Landslide hazard zonation using logistic regression technique: the case of Shafe and Baso catchments, Gamo highland, Ethiopia

Leulalem Shano (✉ leulalemshano@gmail.com)

Arbamich University

Tarun Kumar Raghuvanshi

Addis Ababa University

Matebie Meten

Addis Ababa Science and Technology university

Research Article

Keywords: Landslide hazard, Logistic regression, Influencing factors, Gamo Highland, Ethiopia

Posted Date: May 26th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-191158/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Landslide hazard zonation using logistic regression model: The case of Shafe and Baso catchments, Gamo highland, Southern Ethiopia

Leulalem Shano^{*1,2}

E-mail: leulalemshano@gmail.com;

Phone no: +251-912759892

Tarun Kumar Raghuvanshi^{PhD}³

PO Box 1176,

Addis Ababa, Ethiopia

E-mail: tkraghuvanshi@gmail.com;

Phone no: +251-911-875983

Matebie Meten^{PhD}¹

E-mail: matebe21@gmail.com;

Phone no: +251-911899279

Acknowledgements

The first author expresses sincere thanks to the Addis Ababa Science and Technology University, Arba Minch University and IUC coordination office of Arba Minch University for supporting this research by providing vehicles and field equipments. Also, the authors have special thanks to Mr Abdulsalam for the detailed editing of this research paper. We would like to thank further also the anonymous reviewers for their constructive comments, suggestions and questions which helped us for the improvement of our manuscript. Last but not least, Mr Belachew Moges, Tadele Dagne and Tilahun Mersha for their support during research work.

* Corresponding author, Address: E-mail: leulalemshano@gmail.com; Phone no: +251-912759892

1 Department of Geology, College of Applied Sciences, Addis Ababa Science and Technology University, Ethiopia

2 Department of Geology, College of Natural Sciences, Arba Minch University, Ethiopia

3 School of Earth Sciences, College of Natural Sciences, Addis Ababa University, Ethiopia

Landslide hazard zonation using logistic regression model: The case of Shafe and Baso catchments, Gamo highland, Southern Ethiopia

Abstract

Landslide hazard zonation plays an important role in safe and viable infrastructure development, urbanization, land use, and environmental planning. The Shafe and Baso catchments are found in the Gamo highland which has been highly degraded by erosion and landslides thereby affecting the lives of the local people. In recent decades, recurrent landslide incidences were frequently occurring in this Highland region of Ethiopia in almost every rainy season. This demands landslide hazard zonation in the study area in order to alleviate the problems associated with these landslides. The main objectives of this study are to identify the spatiotemporal landslide distribution of the area; evaluate the landslide influencing factors and prepare the landslide hazard map. In the present study, lithology, groundwater conditions, distance to faults, morphometric factors (slope, aspect and curvature), and land use/land cover were considered as landslide predisposing/influencing factors while precipitation was a triggering factor. All these factor maps and landslide inventory maps were integrated using ArcGIS 10.4 environment. For data analysis, the principle of logistic regression was applied in a statistical package for social sciences (SPSS). The result from this statistical analysis showed that the landslide influencing factors like distance to fault, distance to stream, groundwater zones, lithological units and aspect have revealed the highest contribution to landslide occurrence as they showed greater than a unit odds ratio. The resulting landslide hazard map was divided into five classes: very low (13.48%), low (28.67%), moderate (31.62%), high (18%), and very high (8.2%) hazard zones which was then validated using the goodness of fit techniques and receiver operating characteristic curve (ROC) with an accuracy of 85.4. The high and very high landslide hazard zones should be avoided from further infrastructure and settlement planning unless proper and cost-effective landslide mitigation measures are implemented.

Key words: Landslide hazard, Logistic regression, Influencing factors, Gamo Highland, Ethiopia

1. Introduction

In Ethiopia, landslides are common problems in mountainous terrain thereby causing hazards on infrastructures, human life, and their properties. Although landslide hazard is the most common problem in South Ethiopia, not much visible studies have yet been conducted in this area. The current study area, Gamo Highland consists of the mountain Guge in South Ethiopia, which is the highest mountain in South Ethiopia. Various sizes of landslides were repeatedly occurring in this mountain chain toward both sides of mountain slopes. Specifically, the current study area is some parts of this Gamo Mountain on the eastern side of the mountain chain. Landslide hazards of the Shafe and Baso catchments are frequently affecting people's life, properties, and their day-to-day activities. Last 30 years, more than 800 people were displaced and several properties were partially or completely damaged. From the total study area of 284km², about 71.53km² has been affected by landslide hazards. The main aims of the current research are to zone the landslide hazard levels of the area; to identify the intensity of the land sliding, and to predict future landslide hazards based on the past and present landslide evidence. The importance of landslide hazard zonation is to manage the landslide hazard of the area; to prepare the future environmental plan of the area and to recommend a cost-effective technique for landslide management which can be selected according to international landslide hazard mapping guidelines (AGS 2000; Fell et al. 2008; Dahal and Dahal 2017).

For landslide hazard studies, the researchers should identify the types of landslides based on the various classification schemes; spatial-temporal landslide inventories; and different types of influencing factors. The landslide classifications depend on the material involved, types of movement, depth of failure surfaces, and moisture contents of the moving materials if it is earth (soil). A general definition of landslide is it consists of almost all varieties of mass movements on the slope, including some, such as rockfalls, topples, and debris flows, that involve little or no true sliding (Varnes et al. 1984; Fell et al. 2008; van Westen et al. 2008; Reichenbach et al. 2018; Shano et al. 2021). To characterize or model the above-defined landslide types, there are different types of techniques such as heuristic, statistical, machine learning and deterministic approaches (Raghuvanshi et al. 2015; Zumpano et al. 2018; Tiranti & Crenonini 2019; Shano et al. 2020). These approaches have advantages and disadvantages as well as purpose and application scale ranges when using for landslide hazard zonation or mapping; these were well characterized (Shano et al. 2020). A landslide hazard zonation is that applies in a general sense to divide the land surface into areas and the ranking of these according to degrees of the actual or potential hazard for landslides or other mass movements on slopes (Guzzetti et al. 1999; Chen and Wang 2007; Wang and Peng 2009; Davies 2015; Raghuvanshi et al. 2015; Thennavan and Pattukandan Ganapathy 2020). In the present study, a binary logistic regression was applied to characterize the landslide hazard level in Baso and Shafe catchments. Logistic regression is one of the multivariate statistical approaches which is often used for landslide hazard zonation or modelling (Lee & Pradhan 2006; Pradhan & Lee 2009; Chen & Wang 2015; Meten et al. 2015; Wubalem and Meten 2020). Landslide modelling by logistic regression is paramount important to identify the influencing factors and to predict the landslide occurrence. The statistical and machine learning methods need statistical validation and confirmation in the field (Tien Bui et al. 2015; Duric et al. 2019; Shano et al. 2020).

