Landslide-dammed Mapping and Logistic Regression Modeling Using GIS and R Statistical Software in The Northeast of Afghanistan

Mohammad Kazem Naseri [1] and Kang Dongshik [2]
[1] Graduate School of Engineering and Science, University of the Ryukyus, Nishihara-Cho, 903-0129, Okinawa, Japan
[2] Department of Information Engineering, University of the Ryukyus, Nishihara-Cho, 903-0213, Okinawa, Japan
E-mail: [1]kanaseriy@gmail.com, [2]kang@ie.u-ryukyu.ac.jp

Abstract: The landslide-dammed have been created in northeastern of Afghanistan, were all due to a complex topography, debris flow, rock falls, avalanches, and landslide occurances. It is a very rare case compare to other provinces because of a high displacement and mass movement by glacier activities and seismic ground amplification, in the study area. In this paper, we accomplished a logistic regression modeling in order to map and predict the probability of occurrences of lakes by landslides events. A total of 361 lakes were mapped in the area which all of them associated with a very fast morphological activities in the past. Moreover, as well as these 361 lakes point locations, 361 non-lakes location points were used for the model validation. They were randomly selected by creating a fishnet for the study area using Arctoolbax in GIS. Three independent variables which are mostly contributing for landslide lakes occurences were extracted from digital elevation model (DEM) data by using GIS. They were slope degrees, distance to mean water sources, and relief classes. The result is a grid map that shows tree independent variables which were how statistically significant to the model.

Keywords: GIS and R, Logistic Regression, Landslide-dammed mapping, in Northeast of Afghanistan

1. Introduction

Northeast of Afghanistan is a very susceptible region for landslides, debris flow, earth flow, earthquake, and landslide-dammed occurrences. Due to rapid glacier morphological activities, ground instability, enormous seismic activities, the landslide-dammed are very dangerous and could be a big threat for the local community and their lives. Therefore, there is an urgent need for recognizing, mapping, and analyzes of these kind of problems. Fortunately, GIS and R software are very useful and suitable for performing logistic regression modeling to know and predict an unknown probability of independent variables that are incorporated into the situation.

The Usoi Dam is a natural landslide dammed along the Murghab River in Tajikistan. At 567 meters high, it is the tallest dam in the world, either natural or man-made. The dam was created on February 18, 1911, when the magnitude of 7.4 Sarez earthquake caused a massive landslide that blocked the flow of the river [1]. The dam is formed of 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 is 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 [2]. 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 dissipates about 250 megawatts [3]. 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 [4].

As well as the larger rockslides and rock slide complexes, there are categories of fine-grained mass movement in northeastern Afghanistan that occur where either water or seismic acceleration are the primary causes of slope failure(Shroder). On early 12 April 2014, a land sliding caused by the massive earthquake, in Rustaq district of Takhar province, Afghanistan. At least four people have been found dead and more than 100 houses were damaged (Local officials and news). On 2 May at around 11.00 a.m. local time in Afghanistan, 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 more than 350 and 100s was missed, According to local officials quoted by OCHA. The local estimations say more than 2000 people are 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 a landslide movement also can break a dam which is created by a landslide. Therefore, it is very necessary to recognize and map them in order to mitigate or predict the probability of future occurrences based on causative parameters.

The total study area is covered 22173.302km2 with high slope angle ranges from 0-80 degrees. The elevation of the study area ranging from 1,415-6,585m. The landslide-dammed are formed quite notable because of their areas extend from 0.002869 to 13.629233 km2. The area is very active seismically and tectonically. According to the USGS-survey in 2000, there has been 2,031 earthquake happened in the area, magnitudes of ranging from 0 to 6.6. Among the all landslide-dammed, a few of them are very catastrophic for the local community. The example is Lake Shiveh in Badakhshan province of Afghanistan (Figure.1)

The study area incorporates with many parameters which are statistically significant to the probability of future occurrences. Therefore, logistic regression is a useful method for modeling because it can handle both variables with a continued value such as raster grids and categorical value like polygons. But the dependent variable is 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). And independent variables were used in the model are slope degrees, relief and distances to major water sources.

