Lowland Technology International 2018; 20 (3): 401-412
International Association of Lowland Technology (IALT): ISSN 1344-9656

Technical Note

Comparison of major statistical methods and their combination using matrix validation for landslide susceptibility mapping

A.Q. Akbar[1] and G. Chen [2]

ARTICLE INFORMATION

Article history:

Received: 26 November, 2017
Received in revised form: 12 July, 2018
Accepted: 15 July, 2018
Publish on: 07 December, 2018

Keywords:

Kabul City
Landslide
Statistical Methods
Combination
Matrix Validation

ABSTRACT

Landslide risk exists with the mountain regions and every year creates a great life and financial losses. To prevent the disaster, numbers of statistical methods have been proposed, but it is still unclear which one is more accurate and yet very few studies proposes a reliable method. Therefore, this study aims to compare the commonly used bivariate statistical method and multivariate statistical methods and their combination to achieve higher accuracy for landslide susceptibility map. Moreover, the classification used for landslide susceptibility mapping is associated with errors, which affects the accuracy of the analysis. In this study, new tool was designed to reduce the classification. To implement this study, a landslide susceptibility maps were created Kabul city. The result proposes that the new designed tool is a good way not only to reduce the classification error by defining the critical thresholds for the classifications. Moreover, all of the statistical methodologies is giving and acceptable result but the combination bivariate and multivariate statistical methods increase the accuracy of the analysis and they are complimentary to each other.

1. Introduction

Landslide risk exists with the mountainous regions and every year causes a great life and financial losses (Dhital and M.R. 2017). Worldwide around 300 million people are at the risk of slope failure, which about 66 million of those people are living in the areas with the high risk (Dilley et al. 2005), however, from 2003 to 2010 the phenomenon had around 70000 fatalities and year by year the fatal landslides are increasing (Petley and Dave 2011).

To prevent the disaster, numbers of statistical methods have been proposed, but it is still not clear if you want one specific and reliable method for the landslide susceptibility mapping and yet very few studies have been done to compare the existing methods (Yilmaz and Işık 2009), (Isik and Yilmaz 2010), the proposed methodologies has its merits and demerits, therefore, this study is to compare the commonly used statistical methods “Frequency Ratio (FR), Weight of Evidence (WOE), Logistic Regression (LR)" (Lee, Saro , 2007) (Menard 1995)(Yilmaz and Işık 2009)., And their combination to achieve higher accuracy for the landslide susceptibility mapping.

Only finding one specific reliable method is not enough, the accuracy of an analysis not always depend on the methodologies or the quality of the raw data used in the study but it is the classification, which plays an important role in displaying a susceptibility mapping. Differentiating hazardous classes from the non-hazardous class normally natural break or other classification method

[1] Department of Civil and Structural Engineering, Kyushu University, Fukuoka 819-0395,JAPAN, qasim1368@live.com
[2] Department of Civil and Structural Engineering, Kyushu University, Fukuoka 819-0395,JAPAN, chen@civil.kyushu-u.ac.jp
Note: Discussion on this paper is open until June 2019

discussed in this paper are used which some unintentional errors are associated while selecting the hazardous classes, therefore, to overcome with the problem a new tool was designed to reduce the classification error.

To implement this study, Kabul city, the world's fastest growing city and the capital of Afghanistan is chosen. The city has the area of 275 km^2 and is the home for around 4.6 million people (GeoHive 2015). Due to unplanned housing, improper drainage system, overloading, cutting, saturating, and removing the vegetation, the risk of slope failure has increased in the mountains (Dhital and M.R. 2017). As an example, the February 2015 rock fall blocked Kabul-Jalalabad highway choked for the traffic(KhaamaPress 2015). Furthermore, unplanned housing in the hillside (**Fig. 1**) exposed the people to the risk which even the movement of a boulder can be catastrophic.

Fig. 1. The Kabul Residential areas (Wikipedia).

A landslide inventory map was created using the visual interpretation technique and the landslide susceptibility map using the "Frequency Ratio (FR), Weight of Evidence (WOE), Logistic Regression (LR)" and the combination of these methods considering the landslide factors such as "elevation, slope angle, geology, etc." was created in the GIS platform.

To validate the result number of control points from the landslide and randomly from the non-landslide areas are selected and a contagious table is built, and since the matrix validation result depend on the threshold values therefore a new tool was designed to specify the accuracy. The result from the matrix validation proposes that combination of the bivariate statistical methods "Frequency Ratio, Weight of Evidence" and multivariate statistical method "Logistic Regression" increase the accuracy of the analysis and it is more reliable than the methods alone. Moreover, because of 78.511%of the success rate, the combined method of Frequency Ratio (FR) and Logistic regression (LR) is reliable for the study area.

Furthermore, to investigate the classification error the combined method of Frequency Ratio (FR) and Logistic regression (LR) is used. The result shows that the natural break classification used to classify the landslide susceptibility map not only does not display the highest accuracy of the method but it has about 32.654% miss alarm rate. In the other hand, normally in a regression analysis 0.5 is the critical boundary to separate the landslide from the non-landslide, the classification represent the highest accuracy for the analysis but it has around 16.097% of miss alarm rate which is problematic. The designed tool to define the classification accuracy not only the miss alarm rate was reduced to 10.968% but it gives a numerical argument to classify the map in an acceptable way for an area.

2. Relevant Data

Mainly, two types of data are used in this study. 1: Categorical such as (lithology, faults, slope aspect, roads, rivers, and landcover). 2: Scale such as (elevation, slope angle, curvature, stream power index (SPI), topographic wetness index (TWI), and precipitation).

