Validation of Spatial Prediction Models for Landslide Susceptibility
Maps

S.B. Bai 1, J. Wang 1, A. Pozdnoukhov 2 and M. Kanevski 2

1 National Education Administration Key Laboratory of Virtual Geographic Environments, Nanjing Normal University, Nanjing, 210046, China

2 Institute of Geomatics and Analysis of Risk, University of Lausanne, Amphipole, 1015 Lausanne, Switzerland

Abstract. A wide range of numerical models and tools have been developed over the last decades to support the decision making process in environmental applications, ranging from physical models to a variety of statistically-based methods. In this study, a landslide susceptibility map of a part of Bailongjiang River, in northwest China was produced, employing binary logistic regression analyses. The available information includes the digital elevation model of the region, geological map and different GIS layers including land cover data obtained from satellite imagery. The landslides were observed and documented during the field studies. To achieve the most appropriate results some sensitivity analyses were also carried out. To validate the quality of mapping, the studied area was divided into training part (3 sub-basins) and validation part (2 sub-basins). Correct classification percentage and Root Mean Square Error (RMSE) values for the validation data for that case were estimated as 76.6% and 0.432, respectively.

Keywords: landslide susceptibility, GIS, binary logistic regression, validation

1. Introduction

The logistic regression (LR) approach is further elaborated on by crosstabs method, which is used to analyze the relationship between the categorical or binary response variable and one or more continuous or categorical or binary explanatory variables derived from samples. It is an objective assignment of coefficients serving as weights of various factors under considerations while expert opinions make great difference in heuristic approaches. Different from deterministic approach, it is very applicable to regional scale.

2. Geological environment of the studied area

The site lies in the middle south of the west wing of Qinling orogen. The area is controlled by Qinghai-Tibet tectonic belt and Wudu arc structure and is affected by uplift of the Qinghai-Tibet plateau. When structure movement shift its forms from main horizontal movement to main vertical movement, it is a very profound effect to the wings of Qinling in this location. The unbalanced vertical movement creates the extreme development of folds, crushs, faults and joints in the location. Lithology is mainly phyllite, schist, slate, carbonate rocks and all kinds of clastic rocks. Today in China tectonic units can be divided into three tectonic systems: Alps - Himalaya tectonic system, shortly as Tethys tectonic domain; the Pacific tectonic system and the new generation of Central Xibailie tectonic system (or ancient Asia tectonic domain). Two tectonic belt: Helan-Sichuan-Yunnan north and south of the belt and Kunlun-Qilian-Qinling-Dabieshan east and west of tectonic belt (central orogenic system).

2.1. Lithology of stratum

The case study area distributes much of loose soil layer due to all kinds of causes, such as the Silurian, Devonian, Carboniferous, Permian and Triassic in Paleozoic and the Triassic of Mesozoic and Quaternary. The oldest stratum is the Silurian in the area which mainly distributes in the two shores of Bailongjiang
River and forms multiple anticlines. The stratum constructed mainly by shallow marine sedimentary metamorphic rocks and carbonate rocks. From east to west, the sediment thickness becomes from thin to thick, sediment is from coarse to fine and carbon capacity is from low to high. Lower Silurian is composed by carbonate rocks and clastic rock and is the nuclear of multiple anticline of Bailongjiang river, at the same time, it lies in the south of Bailongjiang river. Between middle and upper Silurian is called Bailongjiang group made up of clastic and carbonate rocks. It mainly distributes in the Bailongjiang northern shore of the east of two estuary and southern shore from the west of two estuaries to Zhouqu. Devonian develops greatly in the area and mainly located in two wings of Bailongjiang multiple anticlines. Middle Devonian and upper Devonian have a close connection with landslide, while as we know Middle Devonian mainly distributes in the north wing of Bailongjiang multiple anticlines.

2.2. Geological structure
The study area is located in the intersection of the new regional structure north-south and east-west tectonic activity zone in the north edge of the Qinghai-Tibet Plateau. In the tectonic, it is the West Qinling orogen of a Qinling micro-plot. With the long-term impact of tectonic activity, displayed extrusion zone along in the western; It is along a EW trend in the eastern, including folds and faults, while the stratum also distribute to the NW-NWW as a band Since the Cenozoic, Indian plate and Eurasian plate collision developed mountains, the Qinghai-Tibet Plateau uplift, and large-scale strike-slip of alitplano crust result in stress-strain field and tectonic activity exceptionally complex. Strong seismic activities and water system development of Bailong River drainage are the main reasons for the landslide and damming disasters.

