0.0673	0.0689	0.0636	0.0050	1.123
Lithology	0.0967	0.0892	0.0931	0.0985	0.0967	0.0989	0.0991	0.0994	0.0873	0.0864	0.0945	0.0051	1.670
Land-cover	0.0671	0.0567	0.0692	0.0593	0.0488	0.0461	0.0497	0.0599	0.0502	0.0592	0.0566	0.0078	1.000
Soil types	0.0988	0.0985	0.0921	0.0953	0.0941	0.0994	0.0878	0.0863	0.0884	0.0811	0.0922	0.0061	1.629
NDVI	0.0892	0.0873	0.0671	0.0673	0.0668	0.0732	0.0715	0.0879	0.0729	0.0771	0.0760	0.0090	1.343

classes to be defined in relation to their distribution in the corresponding domain of each data source (Zhou, 1999), and therefore the integration of remote sensing or GIS data is very convenient. An artificial neural network is a "computational mechanism able to acquire, represent, and compute a mapping from one multivariate space of information to another, given a set of data representing that mapping" (Atkinson and Tatnall, 1997). An ANN is composed of a set of nodes and a number of interconnected processing elements. ANNs use learning algorithms to model knowledge and save this knowledge in weighted connections, mimicking the function of a human brain (Turban and Aronson, 2001). The Multi-Layer Perceptron (MLP) trained with the back-propagation algorithm (BPA), the most frequently used neural network method, was adopted in this study (Fig. 5). The MLP with the BPA was trained using a set of examples of associated input and output values. The purpose of an artificial neural network is to build a model of the data-generating process, so that the network can generalize and predict outputs from new inputs. The MLP consists of an input layer, a hidden layer, and an output layer. The hidden and the output layer neurons process their inputs by multiplying each input by a corresponding weight, summing the product, and then processing the sum using a nonlinear transfer function to produce a result. An artificial neural network "learns" by adjusting the weights between the neurons in response to the errors between the actual output values and the target output values. At the end of this training phase, the neural network provides a model that should be able to predict a target value from a given input value (Fig. 5).

4.2. Preparation of training and testing data

When developing an artificial neural network, the data is commonly partitioned into at least two subsets such as training and test data. Before running the artificial neural network program, the training site should be selected (Nefeslioglu et al., 2008; Caniani et al., 2008). It is expected that the training data include all the data belonging to the problem domain. Certainly this subset is used in the training stage of the model development to update the weights of the network. On the other hand, the test data should be different from those used in the training stage. The main purpose of this subset is to check the network performance using un-trained data, and to confirm its accuracy. No exact mathematical rule to determine the required minimum size of these subsets exists (Nefeslioglu et al., 2008). However, some suggestions for the portions of these samplings are encountered in the literature (Basheer and Hajmeer, 2000). Considering these suggestions, it is revealed that approximately 80% of whole data is commonly enough to train the network, and the rest of it is usually handled to test the final architecture of the model (Swingler, 1996). In this study, the landslide-prone (occurrence) area and the landslide-not-prone area were selected as training sites. Cells from each of the two classes were randomly selected as training cells, with 398 cells denoting areas where landslide occurred or not. Among the 398 cases of landslide occurrences, 318 cases (80%) were selected for calibrating the ANN and the remaining 80 cases (20%) were used for validation testing. During the training phase, areas not affected by landslides were classified as "areas not prone to landslide", while areas affected by landslides were assigned to the "areas prone to landslide" training set. The distribution of the training pixels for eleven factors is shown in Fig. 6 which demonstrates the training results using MLP by randomly selecting about 80% landslide locations from the entire datasets (318 records).

4.3. Determination of weight of factors

The weight between the layers was acquired by training the neural network so that the contribution or importance of each predictor variable was calculated. The back-propagation algorithm was then applied to calculate the weights between the input layer and the hidden layer, and between the hidden layer and the output layer, by modifying the number of hidden node and the learning rate. A three-layered feed-forward network was implemented using the MATLAB software package. The program, developed by

Table 3
Weights of each factor in case 2.