The data layers such as slope aspect, elevation, slope angle, curvature, stream power index (SPI), topographic wetness index (TWI) layers were extracted from a digital elevation model (DEM) and the landcover layer Landsat imagery which both of the images are with 30m of resolution, are taken from USGS Earth Explorer archive (Earthexplorer, n.d.). The lithology (rocks), faults, roads, and rivers layers were taken from the open file report available in Afghan Geological Survey (Operations et al. 2014). Moreover, the precipitation layer was downloaded and used from the Global Precipitation Climatology Centre ("Precipitation Data," n.d.).

To proceed the analysis, a landslide inventory map was created using the visual interpretation of the high-resolution GIS base map and validated with Google Earth owing to its 3D view and high resolution. Considering the landslide indicators such as differences in sediments color, the roughness of the structure, and sharp contacts, about 413 landslides were detected in the area during the interpretation.

2.1 Categorical data:

It is accepted that the local geology and topography can increase the possibility of slope failure because both the mention characteristics of site has significant influence on the earthquake ground motion in a particular place (Sharma, 2017).

Rocks and Sediments in the study area show three different class such as sedimentary rocks "conglomerate, sandstone, limestone, gravel stone, sand, clay, loess, etc"

however, the sedimentary rocks are divided into two subclasses, loose sedimentary rocks, and compressed sedimentary rocks. Metamorphic rocks “marble, gneisses, mica, schist, biotite, and quartzite”, and igneous rocks “granite, gabbro, peridotite, phyllite, and diorite”.

Geological faults are considered as main triggering factors in slope failure, and because of tectonic activity of Indian plate, Arabian plate and Eurasian plate the area is suffering from various numbers of normal faults, buried and proven. The area around faults are classified into five different classes (≤500m, 500 – 1000m, 1000 – 1500m, 1500 – 2000m, and ≤2000m).

Indirectly, slope aspect indicates the slope instability based on the influence of the related factors such as exposure to the sunlight, exposure to the wind and soil moisture which can cause a landslide (Zelano, Magliulo, and Paolo 2008)(Dick et al. 2011). The aspect is divided clockwise into nine classes (flat, north – east1, north – east2, south-east1, south-east2, south-west1, south-west2, north-west1, and north-west2).

Constructing a road normally carries extensive excavation, overloading or removing vegetation which most of the times unstablizes the slope and causes slope failure (Highland and Bobrowsky 2008). The area around the roads are divided into five different classes (≤20m, 20 - 50m, 50 – 100m, 100 - 150m, 150 - 200m, and ≥200m).

Landslide and flood have a close relationship because both are related to the precipitation, surface runoff, and the amount of water in the river. Basically, slope saturation increases the possibility of the landslide (Highland and Bobrowsky 2008). The area around the river is buffered based on its distance from the river and divided into six different classes (≤50m, 50–100m, 100–150m, 150–200m, 200–250m, and ≥250 m).

The land cover for the study area is created using the Normalized Difference Vegetation Index (NDVI) band combination in GIS platform and the image resolution. It was classified into six classes (water, vegetation, bare ground, urban area, wetlands, and the snow cover).

2.2 Scale data

Elevation and relief illustrate the potential energy for the mass wasting (Ghimir. and M. 2001)(Oguchi Takashi 1997). The elevation is extracted from DEM and divided into eight different classes (≤1000m, 1000–1500m, 1500–2000m, 2000–2500m, 2500–3000m, 3000–3500m, 3500 – 4000, and ≥4000m).

The slope is a primary factor in the dynamics of processes governing land evolution and landslide and it is used as the main triggering factor of the landslide (Bourenane et al. 2015)(Sharma et al. 2017). Based on the physical property of sediments, different sediments react differently to the slope angle but in general, as much as the slope angle raises, the possibility of slope failure rises. The slope is extracted from the DEM and divided into seven different classes (≤5^0, 5^0–10^0, 10^0–15^0, 15^0–20^0, 20^0–30^0, 30^0–45^0 and ≥45^0).

Curvature represents the morphology of an area. Generally, it is classified into three classes: 1, convex; 2, concave; 3, planar (straight). The concave is considered as a potentially unstable, unlike the convex and planar which is more stable for the sliding (Stocking and M.A 1972).

Stream Power Index (SPI) is the measure of erosive power associated with flowing water and it can be calculated using (**Eq.1**).

$$SPI = {A \cdot tan\beta}/{b} \quad [1]$$

A is the flow accumulation, β is the slope angle, and b is the width of a cell through which water flows. Higher SPI value should correspond to a higher likelihood of erosion on the landscape (Wilson et al. 2000). The SPI values are divided into three classes (low, moderate, high).

The Topographic Wetness Index (TWI) is a steady-state wetness index. In some areas, TWI predicts solum depth (Gessler et al. 1995).

$$TW = ln\left({A}/{tan\beta}\right) \quad [2]$$

A is flow accumulation and β represents the slope. Higher TWI values represent drainage depressions however lower values represent crests and ridges. The values are classified into three different classes (low, moderate, high)

Precipitation is considered as the primary triggering factor for the landslides (Bourenane et al. 2015). The more precipitation the more saturation, which leads a slope to the failure. The values are divided into five different classes (very low, low, moderate, high, and very high)

3. Methodologies

3.1 Frequency Ratio (FR) Approach

Frequency Ratio (FR) method is based on the distribution of landslide in each factor’s class (Lee et al. 2007). It is normally using the ratio of landslide area in a class to the total landslide ratio in the area. To find the relation between landslide occurrences in each factor for each factor’s class, a database was developed and FR method was applied (**Eq.3**).

$$Fr = ln\left(\frac{Con - prob}{Prior - prob}\right) = ln\left(\frac{\frac{LPC}{TPC}}{\frac{TLP}{TP}}\right) \quad [3]$$

(LPC) is the number of landslide cell in each class. (TPC) is the total number of the cell in the class, (TLP) is the total number of the landslide cell in all the area and (TP) is the total number of the cell.