2.3. Landform
In the landform, the study area is located in the southern of the Qinling Mountain, and has tall upright mountains, steepness terrain, deep valleys, fast-flowing river and presents V-shaped valley or canyon terrain features. It is the erosion middle and high mountain landform.

2.4. Earthquake and the new tectonic movement
Earthquake is an important reason for landslide damming. The study area lies in the north-south seismic zone of the Qinghai-Tibet Plateau northern earthquake zone. From west to east, there are three about north-south gravity steps zone. The middle of Bailong River is in the Wuwei to Wudu zone of the western border of the zone. This zone begins with Wuwei of the north extending to the SSE, and passing though the Yongdeng, Lanzhou, E County, Dangchang to the south of Wudu, it is about 600 km long and 50-60 km width with a NW direction, and the range of gravity is more than 60 mgal. The area is also one of the strongest new tectonic movement regions. Steep and towering mountains, the terraces of the valley, narrow and steep valley shows strong crustal uplift.

2.5. Hydrology and climate characteristics
The study area is the drainage of Bailong River which is a tributary of the Changjiang River. The water system of Bailong River shows plume-shaped and about NW-SE. The mainstream of Gansu is 475 km long and the both sides of river are asymmetry, South wide and North narrow. Bailong River has many tributaries. The larger length tributaries are Min River, Beiyu River and Baishui River.

The rainfall was not balanced in the time and space. In time, the rainfall mainly concentrated in the 4-10 months, from 6-9 months, Dangchang rainfall is 59.9 percent of the annual rainfall, wudu is 65.5%, Wenxian is 62% and Zhouqu is 61.7%. The rainfall relative concentrates in the time, and has short and heavy rain at times.

In the spatial distribution, rainfall trends to decrease from south to north such as from 900mm of the Wudu southeast to 500mm of the Wudu northwest. In the vertical, rainfall increases with height increasing. As for as the average annual rainfall is 400-500 mm below 1500m elevation, 500-600 mm between 1500m and 2000m and more than 600 mm above 2000m.

The level and vertical zones of the climate are obvious in this region. In the vertical direction, the temperature decreases with altitude increasing. At an altitude of more than 2500m, the annual average temperature is below 5°C, there are very widely differences between the extreme maximum and minimum
temperatures. In the horizontal direction, the temperature trends to gradually decrease from the south to the north.

3. Susceptibility analyses

3.1. Logistic regression approach

Logistic regression allows forming a multivariate regression relation between a dependent variable and several independent variables (Atkinson and Massari 1998). The advantage of logistic regression is that, through the addition of an appropriate link function to the usual linear regression model, the variables may be either continuous or discrete, or any combination of both types, and they do not necessarily have normal distributions. The algorithm of logistic regression applies maximum likelihood estimation after transforming the dependent into a logit variable (the natural log of the odds of the dependent occurring or not). In this way, logistic regression estimates the probability of a certain event occurring (Atkinson and Massari 1998; Dai et al., 2001; Lee and Min 2001).

In the present situation, the dependent variable is a binary variable representing the presence or absence of landslides. The logistic model can be expressed in its simplest form as:

\[ P = \frac{1}{1 + e^{-z}} \]  

where \( P \) is the probability of an event occurring and also is the estimated probability of landslide occurrence.

The value \( P \) varies from 0 to 1 on an s-shaped curve and \( z \) (linear logistic model) varies from \(-\infty \) to \(+\infty\). And where \( z \) is defined as:

\[ z = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \]  

where \( b_0 \) is the intercept of the model, \( n \) is the number of independent variables, \( b_i (i = 1, 2, 3, \ldots, n) \) is the slope coefficient of the model and \( x_i (i = 1, 2, 3, \ldots, n) \) is the independent variable. The linear model formed is then a logistic regression of presence or absence of landslides (present conditions) on the independent variables (prefailure conditions).

In the application of Logical Regression Model for landslide susceptibility mapping, some authors have tried to exploit the category variable via using dummy binary variables for each class of an independent parameter (Guzzetti et al., 1999; Lee and Min, 2001; Dai et al., 2001; Dai and Lee, 2002; Ohlmacher and Davis, 2003). And in this study, we extended the application of logistic regression. The new approach is to use the continuous data as it is, in order not to alter the state and information present in the parameter maps. It solves the problem when quantitative variable is classified and is treated as qualitative variable in conventional method.