Factor	No.												
	1	2	3	4	5	6	7	8	9	10	Average	SD	Weight
Slope angle	0.2827	0.2821	0.2981	0.2983	0.2975	0.2948	0.2989	0.2961	0.2972	0.2975	0.294	0.0064	3.034
Slope aspect	0.122	0.102	0.0987	0.0985	0.0969	0.0932	0.0926	0.0823	0.0898	0.0943	0.097	0.0104	1.000
Curvature	0.1119	0.123	0.1141	0.1143	0.119	0.1265	0.1232	0.1235	0.1345	0.1311	0.122	0.0074	1.259
Altitude	0.1124	0.1103	0.1087	0.1135	0.1107	0.1154	0.1161	0.1145	0.1279	0.1121	0.1142	0.0051	1.177
Distance to roads	0.1409	0.1412	0.1367	0.1421	0.1419	0.143	0.1443	0.1464	0.1521	0.1487	0.144	0.0044	1.482
Soil types	0.1007	0.1098	0.1023	0.0984	0.0965	0.0963	0.0971	0.1183	0.0869	0.0989	0.101	0.0085	1.036
Lithology	0.1265	0.1398	0.1385	0.1376	0.1352	0.1345	0.1269	0.1338	0.1129	0.1212	0.131	0.0087	1.347

Table 4
Weights of each factor in case 3.

Factor	No.												
	1	2	3	4	5	6	7	8	9	10	Average	SD	Weight
Slope angle	0.4478	0.4419	0.4411	0.4318	0.4389	0.4375	0.4412	0.4407	0.4428	0.43009	0.4394	0.0052	3.3387
Curvature	0.1329	0.1325	0.1348	0.1311	0.1329	0.1289	0.1287	0.1264	0.1304	0.1372	0.1316	0.0032	0.9998
Distance to roads	0.2402	0.2417	0.2467	0.2571	0.2486	0.2474	0.2412	0.2519	0.2518	0.2532	0.2480	0.0057	1.8843
Lithology	0.1834	0.1882	0.1854	0.1863	0.1869	0.1884	0.1862	0.1787	0.1783	0.1778	0.1840	0.0042	1.3979

Hines (1997) by using MATLAB, was partially modified for landslide analysis by adapting the input and output routines for the use of GIS data. Here, "feed-forward" denotes that the interconnections between the layers propagate forward to the next layer. The number of hidden layers and the number of nodes in a hidden layer required for a particular classification problem are not easy to deduce. As mentioned above, in this study, three cases were analyzed. First, all 11 factors were used for calculating the weights. Then, among these 11 factors, the 4 factors that showed the smallest weight values were deleted, and 7 factors were used for calculating the weights. Third, among the 7 factors, the 3 factors that showed the smallest weight values were deleted and only 4 factors were used for calculating the weights. Therefore, the structures 11 (inputs) × 24 (hidden neurons) × 2 (outputs) for 11 factors used, 7 (inputs) × 16 (hidden neurons) × 2 (outputs) for 7 factors used, and 4 (inputs) × 10 (hidden neurons) × 2 (outputs) for 4 factors used, were selected for the network, with input data normalized in the range 0.1–0.9. The reason for data normalization is that the eleven variables are different in dimensions and are not suitable for direct input for the ANN model. The common method (Choi et al., 2009) for the data processing is to transform the data to the values between 0 and 1. For example, for $y_i (i = 1, 2, \ldots, n)$:

$$Y_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \quad (1)$$

where, Y_i means the normalized values of y_i, $y_{\min}$ and $y_{\max}$ represent the minimum and maximum value of y_i respectively. In this way, the nominal and interval class group data were converted to continuous values ranging between 0.1 and 0.9. Therefore, the continuous values were not ordinal data, but nominal data, and the numbers denote the classification of the input data.

The learning rate was set to 0.01, and the initial weights were randomly selected between 0.1 and 0.3. To test whether the variation of weights is dependent on the initial weight guess, weights calculated from ten cases were compared. Ten training sites were randomly selected to deduce the influence of the training site. The selected landslide locations were assigned (0.1, 0.9) and the non-landslide-prone locations were assigned (0.9, 0.1). The back-propagation algorithm was used to minimize the error between the predicted output values and the calculated output values. The algorithm propagated the error backwards, and iteratively adjusted the weights. The number of epochs was set to 6000, and the root mean square error (RMSE) value used for the stopping criterion was set to 0.01. Most of the training datasets met the 0.01 RMSE goal. However, if the RMSE value was not achieved, then the maximum number of iterations was terminated at 6000 epochs. When the latter case occurred, then the maximum RMSE value was 0.021.

The final weights between the layers were acquired during training of the neural network and the contribution of each of the eleven factors was used to predict landslide susceptibility. After training, the weights were determined as shown in Tables 2–4 (for 11, 7, and 4 employed factors, respectively). The results revealed that the initial weight did not influence the final result. From each of the two classes (landslide and no-landslide), 318 pixels per class were randomly selected as training pixels. The calculation was repeated ten times by randomly assigning the initial weights to determine how the extracted sample represented each class. Although the final results were not the same, there were no large differences. The standard deviation was distributed from 0 to 0.02, indicating that random sampling does not have a large effect on the result.