To Calculate the Fr weight for each Class of the Factors a code was designed in the Matlab Environment and the result is given as (**Table 1**) below.

Table 1. Result of the Fr weight calculation using the designed MATLAB code.

Factor	Classes	Total PIX	Landslide	Fr Weight
Elevation (m)	< 1000	18981	569	0.282431694
	1000 - 1500	460398	9914	0.138754003
	1500 - 2000	2436862	14158	-0.430190511
	2000 - 2500	1574604	27752	0.051761454
	2500 - 3000	678557	17739	0.222981316
	3000 - 3500	272715	13300	0.493781132
	3500 - 4000	92574	3416	0.37266735
	4000 <	21897	82	-0.620932249
Slope (Degree)	< 5	1635442	1952	-1.117516816
	5 - 10	949008	2527	-0.769026115
	10 - 15	668250	3532	-0.471279755
	15 - 20	584011	5679	-0.206510641
	20 - 30	964208	21051	0.144740518
	30 - 45	690192	40398	0.573028461
	45 <	65477	11791	1.061100384
Aspect	Flat	56076	136	-0.809599602
	North - East 1	750569	9479	-0.092989583
	North - East 2	766719	6658	-0.255653918
	South - East 1	790834	5442	-0.356688278
	South - East 2	678786	7783	-0.134947329
	South - West 1	606647	12800	0.129912432
	South - West 2	594919	16843	0.257600126
	North - West 1	666770	15719	0.17808738
	North - West 2	645268	12070	0.077605659
Curvature	Concave	887475	24330	0.243624498
	Convex	3780092	37303	-0.200120093
	Planar	889021	25297	0.259795518
Distance to the River (m)	< 50	123069	1068	-0.255941638
	50 - 100	121635	1036	-0.264063023
	100 - 150	118764	1063	-0.24251577
	150 - 200	115967	1096	-0.218888087
	200 – 250	112899	1085	-0.211624576
	250 <	4964219	81582	0.021379197
Precipitation	Very Low	2053416	16884	-0.279365113
	Low	1381316	20017	-0.033257568
	Moderate	1180976	21512	0.066076202
	High	707240	22723	0.312535338
	Very High	233614	5794	0.20011611
SPI	Low	2137072	23512	-0.152891064
	Moderate	2074691	24054	-0.130127599
	High	1344825	39364	0.272071966
TWI	Low	2467942	60330	0.193836893
	Moderate	2408702	22181	-0.230163434
	High	679943	4419	-0.381509988
	No Data	1	0	0
Lithology	Metamorphic Rocks	1434876	30624	0.134883326
	Loose Sedimentary Rocks	2094525	5655	-0.763017158
	Igneous rocks	591616	10421	0.051505267
	Compressed Sedimentary Rocks	1435536	40230	0.253171726
Distance to the Fault (m)	< 500	419595	10919	0.220988343
	500 - 1000	399770	12446	0.298855378
	1000 - 1500	391154	10112	0.218125067
	1500 - 2000	370844	6543	0.052221445
	2000 <	3975190	46910	-0.122456677
Land Cover	Water	22129	111	-0.495008143
	Vegetation	684056	240	-1.650249748
	Bare Ground	4582208	84882	0.072371472
	Urban Area	229203	1520	-0.373746043
	Wetlands	1189	54	0.461842572
	Ice Cover	24923	123	-0.502064539
Distance to the Road (m)	200<	4719431	83129	0.051837618
	<20	98259	153	-1.001706206
	20-50	140347	300	-0.864107176
	50-100	217154	832	-0.610669804
	100-150	200262	1167	-0.42855299
	150-200	185438	1349	-0.332212082

Once the Fr weight is calculated for each Factors class, it would be taken into GIS Platform to display each factors susceptibility to the landslide and each factors class has its specific weight as (**Table 1**). Furthermore, to create the landslide susceptibility map, the calculated Fr weights are summed (**Eq.4**).

$$LS_{FR} = \sum Fr_1 + Fr_2 + Fr_3 + \cdots + Fr_n \quad [4]$$

LS is the Landslide Susceptibility Index, (Fr) is the Fr Weight of each factor's classes. LS is representing the relative hazard of landslide occurrence.

The higher result values, the higher risk of slope failure (Lee et al. 2007), therefore, the results were classified into five different classes "Very Low, Low, Moderate, High, Very High" which represents the level of unstable locations in the maps. This information is applicable to all the methods (**Fig. 2**).

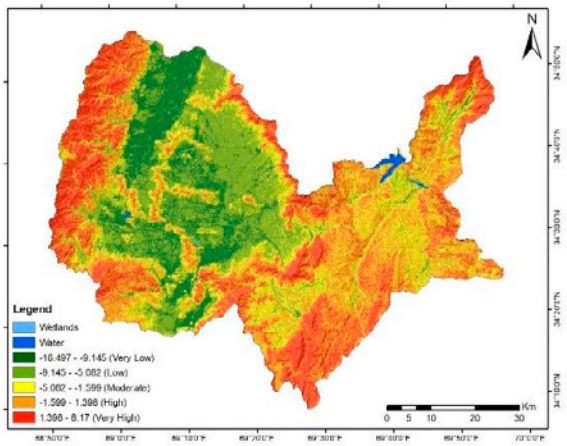

Fig. 2. Result of the frequency ratio method for the landslide susceptibility mapping of Kabul, Afghanistan.