3.2. Split the area and sampling

In this study, the study area was subdivided into calibration area (3 subbasins of the study zone, 924 km²) and validation area (2 subbasins of the study zone, 437 km²). The landslide area covers an area of 14.78 km². Landslide presence represents the 1.6% of the study area (thousands of times fewer than their absence) and they could be considered as rare events (King and Zeng, 2001; Van Den Eeckhaut et al., 2006). In logistic regression there are many works with unequal proportions of them (Atkin and Massari, 1998; Guzzetti et al, 1999; Dai and Lee, 2002; Ohlmacher and Davis, 2003; Ayalew and Yamagishi, 2005; Domínguez-Cuesta, M.J.et al, 2006; Van Den Eeckhaut et al., 2006; Bai et al, 2007; Bai et al, In press a; Bai et al, In press b), but as occurred in our study, it is recommended to use equal proportions of landslide and no landslide samples. And all locations of the landslide cells were thus used to extract the physical parameters (independent variables) automatically from the existing data layers. Moreover, 16442 random sample locations were chosen from the landslide free area as samples representing the absence of landslide and are presented in Fig.1. The landslide cells and the set database were merged and a new column of a binary variable indicating the presence and absence of the landslides were added. This stage is repeated 4 times and we can find if there is any convergence in the success of logistic regression analyses by the 4 different random sample locations sets. As a landslide prediction model does not have a scientific significance without measuring the validity of the results (Chung et al., 2002; Chung and Fabbri, 2005) the sample of 32884 cells
was subdivided into calibration and validation datasets. The calibration dataset contained 80% of the cells, and the validation set contained the remaining 20% of the cells.

### 3.3. Multicollinearity diagnosis

Model fitting via logistic regression is sensitive to collinearities among the independent variables (Hosmer and Lemeshow, 1989). Tolerance (TOL) and the Variance Inflation Factor (VIF) are two important indexes for multicollinearity diagnosis. Tolerance smaller than 0.2 is one indicator for multicollinearity and smaller than 0.1 means that there is serious multicollinearity between independent variables (Menard, 1995). The both indexes were calculated (Table 1) and variables with VIF of >2 and TOL of <0.4 were excluded from the logistic analysis (Allison, 2001).

![Fig. 1 The positions of selected 16442 random landslide free grids, 16442 landslides grids](image)

**Table 1 The multicollinearity diagnosis indexes for variables**

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>TOL</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect</td>
<td>0.928</td>
<td>1.077</td>
</tr>
<tr>
<td>Dem</td>
<td>0.284</td>
<td>3.522</td>
</tr>
<tr>
<td>Earthquake</td>
<td>0.723</td>
<td>1.382</td>
</tr>
<tr>
<td>Fault distance</td>
<td>0.775</td>
<td>1.291</td>
</tr>
<tr>
<td>Inhabited distance</td>
<td>0.313</td>
<td>3.193</td>
</tr>
<tr>
<td>Land cover</td>
<td>0.936</td>
<td>1.068</td>
</tr>
<tr>
<td>Plan curvature</td>
<td>0.780</td>
<td>1.282</td>
</tr>
<tr>
<td>Profile curvature</td>
<td>0.774</td>
<td>1.292</td>
</tr>
<tr>
<td>Day average rainfall</td>
<td>0.599</td>
<td>1.669</td>
</tr>
<tr>
<td>River distance</td>
<td>0.556</td>
<td>1.799</td>
</tr>
<tr>
<td>Road distance</td>
<td>0.425</td>
<td>2.354</td>
</tr>
<tr>
<td>Slope</td>
<td>0.946</td>
<td>1.057</td>
</tr>
<tr>
<td>Summer average rainfall</td>
<td>0.394</td>
<td>2.540</td>
</tr>
<tr>
<td>Lithofacies</td>
<td>0.803</td>
<td>1.246</td>
</tr>
</tbody>
</table>

### 3.4. Implementation of LR models

After exclusion of the highly correlated dependent variables, the sample datasets were then used to input to the logistic regression algorithm within the Statistical Package for Social Science (SPSS) to calculate the correlation of landslide to each factor. The forward stepwise logistic regression was carried in order to incorporate only the predictor variables with an important contribution to the presence of landslides. In this
study the significance level of the score Chi² for entering the model was set at 0.15. The significance level of the Wald Chi² for a variable to stay was set at 0.05. Variables such as distance to fault, distance to river, Day average rainfall, Lithofacies, Land cover and aspect were selected for being statistically significant. Hosmer-Lemeshow test showed that the goodness of fit of the equation can be accepted because the significance of chi-square is larger than 0.05. The value of Cox and Snell R² and Nagelkerke R² showed that the independent variables can explain the dependent variables in a way. The Relative Operating Characteristic (ROC) value of 0.8901 is obtained in this study, which can be taken as a sign of good correlation between the independent and dependent variables.