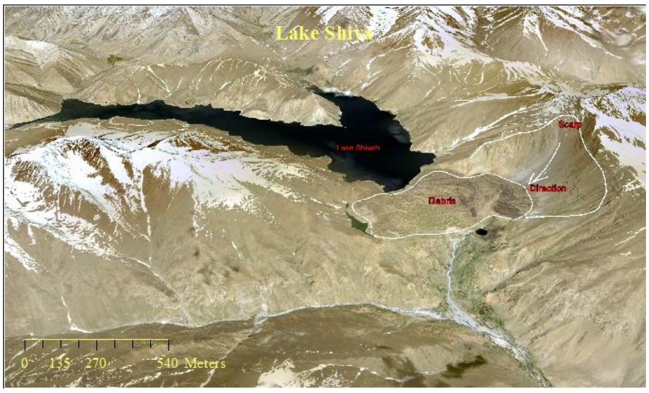


Figure 1: Landslide-dammed (Lake Shiveh) at Badakhshan Province of Afghanistan

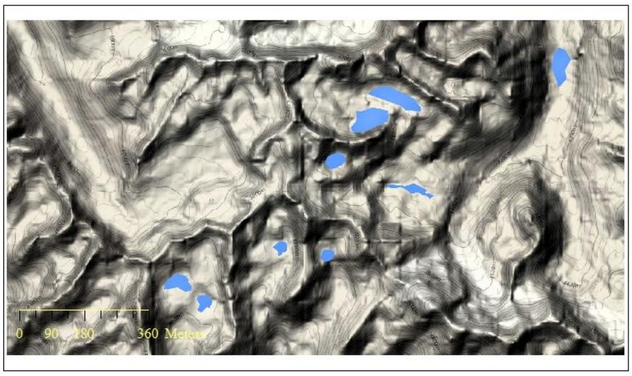


Figure 2: The lakes formed in the high elevation area

Table 1: Lake Shoveh specification

Lake Shiveh Specification	
Type	Landslide-dammed
Elevation	3,130 Meter
Area	13.629233 Square kilometer
Volume	4.1257597756 Cubic kilometer
Location	Badakhshan Afghanistan

2. Methodology and logistic regression

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 any given linear combination of the independent variables. Therefore, we need to link together our independent variables to essentially the Bernoulli distribution; that link is called the logit. That most 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:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\chi_1 \qquad (1)$$

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1} \qquad (2)$$

$$(1-p).e^{\beta_0+\beta_1\chi_1} = p \qquad (3)$$

$$e^{\beta_0+\beta_1\chi_1} = p + p.e^{\beta_0+\beta_1\chi_1} \qquad (4)$$

$$e^{\beta_0+\beta_1\chi_1} = p(1+e^{\beta_0+\beta_1\chi_1}) \qquad (5)$$

$$p = \frac{e^{\beta_0+\beta_1\chi_1}}{1+e^{\beta_0+\beta_1\chi_1}} \qquad (6)$$

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

- Selection of independent variables(Slope, distance to water sources and reilef) 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 generalized linear model(GLM),
- Taken back the intercept and coefficients from the R result to the GIS,
- And finally, exponential and probability calculation and classification of the raster grids.

3. Input Variables

Three independent variables such as slope, releif and distance to major water sources have been used in the model. They were extracted from the digital elevation model (DEM) with 85-meter resolution. Releif has been classified from DEM in six classes, start from minimum 2,509m to the maximum 6,585m (Figure 3). The slope layer also has been classified in six classes at 10, 20,30,40,50, and 80 degrees (Figure 5). For the distance to the major water sources, first of all, the basin and water sheet extracted from DEM data and then it was classified as polyline shapefile which distances generated based on buffer analyzes in different buffer zones at 100, 500, 1000, 2000, and 3000, respectively in GIS (Figure 4). The dependent variables used in the model were 361 lakes point shapefile as well as the randomly selected 361 background points (non-lakes point). For binomial analyzes, the values assigned are *"1"* for the actual location and *"0"* for the background points. The raster values of three independent variables extracted for the particular sites location and exported as dbf file format and changed to CSV extension (comma delimited file) in excel for further modeling in R statistical software.