3.2 Weight of Evidence (WOE) Approach

To evaluate the contribution of each factor towards landslide hazard, the existing landslide distribution, data layer was compared to various thematic data layers separately. (Netra et al. 2009).

$$W_i^+ = ln\left(\frac{\frac{NPix_1}{Npix_1 + Npix_2}}{\frac{NPix_3}{Npix_3 + Npix_4}}\right) \quad [5]$$

$$W_i^- = ln\left(\frac{\frac{NPix_2}{Npix_1 + Npix_2}}{\frac{NPix_4}{Npix_3 + Npix_4}}\right) \quad [6]$$

$$(\text{Weight} = W_i^+ - W_i^-) \quad [7]$$

where Npix1 is the number of pixels representing the presence of both potential landslide predictive factor and landslides, Npix2 is the number of pixels representing the presence of landslides and absence of potential landslide predictive factor, Npix3 is the number of pixels representing the presence of potential landslide predictive factor and absence of landslides, Npix4 is the number of pixels representing the absence of both potential landslide predictive factor and landslides.

Considering the equations above for the weight of evidence and to calculate the weight for each class of the Factors a code was designed in the Matlab environment and the result is given as (**Table 2**) below.

Table 2. Result of the WOE weight calculation using the designed MATLAB code.

Factor	Classes	Count	Landslide	Weight
Elevation (m)	< 1000	18981	569	0.668186
	1000 - 1500	460398	9914	0.360633
	1500 - 2000	2436862	14158	-1.40779
	2000 - 2500	1574604	27752	0.173343
	2500 - 3000	678557	17739	0.623628
	3000 - 3500	272715	13300	1.288664
	3500 - 4000	92574	3416	0.903583
	4000 <	21897	82	-1.44482
Slope (Degree)	< 5	1635442	1952	-2.91978
	5 - 10	949008	2527	-1.94435
	10 - 15	668250	3532	-1.18372
	15 - 20	584011	5679	-0.5257
	20 - 30	964208	21051	0.4276
	30 - 45	690192	40398	1.862486
	45 <	65477	11791	2.761952
Aspect	Flat	56076	136	-1.88623
	North - East 1	750569	9479	-0.24731
	North - East 2	766719	6658	-0.66564
	South - East 1	790834	5442	-0.92053
	South - East 2	678786	7783	-0.35205
	South - West 1	606647	12800	0.349045
	South - West 2	594919	16843	0.709764
	North - West 1	666770	15719	0.490875
	North - West 2	645268	12070	0.208258
Curvature	Concave	887475	24330	0.729589
	Convex	3780092	37303	-1.05898
	Planar	889021	25297	0.783336
Distance to the River (m)	< 50	123069	1068	-0.60658
	50 - 100	121635	1036	-0.62555
	100 - 150	118764	1063	-0.57464
	150 - 200	115967	1096	-0.51882
	200 - 250	112899	1085	-0.50149
	250 <	4964219	81582	0.606446
Precipitation	Very Low	2053416	16884	-0.90057
	Low	1381316	20017	-0.10223
	Moderate	1180976	21512	0.200833
	High	707240	22723	0.905817
	Very High	233614	5794	0.496563
SPI	Low	2137072	23512	-0.52983
	Moderate	2074691	24054	-0.44966

	High	1344825	39364	0.970701
TWI	Low	2467942	60330	1.059366
	Moderate	2408702	22181	-0.81517
	High	679943	4419	-0.96736
	No Data	1	0	0
Lithology	Metamorphic Rocks	1434876	30624	0.453983
	Loose Sedimentary Rocks	2094525	5655	-2.18382
	Igneous rocks	591616	10421	0.135954
	Compressed Sedimentary Rocks	1435536	40230	0.922458
Distance to the Fault (m)	< 500	419595	10919	0.576013
	500 - 1000	399770	12446	0.785072
	1000 - 1500	391154	10112	0.564129
	1500 - 2000	370844	6543	0.131602
	2000 <	3975190	46910	-0.7767
Land Cover	Water	22129	111	-1.15334
	Vegetation	684056	240	-3.94642
	Bare Ground	4582208	84882	2.179522
	Urban Area	229203	1520	-0.89471
	Wetlands	1189	54	1.09452
	Ice Cover	24923	123	-1.17004
Distance to the Road (m)	200<	4719431	83129	1.374068
	<20	98259	153	-2.33703
	20-50	140347	300	-2.02576
	50-100	217154	832	-1.44874
	100-150	200262	1167	-1.02022
	150-200	185438	1349	-0.79197

Once the weight of each factor was calculated using the above equation, with the simple summation of all the factors the Landslide Susceptibility indexation map would be extracted using the (**Eq.8**).

$$LS_{WOE} = Wc_1 + Wc_2 + Wc_3 + \cdots + Wc_n \qquad [8]$$

LS is the Landslide Susceptibility Index, (Wc) is the weight of each factor's classes. LS is representing the relative hazard of landslide occurrence (**Fig. 3**).

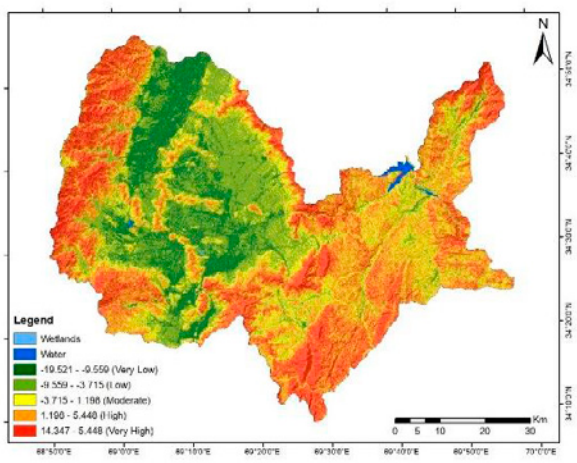

Fig. 3. Result of the weight of evidence method for the landslide susceptibility mapping of Kabul, Afghanistan.

