Regional simulation and landslide risk prediction based on bivariate logistic regression (A case study: Pahne Kola watershed in northern Iran)

Elyyeh Moafi Madani a, Hossein Hasani a, Mehdi Sadighi Hashthijn b, Marzieh Hosseini-Nasab c,*

a Department of Mining and Metallurgy Engineering, Azirkaib University of Technology, Tehran, Iran
b Islamic Azad University of Noor, Science and Research Branch Tehran, Iran
c Department of Mining Engineering, University of Sistan and Baluchestan, Zahedan, Iran

ABSTRACT

This study aims to assess landslide susceptibility in Pahne Kola watershed located in south of Sari, based on bivariate logistic regression. For this purpose, the landslides distribution map of the area was firstly prepared in ArcGIS software, and then, eight effective factors on landslide events such as elevation, slope, slope aspect, rainfall, land use, distance from road, soil and geology were considered as independent variables. PGA was the same over the area, because the study area was small. The independent variables, including eight effective factors, included 61 sliding points as dependent variables, and accordingly, one was devoted to the presence and zero was devoted to the absence of landslide. After quantitative analysis, the relevant data were transferred to SPSS and after interpreting the coefficients, only the distance from road was recognized as a significant variable influencing the final equation and the other independent variables were omitted from the final equation due to lack of statistical correlation. After transferring the final probability equation to ArcGIS software, the landslide hazard map was prepared. Statistical accuracy of model was evaluated and approved by omnibus test, model summary table, classification graph and table. Statistical evaluation of the model showed that the overall accuracy of prepared map was 85.2%.

Keywords: ArcGIS software, PGA, Pahne Kola, omnibus test

1. Introduction

Landslide is a type of mass movement which is highly effective in changing the natural slopes, and when affected by human activities, it will be one of the most dangerous phenomena. The global annual casualties of landslide events are 1000 people with 4 million dollars economic loss [1]. Considering the casualties, economic loss, and environmental impact, landslide in Iran is one of the most important natural disasters having an effective role in destroying roads and producing sediment in the country's watersheds every year [2]. Therefore, it is crucial to estimate the regional landslide susceptibility in Iran [3]. In Pahne Kala watershed, several various landslides have occurred which have endangered many forest-industrial species and roads in the region. Therefore, it is necessary to identify not only the areas that are in crisis, but also the areas that are susceptible to landslide and to take effective measures to stabilize them.

Many efforts have been made by different researchers around the world and in Iran to offer proper models and methods for evaluating the landslide hazard. Moreover, different methods have been tested as landslide hazard zonation to study the landslide areas. Landslide hazard zonation mapping requires various qualitative and quantitative approaches. Some qualitative approaches are based on classification and weighting of the factors which can change to semi-quantitative approaches in nature. Since qualitative and semi-quantitative approaches vary based on the experts’ opinions, they are often appropriate for regional studies [4]. Qualitative approaches are based on numerical expression of the relationship between effective factors and landslides. Statistical approaches are among the quantitative approaches which include bivariate and multivariate statistical methods. In bivariate statistical methods, the importance of each factor in relation to the instability of slopes is analyzed independent of other factors [5-8] and data frequency (e.g. sliding surface and number of landslides occurred in an area) is used to calculate the occurrence probability [9, 10]. Application of multivariate statistical models are useful, because such models simultaneously analyze several variables, and natural phenomena like landslides are resulted from simultaneous performance of several variables, as well [11]. Among the multivariate statistical models for landslide susceptibility mapping, discriminant analysis, artificial neural network, fuzzy logic, and logistic regression could be referred to.

Khameneh et al. [12] mapped the landslide hazard zonation in the Seidfar Galle region in Semnan Province through logistic regression model by means of effective factors such as slope, lithology, slope aspect, digital elevation model, land use, rainfall, and distance from fault. Results showed that rainfall and the distance from fault were omitted from the model in final analysis due to their low impacts. On the other hand, slope and lithology were considered as the most important variables for estimating landslide hazard in the region. Shirzadi et al. [3] mapped the landslide hazard zonation using the logistic regression analysis and five effective factors, including distance from fault, lithology, shape, slope, slope angle, and distance from road which affects the rock fall. Mousavi Khatir et al. [13] mapped landslide hazard zonation in Sajandroud watershed. They referred to six factors including distance from road, slope aspect, lithology, and distance from drainage, land use, and slope angle as the most effective factors on landslide occurrence, and offered a regional model for the area of interest. Chen and Wang [14] mapped landslide hazard zonation in McKenzie Valley in Canada using logistic regression analysis in GIS environment and
effective factors such as geology, surface materials, land cover, and topography. Through interactive interference between geological and geomorphological conditions, they concluded that compared to other methods, the logistic regression model predicts the future landslides and the location of hazardous areas more precisely and prepares more accurate hazard zonation maps. Moreover, logistic regression model has been greatly noticed by researchers because of its tremendous capabilities in landslide modeling, in a way that the foundation of studies by Dai et al. (2001), Ohlamer and Davis (2003), Lee (2004, 2007), Ayalive and amagishi (2005), Chung (2006), Mathew et al. (2009), Pradhan (2011), Yalcin et al. (2011) on landslide were the logistic regression principles [1, 15-22]. In spite of numerous Iranian researches on landslide hazard modeling, application of a new statistical model such as logistic regression has not been properly examined. This research aims to study the factors affecting landslide, to model landslide by means of modern logistic regression method, and to map landslide hazard zonation based on regional model.

