GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China

Shi-Biao Bai a,⁎, Jian Wang a, Guo-Nian Lü a, Ping-Gen Zhou b, Sheng-Shan Hou b, Su-Ning Xu b

a College of Geographical Sciences, Key Laboratory of Virtual Geographical Environment (Ministry of Education), Nanjing Normal University, Nanjing 210046, China
b China Institute of Geo-Environment Monitoring, Beijing 100081, China

1. Introduction

Historically the Three Gorges area of China is characterized by high landslide hazard. The area has received a large amount of attention because of the Three Gorges Dam and Reservoir Project and its potentially strong impact on the environment, geo-hazards, and socioeconomy (Zhou, 1996). More than 2500 known localities of slope instability exist in this area (Liu et al., 2004; Bai et al., 2005, 2009), and the construction of the dam dramatically increases the landslide hazard in this area. The process of creating landslide susceptibility maps can utilize a qualitative or quantitative approach (Soeters and van Westen, 1996; Aleotti and Chowdhury, 1999; Guzzetti et al., 1999). Qualitative methods were widely used during the late 1970s by engineering geologists and geomorphologists, whereas quantitative methods, which are based on numerical expressions of the relationship between controlling factors and landslides, have become popular in the last few decades, assisted by the developments of computer and geographic information system (GIS) technology (Zhou et al., 2002; Chung et al., 2002; Van Westen et al., 2003, 2008; Glade, 2005).

In general, the limitation of logistic regression modeling compared to the other multivariate statistical techniques (multiple regression analysis and discriminant analyses) is that the dependent variable can have only two values (a dichotomous outcome), and that the predicted values can be interpreted as probability because they are constrained to fall into an interval between 0 and 1 (Kleinbaum, 1991). However, the advantage of this binary logical regression is free of data distribution. Logical regression has been applied to susceptibility mapping extensively (Wieczorek et al., 1996; Atkinson and Massari, 1998; Affi and Clark, 1998; Guzzetti et al., 1999; Gorsevski et al., 2000; Dai and Lee, 2002; Ohlmacher and Davis, 2003; Súzen and Doyuran, 2004b; Ayalew and Yamagishi, 2005; Can et al., 2005; Chau and Chan, 2005; Wang and Sassa, 2005; Duman et al., 2006; Greco et al., 2006; Domínguez-Cuesta et al., 2007; Chang et al., 2007; Nefeslioglu et al., 2008; García-Rodríguez et al., 2008).

The main objective of this paper is to produce a landslide susceptibility map for a part of the Three Gorges area in China, using a logistic regression approach with landslide and landslide causative factor databases developed with GIS. To achieve this, the conventional logistic regression landslide susceptibility mapping method was refined by using neighborhood analysis (“seed cells”) and using the landslide density to transform these nominal variables to numeric variable.

2. Description of the study area

The Three Gorges have been formed by severe incision of massive limestone mountains, of lower Palaeozoic and Mesozoic ages (Ji Jialinjiang Group), along narrow fault zones, in response to Quaternary uplift (Li et al., 2001; Huang et al., 2001). Steep slopes develop on...
outcrops of easily erodible or ‘soft’ materials, which are extensive, and landslides are common in these areas (Wu et al., 2001). The average precipitation in this part of China is 100–150 mm per month and the spring–summer (March–August) average can be as high as 200–300 mm per month. Our study area is located in the Zhongxian–Shizhu segment (Fig. 1), which is known as one of the most landslide-prone areas in the Three Gorges Reservoir region of China. This area was selected to evaluate the frequency and distribution of landslides. The site lies between the latitudes 30.11° N and 30.37° N and the longitudes 107.86° E and 108.15° E, and covers an area of 260.9 km².