For easy interpretation, average values of the weights for ten iterations were calculated; these values were divided by the average of the weights of the factor that had a minimum value. For example where 11 factors were used (Table 2), the landcover had the minimum average value of 0.0566 which is normalized to 1.0000, and the slope had the maximum average weight value of 0.1644 which is normalized to 2.905 (0.1644/0.0566) with respect to the landcover value. Similarly, where 7 factors were used (Table 3), the aspect value had the minimum weight of 1.0000 and the slope had the maximum weight of 3.034 and where 4 factors were used (Table 4), the curvature value had the minimum weight of 1.000 and the slope value had the maximum weight of 3.3387. The calculated values were used in the landslide susceptibility analysis, as weights improved the accuracy of the analysis.

4.4. Generation of LSI classification

Once the networks were successfully trained and the weights computed, the trained network with the highest accuracy was used to categorize each and every pixel of the whole dataset to one of the landslide susceptibility zonation classes to produce the LSI classification. In other words, the landslide susceptibility index value was calculated from the weights determined from the back-propagation and the spatial datasets. The index values were between 0.1 and 0.9 for each pixel. The output indices were converted to GIS grid data. Using such values, the landslide susceptibility indices (LSI) were determined and used to create the landslide susceptibility maps. Three landslide susceptibility maps were prepared (for illustration, see Fig. 7), by using n.11 factors (cf. Fig.7a), n.7 factors (cf. Fig.7b), and n.4 factors (Fig.7c). The value of susceptibility was classified by equal area and grouped into five classes for easy and visual interpretation. With an increase in the index, the landslide susceptibility also increases for all the cases. The patterns of the three cases were very similar, but there is some difference in the distribution of index values.

Also, the quality of the LSI classification was assessed against earlier GIS derived maps, based on frequency ratio and bivariate logistic regression methods, and also on field data-as described in the next section.

5. Validation of landslide susceptibility maps

5.1. Validation of LSI classification with landslide testing dataset

The LSI classification produced from ANN for the 3 cases shown in Fig. 7 were validated with the aid of existing landslide test dataset. All the landslides that were not used in the training phase