The study applied a binary logistic regression model because the collected data were a combination of continuous and discrete and also the dependent variable is binary output. The working scale, the data type, and the characteristics of the dependent variable are well described (Chen and Wang 2007; Lee 2015; Shano et al. 2020). For landslide hazard analysis this study used different thematic layers or causative factors. These factor maps were collected from archived (geological & meteorological) data, remote sensing data, and field observation. The data are grouped into different categories such as DEM derivatives (slope, aspect, curvature, and elevation), geological units (lithology & soils), hydrological (groundwater and rainfall), distance (proximity to streams and geological lineaments), and land use/land cover. These causative factors were integrated with spatiotemporal landslide inventory data in the GIS environment and SPSS20 were used for analysis. The major concept of logistic regression used for landslide hazard analysis, the same independent factors data as used for landslide susceptibility, but different landslide inventory data and variant or dependent variable. Landslide inventory data for landslide hazard analysis must contain spatial and temporal components or based on rainfall events. The variant of landslide hazard is the presence of the hazard designated as (1) and non-happening of the hazard indicated (0) (Cheng et al. 2012). If the dependent variable is binary and the independent factors are containing continuous and discrete variables, the logistic regression method is a widely used statistical approach either for landslide susceptibility or hazard mapping. It is most effective in the different studies related to landslide hazard zonation (van Westen et al. 2006; Lee & Pradhan 2006; Pradhan & Lee 2009; Pradhan et al. 2011; Lee 2014; Bourenane et al. 2016; Sun et al 2018).

Not only using reliable input data is important but also cross-checking and validating the reliability of the statistical model used to produce the landslide susceptibility maps. Therefore, this study also checked the model reliability

apart from field validation using various methods such as goodness of fit, pseudo R square, multicollinearity and ROC to evaluate each step in logistic regression analysis.

2. Input data and methods

2.1 Description of the study area

Geographically, the study area is located between latitudes 6843000mN to 7070000mN; and longitudes 343000mE to 361500mE in the UTM Zone 37N with an area extent of about 284 sq. Km (Fig.1). The topography of the area is characterized by the rugged topography and mountainous landscape that shows various geomorphic features, soil types, ecological zones, and steep to very steep slopes. The hilly and mountainous areas cover most of the study area particularly in the eastern and southeastern part of the catchment. The altitudes of the area reach up to a maximum of 3046m in the northwestern part of the study area.

Fig. 1 Location map of the study area

The rainfall distribution of the area is bimodal (i.e. from March up to May and from August up to October). Rainfall in the first category is more torrential and initiates more flooding in the area. In the second category, extensive rainfalls with wide areal coverage are occurring in the study area.

2.2 Input data types

Data used for landslide hazard zonation are landslide inventory data, environmental and triggering factors. The first category of data was collected for this study using Google Earth image because there was no archived landslide historical data; the second category was collected from different data sources as described in Table 1 and the field visits.

Table 1 Data sources and detail of data used in the study

Data type	Source	Spatial resolution	Variables extracted
Satellite imagery/ environmental factors	USGS-ASTER	30m	DEM, slope angle, aspect, curvature, elevation and drainage net work
Google Earth image	-	-	Landslide inventory
Geological map	published by JICA, 2012	1:250,000	Used as supportive data

Environmental factors/causal factors such as lithology, structure, groundwater and soil type	Field survey	1:100,000	Geological map of area and field check for verification of land use/land cover and landslide location
Rainfall	Meteorological agency of Ethiopia, 1983-2018	-----	Annual and monthly rain fall intensity
Land use/ land cover	Sentinel satellite image	Resampled to 1:100,000	Land use/land cover

2.3 Landslide inventory

The landslide inventory map preparation is a very important component of landslide hazard evaluation or analysis (Guzzetti et al. 2006; Roy and Saha 2019). Statistical-based landslide study cardinaly needed landslide inventory data for landslide hazard zonation (Pradhan and Lee 2009; Pourghasemi et al. 2012; Reichenbach et al. 2018). This study collected landslide inventory data from field surveys, archives and Google Earth image analysis approaches. A total of 1554 landslides were collected from Google Earth imagery and intensive field visits. The landslide inventory map was prepared in the ArcGIS environment for dividing landslide training and validating data sets. Spatial and temporal landslide inventory data represent the main component of landslide hazard assessment to predict the probability of future landslide hazard occurrences (Guzzetti et al. 2005; Althuwaynee et al. 2012; Guzzetti et al. 2012; Valenzuela et al. 2017). Both temporal and spatial landslide inventory data were collected from the archives of different offices, aerial photo interpretation, and Google Earth image analysis. The spatial component was collected mainly by using Google Earth image analysis.

Fig. 2 Landslide inventory map

These landslide inventories are displaying not only the spatial and temporal distribution of the landslide but also the size and their distributions. One of the basic characteristics of landslide hazard analysis is to reveal the size distribution of the landslide in this specific area. As indicated (Fig. 2) the size of the landslides varies from 89m² to 170459m² in their temporal and spatial distributions.

Relative landslide frequency analysis was recommended by Guzzetti et al. (2005), Remondo et al. (2005), Corominas and Moya (2008), Corominas et al. (2014), Tanyas et al. (2018a) and Tebbens (2020) for landslide

zonation of large areas to calculate the number of landslide/Km²/year or a number of landslides/number of pixels/year. But this research applied the second analysis technique to evaluate the landslide frequency of the current study area. All the landslide types together as well as all the landslide frequency were considered.

2.3.1 Landslide type, size, and frequency

As highlighted in the introduction section, landslides are grouped based on different criteria, such as material involved, and depth of failure surface, activities, and movement. This study used the first three criteria only. Based on the material involved, landslides in the current study area are debris, earth, and rock slides. From 1554 landslides, only 65 were used for classification or type analysis because these 65 landslides are more or less active to identify the types of landslides. From these 65 active landslides are debris slide, earth slide, and rock slides 60% (39), 25% (16) and 15% (10) respectively (Fig. 7). However, depending on the movement of material, there are no clear shapes of landslide failure surfaces because most of the landslide failure surfaces were concealed by the recently involved mass, agricultural activities, and erosion. Based on the depth of the failure surface, also relatively older landslides have no clear depth because some parts were filled by other small landslides and agricultural activities. But the most recent landslides were measured; the majorities of the depth of failure surfaces are greater than 5m and a few of them are less than 5m.

Fig. 7 Landslide types (a) rock fall; (b) rock slide; (c) debris flow and (d) earth flow

Landslide size is one of the important parameters for landslide hazard zonation or mapping. According to (Guthrie 2002; Medwedeff et al. 2020) the landslides based on sizes are grouped into four namely; very small (<200m²), small (200-2000m²), medium (2000-10,000m²), and large (>10,000m²). Founded on this categorization, the landslides that were collected from the current study area are grouped as small, medium, and large landslides. However, very small landslides which are nearby starting points of the gully erosion sites and on agricultural land were excluded in this study. This is because their sizes are being much smaller when compared to the scale of the map. This research, however, used only active landslides which can be mapped through field observation and Google Earth image analysis. As described in Table 2, these active landslides are 65 (7%) of the total landslide 1554. Of 65 active landslides, about 23% (15) are small, 60% (39) are medium and 17% (11) are large landslides. Landslide frequency analysis is a very important component of landslide hazard zonation especially to predict the probability of future landslides (Corominas et al. 2014; Taynas et al. 2018b). However, conducting the tempo-spatial frequency analyses is difficult in this area as there is no archived landslide data to be used for this large study area. The researcher did not apply landslide size-frequency because of the wider areal extent. But this research preferred to use the mass inventory of active landslides in different year intervals. Relative landslide frequency analysis as recommended (Guzzetti et al. 2005; Remondo et al. 2005; Corominas and Moya 2008; Corominas et al. 2014; Taynas et al. 2018b), landslide zonation for the large area was carried out using the number of landslide/Km²/year or the number of landslides/number of pixels/year. But this researcher used the latter one to analyze the landslide frequency of the current study area.