2. Study area and samples

The studied area is located in south of Sari in northern Iran. It is a branch of great basin of Tajan. The area of the region is 51.5 km², it is stretched and its average height is 430.49 m with an average slope of 11.4% [14]. In general, this region has sensitive formations which are mainly made up of Miocene marine and M₂₄m.s.l unit and in a small part of the region Q² (Alluvial Terraces) are seen in west and northwest of the region (A geological map of the area is shown in Fig. 1). According to Dommarten model, it has a humid climate and the average precipitation in the region is 780 mm. It should be added that snow comprises about 10-15% of total annual precipitation especially in altitudes more than 600 m above the sea level and at the end of fall and winter. A total of 9.83% of the studied area is covered by forests, and 83.61% by agricultural lands and 6.56% by gardens and destroyed forest lands [14]. The position of Pahne Kola watershed is shown in Fig. 2. At first, 61 landslides were identified and recorded through field operations and interviews with local people in Pahne Kola watershed of Mazandaran Province, Iran. Then the coordinates of landslide points were given to ArcGIS software and the landslide distribution map of the region was prepared by software (Fig. 2).

Fig. 1. The geological map of the area.

Fig. 2. Location of the study area and map of landslides distribution in Pahne Kola watershed.
3. Methodology

In this study, all effective factors on landslides were identified through fields studies and finally eight factors including elevation, slope, slope aspect, rainfall, land use, distance from road, soil and geology were considered as the final effective factors in landslide occurrence to be prepared in ArcGIS software. It should be noted that the raster map of digital elevation model was used to prepare the maps of elevation, slope percentage, slope aspect, and precipitation. Moreover, the geological maps were prepared through the 1:100000 geological map of Pol Sefid, the land use map was prepared by the 1:100000 land use map prepared by the former ministry of agriculture in 1996, the soil map was prepared through the 1:25000 soil map prepared by Mazandaran Wood and Paper Industry factory and the road network map was prepared by the 1:25000 topography map of sculpture and Lajym in ArcGIS software environment with different classifications (Fig. 3a-3h).

3.1. Landslide Hazard Modeling by Bivariate Logistic Regression Model

Logistic regression is a statistical approach of generalized linear statistical models which predicts the probable occurrence of an event by means of independent variables. The main point in logistic regression is that the dependent variable is a binary variable; that is, it could be just 0 which means failure and 1 which means occurrence of the event [23]. Two-dimensional nominal regression analysis is used when the dependent variable at nominal level is two-sided (two-alternative) and thus the presence or absence of a trait is predicted based on a set of dependent variables [24]. Considering the landslide susceptibility land, the aim of logistic regression is to find the best model to explain the relationship between the presence or absence of dependent variable (landslide) and a set of independent variables which affect the landslide occurrence.
Fig. 3d. Map of factors influencing landslide occurrence (soil map).

Fig. 3e. Map of factors influencing landslide occurrence (slope map).

Fig. 3f. Map of factors influencing landslide occurrence (slope aspect map).

Fig. 3g. Map of factors influencing landslide occurrence (rainfall levels map).
Table of variables in equation is the most important table in interpreting the results of the significance and the effect of each independent variable on dependent variable. There are several important statistics whose nature and functions are necessary to be known for interpreting the results:

B: This statistic is the estimated regression coefficient with standard error.
SE: This statistic is the standard error.

Wald: is used in interpreting the statistic results. If the statistic of each variable is less than 0.05 percent of probability error levels, then it is concluded that the variable is useful in equation and has a significant effect.