The geological formations of the study area are primarily Quaternary deposits and a Jurassic system. Seven geologic formations are recognized based on the similarity of their engineering properties (China Geological Survey, 1988; Fig. 2). They are loose Quaternary deposits (Q₄), Penglaizhen formation (J₃p), Suining formation (J₃s), Shangshaximiao formation (J₂s), Xiashaximiao formation (J₂xs), Zhenzhuchong formation (J₁z), and Xujiahe formation (T₃x). The Penglaizhen formation consists of purple, brown-red mudstones, sandy mudstones, and grey, greenish-grey–purple medium, and fine-grained feldspathic quartz–sandstones, occasionally intercalated by conglomeratic quartz–sandstones. The Suining formation mainly contains greenish-grey and purple fine-grained feldspathic quartz–sandstones, intercalated by purple, brown-red mudstones, and sandy mudstones. The Shangshaximiao formation is similar to the Xiashaximiao formation in lithologic characteristics, but has a coarser grain size of the sandstone and higher feldspar contents. The Xianshaximiao formation consists of purple mudstones interbedded with yellowish-grey medium and fine-grained feldspathic quartz–sandstones, and it is intercalated by yellowish-grey and greyish-purple fine-grained sandstones within the mudstones. The Zhenzhuchong formation has fine-grained quartz–sandstones, siltstones, mudstones, and shale. Finally, the Xujiahe formation contains fine-grained quartz–sandstones, feldspathic quartz–sandstones, and siltstones.

3. Spatial database design and construction

The identification and mapping of a suitable set of parameters related to landslide require a prior knowledge of their main causes (Guzzetti et al., 1999). The methodology applied in this study is based on the well-known principle of “present and past are keys to the future”. The fundamental principle of this approach to landslide susceptibility mapping is to use the characteristics of existing landslides to evaluate the possible areas of future landslides. For this reason, landslide and landslide causative factor databases were constructed. To store the information of these parameter maps in a uniform thematic database, the size of each pixel or cell for all the products was set to 25 × 25 m.

3.1. Landslide causative factors databases

The first dataset is the geological map. Geological paper maps at 1:10000 scale covering the study area were digitized and the seven geological formations and three petrofabric types were identified. The two largest datasets are topographical parameters that were collected from the 1:50000-scale paper topographic maps. A digital elevation model (DEM) was generated from a triangulated irregular network grid (TIN map) and used as the background layer for the digital elevation model. The topographical parameters included slope, aspect, elevation, aspect disparity, concavity, and ruggedness indices.

Fig. 1. The study area and its location.
(TIN) model that was derived from digitized contours with a contour interval of 25 m. The elevation, slope angle, aspect, terrain roughness, shape of the slope parameters, plan curvature and profile curvature (Fig. 3a–f) were obtained from the DEM using Surfer 8.0. The third dataset is land cover, which was interpreted from Landsat TM 5 (Path 127/Row 39, dated 7/2000) satellite imagery using various image processing and enhancement techniques. The interpreted images were then digitally processed to further modify the boundaries by supervised classification with ERDAS (Earth Resource Data Analysis System) software. The accuracy of the land-use interpretation was checked in the field. Eleven vegetation types were recognized. The fourth dataset is soil type. The dataset was digitized from the 1:50,000-scale paper soil maps. The fifth dataset includes distances from drainages, roads, and settlements. They were digitized from the 1:50000 topographic maps.

Average precipitation in this part of China is 100–150 mm per month and the spring–summer (March–August) average can be as high as 200–300 mm per month. Based on historic documents, earthquakes having a magnitude greater than 4 have never occurred in and around the study area. Furthermore, due to the size of the study area, the effects of both precipitation and seismicity can be regarded as uniform throughout the area. Therefore, they were not considered for the analyses.

3.2 Landslide databases

A detailed landslide-inventory map of the study area was constructed by visual interpretation of 1:20000 color aerial orthophotographs. The typical landslide morphology of visible scarps, hummocky topography and landslide dams can be recognized in the field. Extensive field studies were used to check the size and shape of landslides, to identify the type of their movements and the materials, and to determine the state of activity (active, dormant, etc.). A total of 142 landslides were identified (Fig. 1), which covered an area of 5.32 km², accounting for 2.04% of the study area. The minimum, mean and the maximum landslide areas are 0.003, 0.037 and 0.506 km² respectively. The properties of the landslides were recorded on a standard landslide-inventory data sheet. According to the landslide classification by Varnes (1978), Keefer (1984) and Cruden and Varnes (1996), the landslide inventory contains 21 shallow and 121 deep-seated landslides, including 103 earth slides and 39 rock slides. A digitized map of landslide boundaries was produced using ArcGIS 9.2. A vector-to-raster conversion was performed to provide a raster layer of the landslide areas.