Table 2. Estimated size of active landslide

No	Year	Area (m ²)	year	Area(m ²)	Year	Area (m ²)
1	2003 active	2639.689	2013active	2211.183	2018 active	528.2831
2	2003 active	3076.811	2013active	2227.743	2018 active	749.9144
3	2003 active	3324.521	2013active	2648.039	2018 active	845.6635
4	2003 active	3324.669	2013active	2839.16	2018 active	847.9622
5	2003 active	3520.886	2013active	2961.179	2018 active	954.8659
6	2003 active	4181.356	2013active	2965.086	2018 active	1105.595
7	2003 active	4272.964	2013active	3109.883	2018 active	1118.703
8	2003 active	4794.906	2013active	3156.735	2018 active	1384.665

9	2003 active	5408.985	2013active	3598.719	2018 active	1476.906
10	2003 active	5482.244	2013active	3973.931	2018 active	1495.95
11	2003 active	6224.927	2013active	4632.657	2018 active	1500.607
12	2003 active	6390.273	2013active	5080.663	2018 active	1657.215
13	2003 active	7340.698	2013active	5328.575	2018 active	1806.776
14	2003 active	7550.639	2013active	5485.316	2018 active	1808.325
15	2003 active	9110.146	2013active	8924.702	2018 active	1870.125
16	2003 active	9121.425	2013active	11201.1	2018 active	2209.325
17	2003 active	11294.27	2013active	12431.3	2018 active	2219.586
18	2003 active	13880.82	2013active	14496.79	2018 active	2289.252
19	2003 active	15715.03	2013active	16643.83	2018 active	2672.387
20	2003 active	20418.28	2013active	39813.47	2018 active	4268.543
21	2003 active	22759.12			2018 active	4648.445
22	2003 active	33346.48			2018 active	5370.6
23					2018 active	5821.757

On basis of this principle, the current study applied the pixels of the active landslides to estimate the frequency of the landslides in the study area. This helps to identify how much area is affected by landslide hazards in different parts of the study area. The total pixels under the very high and high hazard zones are 79477 (71.53km²) and from that, the pixels affected by active landslides accounted for 619 (0.56 km²); whereas their difference is 70.97km². Therefore, the frequencies of the average landslide numbers for estimated in 25, 50, and 100 years are presented as follows in Table 3.

Table 3 Landslide frequency analysis in the Baso and Shafe catchments

Years	Number	Percent	Average no of landslide	Time after(years)	Will be affected pixels
2003	22	34	22	25	5225
2013	19	29		50	10450
2018	24	37		100	20900
Total	65	100			18.86 km ² (100 years)

3. Analysis

In this study, a binary logistic regression model was applied to prepare the landslide hazard zonation map. Logistic regression is one of the most widely used statistical methods for landslide hazard probability of occurrence prediction (Chen and Wang 2007; Devkota et al. 2013; Nolasco-Javier and Kumar 2020). It is used to establish a functional relationship among independent and dependent variables. Logistic regression is used to predict a categorical (usually dichotomous) variable from a set of predictor variables (Hosmer & Lemeshow 2005; Rasyid et al. 2016; Wubalem and Meten 2020). It is often chosen if the predictor variables consist of continuous and categorical variables and/or if they are not nicely distributed but logistic regression makes no assumptions about the distributions of the predictor variables). According to (Lee & pradhan 2006; Shano et al. 2020), calculated the Landslide Hazard Index (LHI) by solving the regression equation. Correlation between landslide event and landslide affecting factors is estimated, and then, an equation predicting the landslide hazard is obtained.

$$P(z) = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}} \dots \dots \dots \text{eq (1)}$$

$$Z = \alpha + \sum \beta_i X_i \dots \dots \dots \text{eq(2)}$$

Where, α is the constant; X_i is the independent variable; and β_i is the corresponding coefficient

After identifying the problem and deciding the working scale of the study, different procedures were followed in the present study starting from office work to the detailed field surveys. One of the desk studies which were later confirmed in the field was landslide inventory collection from different archived sources. The 1554 landslides were collected from both archives and geomorphological sources which were later divided into two groups i.e. training landslides (70%) for modeling and validation landslides (30%) for prediction.

The non-landslide pixels were extracted by using a random point extraction tool in a GIS environment with an equal number of training datasets which is 1088. Then landslide and non-landslide points were copied from both attribute tables to MS excel and merged by their IDs. Then, these data were copied to SPSS20 and adjusted their IDs, measurements, and types of data (numeric or string). In this research, landslides were represented by IDs (1) and non-landslide represented zero IDs (0). Furthermore, in SPSS20 software analysis, regression, and binary regression are the common steps to be followed to calculate coefficients of each independent variable. These are processed in the SPSS software clicking on analyze, regression, and binary logistic regression. Then select the decision variable/landslide in the dependent variable box and other independent variables into the covariate box. Finally, click options and check “Hosmer-Lemeshow goodness fit” and confidence interval (CI) $\exp(\beta_i)$ 95% were checked in the boxes.

All result calculations were performed with the aid of algebraic sum in ArcGIS 10.4 to find out the landslide probability value. In general, this research followed six basic steps to calculate the landslide hazard index based on logistic regression formula in eq. (3). These are: (1) multiplying all significant factor maps with the logistic regression coefficients and sums up by using map algebra in ArcGIS environment; (2) add intercept value (α) on step one parameters; (3) make exponential of value resulted from step two; (4) add one on the results of step three; (5) make ratio results of step three to results of step four; and (6) finally, classify the map would be produced by using all the above steps to create a landslide hazard index map. Step six and map classification in different sub-topics in this research used the Natural break method.

4 Landslide factor analyses

4.1 Conditioning factors

The landslide influencing factors are generally grouped into two, conditioning/predisposing and triggering (Nefeslioglu et al. 2008a; Othman et al. 2018). The factors that make the slope stability to marginal levels are the conditioning factors and those factors that are pushing from marginal to failure are the triggering factors. All conditioning factors which will be described here are morphometric, geological, hydrological, and land use/ land cover. The first categories of conditioning factors are morphometric factors that are derived from DEM data including slope, curvature, aspect, and elevation (Fig. 3a - d). The second groups are geological factors (lithology and distance to fault). In the third category, groundwater conditions and distance to stream were considered as hydrological factors for landslide hazard analysis and these factors are the main initiating factors of landslides. Land use/land cover is also another factor in facilitating slope toe erosion by serving as a zone of rainwater absorption or saturation in the study area.

Fig. 3 Morphometric factors (a) slope gradient (in degree), (b) slope aspect (in degree), (c) curvature and (d) elevation (m)

Geological factors in the present study area consist of lithological distribution and their contacts, soil deposits, and geological structures/buffered to zone the geological structures (Fig. 4a). Obviously, some of the geological factors starting from their initial formation are susceptible to landslides such as pyroclastic materials. The susceptibility of these materials is supported by other environmental factors. The study area is covered by tuff to very strong basalt (Fig. 4b). Geological structures which are included for current landslide hazard analyses are major faults. This is because to include all geological lineaments of the area there was a limitation of the applied scale of the study.