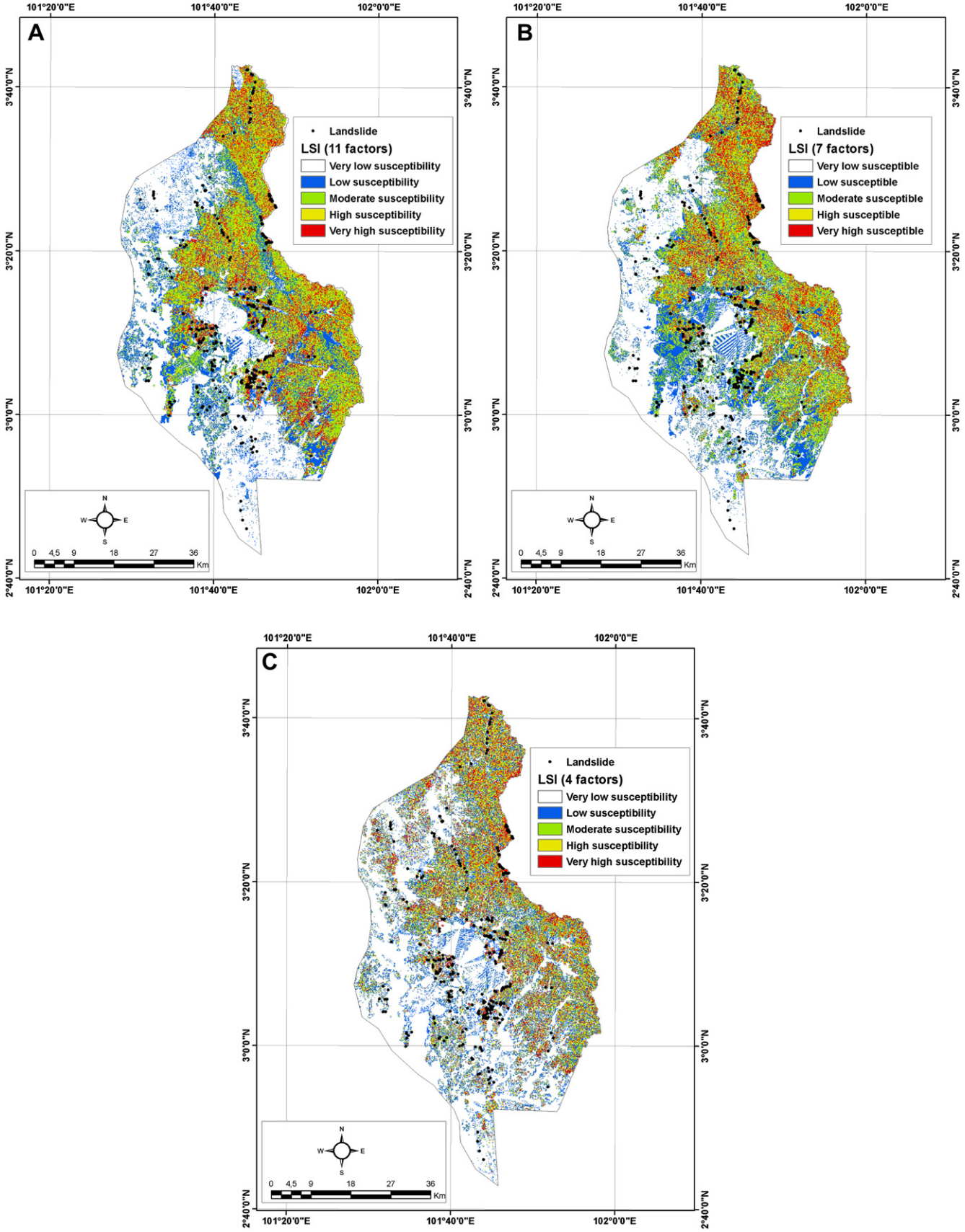

Fig. 7. Landslide susceptibility maps using artificial neural network in the Klang Valley area; (a) using 11 factors; (b) using 7 factors; and (c) using 4 factors.

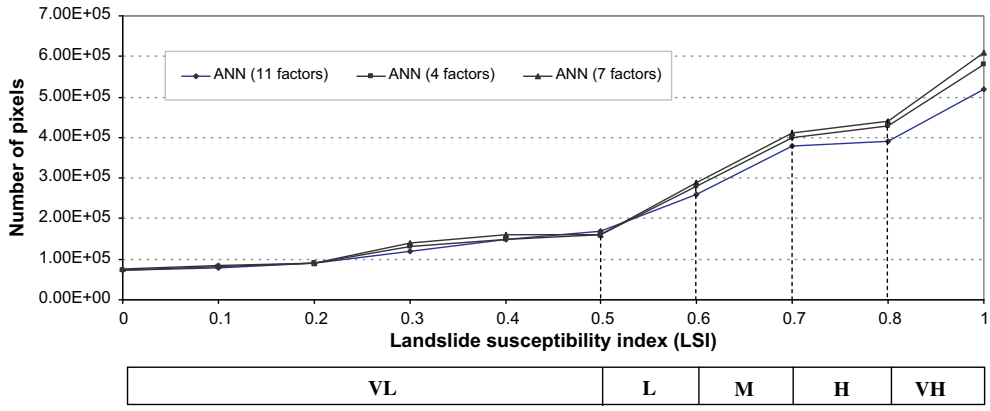

Fig. 8. Cumulative distribution of the susceptibility maps.

were selected as testing sites (80 locations, i.e. 20% of the sample of 398 landslides). Then, the resulted landslide susceptibility maps were reclassified, by defining the limits of the cumulative distribution of LSI values supplied by the ANN. As a result, the LSI were classified into five classes (highest 10%, second 10%, third 10%, fourth 20% and remaining 50%) based on area for visual and easy interpretation.

Relative frequency ratio analysis was performed on the classification results and landslide location test data (Kanungo et al., 2006). The 80 landslide locations were superimposed on the LSI classification obtained from the ANN. In the superimposed image, the boundary of each existing landslide location encloses a number of pixels allocated to five zones of derived landslide susceptibility maps. In the landslide susceptibility maps, the landslide pixels generally coincided with the sites falling in the very high and high susceptibility classes. It can be observed that 90% of landslide pixels were correctly classified (Fig. 8). The total areas and the percentages covered by various landslide susceptibility zones within the boundary of the 80 landslides were determined. Similarly, the total areas with their percentages occupied by various landslide susceptibility zones in the whole image were also determined. Relative frequencies of areas affected by different landslide susceptibility zones were calculated from the ratio. Ideally the frequency ratio value should increase from a very low susceptible zone to a very high-susceptible zone, since the highest landslide susceptible zones are generally more prone to landslides than other zones. On plotting these frequency ratio values (Fig. 8), it can be seen that there is a gradual and smooth increase in the frequency from the not-susceptible zone to the very high-susceptible zone in the three susceptibility maps. It is interesting to note that the analysis carried out by using 7 factors performs better than the 11 and 4 factors, as a number of locations were mapped more accurately in the very high susceptibility zones.