Both of the bivariate statistical methods has some advantages such as, the model can identify the influence of each class within the factor on landslides. Moreover, the method can be used for both scale and categorical factors. But in another hand, it cannot identify the possible relationship between the factors (the relation between the slope angle and lithology, which both factors are really important in slope failure prediction) and all the factor which will be used in the analysis should be conditional independent.

3.3 Logistic Regression (LR) Approach

The principle of logistic regression (LR) rests on the analysis of a problem, in which a result measured with variables such as 0 and 1 or true and false, is determined from one or more independent factors (Menard 1995). In the case of landslide susceptibility mapping, the goal of LR would be to find the best fitting model to describe the relationship between the presence or absence of landslides in a set of independent parameters such as slope angle, aspect, lithology, etc.

LR does not define susceptibility directly like WOE and FR approaches but an inference can be made using the probability. One of the biggest limitations of this method is that the method cannot be calculated for the categorical data. Therefore, the categorical data in this method was removed from the calculation, and however, it has a significant effect in to the result. Generally, LR involves fitting the dependent variable using an equation below.

$$Y = \ Logit(p) = ln\left(\frac{p}{1-p}\right) = C_0 + C_1X_1 + C_2X_2 + \cdots + C_nX_n \quad [9]$$

Where p is the probability that the dependent variable (Y) is 1, $\frac{p}{1-p}$ is the so-called odds or likelihood ratio, $C_0$is the intercept, and $C_1, C_2 \ldots\ldots C_n$ are coefficients, which measure the contribution of independent factors $X_1, X_2 \ldots\ldots X_n$to the variations in Y.

Considering the above equation, a code was designed in the Matlab platform and the C_0 is the intercept, and coefficients for each factor was calculated (**Table 3**)

Table 3. Logistic regression coefficient calculated using the designed MATLAB code.

Independent factors	Coefficients
Curvature	-0.0795
Elevation	0.0034
Faults	0.0001
Precipitation	0.1806
River	-0.0005
Road	0.0612
Slope	0.105
SPI	0.0217
TWI	-0.003
Constant	0.0228

Once we got the Constant and the coefficients with the simple summation and multiplication of the independent factors with its Coefficients, we can get the result where Y is representing the relative hazard of landslide occurrence (**Fig. 4**).

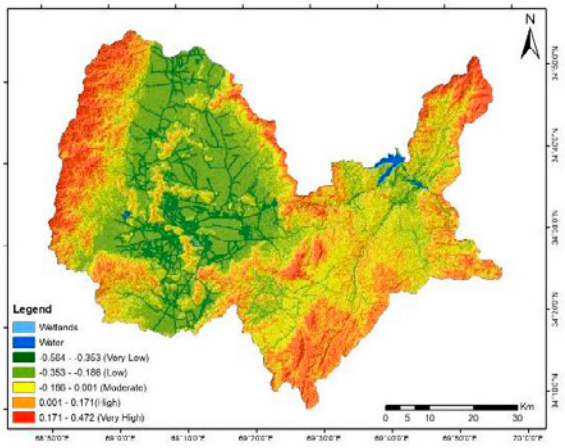

Fig. 4. Result of the logistic regression method for the landslide susceptibility mapping of Kabul, Afghanistan.

The problems on identifying the relationship between factors which existed in bivariate statistical method has been solved in model and the output values represents a meaningful probability in a susceptibility map but the model has some weakness such as, the model cannot identify the influence of each class with a factor on landslide and the categorical data can be calculated or if the categorical data has been changed to scale but still too much of it can create immense problems.

4. Combination of methods

The bivariate analysis is a quantitative method that applies bivariate data and then makes comparisons in order to find any significant relationships. Meanwhile, multivariate analysis is a method that simultaneously observes and analyzes two or more variables of interest (Thi-To-Ngan, Nguyen, and Liu 2014) and to overcome with the strength and weakness of the bivariate and multiversity statistical methods, the combination is used and in the process the above mentioned problems will be solved or in another hand, they can become complimentary for each other. As an example, the multivariate has the limitation of using the categorical data but in bivariate methods such a problem does not exist. Therefore, in this study, we have used the Combination technique to increase the accuracy of the analysis.

4.1 Combination of Frequency ratio method and Logistic regression (FR + LR)

Combination of the methods follows the same path as the main approach follows. In the combination of frequency ration method, firstly the distribution of landslide in each factor's class will be calculated. It is normally using the ratio of landslide in a class to the total landslide ratio in the area (**Eq.3**).

Once the Fr Weight index for each factor's class is calculated then sample points will be taken from the data for the logistic regression method. The combined method is following the same role as the Logistic regression method.

$$Y_{FR} = C_0 + C_1Fr_1 + C_2Fr_2 + \cdots + C_nFr_n \quad [10]$$

C_0 is the constant, and $C_1, C_2 \ldots\ldots C_n$ are coefficients, which measure the contribution of independent factors $Fr_1, Fr_2 \ldots\ldots Fr_n$to the variations in Y. To get the constant and the coefficient for the equation, the designed code

which was used to calculated the logistic regression approach is used (**Fig. 5**).

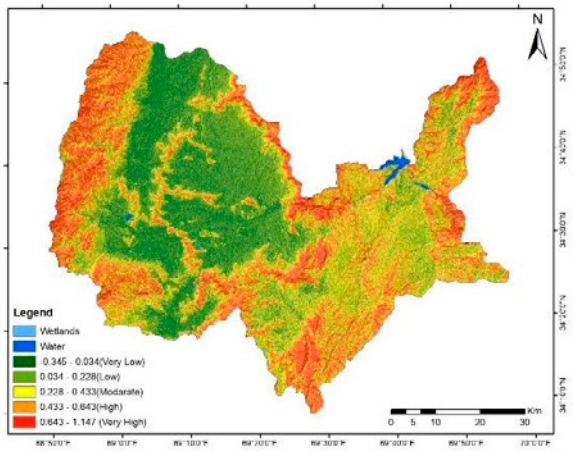

Fig. 5. The combination of frequency ratio and logistic regression methods for the landslide susceptibility mapping of Kabul, Afghanistan.