Hydrological causative factors recognized in the present study are groundwater and rivers/streams. Groundwater and the proximity to streams are the major influencing factors of landslide occurrence in the area (Fig. 4c). For this study researchers used to characterize groundwater based on only surface manifestation due to the subsurface data limitation. Groundwater condition of the area depends on the prolonged rainfall and hydraulic conductivity of geological materials and river/stream erosion depends on the shape of the catchments, human activities of the catchment, geology of the area, and the amount of rainfall (Fig. 4d).

Fig. 4 Geological and hydrological factors (a) lithology and soil types, (b) proximity to lineament (m), (c) groundwater and (d) proximity to streams (m)

Land use is one of the most important conditioning factors in slope stability analysis (van Westen & Leulie 2003; Kayastha et al. 2012; Gariano et al. 2018). Land use/land cover has either a positive or negative influence on soil erosion, landslide, and groundwater occurrences. For landslide hazard analysis this thematic layer was classified as agricultural land, bare land, rangeland, settlement, sparse forest, and moderate forest areas. Then, the agriculture and socio-economic development of Ethiopia is based on land and water resource management and development. These resources are overstretched due to an increase in population and misusing of resources, often leading to resource depletion. This research was basically focused on landslide hazard zonation. During field surveys, different landslides were identified due to soil erosion which was in turn related to poor land use/land cover practices. Generally, land cover is very important to balance landscape dynamics evolutions.

Fig. 5 Land use/land cover map

The soil erosion has a direct relation with land use/land cover. The bare soil is more susceptible to landslides than the soil with vegetation cover. Large percentages of the areas such as settlement, agricultural land, rangeland, bare land, and sparse forest are affected by erosion and shallow landslides. These categories of land use/land cover about 66.79 % of the study area. In this percentage of areal coverage, the daily activity has a high facilitating factor for landslide hazard occurrence. As indicated above (Fig. 5) the area with the second rank is moderate forest land which is about 33.2 % of the area.

4.2 Triggering factor

Rainfall is the major triggering factor of landslide occurrence in the study area. Even though there were no recorded landslide data to know the exact time concerning rainfall due to lack of knowledge or experience of the local people, most landslides occurred during the rainfall season that assured from interviews of local people during field surveys. The second possible triggering factor but no recorded landslide catalog data in this particular study area is an earthquake. Therefore, in this research precipitation is considered as the only triggering factor for landslide hazard analysis. The precipitation data for a time period of 36 years was collected from three stations i.e. Chench, Mirb Abaya, and Arba Minch. In this study area usually the maximum rainfall taking place in two seasonal folds, the first one is April to May 180mm and 156 mm of annual precipitations respectively (Fig. 6a & b). The second rainy

season is September and October with a long annual rainfall of 120mm and 138mm respectively. But in the second season sometimes rainfall continuously increases from July to October in which a high amount of landslides were recorded. In two maximum rainfalls, various landslides and erosions are occurring and affecting human lives, properties, and hindering their activities. A landslide is taking place either during prolonged or torrential rainfalls of the area.

Fig. 6 Long year monthly rain fall (mm) (a) Chench and (b) Mirab Abaya (1983-2018)

5. Results

Different independent factors were used to predict a dependent variable. In the landslide study, the common dependent variable is a landslide. The logistic regression method is used to characterize each thematic layer in a holistic manner (Lee 2005; Mathew et al. 2007,2008; Talaei 2014; Wubalem and Meten 2020). In this study, either of the variables i.e. continuous or non-continuous is considered in a general way, and they have no separate coefficients. As listed in Table 4, nine independent variables were used to predict a dependent variable in the study area. The logistic regression approach is very important to pinout which independent variable would have more influence on landslide occurrence in the study area. As described in the methodology section, these nine independent variables were selected as landslide influencing factors based on field observation and remote sensing image interpretation.

Table 4. Coefficients and statistical parameters of the logistic regression model

Variables	β	SE	Wald	p-value	Exp(β)	Confidence level 95%	
						Lower	Upper
Land use/land cover	-0.132	0.089	2.161	0.142	0.877	0.736	1.045
Distance to stream	0.715	0.127	31.751	0.000	2.043	1.594	2.620
Aspect	0.101	0.088	1.320	0.251	1.107	0.931	1.316
Curvature	-0.331	0.180	3.404	0.065	0.718	0.505	1.021
Elevation	-0.059	0.252	.055	0.815	0.943	0.575	1.546
Lithology	0.079	0.149	.281	0.596	1.082	0.808	1.449
Groundwater	0.451	0.187	5.848	0.016	1.570	1.089	2.263
Slope	-0.407	0.217	3.512	0.061	0.666	0.435	1.019
Distance to faults	1.220	0.114	113.549	0.000	3.386	2.705	4.237
Constant	-4.769	1.052	20.552	0.000	0.008	0.000	0.000

Based on the sign of coefficients, the independent or predictor variables are classified into two; either with a negative or positive coefficient. The variables such as rivers/streams, aspect, geological units, and lineaments have positive coefficients. The positive coefficients indicate that the values of the independent variable increase, the probability of a landslide assuming that the other variables in the model held constant are increasing (Chen & Wang 2007). The others, for example, land use/land cover, curvature, elevation, and slope have negative coefficients. The independent variables with negative coefficients are also important while their negative coefficients don't mean that

those variables do not affect landslide occurrence. Generally, they are indicating that as the independent variables with negative values increase, the occurrence of a dependent variable decreases whereas other variables are constant (Peng & So 2002; Rasyid et al. 2016). For example, for some arbitrary reference point, as the slope increases, then the landslide occurrence decreases. Not only arbitrary reference but also in logistic regression results, if the correlation coefficient is negative, it provides statistical evidence of a negative relationship between the variables. The large coefficient of the confounding variable will cause a decrease in the second variable or make it negative. For example, land use/land cover and distance to stream, when the coefficient of stream erosion becomes larger, the coefficient of land use/land cover is negative. In the literal meaning when the landslide progresses to agricultural land due to the stream erosion, the agricultural land decreases but does not initiate landslide occurrence in an equal manner as stream erosion does. The last issue which is not considered here but in the model validation section is multi-collinearity problems of independent variables. The multi-collinearity problems within independent variables may affect the sign and/magnitude of coefficients. In Table 4 above, the coefficient of each thematic layer indicates the influence of each independent variable on a dependent variable in different degrees of influence.

$$Z = -0.132\text{LULC} + 0.715\text{River} + 0.101\text{Aspect} + -0.331\text{Curvature} + -0.059\text{Elevation} + 0.079\text{Geology} + 0.451\text{Groundwater} + -0.407\text{Slope} + 1.22\text{Lineament} - 4.76 \dots \dots \dots \text{eq. (3)}$$

Fig. 8 Landslide hazard zonation map

6. Data fitness and model validation

Every data and model from the statistical, multi-criteria decision, and machine learning approaches should be checked before interpretation because the processes are invisible during analysis and modelling (Shano et al. 2020). The data fitness and overall model validations are commonly checked by using the statistical evaluation approaches. These are Multi-collinearity, Nagelkerke R^2 and Hosmer-Lemeshow test for data fit; and Omnibus test and ROC curve for overall model validation.