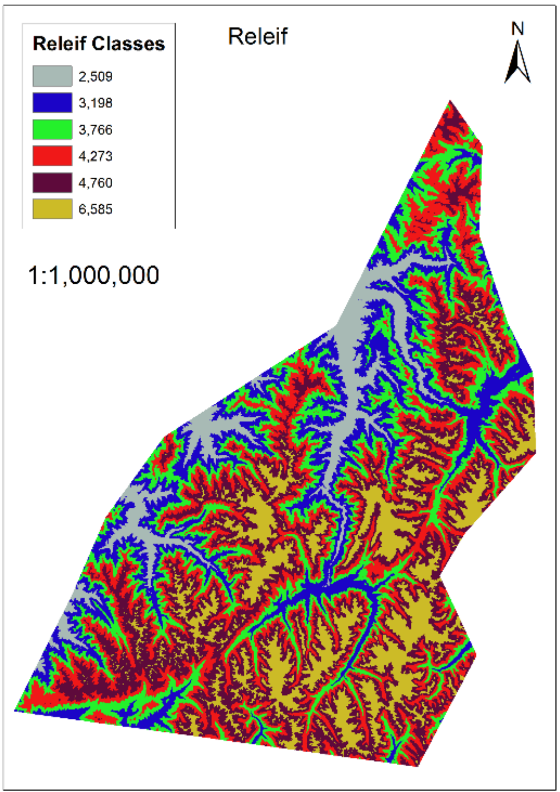


Figure 3: the independent Variable releif

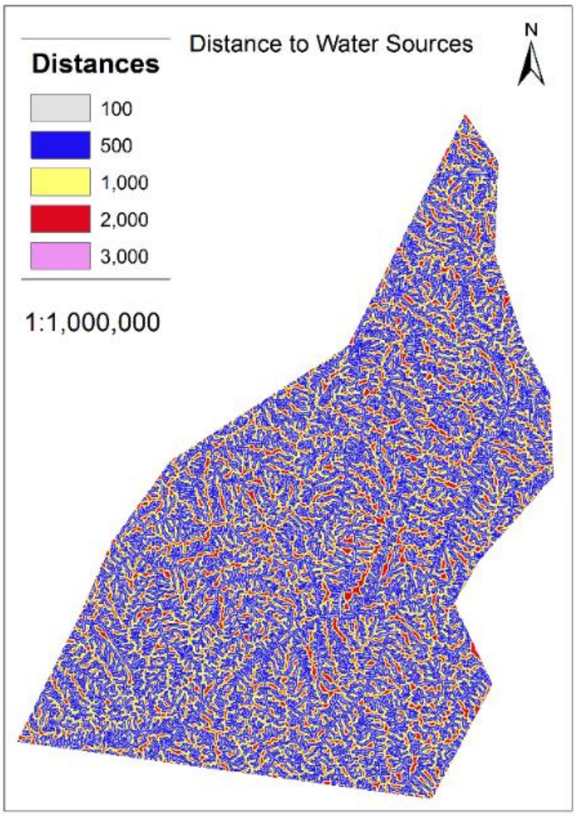


Figure 4: distances to the water sources

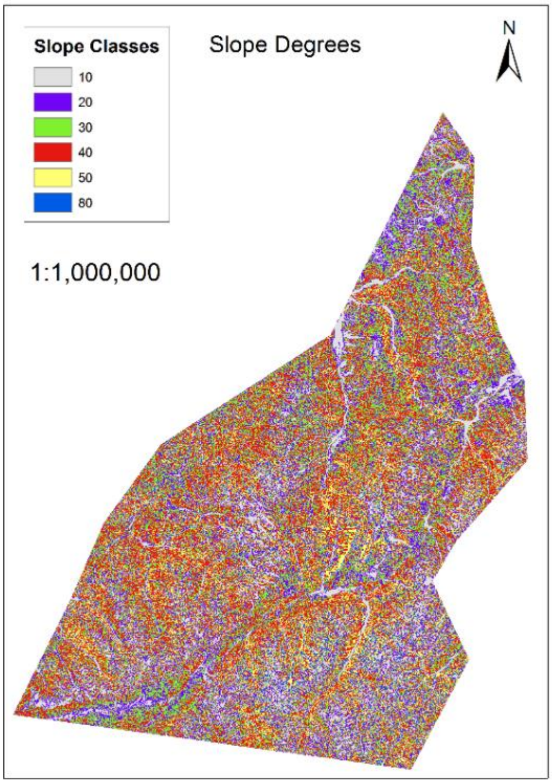


Figure 5: the slope classes

4. Model Result

The R statistical software result showed that three of the independent variables are statistically significant to the model, each with three asterisks (***). In generalized linear model (GLM) as much is the asterisk are more, the variables are more significant. They coded like 0 “***”, 0.001 “**”, 0.01“*”, 0.05“.”, 0.1“” and “1”. In this model the coefficients for the distance to water sources is (-0.0022362), for slope (-0.1827496) and for the releif is (0.0010039), respectively (Figure 7). And the predicted intercept value is (-0.04341865) in the model. For the slope and distance to water sources coefficients, because of their negative values mean that if the grid values increase the chance of being a site is a decrease