Landslide-Dammed Mapping and Logistic Regression Modeling Using GIS and R Statistical Software in the Northeast Afghanistan

Mohammad Kazem Naseri¹ and Dongshik Kang²

1. Graduate School of Engineering and Science, Faculty of Engineering, University of the Ryukyus, Okinawa 309, Japan
2. The Department of Information Engineering, Faculty of Engineering, University of the Ryukyus, Okinawa 309, Japan

Abstract: A complex terrain and topography resulted in an enormous landslide-dammed area northeast of Afghanistan. Moreover, debris, rock avalanches, and landslides occurrences are the primary source of lakes created within the area. Recently, instances have increased because of the high displacement and mass movement by glacial and seismic activities. In this study, using GIS and R statistical software, we performed a logistic regression modeling in order to map and predict the probability of landslide-dammed occurrences. Totally, 361 lakes were mapped using Google Earth historical imagery. This total was divided into 253 (70%) lakes for modeling and 801 (30%) lakes for the model validation. They were randomly selected by creating a fishnet for the study area using Arc toolbox in GIS. Four independent variables that are mostly contributed to landslide-dammed occurrences consisting of slope angles, relief classes, distances to major water sources and earthquake epicenters, were extracted from DEM (digital elevation model) data using 85-meter resolution. The result is a grid map that classified the area into Low (16,834.98 km²), Medium (2,217.302 km²) and High (2,013.55 km²) vulnerability to landslide-dammed occurrences. Overall, the model result has been validated by using a ROC (receiver operator characteristic) curve available in SPSS software. The model validation showed a 95.1 percent prediction accuracy that is considered satisfactory.

Key words: Landslide-dammed area mapping, Northeast Afghanistan, logistic regression modeling, GIS and R.

1. Introduction

Afghanistan forms part of land characterized by desert plains and mountainous landscape, the latter accounting for more than 70% of its territory [1]. The country is located at the junction of the Afghano-Iranian highland and mountain systems of the High Central Asia [1]. Northeast Afghanistan is a very susceptible region for landslides, debris flow, earthquake, and landslide-dammed occurrences. Due to rapid glacier morphological activities, ground instability, enormous seismic activities, the landslide-dammed areas are very hazardous and could threaten the local community lives and properties.

Therefore, there is an urgent need for recognizing, mapping, and mitigating this kind of natural hazards. Fortunately, GIS and R software are very useful and suitable for performing logistic regression modeling to know and predict an unknown probability of four independent variables which were used in the model.

Areas around the study area are susceptible to landslide-damming too. The Usoi Dam is an example of a natural landslide-dammed area along the Murghab River in Tajikistan. It has 567 meters elevation and is the tallest dam in the world, either natural or man-made. The dam was created on February 18, 1911, when the magnitude 7.4 Sarez earthquake caused a massive landslide that blocked the flow of the river [2]. The dam was formed by the approximately 2 cubic kilometers of rock dislodged from the steeply sloped river valley of the Murghab, which cuts from east to
west through the high and the rough Pamir Mountains. It was named after the village of Usoi, which was completely buried by the 1911 landslide. The dam rises to a height of 500 to 700 meters from the original valley floor [3]. The basin formed by Usoi Dam now holds Sarez Lake, a 55.8-kilometre long lake holding 16.074 cubic kilometers of water. Water does not flow over the top of the dam, which would quickly cause it to erode away; instead, water seeps out of the base of the dam at a rate which approximately matches the rate of inflow, maintaining the lake at a relatively constant level. The flow averages about 45 cubic meters per second and produces about 250 megawatts of hydro power [4]. Geologists are concerned that the Usoi Dam may become unstable during future large magnitude earthquakes, which are relatively common in the seismically active Pamirs, and might collapse due to liquefaction or subsequent landslides during such an event.

As well as the larger rockslides and rock slide complexes, there are categories of minor mass movement in northeastern Afghanistan that occur where either water or seismic acceleration are the primary causes of slope failure [5]. On early 12 April 2014, land sliding occurred caused by the massive earthquake, in Rustaq district of Takhar province. At least four people were found dead and more than 100 houses were damaged (from local officials and news). On 2 May midday another landslide occurred in the Nowbad area of Abi-Barik village. The village is located in the Argo district of the eastern province of Badakhshan. The number of people killed by the massive landslide was more than 350 with 100 missing, according to local officials quoted by OCHA. The local estimate has more than 2,000 people presumed dead. The landslide destroyed around 300 houses in the village of Abi-Barik and affected approximately 1,000 families, 700 of which were displaced. If an earthquake can cause landslide movement, then it can also break a dam previously created by a landslide. Therefore, it is very necessary to recognize and map these areas in order to mitigate or predict the probability of future occurrences based on causative parameters. The purpose of this study is to create a vulnerability map of the area, through a logistic regression using GIS and R statistical software.

