Research Article

Assessment of Landslide Susceptibility by Decision Trees in the Metropolitan Area of Istanbul, Turkey

H. A. Nefeslioglu,[1] E. Sezer,[2] C. Gokceoglu,[3] A. S. Bozkir,[2] and T. Y. Duman[1]

[1] Department of Geological Research, General Directorate of Mineral Research and Exploration, 06520 Ankara, Turkey

[2] Department of Computer Engineering, Hacettepe University, Beytepe, 06800 Ankara, Turkey

[3] Department of Geological Engineering, Hacettepe University, Beytepe, 06800 Ankara, Turkey

Correspondence should be addressed to C. Gokceoglu, cgokce@hacettepe.edu.tr

The main purpose of the present study is to investigate the possible application of decision tree in landslide susceptibility assessment. The study area having a surface area of $174.8\,km^2$ locates at the northern coast of the Sea of Marmara and western part of Istanbul metropolitan area. When applying data mining and extracting decision tree, geological formations, altitude, slope, plan curvature, profile curvature, heat load and stream power index parameters are taken into consideration as landslide conditioning factors. Using the predicted values, the landslide susceptibility map of the study area is produced. The AUC value of the produced landslide susceptibility map has been obtained as 89.6%. According to the results of the AUC evaluation, the produced map has exhibited a good enough performance.

1. Introduction

The landslide susceptibility and hazard assessments can be carried out either by using direct mapping techniques or by using indirect mapping techniques. Direct hazard assessment, in which the degree of hazard is determined by the mapping geomorphologist, based on his/her experience and knowledge of the terrain conditions [39]. However, indirect hazard assessment, in which either statistical models or deterministic models are used to predict landslide prone areas, based on information obtained from the interrelation between landscape factors and the landslide distribution [39]. In the recent years, many studies on the indirect landslide susceptibility assessment have been published depending on the developments of Geographical Information Systems (GISs) and digital cartography. It is

possible to produce a landslide susceptibility map employing various indirect mapping techniques such as combination of index maps (i.e., [1, 3, 5, 26]), bivariate and multivariate statistical analyses (i.e., [4, 6–15, 17–19, 21, 25, 46]), neural networks (i.e., [2, 19, 23, 24, 27, 32, 43]) and fuzzy approach (i.e., [28–31, 37]). It is evident that each landslide susceptibility assessment method considered by landslide community has some advantages and drawbacks.

According to Miller and Han [49], similar to many research and application fields, geography has moved from a data-poor and computation-poor to data-rich and computation-rich environment. The scope, coverage, and the volume of digital geographic datasets are growing rapidly. Moreover, new high spatial and spectral resolution remote sensing systems and other monitoring devices are gathering vast amounts of georeferenced digital imagery, video, and sound. Traditional spatial analytical techniques cannot easily discover new and unexpected patterns, trends, and relationships that can be hidden deep within very large diverse geographic datasets [49]. At this point, data mining which encompasses variety of statistical analysis, pattern recognition and machine learning techniques can be used to overcome the problem when processing very large datasets. When producing landslide susceptibility maps, very large datasets should be processed. As mentioned above, landslide susceptibility maps have been produced by using several methods. The investigation of the decision tree in landslide susceptibility assessment constitutes the main purpose of this study. For the purpose of the study, decision tree is used to analyze the geographic data to predict landslide susceptible zones on map for the study area, Cekmece area of Istanbul, Turkey (Figure 1). The study also includes assessment of the landslide conditioning factors, some the oretical information on the decision tree technique, and application of the decision trees to landslide susceptibility assessment.

2. General Characteristics of the Study Area

The study area having a surface area of $174.8\,km^2$ locates at the northern coast of the Sea of Marmara and western part of Istanbul metropolitan area (see Figure 1). The Buyuk Cekmece Lake, and the Kucuk Cekmece Lake and Dikilitas creek are the western and eastern borders of the study area, respectively. The area has a high seismicity. In the last decade, Turkey has experienced some large earthquakes. More than 300 earthquakes in the region have been reported to have occurred between 2100 BC and AD 1900 [55]. The active northern branch of the North Anatolian Fault Zone (NAFZ) passes through approximately 9 km from south of the application site. In the last 20 centuries, between Izmit and Gulf of Saros, 29 historically large (between 6.3 and 7.4 Ms) earthquakes occurred along the northern branch of the NAFZ [35]. Actually, it may be effortlessly remarked that there are two main landslide triggers, one of which is earthquake and the other is heavy rainfall. If the characteristics of the region are taken into consideration, it is possible to say that the landslide triggers exist in the application area.

In the study area, summers are hot and slightly rainy while winters are warm and rainy. The topography of the region and presence of lakes and dams also affect the weather conditions (http://istanbul.meteor.gov.tr/). The region receives 85% of the total annual precipitation in the period of September-May (http://istanbul.meteor.gov.tr/). In this study, the data of Florya Meteorology Station, the nearest station to the application site, was assessed. According to the meteorological data of the period of 1937–1990, the average monthly rainfall varies between 20.5 mm and 102.0 mm. The annual precipitation
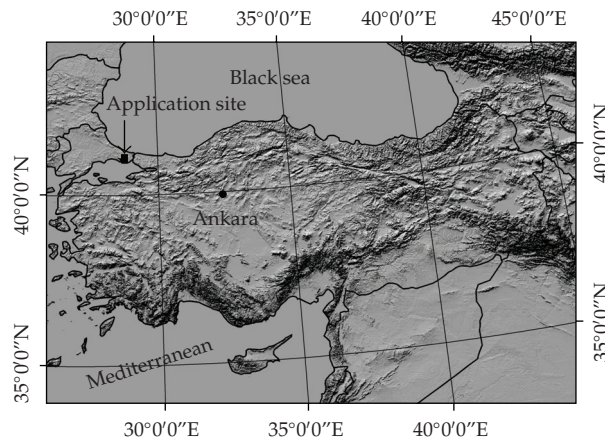
Figure 1: Location map of the study area; rectangle in black pointed out by the arrow in the figure covers the study area.

varies between 500 mm and 1000 mm in the region while average annual precipitation of long period of the Florya Meteorology Station is 642.4 mm. The maximum daily precipitation was observed in Istanbul as 88.9 kg/m^2 (http://www.meteor.gov.tr/). As will be explained in the latter sections, the precipitation is highly effective on the landslide occurrence in the study area.