4.2 Combination of Weight of Evidence and Logistic regression (WOE+LR)

Similarly, the combine method of weight of evidence and logistic regression method follows the same role. First, the weight will be calculated from each factors class, same as the weight of evidence method (**Eq. 5, 6, 7**). Once the weight is calculated (**Table 2**), then the result of the weight of evidence method will be used in the logistic regression method to find the best fit model to separate the landslide from non-landslide in a set of independent parameters.

$$Y_{WOE} = C_0 + C_1\text{Weight}_1 + C_2\text{Weight}_2 + \cdots + C_n\text{Weight}_n \quad [11]$$

C_0 is the intercept, and $C_1, C_2 \ldots\ldots C_n$ are coefficients, which measure the contribution of independent factors $\text{Weight}_1, Weight_2 \ldots\ldots C_n Weight_n$ to the variations in Y(**Fig. 6**).

The higher result values, the higher risk of slope failure (Lee et al. 2007), therefore, the results were classified into five different classes “Very Low, Low, Moderate, High, Very High” which represents the level of unstable locations in the maps. Moreover, visually looking at the maps, it shows that all past occurred landslides were located in the two “high and very high risk” classes which indicate the high accuracy of the analysis.

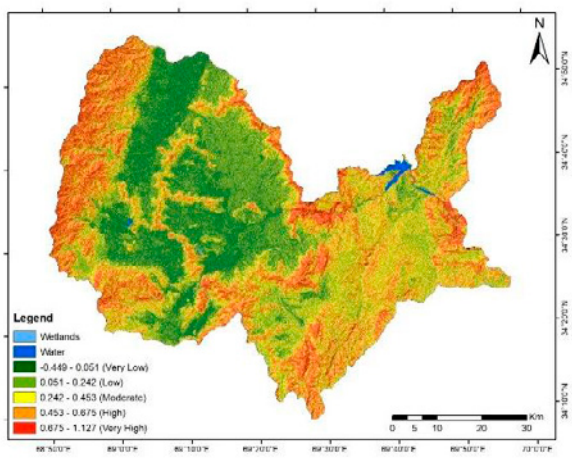

Fig. 6. The combination of weight of evidence and logistic regression methods for the landslide susceptibility mapping of Kabul, Afghanistan.

5. Validation Method and New Designed Tool

To define the accuracy of the result, the matrix validation method is used, this method not only gives the success rate of a methodology but it gives the false alarm rate and miss alarm rate. To operate the matrix validation method uses a contingency table (**Table 4**) build from the control point as below.

Table 4. Contingency table used to validate the result.

		Predicted	
	Total	Landslide	No landslide
Actual	Landslide	Success(A)	Miss-alarm(B)
	No landslide	False-alarm (C)	Success(D)

Considering the contagious table, three indexes (1) success rate (**Eq.12**), (2) miss alarm rate (**Eq.13**) and (3) false alarm rate (**Eq.14**) can be evaluated for efficiency.

$$Success\ Rate = \frac{A+B}{A+B+C+D} \quad [12]$$

$$Miss\ Alarm\ Rate = \frac{B}{A+B} \quad [13]$$

$$Fale\ Alarm\ Rate = \frac{C}{A+C} \quad [14]$$

Success rate shows the percentage of the points which are correctly classified, miss alarm rate shows the percentage of the points which are landslide occurrence but predicted as a non-landslide which is an important rate for landslide hazard mapping. The higher the miss alarm rate values, the higher number of landslide points are predicted as non-landslide. On the other hand, false – alarm rate shows the percentage of the non-landslide

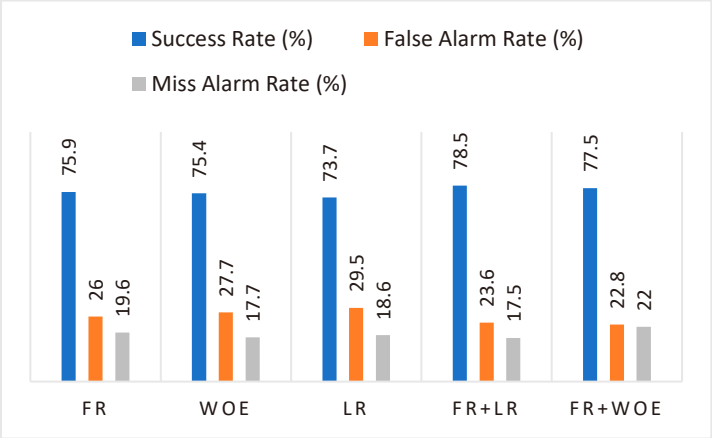

Fig. 7. Validation result of analysis using the new designed tool.

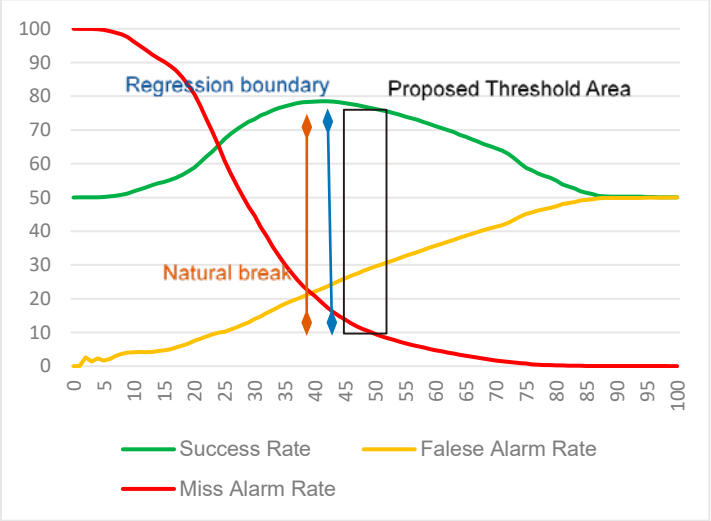

Fig. 8. The influence of threshold value on accuracy of an analysis.

points incorrectly classified as a landslide. The higher the false alarm rate, the higher false information in the landslide hazard prevention.