5.2. Comparison of ANN-derived LSI classification with frequency ratio and logistic regression-based models

For an effective comparison of the ANN-derived landslide susceptibility maps (three cases) with that obtained by using frequency ratio and bivariate logistic regression (Lee and Pradhan, 2007), all maps were evaluated by comparing them separately with the landslide data. Therefore, five landslide susceptibility maps (3 ANN-based susceptibility maps, 1 frequency ratio, and 1 bivariate logistic regression susceptibility maps) were validated by using known landslide locations. The accuracy assessment method was performed by comparing the existing landslide dataset with landslide susceptibility results, as shown in Fig. 8. In terms of rate curves and areas under the rate curve (Chung and Fabbri, 1999). The rate explains how well the model and predictor variables predict the landslide. So, the area under the rate curve can be used to assess the prediction accuracy qualitatively (Lee et al., 2003b; Chung and Fabbri, 2003; Godt et al., 2008). The "areas under curve" constitutes

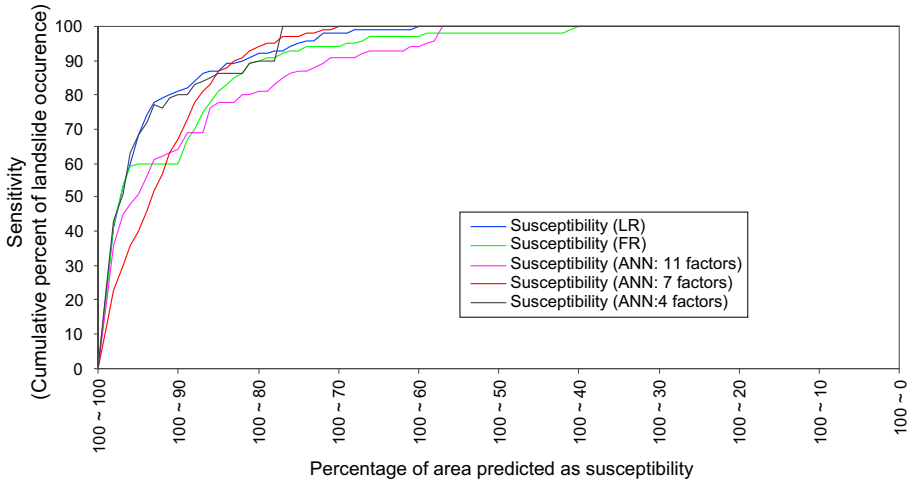

Fig. 9. Illustration of cumulative frequency diagram showing landslide susceptibility index rank (x-axis) occurring in cumulative percent of landslide occurrence (y-axis).

one of the most common used accuracy statistics for the prediction models in natural hazard assessments (Begueria, 2006). To obtain the relative ranks for each prediction pattern, the calculated index values of all cells in the study area were sorted in descending order. Then, the ordered cell values divided into 100 classes was set on the *y*-axis, with accumulated 1% intervals in the *x*-axis. The accuracy assessment results and success rate curve appear as a line in Fig. 8. For example, the 10% (100–90%) class contains about 78% of the study area in the success rate in the case where 7 factors were used. In addition, the 20% (100–80%) class occupies 90% of the study area in the success rate. According to the success-rate verification result, shown in Fig. 8.

In the case of susceptibility map obtained by using weights of 7 factors, the area ratio was 0.94 and the prediction accuracy was 94%. In the case of 11 factors, the area ratio was 0.91 and the prediction accuracy was 91%. Lee and Pradhan (2007) reported that, when frequency ratio and bivariate logistic regression models were used, the prediction accuracy was 90% and 93%, respectively, which is still lower than the accuracy of susceptibility map obtained with the neural network method described in this paper. From the prediction accuracy graphs (Fig. 9), it is quite evident that the susceptibility map with 7 factors shows the best prediction accuracy of 94%. Therefore, the criteria of selection of the parameters and of their weight have an impact on the accuracy of the landslide susceptibility analysis.