Various lithological units from Middle-Late Eocene to Quaternary crop out in the region. The 1/25000 scaled geological map of the study area was prepared by Duman et al. [36]. Areal distributions of the geological formations of the study area are given in Table 1 and Figure 2. The Kirklareli limestone is the oldest rocks of the study area from the Middle-Late Eocene while the youngest unit in the study area is actual alluviums. In the area, some inactive normal faults are typical. The dip values of the beddings of the sedimentary units in the area are rather low, 5–15 degrees. For this reason, the strikes and the dip directions exhibit a high variation in short distances. However, there is no considerable folding in the study area [17].

The altitude values in the study area are between 0 and 200 m while the dominant altitude ranges are 75–100 and 100–125 m (Table 2). The study area has a dendritic drainage pattern, because of presence of soft lithologies and low slope angles. The general physiographic trend of the area is NW-SE. The range of the slope gradient values is from 0 to 57 degrees. Finally, these slope gradient values indicate that the majority of the region has gentle slopes.

3. Assessment of Landslide Conditioning Factors

In this section, the landslide conditioning factors observed in the study area are explained. Before the explanations, the data used is given. In the present study, the digital elevation model (DEM) produced by Duman et al. [17] is used. The DEM was prepared by digitizing 10 m altitude contours of the 1/25,000 scale topographical map [17]. Maps of slope gradient, heat load, altitude, stream power index, plan curvature, and profile curvature are produced using the DEM in raster format with a pixel size of 25 × 25 m^2. The lithology map taken into consideration in this study was prepared by Duman et al. [36]. This vector map
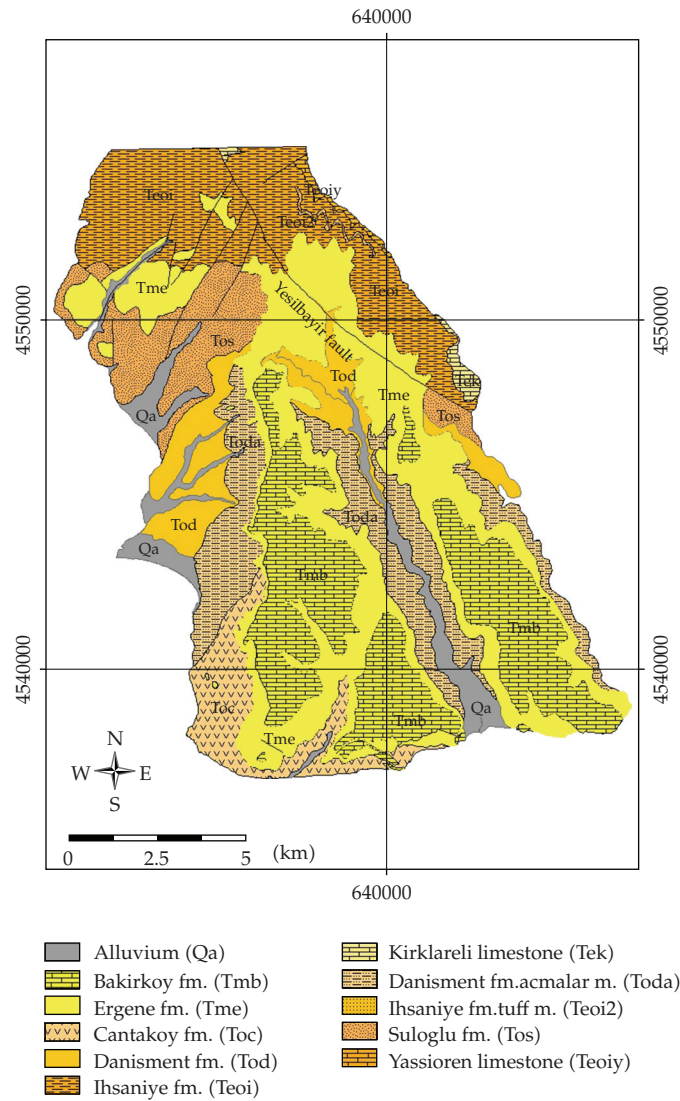
Figure 2: Geological map of the study area [36].

was converted to a raster map with a pixel size of $25 \times 25\,\text{m}^2$ by Duman et al. [17]. A reliable landslide inventory defining the type and activity of all landslides, as well as their spatial distribution, is essential before any analysis of the occurrence of landslides and their relationship to environmental conditions undertaken [34]. Therefore, it is possible to say that a reliable landslide inventory is a crucial part of a landslide susceptibility map among the parameters employed, because it is the fundamental component of the assessments. In Turkey, a landslide inventory project at national scale has been conducted by the Geological Research Department of the General Directorate of Mineral Research and Exploration (MTA). When preparing the landslide inventory, to identify the landslides, the vertical black-and-white aerial photographs of medium scale (1 : 35,000), dated 1955–1956, were used [17]. When describing the type and activity of the landslides in the project [38], mass movements

Table 1: The distribution of the geological formations with respect to landslide in the study area.