Multi-collinearity is a statistical environment that refers to the numbers of simulated factors in a logistic regression model which is highly correlated in such a way that one can be linearly predicted from the others with a non-trivial degree of accuracy. In this study, multi-collinearities among the determining factors of landslide hazard have been identified using tolerances (Table 5). As listed in the following table there is no multi-collinearity among the independent variables. As described in (Davidson et al. 1981; Midi et al. 2010), these two parameters are calculated by using the following equations, eq(4) and eq(5).

$$\text{Tolerance} = \text{tol} = 1 - C^2 \dots \dots \dots \text{eq}(4)$$

Where C^2 is the coefficient of determination for the regression of that explanatory is on all remaining independent variables. The variance inflation factor (VIF) is defined as the reciprocal of tolerance as

$$\text{VIF} = \frac{1}{\text{tolerance}} \dots \dots \dots \text{eq}(5)$$

Table 5 Data fit and model validation statistical parameters

S. no	Parameters	Collinearity statistics		Pseudo R-squares and Hosmer-Lemeshow test		
		Tolerance	(VIF)	Nagelkerke R Square		0.572
1	Land use/land cover	0.764	1.310	Hosmer-Lemeshow test		
2	Distance to stream	0.897	1.115	Chi-square	df	Sign.

3	Aspect	0.990	1.0110	5.038	8	0.753	
4	Curvature	0.989	1.011	Omnibus test			
5	Elevation	0.799	1.251	Name	X ²	df	Sign.
6	Lithology	0.942	1.061	Model	295.425	9	0.000
7	Groundwater condition	0.925	1.081				
8	Slope	0.985	1.015	ROC	85.4		
9	Distance to faults	0.879	1.137				

According to (Saha 2017), tolerance is less than 0.10 or variance inflation factor (VIF) is greater than 10 indicates the presence of multicollinearity problems. For the test of multi-collinearities, 1554 landslide points have been selected on a random basis and data have been extracted from 9 landslide influencing thematic layers for these randomly selected points. The result showed that all tolerance values were greater than 0.10 which indicates no collinearity among the nine landslide hazard determining factors. In turn, the values of VIF are less than 10 which indicated in the above Table (5), with no collinearity among the 9 independent factors. From this pre-modelling analysis, it can be concluded that there is no collinearity problem within independent variables.

The Omnibus test is a simple type of statistical test that is used to check whether the explained variance in a set of data is significantly greater than the unexplained variance in the overall model (Hosmer & Lemeshow 2000; Bewick et al. 2005; Chen et al. 2018). In addition, the Omnibus test as a general name refers to an overall or a global test of a model. The chi-square statistics under the omnibus test is very important to characterize the overall statistical reality of logistic regression models (Hosmer & Lemeshow 2000; Chen et al. 2018). The main purpose of this test is to determine if the significance value is the maximum cut value. In this case, the researcher should model again because there may be unimportant variables included in that model. In the current research, this condition did not happen as the validated Omnibus test showed the statistical result of $\chi^2=295.425$, $df=9$ and $p<0.00$. The Omnibus Tests of model coefficients is used to check whether the new model (with explanatory variables included) showed an improvement over the baseline model that is mentioned in the stepwise model (Peng & So 2002; Chen et al. 2018). This model used a baseline or first step model because of its fitness with all parameters.

The Hosmer-Lemeshow test is commonly used for checking the goodness of fit of a model (Hosmer & Lemeshow 2000; Paul et al. 2013). Its advantage is applicable to whether the predictor variables are categorical or continuous (Biweck et al. 2005). The Hosmer and Lemeshow test ($P = 0.753$) indicates that the numbers of landslide occurrences are not significantly different from those predicted by the model and that the overall model fit is very good. In simple expression when the significance (P) value increases the significance of variation in the observed and expected value becomes less.

Most of the statistical analyses afford further statistical validations that may be used to measure the usefulness of the model and that is similar to the coefficient of determination (true R^2) in linear regression (Biweck et al. 2003; Cheng et al. 2012; Nattino et al. 2020). Nagelkerke R^2 is used for logistic regression. The value for this pseudo-R-square is 0.572 (Table 5). Therefore, the researchers used it for interpretation the Nagelkerke R^2 as it is an adjusted version of the Cox & Snell R^2 and covers the full range from 0 to 1. Consequently, it is often preferred. The value of 0.572 indicates that the model is useful in predicting the occurrence of landslide hazards. The factor influencing the landslide in the study area 57.2% is predicted in this model while 42.8% remained unpredicted.

The overall model validation is the most important step which is applied before interpretation of the model. A receiver operating characteristic (ROC) curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity). Generally, the sensitivity increases, specificity also increases. However, the increment of specificity from some limits decreases the quality of the model. Without a clear cut point but ROC is well described in (Zhou et al. 2002; Krzanowski and Hand 2009; Nattino et al. 2020).

In the present study, the ROC curve is used to evaluate the models and its result is very good for logistic regression as indicated in Fig.9. The model is performed well as indicated in ROC evaluation, the value of logistic regression is 85.4 as measured in real statistical software. According to (Bradley 1997; Fawcett 2005), random guessing of things fails on the diagonal line between (0, 0) and (1, 1), which has an area of 0.5; no realistic classifier should have an AUC less than 0.5.

The ROC curve value from logistic regression is 85.4 indicating that a model has a good prediction power. Thus, the result of this model is reliable to be replicated in other parts of the world in order to highlight the landslide hazard zones which will, in turn, enable us to take proper remedial measures.

Fig. 9 ROC graph of logistic regression

7. Discussion

The causative factors selected for landslide hazard zonation include slope gradient, slope aspect, curvature, elevation, groundwater, distance to faults, lithology, distance to streams, and land use/land cover as predisposing factors, and rainfall as triggering factors. The binary logistic regression model has been interpreted based on the coefficients of independent variables and different parameters listed in Table 5. Most papers in the literature indicate that they used odds of the logistic regression. However, other statistical parameters are also important to interpret whether the predicting/independent factors have an influence on a dependent variable. Using the odds of logistic regression for model interpretation may cause a rejection of other important factors because the odds depend on only the coefficient of independent variables. It is also better to use Wald statistics, p-value, and confidence intervals. The importance of the Wald test is to identify the ratio of coefficient (β_i) to standard error (S.E) which is then squared. It is known that the coefficient of binary logistic regression does not interpret straightforwardly as the coefficient of linear regression. The better way to interpret binary logistic regression is to use the odds ratio which is the $\exp(\beta_i)$ which represents the ratio change in the odds of events (E) for a unit change of the value of the respective predictor variable while keeping all other things being equal. For example, in this research, the estimated coefficient of the parameter “stream erosion” is 0.715 and the exponentiated value is 2.043. So if there is a pixel with 0.5 probabilities (p) of a slide at a certain distance to stream, the corresponding odds of the slide [O (E)] are $1/[O(E) = p/(1-p)]$ for that pixel. In the view of the above expression, the value of distance to stream is 2.043 that means the odd of distance to stream is 2.043. Now, the probability of distance to stream is 0.67 by using $(2.043 = p/(1-p))$. To check the increment of the probability of landslide chance (0.5), by 34% reveals how far from the baseline. All other parameters that were used to develop the model such as distance to fault, groundwater, aspect, and lithology were also interpreted in a similar way with stream erosion.