6. Conclusions and discussion

In this study, a neural network approach for assessing landslide susceptibility was applied three different ways compared with previous analyses based on frequency ratio and bivariate logistic regression models using GIS and remote sensing data. An artificial neural network approach was used to estimate areas susceptible to landslides using a spatial database for the Klang Valley. Through multiple fieldwork and aerial photo interpretation, a landslide inventory map was created. Of the 390 landslides identified from historical records in a 23 year period, 318 (80%) locations were used for the training of the model, while the remaining 80 (20%) cases were used for the model validation. Among the landslide-related factors, slope angle, slope aspect, curvature, altitude, distance to roads, distance to rivers, lithology, distance to faults, soil, landcover and NDVI were used for the determination of the weights. These latter were determined in three different ways, by considering 11, 7, and 4 factors, respectively. The weights were applied to the entire study area, and three landslide susceptibility maps were obtained.

By comparing landslide occurrence location and susceptibility maps, the accuracy of the susceptibility assessment showed a satisfactory agreement when only 7 factors were used, the best result of three experiments was obtained; when either 11 factors or 4 factors were used, similar but worse results were obtained for the same study area. It is inferred that the ANN-based methodology for integration of various topographic, geological, structural, landcover and the other datasets is quite suitable for developing an LSI classification. The main advantage of utilising ANNs is their objectively in assigning weights to different causative factors, as they involve a minimum of human interference.

Furthermore, the susceptibility maps were compared with results of previous analyses, performed by adopting frequency ratio and logistic regression model of the same area, as reported in Lee and Pradhan (2007). Again, the results obtained with 7 factors (94%) were better than those of frequency ratio (93%) and logistic regression model (90%). Therefore, the study shows that results of ANN-derived LSI classification are comparable and slightly better than those obtained from frequency ratio and logistic regression model. The relative importance of weight was calculated using the artificial neural network method. In the case of 7 factors used, the result shows that the slope is the most important factor for landslide susceptibility mapping. The average normalized value, in the case where 7 factors was used, shows that slope has the highest weight values (3.034) followed by distance to roads (1.482) and then lithology (1.347).

The scientific weights and ratings are essential to landslide susceptibility mapping. The back-propagation training algorithm presents difficulties when trying to follow the internal processes of the procedure. The method also involves a long execution time with a heavy computing load. Therefore, the thematic data layers were converted into ASCII format to speed up the computing process. Computation of the weight of the factors and artificial neural network modeling was performed in MATLAB; the outputs were exported to GIS for map production and visual interpretation. Landslide susceptibility maps were analyzed qualitatively using equal area classification schemes.

These results can be used as basic data to assist slope management and landuse planning. The methods used in the study can also be used for generalized planning and assessment purposes, although they may be less useful on the site-specific scale, where local geological and geographic heterogeneities may prevail. Landslide susceptibility maps are of great help to planners and engineers for choosing suitable locations to implement development action plans. In spite of a number of weaknesses in the database, the ANN modeling approach, combined with the use of remote sensing and GIS spatial data, yields a reasonable accuracy for the landslide prediction. In order to obtain a higher prediction accuracy it is recommended to use a suitable dataset of landslide data.

Acknowledgment

First author would like to thank the Alexander von Humboldt Foundation, Germany, for awarding a visiting scientist position at Dresden University of Technology, Germany. This article is greatly benefited from very helpful reviews by Candan Gokceoglu, five anonymous reviewers and editorial comments by G. Iovine.

References

Akgün, A., Dag, S., Bulut, F., 2008. Landslide susceptibility mapping for a landslide-prone area (Findikli, NE of Turkey) by likelihood-frequency ratio and weighted linear combination models. Environmental Geology 54, 1127–1143.

Akgün, A., Bulut, F., 2007. GIS-based landslide susceptibility for Arsin-Yomra (Trabzon, North Turkey) region. Environmental Geology 51, 1377–1387.

Arora, M.K., Gupta, A.S.D., Gupta, R.P., 2004. An artificial neural network approach for landslide hazard zonation in the Bhagirathi (Ganga) Valley, Himalayas. International Journal of Remote Sensing 25 (3), 559–572.

Atkinson, P.M., Tatnall, A.R.L., 1997. Neural networks in remote sensing. International Journal of Remote Sensing 18, 699–709.

Ayalew, L., Yamagishi, H., 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology 65, 15–31.

Basheer, I.A., Hajmeer, M., 2000. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 43, 3–31.

Begueria, S., 2006. Validation and evaluation of predictive models in hazard assessment and risk management. Natural Hazards 37, 315–329.

Caniani, D., Pascale, S., Sdao, F., Sole, A., 2008. Neural networks and landslide susceptibility: a case study of the urban area of Potenza. Natural Hazards 45, 55–72.

Catani, F., Casagli, N., Ermini, L., Righini, G., Menduni, G., 2005. Landslide hazard and risk mapping at catchment scale in the Arno River Basin. Landslides 2 (4), 329–343.