Formation	Symbol	Grid cells with landslides		All grid cells		LandslideDensity (%)
		Frequency	%	Frequency	%	
Quaternary	Qa	1127	2.1	15495	5.54	7.27
Bakirkoy fm.	Tmb	4473	8.34	56231	20.1	7.95
Ergene fm.	Tme	13617	25.37	71287	25.49	19.1
Cantakoy fm.	Toc	5706	10.63	14210	5.08	40.15
Danisment fm.	Tod	4728	8.81	20650	7.38	22.9
Ihsaniye fm.	Teoi	1810	3.37	43337	15.49	4.18
Kirklareli limestone	Tek	0	0	1613	0.58	0
Danisment fm. Acmalar m.	Toda	17151	31.95	34213	12.23	50.13
Ihsaniye fm. Tuff m.	Teoi2	66	0.12	478	0.17	13.81
Suloglu fm.	Tos	4996	9.31	21166	7.57	23.6
Yassioren limestone	Teoiy	0	0	1035	0.37	0

were classified according to the cinematic types of classification proposed by Varnes [16], that is, flows, falls, and slides. The landslides are also classified according to their relative depths, as shallow (depth < 5 m) and deep seated (depth > 5 m). Landslide activities were classified into two groups as active and inactive by Duman et al. [38]. Active landslides are defined as those currently moving; whereas inactive ones are as relict according to Working Party on World Landslide Inventory WP/WLI [40]. Shallow landslides are classified as active because of their ongoing observed movements [41].

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 Duman et al. [17] considering the logistic regression technique. Duman et al. [17] have employed geological formations, geomorphological units, and relative permeability of different lithological units, slope gradient, slope aspect, altitude, plan and profile curvatures, and stream power index parameters as the landslide conditioning factors. However, in the present study, some parameters such as geomorphological units and relative permeability of the lithological units have been eliminated. Furthermore, several topographic parameters used in the study reflect geomorphological characteristics of the study area. To abstain a redundancy, the geomorphologic units are eliminated from the model. The other eliminated parameter used by Duman et al. [17] is the relative permeability of the lithological units. Determination of this parameter is too difficult, sometimes impossible. As a result, the relative permeability is also excluded from the model. One of the topographic parameters used by Duman et al. [17] is the slope aspect. Aspect of a slope is the direction or azimuth of a slope faces. It strongly influences potential direct incident radiation and temperature. Untransformed slope aspect is a poor variable for quantitative analysis, since 1 degree is adjacent to 360 degrees; the numeric values are very different even though the slope aspect is about the same. Hence, slope aspect values need to be transformed in one of several ways, depending on the precision with which it was measured and the environmental factor(s) the analyst would like it to represent [42]. In this study, the heat load index [42] has been used instead of the slope aspect. According to the descriptive statistics of the parameter, the mean heat load index value on the grid cells with landslides is observed as 0.529 (±0.338) (Table 2). The other conditioning parameters

Table 2: General descriptive statistics of topographical variables with respect to landslides.

Data	Variable	Min.	Max.	Mean	Std. deviation
Grid cells with landslides ($N = 53,674$)	Altitude (m)	0.000	194.680	85.009	45.165
	Slope gradient (°)	0.000	57.950	7.966	5.307
	Plan curvature	−3.130	2.870	−0.013	0.238
	Profile curvature	−2.930	3.080	0.023	0.303
	Heat load	0.000	1.000	0.529	0.338
	Stream power index *(SPI)*	0.000	8.330	0.724	0.934
Grid cells without landslides ($N = 226,041$)	Altitude (m)	0.000	200.000	98.229	52.370
	Slope gradient (°)	0.000	55.230	4.642	3.924
	Plan curvature	−4.500	2.690	0.004	0.137
	Profile curvature	−2.920	3.430	−0.004	0.188
	Heat load	0.000	1.000	0.524	0.342
	Stream power index *(SPI)*	0.000	8.670	0.453	0.775

used in the study are slope gradient, stream power index, plan and profile curvatures, and altitude. These parameters can also be obtained easily from the digital elevation model of the study area.

The characters of landslides identified in the region are mainly deep seated and active. They are generally located in the lithologies including the permeable sandstone layers and impermeable layers such as claystone, siltstone, and mudstone layers. This is typical for the landslides identified in the study area. When considering this finding, it may be said that one of the main conditioning factors of the landslides in the study area is lithology [17]. This can be seen clearly in Table 1. As can be seen in Table 1, the majority of the landslides (approximately 60%) occurred in two formations, namely, Danisment formation-Acmalar member (Toda) and Ergene (Tme) formation. According to Duman et al. [17], another factor governing the landslides in the study area is the sandstone bedding planes. If the orientation of slope and bedding plane is roughly similar, some large landslides can occur. In these areas, initiation of the landslides is controlled by the bedding planes as planar failure, and then in the displaced and accumulated material, some rotational landslides are observed [17]. Rarely, in this material, some earth flows may occur depending on the heavy precipitation [17]. Besides, permeability differences between sandstones and claystones are highly important on landslide occurrence in the study area. The sandstones at upper levels are saturated by the surface waters because the surface waters do not infiltrate to claystones at lower levels.

One of the most important topographical factors conditioning landslides is the slope gradient. In the regional landslide susceptibility or hazard assessments, several researchers (i.e., [7, 8, 33, 44, 45]) took into consideration statistical techniques for the assessment of slope gradient in terms of landslide activity. In the present study, the slope gradient is considered as a conditioning factor during the analyses. Descriptive statistics revealed that the landslides in the application site typically occur on the gentle slopes (Table 2). Some authors (i.e., [26,

30]) pointed out that the altitude is a good indicator for the landslide susceptibility. For this reason, the altitude has been accepted as a conditioning factor, as well. The mean altitude value on the grid cells with landslides is observed as 85.1 m (±45.2 m) (Table 2).