In addition, the accuracy of a landslide susceptibility map also depends on the selection of the critical values, has equally important role. Generally, commonly classification method used in landslide susceptibility mapping classification are associated with errors that affects the accuracy of the analysis. The mentioned classifications cannot clarify the information needed for the susceptibility map classification. To reduce the error and define the critical boundary for the classification a tool was designed in the Matlab program. The new tool needs the landslide susceptibility map values in a set of the point equally divided in landslide and randomly from non-landslide. The tool naturalize the values and calculated the three above rate for all of threshold values from 0% to 100%.

6. Result and Discussion

To evaluate the accuracy of the analysis number of control points equally divided from the landslide and randomly from the non-landslide area are chosen and with the use of new tool the above three rate was calculated. To compare the methodologies the highest success rate of all of the methodologies were considered (**Fig. 7**).

The result (**Fig. 7**) show that all the methodologies used in this study is giving an acceptable accuracy but the combination of the bivariate and multivariate statistical method increases the accuracy of the analysis and they are more reliable than the methods alone. From the result

it has been clarified that the combination method of Frequency Ratio (LR) and Logistic Regression method (LR) is more reliable because of its higher success rate and lower miss alarm rate.

Furthermore, the used classifications for a susceptibility maps are not enough because they are associated with the higher miss alarm rate and it does not give the critical boundaries for the classification. The classification has to be with an acceptable success rate, miss alarm rate, and false alarm rate. The new designed tool allows the user to define suitable classification thresholds for the susceptibility map (**Fig. 8**).

From the **Fig. 8** It is clear that the threshold used in the natural break classification of GIS not only it is not displaying the high accuracy but has 32.654 % of miss alarm rate. Similarly, the 0.5 threshold which is the regression critical boundary to separate the landslide from the landslide is giving a higher accuracy than the natural break classification but it is still not the highest accuracy and the boundary is associated with around 16.097 % of miss alarm rate. In a landslide susceptibility mapping the higher the miss alarm rate the higher miss prediction which still do not answer the need of a landslide susceptibility map by not having low miss alarm rate and not specifying the rest of the classification boundary.

The result of designed tool, not only the critical boundary can be selected based on the study propose in an area but it is giving an idea how to select the reset of classification boundaries, as an example If the study area is in a city with a high population then the miss alarm rate has to lower. On the other hand, for a rural area the higher miss alarm rate is acceptable unless there is no big financial threat.

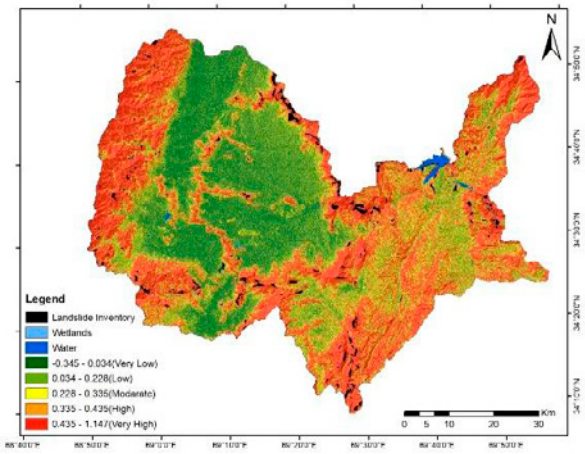

Fig. 9. Modified Map of Combined Frequency Ratio and Logistic Regression Methods for the Landslide Susceptibility Mapping of Kabul, Afghanistan.

Figure 9 is reclassified map considering the miss alarm rate. The miss alarm rate used up to 10% is considered as a very high risk zone, 5% as a high risk zone, 1% moderate zone, 0.5% low, and the rest is very low. Visually compering the modified result with the (**Fig. 5**) the different can easily been seen because some of the areas which previously was in the moderate risk zone was now in the high and very high risk zone.

7. Conclusion

Every year thousands of people die, injure and lose their assets because of the unexpected landslide and the phenomena not only affect people but also affects the economy, damage buildings, lifeline and it will damage everything that comes to its path. It is impossible to stop a slope from the failure but there are ways to mitigate or reduce the risk of slope failure.

To take a step towards the hazard management and to mitigate the disaster, numbers of statistical methods have been proposed however very few proposed one specific reliable method, therefore, this study was performed in search of one reliable method with high accuracy for landslide susceptibility mapping. From results, it can is concluded that, all of the used methodologies give an acceptable result but the combination of bivariate and multivariate statistical method gives a higher accuracy for analysis and they are complimentary for each other.

From this study, it has been found that the classification boundary (threshold values) is playing an important role in displaying the result on a susceptibility map. The commonly used classification method does not present the best results because high miss alarm rate may be caused. The new developed tool to calculate the success rate, miss alarm rate, and false alarm rate for all of the threshold values, insures to make a desired map by focusing on the miss alarm rate or the false alarm rate in an specific area since higher miss alarm rate for high risk zone of a rural area may be acceptable but the miss alarm rate should be smaller for a city area.

Finally, a practical landslide susceptibility maps for the Kabul city has been made based on the analysis of combination FR and LR method and using the newly developed tool for determining threshold values. It is expected that the map will be helpful in landslide prevention plan of the area in future.