Exp [B]: This statistic which is known as odds ratio is the ratio of probability of occurrence of an event to its failure. When the odds ratio is less than 1, it is said that by increasing the independent variable, the probability of occurrence of the event decreases; and when the odds ratio is more than 1, it is said that by increasing the independent variable, the probability of occurrence of the event increases. Therefore, in logistic regression analysis, the negative effect of each independent variable is known in two ways:

1. Through the negative sign of the statistic.
2. Through the fact that Exp [B] is less than 1 [24].

After applying the obtained coefficients of the model on independent variables, according to Eq. (1), the linear parameter of Z was obtained and by placing it in the main logistic regression Eq. (2) and by transferring it to ArcGIS software, the preliminary map of landslide hazards of the region were prepared. Then, cumulative frequency diagram of pixels was used to classify the landslide hazard zonation map into high risk and low risk classes, so that the points of the curve with natural fracture were considered as boundary between two classes. At last, the final zonation map was prepared in four classes as low risk, medium risk, high risk, and very high risk.

### 3.2 Model Accuracy Evaluation

Omnibus test determines how significant and efficient the model is. Omnibus test as a statistical test is implemented on an overall hypothesis that tends to find general significance between parameters’ variance, while examining parameters of the same type, such as: Hypotheses regarding equality vs. inequality between k expectancies $\mu_j = \mu_j' \ldots = \mu_k$ vs. at least one pair $\mu_j \neq \mu_j'$, where $j, j' = 1 \ldots k$ and $j' \neq j$; in Analysis Of Variance (ANOVA); or regarding coefficients $\beta_1, \beta_2, \ldots, \beta_k$ vs. at least one pair $\beta_j \neq \beta_j'$ in Multiple linear regression or in Logistic regression [24].

According to Omnibus test, the model is properly fit when the error level is less than 0.01 [24]. Table of model summary shows the results of logarithm of likelihood function and Pseudo coefficient of determination (including Cox and Nell coefficient and Nigel Kirk coefficient of determination). Negative factor of doubled likelihood logarithm behaves like $K^2$ and thus, the model is consistent with observed data when it allocates the minimum amount of $-2LL$ factor and the maximum amounts of R$^2$Coax and Snell and Nagelkirk factors to itself [27].

One standard analysis model of logistic regression to examine the model accuracy is the statistical test for model prediction accuracy based on 1 and 0 classification table. In this table, diagonal cells show the number of correct predictions and off-diagonal cells show the number of incorrect predictions. The last column of the table shows the success percentage of independent variables in determining dependent variable changes. Classification graph shows the visual image of classification accuracy in a stacked histogram. According to this graph, the more the 1s in the right side of the graph and the more the 0s in the left side of the graph, the more accurate the model classification and prediction [24].

### 4. Result and Discussion

In this research, the table of variables in equation and the related statistics were obtained as follows (Table 1).
The maps and results of study show that the road depend on landslide and we calculate the variable in logistic for the road.

Table 1 reveals that Wald statistic was only significant for the independent variable of distance from the road at error level of smaller than 0.05 and thus other independent variables with error level of smaller than 0.05 were neutral in landslide occurrence and were consequently omitted. Also, the constant coefficient of the model was smaller than 0.05.

Negative sign of Coefficient B (-1.696), and also the statistic Exp [B] which was less than 1 (0.183) for the independent variable of road indicate the negative effects of these factors on the model. Therefore, as the independent variable increases, the occurrence probability of dependent variable decreases; in other words, as the distance from the road increases, the probability of landslide occurrence decreases.

Considering the obtained coefficients in Table 1, linear parameter [z] and the ultimate logistic regression equation are obtained as follows [25].

\[
Z = 3.275 - 1.696e \\
P = \frac{1}{1 + e^{-(3.275-1.696e)}}
\]

Where;
- \(P\): Probability of landslide occurrence that ranges between 0 and 1;
- \(e\): independent variable of distance from the road;
- Landslide hazard zonation map was prepared by means of Eq. (4) and its transfer to ArcGIS software (Fig. 4).

### 4.1. Evaluation of Final Logistic Regression Model

Considering the results of Omnibus test in Table 2, the error level is 0.000 which indicates that the applied logistic regression model in this research has been highly efficient and powerful in explaining the points.

<table>
<thead>
<tr>
<th>Sig.</th>
<th>df</th>
<th>chi-square</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>1</td>
<td>68.187</td>
</tr>
<tr>
<td>0.000</td>
<td>1</td>
<td>68.187</td>
</tr>
<tr>
<td>0.000</td>
<td>1</td>
<td>68.187</td>
</tr>
</tbody>
</table>

Regarding the model summary table in this research (Table 3), the statistics 1 and 2 were 0.571 and 0.428, respectively, which indicates that the role of independent variable of distance from the road is appropriate in explaining dependent variable variance (landslide occurrence or failure). Also the obtained statistic was equal to 100.