The term curvature is generally defined as the curvature of a line formed by intersection of a random plane with the terrain surface [40]. The influence of plan curvature on the land degradation processes is the convergence or divergence of water during downhill flow. In addition, this parameter constitutes one of the main factors controlling the geometry of the terrain surface where landslides occur [47]. In this study, profile and plan curvature values have been calculated using a script, namely, Digital Elevation Model Analysis Tool (DEMAT) compiled in the Avenue soft computing language of ArcView GIS by Behrens [48] and these parameters have been considered as the conditioning factors. While the positive values of slope curvature define the convexity, the negative ones present the concavity of the terrain surface. The minimum and the maximum profile curvature values were calculated as −2.930 and 3.080 on the grid cells with landslides, and −2.920 and 3.430 on the grid cells without landslides (Table 2). These ranges have been obtained for the plan curvature values on the grid cells with landslides and without landslides as −3.130 to 2.870 and −4.500 to 2.690, respectively (Table 2).

The last parameter considered in the present study is stream power index (SPI). It is a measure of erosive power of water flow based on the assumption that discharge (q) is proportional to specific catchment area (A_s) [20] (3.1). Although, the minimum and the maximum ranges of the SPI values on the grid cells with landslides and without landslides are observed as close (Table 2), the mean SPI value for the grid cells with landslides was calculated more than that for the grid cells without landslides (Table 2). Theoretically, the maximum SPI values are calculated in drainage channels. Therefore, the values of SPI on the grids cells with landslides should not be expected as high. This hypothetical expectation is supported by the observation of the maximum SPI value on the grid cells without landslides (Table 2)

$$\text{SPI} = A_s \tan \beta, \tag{3.1}$$

where A_s is the specific catchment area (m^2m^{-1}) while β is the slope gradient in degree.

4. Modeling Approach

Data mining involves various techniques such as statistics, neural networks, decision tree, genetic algorithm, and visualization techniques that have been developed over the years. Data mining problems are generally categorized as association, clustering, classification, and prediction [50]. Classification refers to finding rules to assign data items into pre-existing classes [49]. Association analysis is used to discover patterns that describe strongly associated features in the data. On the other hand, aim of clustering is to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters [51].

In practice, there are several data mining tools such as Oracle DM, SQL Server Analysis Services, SPSS Clementine, and SAS Enterprise Miner for commercial use. In the present study, the decision tree technique is used to predict the landslide susceptibility classes by employing Microsoft Server 2008 Analysis Services. Decision tree is a data mining approach that is often used for classification and prediction. Although other methodologies such as neural network can also be used for classification, decision tree has the advantages of easy
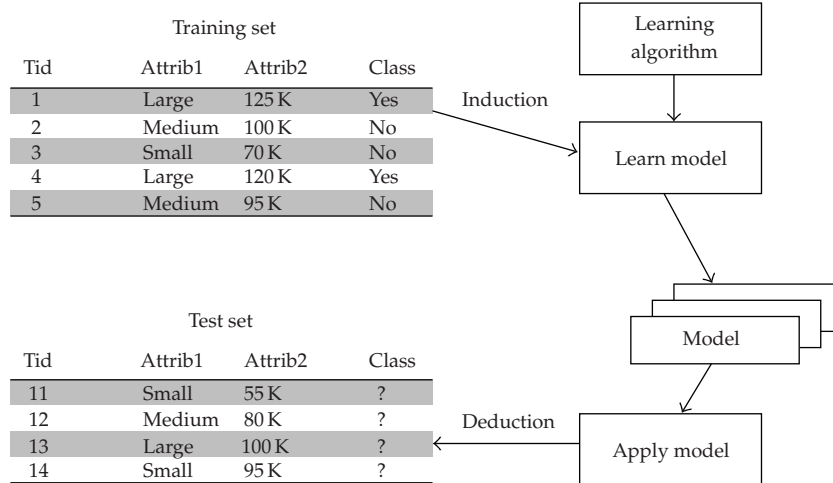
Figure 3: Schematic illustration of the construction of decision tree by using training data set and an example view of a prediction on test data [52].

interpretation and understanding for the decision makers to compare with their domain knowledge for validation and justify their decisions [50]. In addition, there are a few advantages of using decision trees over using other data mining algorithms, for instances, decision trees are quick to build and easy to interpret and prediction based on decision trees is efficient [22].

Decision trees are built through recursive data partitioning, where in each iteration the data is split according to the values of a selected attribute. The recursion stops at "pure" data subsets which only include instances of the same class [53]. In other words, the principle idea of a decision tree is to split your data recursively into subsets so that each subset contains more or less homogeneous states of your target variable (predictable attribute). At each split in the tree, all input attributes are evaluated for their impact on the predictable attribute. When this recursive process is completed, a decision tree is formed [22]. If the predictable target attribute consists of discrete data, the formed decision tree model is called a classification tree. However, if the target attribute is a continuous variable, then the model is called as a regression tree. The process of decision tree building is sometimes called decision tree induction. Many techniques have been developed for decision tree induction up to now. Nevertheless, the general idea of decision tree induction is same on every type of decision tree methods. Each technique employs a learning algorithm to identify a model that best fits the relationships between the attribute set and class label of input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before [51]. An example showing the generation of a decision tree by employing train data and the prediction ability by using the test data is given in Figure 3.

ID3 is a well-known decision tree algorithm proposed by Ross Quinlan of the University of Sydney, Australia. ID3 tree was later enhanced to be C4.5. C4.5 can handle numeric attributes, missing values, and noisy data. Some decision trees can perform regression tasks, for example, to predict continuous variables such as temperature and humidity. The Classification and Regression Tree (CART) proposed by Briemann is a popular decision tree algorithm for classification and regression [22].

5. Landslide Susceptibility Mapping and Results

In order to perform the research reported in the present paper, Microsoft SQL Server 2008 Analysis Services software is chosen as the analyzing platform as it supports decision trees with continuous variables (called as regression trees). High scalability and having support for nested table, automatic feature selection, automatic cardinality reduction features of it are the other reasons for choosing this data mining platform. Additionally, Microsoft Analysis Services allows building data mining applications via the support of Microsoft Visual Studio Integrated Development Environment and ADOMD extensions [22]. Since the purpose of this data-mining study is to develop a model for predicting the landslide susceptible areas, decision tree, a well-known classification technique, has been utilized during the analysis. Microsoft SQL Server Analysis Services employ their own decision tree algorithm that is called Microsoft Decision Trees. This commercial algorithm can handle both discrete and continuous valued variables/attributes and presents useful parameters for configuring tree induction step such as tree splitting options. Another reason of selecting this algorithm and software is that 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).