Choi, J., Oh, H.J., Won, J.S., Lee, S., 2009. Validation of artificial neural network model for landslide susceptibility mapping. Environmental Earth Sciences. doi:10.1007/s12665-009-0188-0 (On-line first).

Chung, C.F., Fabbri, A.G., 1999. Probabilistic prediction models for landslide hazard mapping. Photogrammetric Engineering and Remote Sensing 65 (12), 1389–1399.

Chung, C.F., Fabbri, A.G., 2003. Validation of spatial prediction models for landslide hazard mapping. Natural Hazards 30 (3), 451–472.
Dahal, R.K., Hasegawa, S., Nonomura, S., Yamanaka, M., Masuda, T., Nishino, K., 2008. GIS-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping. Environmental Geology 54, 311–324.
Ercanoglu, M., 2005. Landslide susceptibility assessment of SE Bartin (West Black Sea region, Turkey) by artificial neural networks. Natural Hazards and Earth System Sciences 5, 979–992.
Ercanoglu, M., Gokceoglu, C., 2002. Assessment of landslide susceptibility for a landslide-prone area (north of Yenice, NW Turkye) by fuzzy approach. Environmental Geology 41, 720–730.
Ermini, L., Catani, F., Casagli, N., 2005. Artificial neural networks applied to landslide susceptibility assessment. Geomorphology 66 (1–4), 327–343.
Godt, J.W., Baum, R.L., Savage, W.Z., Salciarini, D., Schulz, W.H., Harp, E.L., 2008. Transient deterministic shallow landslide modeling: requirements for susceptibility and hazard assessments in a GIS framework. Engineering Geology 102, 214–226.
Gokceoglu, C., Sonmez, H., Ercanoglu, M., 2000. Discontinuity controlled probabilistic slope failure risk maps of the Altindag (settlement) region in Turkey. Engineering Geology 55, 277–296.
Gomez, H.T., Kavzoglu, T., 2005. Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. Engineering Geology 78 (1-2), 11–27.
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, 323–351.
Guzzetti, F., Carrarra, A., Cardinali, M., Reichenbach, P., 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study. Central Italy. Geomorphology 31, 181–216.
Hines, J.W., 1997. Fuzzy and Neural Approaches in Engineering. Wiley, New York. 210 pp.
Iovine, G., Di Gregorio, S., Lupiano, V., 2003a. Assessing debris-flow susceptibility through cellular automata modelling: an example from the May 1998 disaster at Pizzo d'Alvano (Campania, southern Italy). In: Rickenmann, D., Chen, C.L. (Eds.), Debris-flow. Hazard Mitigation: Mechanics, Prediction and Assessment, Proc. 3rd DFHM Int. Conference, Davos, Switzerland, September 2003, vol. 1. Millpress Science Publishers, Rotterdam, pp. 623–634.
Iovine, G., Di Gregorio, S., Lupiano, V., 2003b. Debris-flow susceptibility assessment through cellular automata modelling: an example from 15–16 December 1999 disaster at Cervinara and San Martino Valle Caudina (Campania, southern Italy). Natural Hazards and Earth System Sciences 3, 457–468.
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, 347–366.
Lamelas, M.T., Marinoni, O., Hoppe, A., Riva, J., 2008. Doline probability map using logistic regression and GIS technology in the central Ebro Basin (Spain). Environmental Geology 54, 963–977.
Lee, S., 2007. Application and verification of fuzzy algebraic operators to landslide susceptibility mapping. Environmental Geology 52, 615–623.
Lee, S., Chwae, U., Min, K., 2002a. Landslide susceptibility mapping by correlation between topography and geological structure: the Janghung area, Korea. Geomorphology 46, 149–162.
Lee, S., Choi, J., Min, K., 2002b. Landslide susceptibility analysis and verification using the Bayesian probability model. Environmental Geology 43, 120–131.
Lee, S., Dan, N.T., 2005. Probabilistic landslide susceptibility mapping in the Lai Chau province of Vietnam: focus on the relationship between tectonic fractures and landslides. Environmental Geology 48, 778–787.
Lee, S., Pradhan, B., 2006. Probabilistic Landslide risk mapping at Penang Island, Malaysia. Journal of Earth System Science 115 (6), 661–672.
Lee, S., Pradhan, B., 2007. Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. Landslides 4, 33–41.
Lee, S., Ryu, J.H., Min, K., Won, J.S., 2003a. Landslide susceptibility analysis using GIS and artificial neural network. Earth Surface Processes and Landforms 27, 1361–1376.