The odds of coefficients in Table 4 indicate the level influence on landslide occurrence. The first parameter with a higher odds ratio than other parameters was the distance to the fault. This is because it affects stream flow, rock strength, and groundwater occurrence. These lineaments/normal faults have been making different slope heights with various slope gradients. As identified in the field and remote sensing image interpretation, most landslides concentrated along with fault scars. This is because the movement of rocks from faulting causes the crushing of rocks around fault/fault zones. These activities made the rock to be weaker around the fault zone. Along these fault zones, there are plenty of springs that affect the strength of rocks and initiate landslide. In addition, there are some stream flows following these fault zones. The other important factor is groundwater condition. Some areas were covered by fine soils. When the groundwater flows from fractured rock to these fine soils, the pore pressure is increased this will, in turn, causes landslide hazard occurrence. In a sloppy area, this factor is increasing the weight and lubricating the failure surfaces of landslides.

Aspect is one of the most landslide influencing factors as identified in the result section. This factor is causing landslides when there is an increment of moisture content associated with the different sunlight exposure of slopes. If there is an increment of moisture content in geological materials, there is an increment of moisture content

continually which will develop pore water pressure. This situation is clearly observed from the field surveys and the result of logistic regression is also confirming this fact.

The geological/lithologic units of the area include basalt, ignimbrite, tuff, alluvial soil, colluvial soil, and residual soil. From these geological units, basalt, ignimbrite, residual soil, and alluvial soil have positive coefficients indicating a good correlation with landslide occurrence. Relatively basalt has a strong correlation as compared to ignimbrite which was confirmed during the field visit. Although ignimbrite formed a high cliff in the study area, there is no much landslide occurrence in this rock unit because the ignimbrite is less fractured and weathered as compared to basalt. Normally tuff is much weaker than basalt and ignimbrite but in the study area on tuff, almost no/very few landslides were registered because tuff is covered by basalt at top in the flat land. Alluvial soil was more sorted and deposited around flat land when compared with colluvial soils. Due to stream/ river activities in the alluvial soil, a lot of shallow landslides were observed in this unit. Naturally, colluvial soil is more susceptible to landslides but in the study area alluvial soil is more susceptible due to stream erosion after deposition. These stream erosion-associated landslides affect different crops and banana plantations around the rivers or on alluvial deposits.

In general, based on the aforementioned parameters, the landslide hazard zonation map of the study area was produced. Most landslides occurred due to groundwater conditions affecting agricultural land, human life, and the day-to-day activities of the local people. Such types of landslides covered relatively long distances up to downstream slopes. In some parts of the study area, landslides also occurred due to faulting that may offset the geological units. In other places, the agricultural activities caused deep-seated landslides. As indicated in Fig. 8, the landslide hazard map is classified into five hazard zones of very low (0 – 0.2), low (0.2 -0.4), moderate (0.4 – 0.6), high (0.6 – 0.8) and very high (0.8 – 1). The landslide hazard zonation map from logistic regression comprises 13.48%, 28.67%, 31.62%, 18% and 8.2% for very low, low, moderate, high and very high hazard zones respectively.

8. Conclusion

This research concluded that the logistic regression approach is effective to correlate independent variables with a dependent binary variable. Binary logistic regression is very effective to zone landslide hazards in Gamo Highland or western escarpment of the main Ethiopia rift system. In the current study area, this model used the most influencing factors such as lithology, distance to fault, distance to stream, groundwater conditions and aspect as the odds ratio values of these variables are greater than one. This is the main reason why the impact of these factors is much greater than the probability of occurrence by chance or 0.5. Most landslide hazards in the study area have occurred more or less due to the influence of these factors. However, the factors like land use/cover, slope, curvature and elevation have a negative correlation with landslides in this particular area. This is evidenced by a negative coefficient as can be seen in Table 4.

This approach has so many cross-checking methods from the accuracy of data to model validation. Almost all data accuracy is checking and model validation methods were applied in this study using a step-wise data revision. The multi-collinearity, Wald statistics test, pseudo-R square tests, Hosmer-Lemeshow test, the coefficient of each independent variable, odds of each coefficient, and the significance level of each step were thoroughly followed. Lastly, the model performance was evaluated using an ROC value (85.4), and the result indicated that this model has a high prediction performance for the Shafe and Baso catchments. The resulting landslide hazard map was classified into five classes, namely; very low (0 - 0.2), low (0.2 - 0.4), moderate (0.4 - 0.6), high (0.6 - 0.8) and very high (0.8 – 1). The areas of under high and very high hazard zones are difficult for infrastructure development. Hence, the local, zonal and regional governments should take proper and cost-effective remedial measures so as to prevent the impending future landslide occurrence in these potential hazard zones.

Disclosure statement

The authors declare that they have no any competing interests.

Funding

This research didn't receive specific grant from any funding agencies for this research work.

Reference

- AGS (Australian Geomechanics Society) (2000) Landslide risk management concepts and guidelines. Australian Geomechanics Society, Sub-Committee on Landslide Risk Management. Australian Geome 35:49–92
- Althuwaynee O F, Pradhan B, Mahmud A R, Yusoff Z M (2012) Prediction of slope failures using bivariate statistical based index of entropy model. IEEE Colloquium on Humanities, Science and Engineering (CHUSER). <https://doi.org/10.1109/chuser.2012.6504340>
- Bewick V, Cheek L, Ball J (2003) Statistics review 7: Correlation and regression. *Crit Care* 7, 451. <https://doi.org/10.1186/cc2401>
- Bewick V, Cheek L, Ball J (2005) Statistics review 14: Logistic regression. *Critical Care (London, England)* 9(1): 112–118. <http://dx.doi.org/10.1186/cc3045>
- Bourenane H, Guettouche M, Bouhadad Y, Braham M (2016) Landslide hazard mapping in the Constantine city, Northeast Algeria using frequency ratio, weighting factor, logistic regression, weights of evidence, and analytical hierarchy process methods. *Arabian J Geosci* 9(2): 1–24
- Bradley A P (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7): 1145–1159. [https://doi.org/10.1016/s0031-3203\(96\)00142-2](https://doi.org/10.1016/s0031-3203(96)00142-2)
- Chen J, King E, Deek R, Wei Z, Yu Y, Grill D, Ballman K (2018) An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics* 34(4): 643–651
- Chen Z, Wang J (2007) Landslide hazard mapping using logistic regression model in Mackenzie Valley, Canada. *Nat Hazards* 42(1): 75–89. <https://doi.org/10.1007/s11069-006-9061-6>
- Cheng S, Wang Y S, Chen F Y (2012) Geohazard Risk Assessment Method Based on Logistic Regression Model. *Adv Materials Research* 588–589: 1934–1937
- Chung C-J F, Fabbri A G (2012) Systematic Procedures of Landslide Hazard Mapping for Risk Assessment Using Spatial Prediction Models. *Landslide Hazard and Risk* 139–174. <https://doi.org/10.1002/9780470012659.ch4>
- Corominas J, Moya J (2008) A review of assessing landslide frequency for hazard zoning purposes. *Eng Geol* 102(3–4): 193–213. <https://doi.org/10.1016/j.enggeo.2008.03.018>
- Corominas J, van Westen C, Frattini P et al. (2014) Recommendations for the quantitative analysis of landslide risk. *Bull Eng Geo and the Environ*. <https://doi.org/10.1007/s10064-013-0538-8>
- Dahal, B. K., & Dahal, R. K. (2017) Landslide hazard map: tool for optimization of low-cost mitigation. *Geoenviron Disasters*. <https://doi.org/10.1186/s40677-017-0071-3>
- Dahal R K, Hasegawa S, Nonomura A, Yamanaka M, Masuda T, Nishino K (2007) GIS-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping. *Environ Geol* 54(2): 311–324. <https://doi.org/10.1007/s00254-007-0818-3>
- Dai, F, Lee C, Ngai Y (2002) Landslide risk assessment and management: an overview. *Eng Geol* 64(1): 65–87. [https://doi.org/10.1016/s0013-7952\(01\)00093-x](https://doi.org/10.1016/s0013-7952(01)00093-x)
- Davies T (2015) Landslide Hazards, Risks, and Disasters. *Landslide Hazards, Risks and Disasters* 1–16. <https://doi.org/10.1016/b978-0-12-396452-6.00001-x>
- Devkota K C, Regmi A D, Pourghasemi H R et al. (2013) Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling–Narayanghat road section in Nepal Himalaya. *Nat Hazards* 65(1): 135–165. <https://doi.org/10.1007/s11069-012-0347-6>
- Đurić U, Marjanović M, Radić Z, Abolmasov B (2019) Machine learning based landslide assessment of the Belgrade metropolitan area: Pixel resolution effects and a cross-scaling concept. *Eng Geol* 256: 23–38. <https://doi.org/10.1016/j.enggeo.2019.05.007>
- Fawcett T (2005) An introduction to ROC analysis. *Pattern Recognition Letters* 27(8): 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Fell R, Corominas J, Bonnard C, Cascini L, Leroi E, Savage WZ, (2008) Guidelines for landslide susceptibility, hazard and risk zoning for land use planning, on behalf of the JTC-1 Joint Technical Committee on Landslides and Engineered Slopes. *Eng Geol* 102: 85–98
- Gariano S L, Petrucci O, Rianna G, Santini M, Guzzetti F (2018) Impacts of past and future land changes on landslides in southern Italy. *Regional Environ Change* 18(2): 437–449
- Guthrie, R H (2002) The effects of logging on frequency and distribution of landslides in three watersheds on Vancouver Island, British Columbia. *Geomo* 43(3–4): 273–292