In this study, all of the input variables and target output variable are continuous, so resulted tree is a special version of decision tree named regression tree. Regression is similar to classification. The only difference is that regression predicts continuous attributes. Although the basic task of a decision tree algorithm is classification, it can be used for regression as well. Another well-known regression tree algorithm is CART. The Microsoft Decision Trees algorithm adds the support for regression in SQL Server 2005 and 2008. Microsoft Regression Trees contain a linear regression formula at each leaf node. Using a regression tree has its advantages over simple linear regression in that a tree can represent both linear and nonlinear relationships [22].

The data understanding and the data preparation stages are among the most important steps in the data mining applications [54]. At the start-up of this study, entire dataset is converted to Access 2003 format from SPSS format to be used in Analysis Services easily. The dataset used in the analysis consists of 17 input attributes and 1 predictable attribute (landslide; 1 = grid cells with landslide, 0 = grid cells without landslide) and 280,132 records/cases (Table 3). 226,041 cases belong to pixel without landslide. However, 53,674 cases stand for pixel with landslide and 417 missing cases exist. Although predictable attribute (landslide information) is a discrete valued attribute, it is handled as continious attribute to obtain more accurate results such as in the range of 0-1 instead of 0 = false and 1 = true. After data preparation stage, "Microsoft Decision Trees" algorithm is run on training dataset by separating whole dataset to 85% training and 15% test cases of whole dataset. By using the trained model, the landslide susceptibility values of each pixel on the map are determined via prediction tool of software and the landslide susceptibility map is produced with the help of predicted numerical values of each pixel on the map. One of the useful features of "Microsoft Decision Trees" algorithm is building and reporting dependency network graphs. The dependency network graphs display the relationships among attributes derived from decision tree model's content [22]. An example view of decision tree that is derived from our model is presented in Figure 4.

Using the predicted landslide susceptibility values, the landslide susceptibility map of the study area is produced (Figure 5). On the map the reddish tones (close to 1) mean high susceptible areas while the green parts (close to 0) represent more stable zones in the
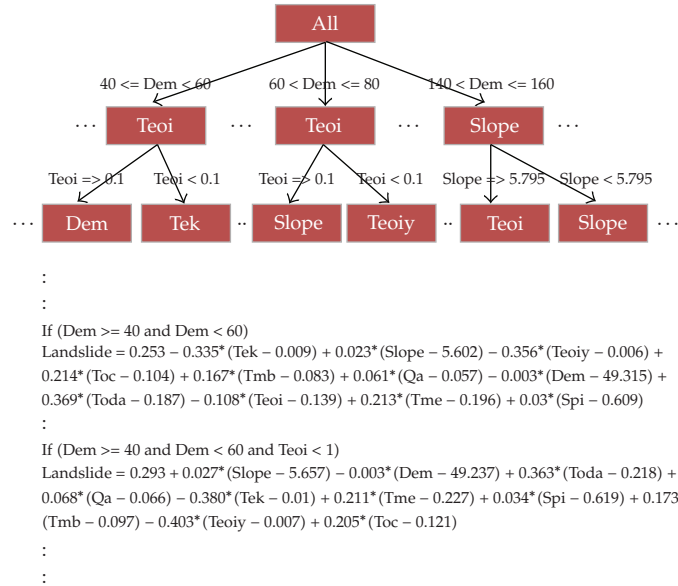
Figure 4: An example view of a part of the decision tree.

Table 3: The attributes considered in the study and the effect importance order on the predictable variable.

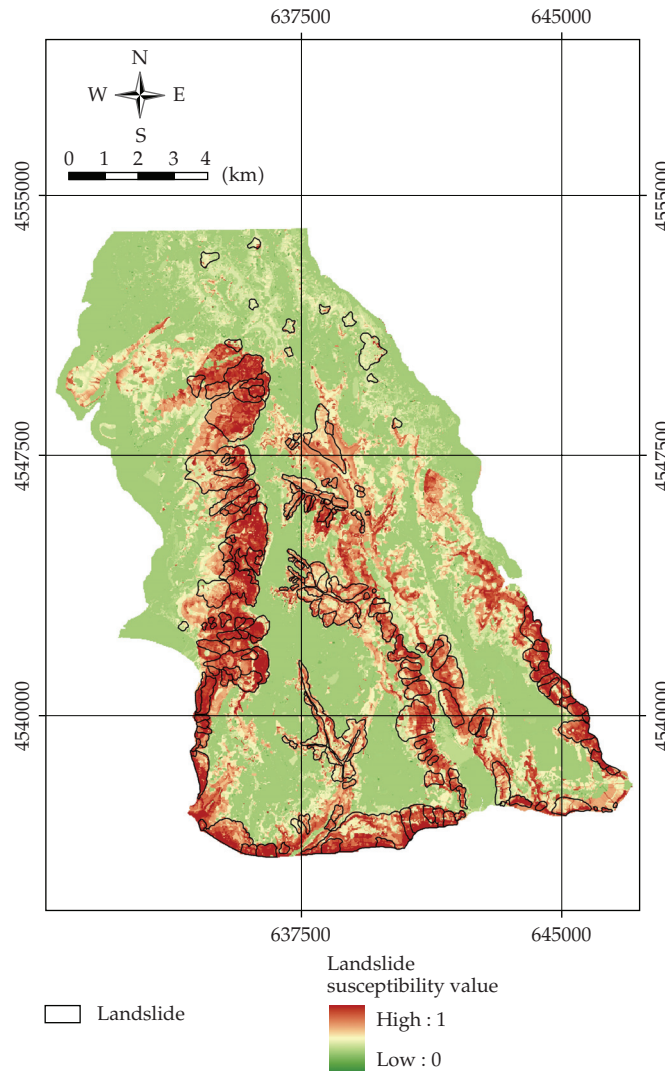| Attribute	Continues/discrete	Usage	Effect importance order on the output
Altitude	Continuous	Input	12
Heat load index	″	″	14
Plan curvature	″	″	15
Profile curvature	″	″	5
Stream power index	″	″	**3**
Slope gradient (°)	″	″	**4**
Alluvium (Qa)	Discrete in nature handled Continuous	″	10
Kirklareli limestone (Tek)	″	″	**2**
Ihsaniye fm. (Teoi)	″	″	6
Ihsaniye fm. Tuff m. (Teoi2)	″	″	13
Yassioren limestone (Teoiy)	″	″	17
Bakirkoy fm. (Tmb)	″	″	16
Ergene (Tme)	″	″	8
Cantakoy fm. (Toc)	″	″	9
Danisment fm. (Tod)	″	″	11
Danisment fm. Acmalar m. (Toda)	″	″	**1**
Suloglu fm. (Tos)	″	″	7
Landslide	″	Output	