Lee, S., Ryu, J.H., Lee, M.J., Won, J.S., 2003b. Landslide susceptibility analysis using artificial neural network at Boun, Korea. Environmental Geology 44, 820–833.
Lee, S., Ryu, J.H., Won, J.S., Park, H.J., 2004. Determination and application of the weights for landslide susceptibility mapping using an artificial neural network. Engineering Geology 71, 289–302.
Lee, S., Sambath, T., 2006. Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. Environmental Geology 50, 847–855.
Lee, S., Talib, J.A., 2005. Probabilistic landslide susceptibility and factor effect analysis. Environmental Geology 47, 982–990.
Lui, Y., Guo, H.C., Zou, R., Wang, L.J., 2006. Neural network modelling for regional hazard assessment of debris flow in Lake Qionghai Watershed, China. Environmental Geology 49, 968–976.
Melchiorre, C., Matteucci, M., Azzoni, A., Zanchi, A., 2008. Artificial neural networks and cluster analysis in landslide susceptibility zonation. Geomorphology 94, 379–400.
Neaupane, K.M., Achet, S.H., 2004. Use of backpropagation neural network for landslide monitoring: a case study in the higher Himalaya. Engineering Geology 74 (3–4), 213–226.
Nefeslioglu, H.A., Gokceoglu, C., Sonmez, H., 2008. An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps. Engineering Geology 97, 171–191.
Pradhan, B., Lee, S., 2007. Utilization of optical remote sensing data and GIS tools for regional landslide hazard analysis by using an artificial neural network model. Earth Science Frontier 14 (6), 143–152.
Pradhan, B., Lee, S., 2009a. Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network model. Environmental Earth Sciences. doi:10.1007/s12665-009-0245-8 (on-line first).
Pradhan, B., Lee, S., 2009b. Landslide risk analysis using artificial neural network model focusing on different training sites. International Journal of Physical Sciences 4 (1), 1–15.
Pradhan, B., Lee, S., 2009c. Regional landslide susceptibility analysis using backpropagation neural network model at Cameron Highland, Malaysia. Landslides. doi:10.1007/s10346-009-0183-2 (on-line first).
Pradhan, B., Lee, S., Buchroithner, M.F., 2009. Use of geospatial data for the development of fuzzy algebraic operators to landslide hazard mapping: a case study in Malaysia. Applied Geomatics 1, 3–15.
Pradhan, B., Lee, S., Mansor, S., Buchroithner, M.F., Jallaluddin, N., 2008. Utilization of optical remote sensing data and geographic information system tools for regional landslide hazard analysis by using binomial logistic regression model. Journal of Applied Remote Sensing 2 (1), 1–11.
Pradhan, B., Singh, R.P., Buchroithner, M.F., 2006. Estimation of stress and its use in evaluation of landslide prone regions using remote sensing data. Advances in Space Research 37, 698–709.
Pradhan, B., Youssef, A.M., 2009. Manifestation of remote sensing data and GIS on landslide hazard analysis using spatial-based statistical models. Arabian Journal of Geosciences. doi:10.1007/s12517-009-0089-2 (on-line first).
Süzen, M.L., Doyuran, V., 2004. A comparison of the GIS based landslide susceptibility assessment methods: multivariate versus bivariate. Environmental Geology 45, 665–679.
Swingler, K., 1996. Applying Neural Networks: a Practical Guide. Academic Press, New York.
Tunusluoglu, M.C., Gokceoglu, C., Nefeslioglu, H.A., Sonmez, H., 2008. Extraction of potential debris source areas by logistic regression technique: a case study from Barla, Besparmak and Kapi mountains (NW Taurids, Turkey). Environmental Geology 54, 9–22.
Turban, E., Aronson, J.E., 2001. Decision Support Systems and Intelligent Systems, sixth ed. Prentice International Hall, Hong Kong.
USGS, 1993. USCS data users guide 5 for DEM's, ftp://mapping.usgs.gov/pub/ti/DEM/demguide (connected: 07.08.2008)
Youssef, A.M., Pradhan, B., Gaber, A.F.D., Buchroithner, M.F., 2009. Geomorphological hazard analysis along the Egyptian red sea coast between Safaga and Quseir. Natural Hazards and Earth System Sciences 9, 751–766.
Zerger, A., 2002. Examining GIS decision utility for natural hazard risk modeling. Environmental Modeling and Software 17, 287–294.
Zhou, W., 1999. Verification of the nonparametric characteristics of backpropagation neural networks for image classification. IEEE Transactions on Geoscience and Remote Sensing 37, 771–779.