on that particular grids. But in the case of releif, it is not the same, as much as releif values get bigger the probability of being a site is an increase. After multiplying each coefficient with its raster grid and combining them as one grid output in GIS raster calculator, we come up with a raster grid and the further process is to add the intercept value for the grids. From the outcome, we calculated the exponential of the grids which considered as odds to probabilities and after that in order to scale it therefore, added a value of *“1”*. The final grid values ranged from *“0”* to *“1”* which has going for away from *“0”* to *“1”*, the chance of a site being present is increase and vise verse from *“1”* to *“0”* the probability of being a site is decreased. The result showed in three classes were broke down by a natural break method and presented the probability classes (Figure 6).

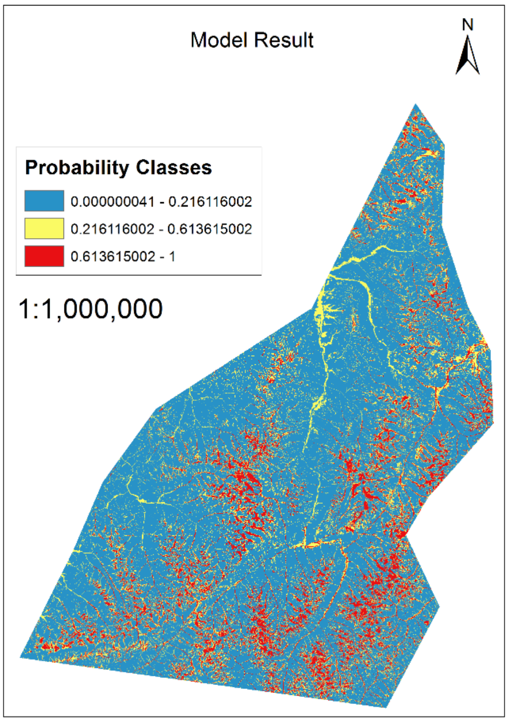
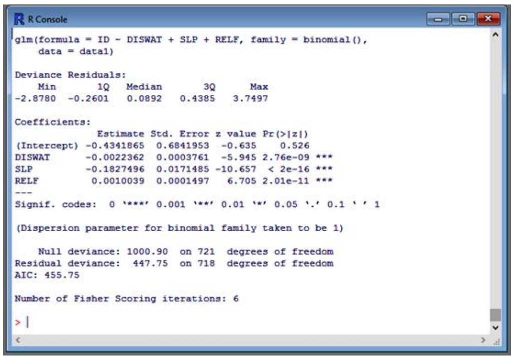