Figure 5: Landslide susceptibility map of the study area produced by using the decision tree technique.

region. The dependent variable is controlled by coefficients of the independent variables in the equations. The effect importance order is obtained from the dependency network graphs by considering the coefficients of the independent variables in the model. According to the dependency network graph, the most effective parameters are the geological formations Danisment fm.-Acmalar m. (Toda) and Kirklareli limestone (Tek), SPI, slope gradient, and profile curvature (Table 3). This result is in a good accordance with the landslide density assessment of the lithological units, because the majority of the landslides occurred in these lithologies. After production of the landslide susceptibility map, the performance of the map is also assessed. To assess the spatial effectiveness of the susceptibility maps using a threshold-independent method, ROC curve was drawn in the present study (Figure 6). Moreover, plotting different accuracy values obtained against the whole range of possible threshold values constitutes the essential of a ROC curve evaluation [51]. The area under
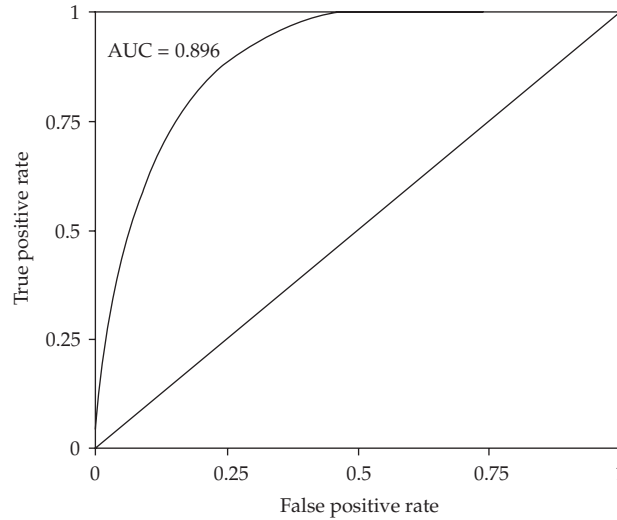
Figure 6: ROC (Receiver-Operating Characteristic) curve evaluation of the constructed model.

ROC curve (AUC) constitutes one of the most common used accuracy statistics for the prediction models in natural hazard assessments [56]. The minimum value of AUC is 0.5 means no improvement over random assignment while the maximum value of that is 1 denotes perfect discrimination. Finally, according to the results of the AUC evaluation (Figure 6), the obtained map exhibited a good performance.

6. Conclusions

Duman et al. [17] produced previously the landslide susceptibility map of the same area by logistic regression analysis. The model produced by Duman et al. [17] is used to generate a landslide susceptibility map that correctly classified 83.8% of the landslide-prone areas. This percentage (83.8%) belongs to a correct classification of the landslide information of being only "one" (landslide presence). When the information regarding being "zero" (landslide absence) is also considered, the overall correct classification value becomes 76.0% [17]. In the present study, to check the performance of the produced map, the ROC curve is drawn and the value of AUC is obtained as 89.6%. This result reveals that the performance of the produced map in this study is obviously higher than that of the map produced by Duman et al. [17]. As a result of the application of decision tree method, two geological formations (Danisment fm.-Acmalar m., Toda; Kirklareli limestone, Tek), stream power index, and slope gradient are obtained as the most effective parameters on the landslide occurrence in the study area. If the field conditions are taken into account, this result is physically plausible. It is possible to produce landslide susceptibility maps by various approaches as statistical methods, artificial neural networks, fuzzy approaches, and so forth. The main difference of the decision trees employed in the present study from the other methods can exhibit the order of the conditioning parameters. This situation provides a comparison between analysis results and field observations for the expert. Especially, visual interpretations of the decision trees are powerful tool when compared with the other approaches. Additionally, decision trees applied in the study shows a high prediction capacity. These interpretations reveal that

the decision tree is a useful tool when producing reliable landslide susceptibility maps. The reliability of landslide susceptibility maps is highly important because when landslide hazard and risk maps are produced, the landslide susceptibility maps should be used. The produced map has a sufficient capacity for the medium scaled and regional planning purposes. This means that according to the results of the present study the decision tree is an efficient tool for medium scaled and regional landslide susceptibility analyses.

As a final point, in the present study, decision tree, one of the data mining methods, is investigated to produce landslide susceptibility map of a landslide-prone area (Cekmece, Istanbul, Turkey). By using the developed decision tree model, two important results can be obtained; the model is used to predict the landslide susceptibility degrees and the effect order of input attributes on landslide occurrence is investigated.