2. Study Area

The study area is located in the northeastern part of Afghanistan that includes parts of Badakhshan, Takhar and Baghlan provinces (Fig. 1). Totally, 549 settlement centers were recorded in the 2004 survey by USGS (United States Geological Survey). The terrain in the area is mountainous to semi-mountainous [1]. It encompasses a very complex geology with intrusive and stratified formations. Landcover types include: irrigated agricultural land (AGI), rain-fed agricultural land (AGR), fruit trees (AGT), barren land (BRS), rangeland (NHS), forest and shrubs (NFS), permanent snow (SNW), built up (URB), and water bodies and marshland (WAT). It covers 21,071.46 km² area with high slope angle ranging from 0-80 degrees. The elevation ranges from 2,000-7,390 m and lakes were formed mostly in the higher elevation areas due to glacier movements and debris flow (Fig. 3). The landslide-dammed areas formed are quite notable because of their large areas extending from 0.002869 to 13.629233 square kilometers. Active tectonic processes in Badakhshan are responsible for the majority of shallow earthquakes registered between 1965 and 1980 at depths of < 70 km. Estimates of peak ground acceleration with an exceedance probability of 10% in 50 years range between 5 to 10 ms⁻² (USGS). The climatic zone of alpine relief in Badakhshan is characterized by a short cold summer and a long rigorous winter, with most precipitation occurring as snow [1]. According to the USGS in 2000, there have been 2,054 earthquakes that occurred in the area with magnitudes ranging from 0 to 7. Among the landslide-dammed areas, a few of were very catastrophic for the local community because of earthquake shaking. The example is Lake Shiveh in
Badakhshan (Fig. 2).

The study area incorporates many parameters which are statistically significant to the probability of future lake formation due to landslides or a debris flow. Therefore, logistic regression is a useful method for modeling because it can handle both variables with continuous values such as raster grids and categorical value like polygons. The dependent variable, however, is either binomial or dichotomist. In this study, the dependent variable is a combination of landslide-dammed points and randomly selected background values (non-landslide-dammed points). Independent variables used in the model are slope angles, relief, distances to major water sources and earthquake epicenters.

3. Methodology and Logistic Regression

Logistic regression is an approach to prediction, like OLS (ordinary least squares) regression. However, with logistic regression, the researcher is predicting a dichotomous outcome [6]. It is modeling the conditional probabilities. But, there are many situations where however we are interested in input-output relationships, as in regression, but the output variable is discrete rather than continuous. In particular there are many situations where we have binary outcomes. In addition to the binary outcomes, we have some input variables, which may or may not be continuous. How could we model and analyze such data? We could try to come up with a rule which guesses the binary output from the input variables. This is called classification, and is an important topic in statistics and machine learning. However, simply guessing “yes” or “no” is pretty crude especially if there is no perfect rule. The dependent variable in logistic regression follows the Bernoulli distribution having an unknown \( p \) (probability). The Bernoulli distribution is just a special case of the binomial distribution where \( n = 1 \) (just one trial). Success is “1” and failure is “0”, so the probability of success is \( p \) and failure is \( q = 1 - p \). In logistic regression, we are estimating an unknown \( p \) for

<table>
<thead>
<tr>
<th>Table 1 Lake Shiveh specification.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
<tr>
<td>Elevation</td>
</tr>
<tr>
<td>Area</td>
</tr>
<tr>
<td>Volume</td>
</tr>
<tr>
<td>Location</td>
</tr>
</tbody>
</table>

Fig. 1 Location of study area.

Fig. 2 Landslide-dammed (Lake Shiveh).