Figure 6: The model result



```
glm(formula = ID ~ DISWAT + SLP + RELF, family = binomial(),
    data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8780  -0.2601   0.0892   0.4385   3.7497

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4341865  0.6841953  -0.635    0.526
DISWAT      -0.0022362  0.0003761  -5.945 2.76e-09 ***
SLP         -0.1827496  0.0171485 -10.657  < 2e-16 ***
RELF         0.0010039  0.0001497   6.705 2.01e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1000.90  on 721  degrees of freedom
Residual deviance:  447.75  on 718  degrees of freedom
AIC: 455.75

Number of Fisher Scoring iterations: 6

>
```

Figure 7: R result(intercept and coefficients)

5. Model Accuracy

In order to check the model accuracy and how it predicted well, we run the model histogram by converting the landslide-dammed point to graphic and extract its value as zonal table in the GIS environment. To see the total actual site points falling in each particular grid cells. As it showed in the below (Figure.8) the highest number of points 286 were filled in the grids closer to value of “1”, 41 points in yellow color grids and 34 of them in blue color grids. Moreover, compare to the area of each grids, we mapped the smallest area with high probability to landslide-dammed occurrences and the biggest area with low probability to landslide-dammed occurrences. Therefore, the model accuracy seems quite reasonable and mapped accurately as it was expected.

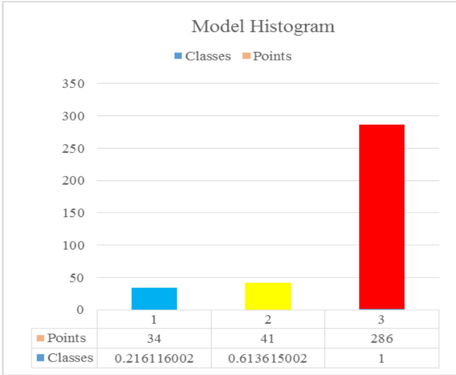


Figure 8: totall damed points in each class

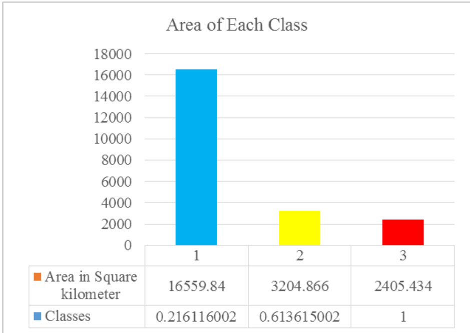


Figure 9: area of classes in square kilometer

6. Conclusion

In conclusion, this result will play a key role in terms of site selection, and developing of an infrastructure in the area. It has been realized by modeling that the area is high susceptible to landslide-dammed occurrences, due to its morphological characteristics and a very fast glacier mass movement. Moreover, the area is very active seismically which made instability and could happen a big disaster if that landslide-dammed burst. Therefore, we highly recommend for material strength and permeability tests for the dam site and investigate more about safety system in order to set up an emergency warning system in the dammed site to optimize its safety.

References

[1] National Geophysical Data Center, “*Significant Earthquake*”, Retrieved June 21, 2012.

[2] Usoi Landslide Dam and Lake Sarez, “*An Assessment of Hazard and Risk in the Pamir Mountains, Tajikistan*” Retrieved August 28, 2012.

[3] Alford, Donald; Cunha, Stephen F; Ives, Jack D, “*Lake Sarez, Pamir Mountains, Tajikistan: Mountain Hazards and Development Assistance, Mountain Research and Development*”, February 2000.