References

[1] M. C. Turrini and P. Visintainer, "Proposal of a method to define areas of landslide hazard and application to an area of the Dolomites, Italy," *Engineering Geology*, vol. 50, no. 3-4, pp. 255–265, 1998.

[2] H. A. Nefeslioglu, C. Gokceoglu, and H. Sonmez, "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*, vol. 97, no. 3-4, pp. 171–191, 2008.

[3] C. Gökceoglu and H. Aksoy, "Landslide susceptibility mapping of the slopes in the residual soils of the Mengen region (Turkey) by deterministic stability analyses and image processing techniques," *Engineering Geology*, vol. 44, no. 1–4, pp. 147–161, 1996.

[4] D. Turer, H. A. Nefeslioglu, K. Zorlu, and C. Gokceoglu, "Assessment of geo-environmental problems of the Zonguldak province (NW Turkey)," *Environmental Geology*, vol. 55, no. 5, pp. 1001–1014, 2008.

[5] L. Donati and M. C. Turrini, "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*, vol. 63, no. 3-4, pp. 277–289, 2002.

[6] A. Carrara, M. Cardinali, R. Detti, F. Guzzetti, V. Pasqui, and P. Reichenbach, "GIS techniques and statistical models in evaluating landslide hazard," *Earth Surface Processes & Landforms*, vol. 16, no. 5, pp. 427–445, 1991.

[7] P. M. Atkinson and R. Massari, "Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy," *Computers & Geosciences*, vol. 24, no. 4, pp. 373–385, 1998.

[8] F. Guzzetti, A. Carrara, M. Cardinali, and P. Reichenbach, "Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy," *Geomorphology*, vol. 31, no. 1–4, pp. 181–216, 1999.

[9] C. Baeza and J. Corominas, "Assessment of shallow landslide susceptibility by means of multivariate statistical techniques," *Earth Surface Processes and Landforms*, vol. 26, no. 12, pp. 1251–1263, 2001.

[10] S. Lee and K. Min, "Statistical analysis of landslide susceptibility at Yongin, Korea," *Environmental Geology*, vol. 40, no. 9, pp. 1095–1113, 2001.

[11] A. Clerici, S. Perego, C. Tellini, and P. Vescovi, "A procedure for landslide susceptibility zonation by the conditional analysis method," *Geomorphology*, vol. 48, no. 4, pp. 349–364, 2002.

[12] S. Lee, "Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data," *International Journal of Remote Sensing*, vol. 26, no. 7, pp. 1477–1491, 2005.

[13] T. Can, H. A. Nefeslioglu, C. Gokceoglu, H. Sonmez, and T. Y. Duman, "Susceptibility assessments of shallow earthflows triggered by heavy rainfall at three catchments by logistic regression analyses," *Geomorphology*, vol. 72, no. 1–4, pp. 250–271, 2005.

[14] K. T. Chau and J. E. Chan, "Regional bias of landslide data in generating susceptibility maps using logistic regression: case of Hong Kong Island," *Landslides*, vol. 2, no. 4, pp. 280–290, 2005.

[15] C. Gokceoglu, H. Sonmez, H. A. Nefeslioglu, T. Y. Duman, and T. Can, "The 17 March 2005 Kuzulu landslide (Sivas, Turkey) and landslide-susceptibility map of its near vicinity," *Engineering Geology*, vol. 81, no. 1, pp. 65–83, 2005.

[16] D. J. Varnes, "Slope movement types and processes," in *Landslides Analysis and Control*, R. L. Chuster and R. J. Krizek, Eds., vol. 176, pp. 12–33, Transportation Research Board, National Academy of Sciences, New York, NY, USA, 1978.

[17] T. Y. Duman, T. Can, C. Gokceoglu, H. A. Nefeslioglu, and H. Sonmez, "Application of logistic regression for landslide susceptibility zoning of Cekmece Area, Istanbul, Turkey," *Environmental Geology*, vol. 51, no. 2, pp. 241–256, 2006.

[18] F. Guzzetti, M. Galli, P. Reichenbach, F. Ardizzone, and M. Cardinali, "Landslide hazard assessment in the Collazzone area, Umbria, Central Italy," *Natural Hazards and Earth System Science*, vol. 6, no. 1, pp. 115–131, 2006.

[19] S. Lee and B. Pradhan, "Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models," *Landslides*, vol. 4, no. 1, pp. 33–41, 2007.

[20] I. D. Moore, R. B. Grayson, and A. R. Ladson, "Digital terrain modelling: a review of hydrological, geomorphological, and biological applications," *Hydrological Processes*, vol. 5, no. 1, pp. 3–30, 1991.

[21] T. Gorum, B. Gonencgil, C. Gokceoglu, and H. A. Nefeslioglu, "Implementation of reconstructed geomorphologic units in landslide susceptibility mapping: the Melen Gorge (NW Turkey)," *Natural Hazards*, vol. 46, no. 3, pp. 323–351, 2008.

[22] Z. Tang and J. MacLennan, *Data Mining with Sql Server*, John Wiley & Sons, New York, NY, USA, 2005.

[23] H. Gómez and T. Kavzoglu, "Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela," *Engineering Geology*, vol. 78, no. 1-2, pp. 11–27, 2005.

[24] D. P. Kanungo, M. K. Arora, S. Sarkar, and R. P. Gupta, "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*, vol. 85, no. 3-4, pp. 347–366, 2006.

[25] H. A. Nefeslioglu, T. Y. Duman, and S. Durmaz, "Landslide susceptibility mapping for a part of tectonic Kelkit Valley (Eastern Black Sea region of Turkey)," *Geomorphology*, vol. 94, no. 3-4, pp. 401–418, 2008.

[26] A. K. Pachauri and M. Pant, "Landslide hazard mapping based on geological attributes," *Engineering Geology*, vol. 32, no. 1-2, pp. 81–100, 1992.

[27] B. Pradhan and S. Lee, "Regional landslide susceptibility analysis using backpropagation neural network model at Cameron Highland, Malaysia," *Landslides*. In press.