Acknowledgments

The authors of this paper sincerely thanks Mr. Mohammad Idrees Ahmadi from Missouri University of Science and Technology for his useful contribution on literature review.

References

Bourenane, Hamid, Bouhadad, Youcef, Mohamed Said Guettouche, and Massinissa Braham. 2015. “GIS-Based Landslide Susceptibility Zonation Using Bivariate Statistical and Expert Approaches in the City of Constantine (Northeast Algeria).” Springer Berlin Heidelberg 74 (2): 337–55.

Dhital, and M.R. 2017. “Geomorphic Approach of Controlling Mass Movements on Tama Koshi Road in Central Nepal, Lowland Technology International.” International Association of Lowland Technology (IALT) 18 (4): 283–96.

Dick, Dieu, Tien, and Bui. 2011. “Landslide Susceptibility Analysis in the Hoa Binh Province of Vietnam Using Statistical Index and Logistic Regression.” Natural Hazards 59 (3): 1413. https://doi.org/10.1007/s11069 – 011 – 9844 – 2.

Dilley, Maxx, Chen, Robert S, Deichmann, Uwe, Lerner-Lam, et al. 2005. “Natural Disaster Hotspots: A Global Risk Analysis.” Washington, DC: World Bank. 2005. http://documents.worldbank.org/curated/en/621711468175150317/Natural-disaster-hotspots-A-global-risk-analysis.

Earthexplorer. n.d. “USGS Earthexplorer Archive.”

GeoHive. 2015. “Kabul Extendet Population Statistics.” Wikipedia. 2015. https://en.wikipedia.org/wiki/Kabul#cite_note-Factbook-1.

Gessler, Paul E, Moore, I.D., McKenzie, Neil James, and P.J.Ryan. 1995. “Soil-Landscape Modeling and Spatial Prediction of Soil Attributes.” Geographical Information Systems 9 (4): 421–32. https://doi.org/10.1080/02693799508902047.

Ghimir., and M. 2001. “Geo-Hydrological Hazard and Risk Zonation of Banganga Watershed Using GIS and Remote Sensing.” Nepal Geological Society 23: 99–110.

Highland, LM, and P Bobrowsky. 2008. “The Landslide Handbook — A Guide to Understanding Landslides.” US Geological Survey Circular, 129. https://doi.org/Circular 1325.

Isik, and Yilmaz. 2010. “A Comparative Study of Frequency Ratio, Weights of Evidence and Logistic Regression Methods for Landslide Susceptibility Mapping: Sultan Mountains, SW Turkey” 61 (4): 821–36.

KhaamaPress. 2015. “Rock Fall Blocks Kabul-Jalalabad Highway.” 2015. https://www.khaama.com/rock-fall-blocks-kabul-jalalabad-highway-29112.

Lee, Saro, Pradhan, and Biswajeet. 2007. “Landslide Hazard Mapping at Selangor, Malaysia Using Frequency Ratio and Logistic Regression Models.” Landslides 4 (1): 33–41. https://link.springer.com/article/10.1007/s10346-006-0047-y.

Menard. 1995. Applied Logistic Regression Analysis” Sage University Paper Series on Quantitative Applications in Social Sciences, Vol. 106. Thousand Oaks, California.

Netra, R.Regmi, John, R.Giardino, and John D.Vitek. 2009. “Modeling Susceptibility to Landslides Using the Weight of Evidence Approach: Western Colorado, USA.” Geomorphology 115 (1–2): 172–87.

Oguchi Takashi. 1997. “Drainage Density and Relative Relief in Humid Steep Mountains with Frequent Slope Failure.” Earth Surface Processes and Landforms 22 (2): 107–20.

Operations, Stability, Report, and Administrative. 2014. “U . S . GEOLOGICAL SURVEY AFGHANISTAN PROJECT PRODUCTS,” 1–70.

Petley, and Dave. 2011. “Global Deaths from Landslides in 2010 (Updated to Include a Comparison with Previous Years), American Geophysical Union.” American Geophysical Union. 2011. https://blogs.agu.org/landslideblog/2011/02/05/global-deaths-from-landslides-in-2010/.

“Precipitation Data.” n.d. Global Precipitation Climatology Center. http://gpcc.dwd.de.

Sharma, K., M. Subedi, R. R. Parajuli, and B. Pokharel. 2017. “Effects of Surface Geology and Topography on the Damage Severity during the 2015 Nepal Gorkha Earthquake.” Lowland Technology International 18 (4): 269–82.

Stocking, and M.A. 1972. “Relief Analysis and Soil Erosion in Rhodesia Using Multivariate Techniques.” Zeitchrift Für Geomorphologie 16: 432–43.

Thi-To-Ngan, Nguyen, and Cheng-Chien Liu. 2014. “Combining Bivariate and Multivariate Statistical Analyses to Assess Landslide Susceptibility in the Chen-Yu-Lan Watershed, Nantou,Taiwan, Sustain.” Environ 24 (2): 257–71.

Wilson, J.P., And, Lorang, and M.S. 2000. “Spatial-Models-of-Soil-Eroision-and-GIS.Pdf.” In New Potential and New Model, 83–86.

Yilmaz, and Işık. 2009. “Landslide Susceptibility Mapping Using Frequency Ratio, Logistic Regression, Artificial Neural Networks and Their Comparison: A Case Study

from Kat Landslides (Tokat-Turkey).” Computers & Geosciences 35 (6): 1125–38. https://doi.org/10.1016/j.cageo.2008.08.007.

Zelano, Magliulo, and Paolo. 2008. “Geomorphology and Landslide Susceptibility Assessment Using GIS and Bivariate Statistics: A Case Study in Southern Italy.” Natural Hazards 47 (3): 411–35. https://doi.org/https://doi.org/10.1007/s11069-008-9230-x Publisher Name Springer Netherlands Print ISSN 0921-030X.