- Guzzetti F, Carrara 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(1-4): 181–216
- Guzzetti F, Reichenbach P, Ardizzone F, Cardinali M, Galli M (2006) Estimating the quality of landslide susceptibility models. *Geomorphology* 81(1-2): 166–184. <https://doi.org/10.1016/j.geomorph.2006.04.007>
- Guzzetti F, Reichenbach P, Cardinali M, Galli M, Ardizzone F (2005) Probabilistic landslide hazard assessment at the basin scale. *Geomorphology* 72(1-4): 272–299. <https://doi.org/10.1016/j.geomorph.2005.06.002>
- Hina S, Kawasaki A, Qasim M (2014) Landslide Susceptibility Analysis Using GIS and Logistic Regression Model A Case Study In Malang, Indonesia. *Asian J Environ and Disas Manag (AJEDM) - Focusing on Pro-Active Risk Reduction in Asia* 06(02): 117–129. <https://doi.org/10.3850/s1793924014000323>
- Hong H, Xu C, Bui D T (2015) Landslide Susceptibility Assessment at the Xiushui Area (China) Using Frequency Ratio Model. *Procedia Earth and Planetary Sci* 15: 513–517. <https://doi.org/10.1016/j.proeps.2015.08.065>
- Hosmer D W, Lemeshow S (2000) *Applied Logistic Regression*. <https://doi.org/10.1002/0471722146>
- Hosmer D W, Lemeshow S (2005) *Logistic Regression, Conditional*. *Encyclopedia of Biostatistics*. <https://doi.org/10.1002/0470011815.b2a10030>
- Kayastha P, Dhital M R, De Smedt F (2012) Evaluation of the consistency of landslide susceptibility mapping: a case study from the Kankai watershed in east Nepal. *Landslides* 10(6): 785–799
- Krzanowski W J, Hand D J (2009) ROC Curves for Continuous Data. <https://doi.org/10.1201/9781439800225>
- Lee C-T (2014) Multi-stage Statistical Landslide Hazard Analysis: Rain-Induced Landslides. *Landslide Sci for a Safer Geoenviron* 291–298. https://doi.org/10.1007/978-3-319-04996-0_45
- Lee S (2005) Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. *Int J Rem Sens* 26(7): 1477–1491
- Lee S, Pradhan B (2006) Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. *Landslides* 4(1): 33–41. <https://doi.org/10.1007/s10346-006-0047-y>
- Mandal S, Maiti R (2014) Application of Analytical Hierarchy Process (AHP) and Frequency Ratio (FR) Model in Assessing Landslide Susceptibility and Risk. *Semi-Quantitative Approaches for Landslide Assessment and Prediction* 191–226. https://doi.org/10.1007/978-981-287-146-6_7
- Mandal S, Mandal K (2018) Modeling and mapping landslide susceptibility zones using GIS based multivariate binary logistic regression (LR) model in the Rorachu river basin of eastern Sikkim Himalaya, India. *Modeling Earth Syst and Environ* 4(1): 69–88. <https://doi.org/10.1007/s40808-018-0426-0>
- Mathew J, Jha V K, Rawat G S (2007) Application of binary logistic regression analysis and its validation for landslide susceptibility mapping in part of Garhwal Himalaya, India. *Int J Rem Sens* 28(10): 2257–2275. <https://doi.org/10.1080/01431160600928583>
- Mathew J, Jha V K, Rawat G S (2008) Landslide susceptibility zonation mapping and its validation in part of Garhwal Lesser Himalaya, India, using binary logistic regression analysis and receiver operating characteristic curve method. *Landslides* 6(1): 17–26. <https://doi.org/10.1007/s10346-008-0138-z>
- Medwedeff W G, Clark M K, Zekkos D, West A J (2020) Characteristic landslide distributions: An investigation of landscape controls on landslide size. *Earth and Planetary Sci Letters* 539: 116203
- Meten M, Bhandary N P, Yatabe R (2015) GIS-based frequency ratio and logistic regression modelling for landslide susceptibility mapping of Debre Sina area in central Ethiopia. *J Mountain Sci* 12(6): 1355–1372
- Midi H, Sarkar S K, Rana S (2010) Collinearity diagnostics of binary logistic regression model. *J Interdisciplinary Math* 13(3): 253–267. <https://doi.org/10.1080/09720502.2010.10700699>
- Mirnazari M, Javad J, Ahmad A et al. (2014) Using Frequency Ratio Method for Spatial Landslide Prediction. *Research J Appl Sci, Eng and Technol* 7(15): 3174–3180. <https://doi.org/10.19026/rjaset.7.658>
- Nattino G, Pennell M L, Lemeshow S (2020) Rejoinder to “Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test.” *Biometrics* 76(2): 575–577
- Nefeslioglu H A, Duman T Y, Durmaz S (2008a) Landslide susceptibility mapping for a part of tectonic Kelkit Valley (Eastern Black Sea region of Turkey). *Geomorphology* 94(3-4): 401–418
- Nolasco-Javier D, Kumar L (2020) Landslide Susceptibility Assessment Using Binary Logistic Regression in Northern Philippines. *ICL Contribution to Landslide Disaster Risk Reduction*. https://doi.org/10.1007/978-3-030-60227-7_20
- Oštir K, Veljanovski T, Podobnikar T, Stančič Z (2003) Application of satellite remote sensing in natural hazard management: The Mount Mangart landslide case study. *Int J Rem Sens* 24(20): 3983–4002
- Othman A A, Gloaguen R, Andreani L, Rahnama M (2018) Improving landslide susceptibility mapping using morphometric features in the Mawat area, Kurdistan Region, NE Iraq: Comparison of different statistical models. *Geomorphology* 319: 147–160