Fig. 3 The lakes formed in the higher elevation area.
any given linear combination of the independent variables. Therefore, we need to link together our independent variables essentially for the Bernoulli distribution. This link is called the logit. Mostly map the linear combination of variables that could result in any value onto the Bernoulli probability distribution with a domain from “0” to “1”. The natural logarithm of the odds ratio is equivalent to a linear function of the independent variables. The antilog of the logit function allows us to find the estimated regression equation as below:

\[
\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1
\]

(1)

and

\[
P = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}
\]

(2)

The step processes of the data calculations performed in GIS and R statistical software are as following:

- Selection of independent variables (slope angle, relief distance to water sources and seismic epicenters). And dependent variables (lake locations and non-lake locations);
- Sampling by the dependent variables from the independent variables in GIS;
- Taking the sampled data to R statistical software for GLM (generalized linear model);
- Taking back the intercept and coefficients from the R result to the GIS; and
- Exponents, probability calculation and classification of the raster grids to different vulnerability zones.

4. Input Variables

Four independent variables such as slope, relief, distance to major water sources and distance to the earthquake epicenters were used in the model. They were extracted from the DEM (digital elevation model) with 85-meter resolution. It was the only available free data source for the study area. Obviously, the more pixel resolution will ensure best model fit and accuracy. But in the case of data limitation that could be a reliable source in order to do modeling and recognize the susceptible zones to the natural hazard. Due to the vertical error of SRTM reflectance based on different landcover classes, there will be some data uncertainty but for this type of study as a regional mapping will not affect the model accuracy more. Because in logistic regression the modeling is based on specific site locations like control points. Relief was classified from DEM into six classes starting from the minimum 2,000 m to the maximum 7,390 m (Fig. 4). The slope layer was also classified into five classes at 11, 22, 32, 44 and 80 degrees (Fig. 5). For the distance to the major water sources, firstly, the basin and water sheet were extracted from DEM data and then classified as polyline shapefile with distances generated based on buffer analyzes in different buffer zones at 0.2, 0.5, 1, and 3 kilometers, respectively in GIS (Fig. 6). For the seismic activities, the buffer distances were generated from the earthquake epicenters as 0.2, 0.5, 1, 2, and 20 kilometers by using GIS (Fig. 7).

The dependent variables used in the model were 253 lakes point shapefiles as well as the randomly selected 253 background points (non-lakes point). For binomial analyses, the values assigned were “1” for the actual location and “0” for the background points. The raster values of four independent variables were extracted for the particular sites location exported as dbf file format and converted to CSV extension (comma delimited file) in excel for further modeling in R statistical software.

Fig. 4 The relief classes.
5. Result and Vulnerability Analyses

The R statistical result showed that all of the four independent variables are statistically significant to the model, each with a specific number of asterisks. In GLM (generalized linear model) as the more the asterisks the more significant are the variables. They are coded as 0 “***”, 0.001 “**”, 0.01“*”, 0.05.”, 0.1“’” and “1”. In this model, the coefficients for the input parameters are distance to water sources (-0.001253), slope classes (-0.2863), relief (0.0009096), and distance to earthquake epicenters (0.00004654), respectively. The predicted intercept value in the model is 0.7364. For the slope and distance to water sources coefficients, because of their negative values, it means that if the grid values increase the chance of a landslide-dammed occurrences decreases on that particular grid. But in case where the relief and distance to the earthquake epicenters are not the same, their values get bigger, and the probability of lake occurrences increase. After multiplying each coefficient with its raster grid and combining these as one grid output in GIS raster calculator, we come up with a raster grid with further processing to add the intercept value for the grids (Table 2). From the outcome, we calculated the exponents of the grids which were considered at odds to the probabilities. Then added a value of “1” in order to scale it. The final grid values ranged from “0” to “1” which goes far away from “0” to “1”. The chance of a landslide-dammed occurrences increase and vice versa from “1” to “0” and the probability of lake occurrences decrease. The result shows three classes broken down by a natural break method and presented as the probability classes of lake occurrences within the area.

In this study the logistic regression is used where the unknown probability of dependent variable can be predicted based on the independent input parameters. All of the four independent variables were inputted to the model, in order to develop the final landslide-dammed occurrences vulnerability map in different vulnerability classes (Table 2 and Fig. 8) including, Low, Medium and High. Most of the vulnerable areas are located in the higher elevation.
Table 2  The vulnerability level and coefficients in R.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>R result and coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>0.7364 Signif.codes</td>
</tr>
<tr>
<td>Relief</td>
<td>0.0009096 0.001</td>
</tr>
<tr>
<td>Slope</td>
<td>-0.2863 0.001</td>
</tr>
<tr>
<td>Distance to water sources</td>
<td>-0.001253 0.01</td>
</tr>
<tr>
<td>Distance to earthquake epicenters</td>
<td>0.00004654 0.05</td>
</tr>
</tbody>
</table>

Vulnerability Classes with their areas in sq/km

<table>
<thead>
<tr>
<th>Area in sq/km</th>
<th>Vulnerability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1683.98</td>
<td>Low</td>
</tr>
<tr>
<td>2217.302</td>
<td>Medium</td>
</tr>
<tr>
<td>2013.55</td>
<td>High</td>
</tr>
</tbody>
</table>

Fig. 8  Vulnerability classes result.