### Table 3. Model Summary

<table>
<thead>
<tr>
<th>Nagelkerk R²</th>
<th>Cox and Snell R²</th>
<th>-2Log likelihood</th>
<th>step</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.571</td>
<td>0.428</td>
<td>100.941</td>
<td>1</td>
</tr>
</tbody>
</table>

Results of classification table (Table 4) indicate that the overall success and classification accuracy in this model is 85.2%. This means that it is 85.2 % possible to determine the dependent variable changes (occurrence or lack of occurrence of landslide) by independent variable of distance from the road.

### Table 4. Classification of 0 and 1 points for observed and predicted spots

<table>
<thead>
<tr>
<th>predicted spots</th>
<th>observed spots</th>
</tr>
</thead>
<tbody>
<tr>
<td>accurate percent</td>
<td>spots value</td>
</tr>
<tr>
<td>86.0</td>
<td>8</td>
</tr>
<tr>
<td>81.6</td>
<td>51</td>
</tr>
<tr>
<td>85.2</td>
<td>Total model accurate percent</td>
</tr>
</tbody>
</table>

Each digit (0 or 1) in classification diagram displays five points (sliding or non-sliding). As mentioned before, the more the number of 1s in the right side and the number of 0s in the left side of the diagram, the more accurate is the model classification and prediction. By examining Fig. 5, the results of Table 4 and the accuracy of model classification and prediction were approved.

As mentioned before and regarding the ultimate model and table of variables in equation (Table 1), it was found that the Wald coefficient was just significant for the variable of distance from the road at error level less than 0.05 in logistic regression output and other effective factors in landslide occurrence of the region (elevation, precipitation, slope, slope aspect, soil, land use, and geology) were omitted from the final regression equation as they were insignificant and the results of the table were displayed in one step. The results indicated that non-
principal road construction mainly for industrial exploitation of forests and timber in the region has the leading role in landslide occurrence in Pahne Kola watershed.

Therefore, it is suggested that maximum caution and care should be exercised in construction of roads and facilities in this region and, if possible, road construction can be prevented in landslide susceptible areas so that the roads could be as short and as practical as possible. In addition, it is recommended to use drains around and beside the roads in order to discharge runoff from rainfall and snowmelt.

Many researchers have emphasized the importance and the effect of distance from road on landslide occurrence such as Ahmadi and Mohammad Khan (2002), Karam and Mahmoudi (2002), Billiardi et al. (2003), Shirzadi et al. (2006), Greco et al. (2007) whose findings are consistent with the results of the research [3, 28–31]. The result about the deletion of elevation in logistic regression model is consistent with the findings of Mousavi Khatir [13].

Rainfall factor was also omitted after implementing logistic regression model since it was insignificant. Ayaliło and Yamagishi [18] ignored the rainfall factor and eliminated it because of uniform rainfall in Kakoda Yahiko in Japan. Khamechian et al. [12] also omitted rainfall factor while preparing landslide hazard zonation map through logistic regression model and the results of the current research in this part is consistent with their findings.

However, considering other effective factors in landslide occurrence (slope angle, slope aspect, soil, land use, and geology), it could be said that the results of this research in this part are not consistent with the findings of Can et al. [32], Garcia et al. [23], Hosseinzadeh et al. [33].

5. Conclusion

In this research, eight effective factors on landslide event including elevation, slope, slope aspect, rainfall, land use, distance from road, soil and geology were considered as independent variables to assess landslide susceptibility in Pahne Kola watershed. But, quantitative analysis and interpreting the coefficients revealed that only the distance from the road was recognized as a significant variable influencing the final equation and the other independent variables were omitted from the final equation due to lack of statistical correlation. After transferring the final probability equation to ArcGIS software, the landslide hazard map was prepared. Statistical model accuracy was evaluated and approved by omnibus test, model summary table, classification graph and table. Statistical evaluation of the model showed that the overall accuracy of prepared map was 85.2%. This risk probability level means approved by omnibus test, model summary table, classification graph and the other independent variables were omitted from the final equation and the other independent variables were omitted from the final equation due to lack of statistical correlation. After transferring the final probability equation to ArcGIS software, the landslide hazard map was prepared. Statistical model accuracy was evaluated and approved by omnibus test, model summary table, classification graph and table. Statistical evaluation of the model showed that the overall accuracy of prepared map was 85.2%. This risk probability level means consistent with their findings.

However, considering other effective factors in landslide occurrence (slope angle, slope aspect, soil, land use, and geology), it could be said that the results of this research in this part are not consistent with the findings of Can et al. [32], Garcia et al. [23], Hosseinzadeh et al. [33].

Acknowledgments

We wish to extend our gratitude to the Amirkabir University of Technology and University of Sistan and Baluchestan for their supports during this research. We are grateful to the reviewers for their useful comments.

REFERENCES