[28] C. H. Juang, D. H. Lee, and C. Sheu, "Mapping slope failure potential using fuzzy sets," *Journal of Geotechnical Engineering*, vol. 118, no. 3, pp. 475–494, 1992.

[29] E. Binaghi, L. Luzi, P. Madella, F. Pergalani, and A. Rampini, "Slope instability zonation: a comparison between certainty factor and fuzzy Dempster-Shafer approaches," *Natural Hazards*, vol. 17, no. 1, pp. 77–97, 1998.

[30] M. Ercanoglu and C. Gokceoglu, "Assessment of landslide susceptibility for a landslide-prone area (north of Yenice, NW Turkey) by fuzzy approach," *Environmental Geology*, vol. 41, no. 6, pp. 720–730, 2002.

[31] M. Ercanoglu and C. Gokceoglu, "Use of fuzzy relations to produce landslide susceptibility map of a landslide prone area (West Black Sea Region, Turkey)," *Engineering Geology*, vol. 75, no. 3-4, pp. 229–250, 2004.

[32] B. Pradhan and S. Lee, "Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network model," *Environmental Earth Sciences*. In press.

[33] R. J. Maharaj, "Landslide processes and landslide susceptibility analysis from an upland watershed: a case study from St. Andrew, Jamaica, West Indies," *Engineering Geology*, vol. 34, no. 1-2, pp. 53–79, 1993.

[34] R. Soeters and C. J. Van Westen, "Slope instability recognition, analysis and zonation," in *Landslides: Investigation and Mitigation*, A. K. Turner and R. L. Schuster, Eds., vol. 247, pp. 129–177, National Academy Press, Washington, DC, USA, 1996.

[35] N. Ambraseys, "The seismic activity of the Marmara Sea region over the last 2000 years," *Bulletin of the Seismological Society of America*, vol. 92, no. 1, pp. 1–18, 2002.

[36] T. Y. Duman, M. Kecer, S. Ates, et al., "Istanbul Metropolu Batisindaki (Kucukcekmece-Silivri-Catalca yoresi) Kentsel Gelisme Alanlarinin Yer Bilim Verileri," *MTA*, no. 3, p. 249, 2004 (Turkish).

[37] B. Pradhan, S. Lee, and M. F. Buchroithner, "Use of geospatial data for the development of fuzzy algebraic operators to landslide hazard mapping: a case study in Malaysia," *Applied Geomatics*, vol. 1, pp. 3–15, 2009.

[38] T. Y. Duman, Ö. Emre, T. Can, et al., "Turkish landslide inventory mapping project: methodology and results on Zonguldak quadrangle (1/500000), working in progress 25 on the geology of turkey and its surroundings," in *Proceedings of the 4th International Turkish Geology Symposium (ITGS '01)*, p. 392, Adana, Turkey, September 2001.

[39] C. J. Van Westen, A. C. Seijmonsbergen, and F. Mantovani, "Comparing landslide hazard maps," *Natural Hazards*, vol. 20, no. 2-3, pp. 137–158, 1999.

[40] J. P. Wilson and J. C. Gallant, *Terrain Analysis Principles and Applications*, John Wiley & Sons, New York, NY, USA, 2000.

[41] T. Y. Duman, T. Çan, Ö. Emre, et al., "Landslide inventory of northwestern Anatolia, Turkey," *Engineering Geology*, vol. 77, no. 1-2, pp. 99–114, 2005.

[42] M. Bruce and K. Dylan, "Equations for potential annual direct incident radiation and heat load," *Journal of Vegetation Science*, vol. 13, no. 4, pp. 603–606, 2002.

[43] C. F. Lee, H. Ye, M. R. Yeung, X. Shan, and G. Chen, "AIGIS-based methodology for natural terrain landslide susceptibility mapping in Hong Kong," *Episodes*, vol. 24, no. 3, pp. 150–158, 2001.

[44] S. Jager and G. F. Wieczorek, "Landslide susceptibility in the Tully Valley Area, Finger Lakes region," U.S. Geological Survey Open-File Report 94-615, 1994.

[45] R. L. Baum, A. F. Chleborad, and R. L. Schuster, "Landslides triggered by the winter 1996 –1997 stroms in the Puget Lowland," Geological Survey Open-File Report 98-239, Washington, Wash, USA, 1998.

[46] H. B. Wang and K. Sassa, "Comparative evaluation of landslide susceptibility in Minamata area, Japan," *Environmental Geology*, vol. 47, no. 7, pp. 956–966, 2005.

[47] I. S. Evans, "What do terrain statistics really mean?" in *Landform Monitoring, Modelling and Analysis*, S. Lane, K. Richards, and J. Chandler, Eds., pp. 119–138, John Wiley & Sons, Chichester, UK, 1998.

[48] T. Behrens, "DEM Analysis Tool," 2005, http://www.esri.com.

[49] H. J. Miller and J. Han, *Geographic Data Mining and Knowledge Discovery*, CRC Press, Boca Raton, Fla, USA, 2001.

[50] C.-F. Chien and L.-F. Chen, "Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry," *Expert Systems with Applications*, vol. 34, no. 1, pp. 280–290, 2008.

[51] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, pp. 1285–1293, 1988.

[52] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson Education, Delhi, India, 2005.

[53] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," *International Journal of Medical Informatics*, vol. 77, no. 2, pp. 81–97, 2008.

[54] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, 2005.

[55] H. Soysal, S. Sipahioglu, D. Kolcak, and Y. Altinok, "Turkiye ve Cevresinin Tarihsel Deprem Katalogu (MO 2100-MS 1900)," TUBITAK project Tbag 341, Istanbul, Turkey, 1981.

[56] S. Beguería, "Validation and evaluation of predictive models in hazard assessment and risk management," *Natural Hazards*, vol. 37, no. 3, pp. 315–329, 2006.