Fig. 9  AUC constructed by ROC.

And will be the most unstable vulnerable zones because of seismicity of the area.

6. Model Accuracy

The correct identification and classification are fundamental in ensuring the accuracy and usability of the susceptibility map [7]. In order to check the model accuracy and its prediction performance, we draw the ROC (receiver operator characteristic) curve. The model accuracy and its performance directly depend on the input parameters used to generate the final vulnerability map. In this study, ROC curve analysis was used to validate the model performance. It is available in the IBM SPSS Statistics 20 software, first released in the USA in 1968. The ROC curve is a graphical plot of false positive rate (1-specificity) on the X-axis and true positive rate (sensitivity) on the Y-axis at different cut-off values. ROC graphs have long been used in signal detection theory to depict the trade off between hit rates and false alarm rates of classifiers. ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems [8]. The accuracy of a test is measured by the area under the ROC curve which is known as AUC [9]. Theoretically, the AUC ranges from 0.5 up to 1. The value of 0.5 is known as the reference line which means as a worthless prediction but when much of the values get closer to the top left corner of the plot near to value of 1 represents the higher prediction accuracy of the model. In the present study, the ROC curve was constructed by using the 108 lake locations that have not been used in creating the model itself but to validate the model. The plot shows the AUC of 0.951 with an Asymptotic 95% Confidence Interval. This is not a 100 percent prediction test result but the model performance is still considered excellent as a good indicator with 95.1 percent prediction accuracy (Fig. 9).

7. Conclusion

The northeast part of Afghanistan is a vulnerable
region for landslide-dammed occurrences. That could facilitate a valley flooding to wash away houses, agricultural lands, as well as taking the local people’s lives. The only way to identify and mitigate this natural disaster is to map them out and conduct planning, purposely. The objective of this study is to create a vulnerability map of the area in order to know hazard level of the area. The primary cause for lakes to form is due to the active seismicity, glacier movements, landslides and debris flows. The superiority of the logistic regression is that it can analyze both categorical and continuous values. Therefore, it is the suitable way of hazard mapping, mitigation and vulnerability prediction. The predictive result depends on, the input data preparation, cell size, and areal extents. In this study, the four independent input parameters were prepared in the same cell size and extended, using 85-meter resolution DEM. Compared to the world seismic stations, the ground amplification is high within the study area and increases continually. Because of the location of the lakes in the high elevation and steep slope areas, there is fear for the landslide-dammed areas to burst suddenly, just through weak ground amplification. The 95.1% accuracy obtained from the model validation using ROC curve indicates the best model performance and prediction in the study area. The result of this study is not a 100% predicted accuracy to be used on area identification and planning, due to the scale and size of the area. But, it can serve as a primary knowledge about glacier process, landslides-dammed study, and natural hazard mitigation.

The input parameters and DEM data are available as open source, therefore, this model is replicable and can be used anywhere with different associated causative parameters to the model. Even, will be the most accurate result due to a high-resolution DEM that might be accessible for the areas. But, primarily, there is a need for the inventory map or the specific sites locations because basically, the GLM modeling predicts the unknown probabilities based on the sites locations. Any objectives can be used. In this study, the inventory was made by using Google Earth imagery and non-sites locations that were selected randomly by GIS. The modeling is performed based on the inventory landslide-dammed locations. It means, in what extend areas will affect the same pixel value and input characteristics. Any other causative parameters can be added to the model as input. But, an overall goal of the study is to evaluate the hazard level of the area based on the landslides and debris flow that are the common phenomenon in the northeast Afghanistan. This study is a regional hazard zonation for vulnerability recognition but cannot serve as a specific site selection and detailed engineering design purposes.

Acknowledgement

We gratefully acknowledge the Graduate School of Engineering and Science (University of the Ryukyus), and JICA-PEAC program for Afghanistan.

References