4. Validation and results

Applying the B coefficients in logistic regression equations, the 3 sub-basins susceptibility map was produced (Fig. 2). According to Can et al. (2005), on a spatially effective landslide susceptibility map, observed landslide areas should coincide with the areas having high susceptibility values (rule-i). In addition, high susceptibility values should cover only small areas (rule-ii). According to the decision rules suggested by Can et al. (2005), when the susceptibility value of 0.5 is taken into consideration as the cut-off value, 87.28% of observed landslides of are located on the areas having high susceptibility values. For this reason, the landslide susceptibility maps produced using these sets are considered as spatially effective for rule-i.

To evaluate the sensitivity of the prediction capabilities of the LR models, correct classification percentages (for the cut-off value of 0.5) and Root Mean Square Error (RMSE) (Eq. (3)) values were calculated for the validation data sets (the remaining 20% of the cells).

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y - y')^2}
\]

where, N is the number of cases, y is the observed presence, and y’ is the predicted presence.

The predicted landslide susceptibility (probability) distributions for these sets are also given in Fig. 3. On the graphs in this figure, the row data of validation sets constitute the x-axes, and y-axes comprise landslide susceptibility values, calculated by the previously constructed models. Correct classification percentages and RMSE values was calculated.

In order to validate the prediction capability of the adopted sampling procedures and to test the representativeness of the selected 3 sub-basins sampling zone, the LR procedure was extended to the whole study area including the 5 sub-basins. Applying the B coefficients in logistic regression equations, the 5 sub-basins susceptibility map was produced (Fig. 4.). Then, the validity was then checked using the data from the whole study area. Our working hypothesis in the validation process was that an active landslide must be triggered in those areas showing at least low values of susceptibility (≥0.288 and <0.545, class 3 in Table 2), and will be more probable for higher susceptibility values (classes 4 and 5 in Table 2). Most of the
landslides in the whole study area (94.28%) are included in areas showing susceptibility values higher than 0.288 or classes 3, 4 and 5.

Fig. 3. Landslide susceptibility distributions with correct classification percentages and RMSE values for the validation data sets

![Fig 3. Landslide susceptibility distributions with correct classification percentages and RMSE values for the validation data sets](image)

**Fig 4. The hazard map as a result of logistic regression analysis**

Table 2. The five susceptibility zones

<table>
<thead>
<tr>
<th>Reclassified value</th>
<th>Hazard class</th>
<th>Grid number</th>
<th>% area covered</th>
<th>Landslide grid number</th>
<th>% landslide area covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0-0.030818</td>
<td>extremely low</td>
<td>18074</td>
<td>1.20</td>
<td>4</td>
<td>0.02</td>
</tr>
<tr>
<td>0.030818-0.28815</td>
<td>Very low</td>
<td>400972</td>
<td>26.58</td>
<td>1247</td>
<td>5.70</td>
</tr>
<tr>
<td>0.28815-0.545486</td>
<td>low</td>
<td>462294</td>
<td>30.65</td>
<td>4124</td>
<td>18.86</td>
</tr>
<tr>
<td>0.545486-0.802821</td>
<td>medium</td>
<td>493547</td>
<td>32.72</td>
<td>11766</td>
<td>53.80</td>
</tr>
<tr>
<td>0.802821-0.934757</td>
<td>high</td>
<td>13348</td>
<td>8.85</td>
<td>4727</td>
<td>21.62</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>1508335</strong></td>
<td><strong>100.00</strong></td>
<td><strong>21868</strong></td>
<td><strong>100.00</strong></td>
</tr>
</tbody>
</table>

5. Conclusions

Landslide susceptibility map is a fundament of disaster risk evaluation. There are many GIS-based qualitative and quantitative techniques useful to analyze the relationship between landslides and landslide related factors. This study extended the application of logistic regression to prepare a susceptibility map based on GIS. The quality of susceptibility mapping was validated and this prediction capability shows that the landslide susceptibility map produced in this research paper can be used for the planning of protective and mitigation measures in the region.
6. References