- Ozdemir A, Altural T (2013) A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan Mountains, SW Turkey. *J Asian Earth Sci* 64: 180–197
- Paul P, Pennell ML, Lemeshow S (2013) Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Stat Med* 32(1):67-80
- Peng C-Y J, So T-S H (2002) Logistic Regression Analysis and Reporting: A Primer. *Understanding Statistics* 31–70. https://doi.org/10.1207/s15328031us0101_04
- Pourghasemi H R, Pradhan B, Gokceoglu C, Mohammadi M, Moradi H R (2012) Application of weights-of-evidence and certainty factor models and their comparison in landslide susceptibility mapping at Haraz watershed, Iran. *Arabian J Geosci* 6(7): 2351–2365. <https://doi.org/10.1007/s12517-012-0532-7>
- Pradhan B (2013) A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comp & Geosci* 51: 350–365. <https://doi.org/10.1016/j.cageo.2012.08.023>
- Pradhan B, Lee S (2009) Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models. *Environ Earth Sci* 60(5): 1037–1054
- Pradhan B, Lee S. (2010) Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environ Modelling & Software* 25(6): 747–759. <https://doi.org/10.1016/j.envsoft.2009.10.016>
- Pradhan B, Mansor S, Pirasteh S, Buchroithner M F (2011) Landslide hazard and risk analyses at a landslide prone catchment area using statistical based geospatial model. *Int J Rem Sens* 32(14): 4075–4087
- Raghuvanshi T K, Negassa L, Kala P M (2015) GIS based Grid overlay method versus modeling approach – A comparative study for landslide hazard zonation (LHZ) in Meta Robi District of West Showa Zone in Ethiopia. *The Egyptian J Rem Sens and Space Sci* 18(2): 235–250
- Rasyid A R, Bhandary N P, Yatabe R (2016) Performance of frequency ratio and logistic regression model in creating GIS based landslides susceptibility map at Lompobattang Mountain, Indonesia. *Geoenviron Disasters*. <https://doi.org/10.1186/s40677-016-0053-x>
- Reichenbach P, Rossi M, Malamud B D, Mihir M, Guzzetti F (2018) A review of statistically-based landslide susceptibility models. *Earth-Sci Revi* 180: 60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- Reid L, Page M (2003) Magnitude and frequency of landsliding in a large New Zealand catchment. *Geomorphology* 49(1-2): 71–88. [https://doi.org/10.1016/s0169-555x\(02\)00164-2](https://doi.org/10.1016/s0169-555x(02)00164-2)
- Remondo J, Bonachea J, Cendrero A (2005) A statistical approach to landslide risk modelling at basin scale: from landslide susceptibility to quantitative risk assessment. *Landslides* 2(4): 321–328
- Roy J, Saha S (2019) Landslide susceptibility mapping using knowledge driven statistical models in Darjeeling District, West Bengal, India. *Geoenviron Disasters*. <https://doi.org/10.1186/s40677-019-0126-8>
- Shano L, Raghuvanshi T K, Meten M (2020) Landslide susceptibility evaluation and hazard zonation techniques – a review. *Geoenviron Disas*. <https://doi.org/10.1186/s40677-020-00152-0>
- Shano L, Raghuvanshi T K, Meten M (2021) Landslide susceptibility mapping using frequency ratio model: the case of Gamo highland, South Ethiopia. *Arabian J of Geosci*. <https://doi.org/10.1007/s12517-021-06995-7>
- Sun X, Chen J, Bao Y, Han X, Zhan J, Peng W (2018) Landslide Susceptibility Mapping Using Logistic Regression Analysis along the Jinsha River and Its Tributaries Close to Derong and Deqin County, Southwestern China. *ISPRS Int J Geo-Information*. <https://doi.org/10.3390/ijgi7110438>
- Talaei R (2014) Landslide susceptibility zonation mapping using logistic regression and its validation in Hashtchin Region, northwest of Iran. *J Geol Soci of India* 84(1): 68–86. <https://doi.org/10.1007/s12594-014-0111-5>
- Tanyaş H, Allstadt K E, van Westen C J (2018a) An updated method for estimating landslide-event magnitude. *Earth Surface Processes and Landforms* 43(9): 1836–1847
- Tanyaş H, Westen C J, Allstadt K E, Jibson R W (2018b) Factors controlling landslide frequency–area distributions. *Earth Surface Processes and Landforms* 44(4): 900–917. <https://doi.org/10.1002/esp.4543>
- Tebbens S F (2020) Landslide Scaling: A Review. *Earth and Space Science*. <https://doi.org/10.1029/2019ea000662>
- Thennavan E, Pattukandan Ganapathy G (2020) Evaluation of landslide hazard and its impacts on hilly environment of the Nilgiris District - a geospatial approach. *Geoenviron Disasters*. <https://doi.org/10.1186/s40677-019-0139-3>
- Tien Bui D, Tuan T A, Klempe H, Pradhan B, Revhaug I (2015) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides* 13(2): 361–378
- Tiranti D, Cremonini R (Eds.) (2019) Landslide Hazard in a Changing Environment. *Frontiers Research Topics*. <https://doi.org/10.3389/978-2-88945-793-9>

- Valenzuela P, Domínguez-Cuesta M J, Mora García M A, Jiménez-Sánchez M (2017) A spatio-temporal landslide inventory for the NW of Spain: BAPA database. *Geomorphology* 293: 11–23
- Van Westen C J, Castellanos E, Kuriakose S L (2008) Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. *Engineering Geology* 102:112–131
- Van Westen C J, Lulie Getahun F (2003) Analyzing the evolution of the Tessina landslide using aerial photographs and digital elevation models. *Geomorphology* 54(1-2): 77–89
- Van Westen C J, van Asch T W J, Soeters R (2006) Landslide hazard and risk zonation—why is it still so difficult? *Bulletin of Engineering Geology and the Environment* 65(2): 167–184. <https://doi.org/10.1007/s10064-005-0023-0>
- Varnes D, IAEG (1984) Landslide hazard zonation: a review of principles and practice. U N Sci Cult Organ, Paris, pp 1–6
- Wang J, Peng X (2009) GIS-based landslide hazard zonation model and its application. *Procedia Earth and Planetary Science* 1(1): 1198–1204. <https://doi.org/10.1016/j.proeps.2009.09.184>
- Wubalem A, Meten M (2020) Landslide susceptibility mapping using information value and logistic regression models in Goncha Siso Eneses area, northwestern Ethiopia. *SN Applied Sciences*. <https://doi.org/10.1007/s42452-020-2563-0>
- Zhou C, Lee C, Li J, Xu Z (2002) On the spatial relationship between landslides and causative factors on Lantau Island, Hong Kong. *Geomorphology* 43(3-4): 197–207. [https://doi.org/10.1016/s0169-555x\(01\)00130-1](https://doi.org/10.1016/s0169-555x(01)00130-1)
- Zumpano V, Pisano L, Malek Ž et al. (2018) Economic Losses for Rural Land Value Due to Landslides. *Frontiers in Earth Science*. <https://doi.org/10.3389/feart.2018.00097>