Since the undisturbed morphological conditions before landslides occurred can be inferred from landforms in the vicinity of the landslide polygons, Süzen and Doyuran (2004a, b) and Nefeslioglu et al. (2008) used an approach called “seed cells” in generating decision rules of landslide occurrence. Seed cells are located in a buffer zone along the crown and flanks of each landslide. In the study area, 6793 seed cells were derived from landslide polygon boundaries (Fig. 4).

4. Mapping of landslide susceptibility

4.1 Logistic regression approaches

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 variable 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 and Lee, 2002).
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 (landslide) occurrence, which varies from 0 to 1 on an s-shaped curve; \( z \) is defined as the following equation (linear logistic model), and its value varies from \(-\infty\) to \(+\infty\):

\[ z = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n \]  

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 a logistic regression representing the presence or absence of landslides (at the present conditions) on the independent variables (pre-failure conditions).

In the application of the logistic regression model for landslide susceptibility mapping, some scientists have tried to exploit the data using dummy binary variables for each class of an independent parameter (Guzzetti et al., 1999; Dai and Lee, 2002; Ohlmacher and Davis, 2003). This increases the complexity of the data structure, and limits the flexibility of the statistical system. If many parameters are included, the regression equation will be very long, and it may even introduce numerical problems.

In this study, we extended the application of logistic regression by using the continuous data as they are, so that changes in state and information present in the parameter maps can be prevented. Using
landslide density to transform the nominal variables to numeric variables can avoid the creation of an excessively high number of dummy variables, and allow consideration of the so-called ‘previous knowledge’ of landslide susceptibility (Carrara, 1983). Landslide density is calculated as:

\[
\text{Landslide density} = \left( \frac{B_i}{A_i} \right) \sum_{i=1}^{N} \left( \frac{B_i}{A_i} \right)
\]

where \( A_i \) is the area of the \( i \)-th type of a certain parameter, \( B_i \) is the landslide area of the \( i \)-th type of the parameter, and \( N \) is the type number of the parameter.

### 4.2. Splitting the dataset and sampling

Landslide presence represents 2.04% of the study area, which is thousands times fewer than their absence, and therefore it is considered as a rare event (King and Zeng, 2001; Van Den Eeckhaut et al., 2006). It is generally recommended that the ratio of landslide presence (1)/landslide absence (0) should be equal to 1 in the training dataset (Süzen and Doyuran, 2004a; Nefeslioglu et al., 2008), but there are still many studies on unequal proportions of them (Atkinson and Massari, 1998; Guzzetti et al, 1999; Dai and Lee, 2002; Ohlmacher and Davis, 2003; Ayalew and Yamagishi, 2005; Van Den Eeckhaut et al., 2006; Domínguez-Cuesta et al, 2007). To overcome this problem, a series of sensitivity analyses were performed using different ratios of the number of landslides to that of no landslide. The values of the ratio were 1, 2, 3, 4, and 5, and we tried to use equal proportions of landslide and non-landslide samples. For this purpose, all locations of the landslide seed cells were thus used to extract the physical parameters (independent variables) automatically from the existing data layers. Moreover, 6793 random sample locations were chosen from the landslide-free area as samples representing the absence of landslide (Fig. 5). The landslide seed cells and the set of data were merged, and a new column of a binary variable was added to indicate the presence or absence of landslides. This stage was repeated four times to find whether there was any convergence in the success of logistic regression analyses by four different random sample location 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), it is commonly suggested that approximately 20% of the parent data selected at random should be used as the test data (Van Den Eeckhaut et al., 2006). Therefore, the sample of 13,586 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.

### 4.3. 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. TOL smaller than 0.2 is an indicator for multicollinearity between independent variables, and TOL smaller than 0.1 suggests serious multicollinearity (Menard, 1995). In this study, both indexes were calculated (Table 1), and variables with VIF > 2 and TOL < 0.4 were excluded from the logistic analysis (Allison, 2001).

### 4.4. 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 landslides to each factor. The forward stepwise logistic regression was carried out to incorporate only the predictor variables with an important contribution to the presence of
In this study, the significance level of the score $\chi^2$ for entering the model was set at 0.15. The significance level of the Wald $\chi^2$ for a variable to stay was set at 0.05. Variables such as petrofabrics type, soil type, slope angle, distance to drainage lines, plan curvature, land cover, elevation and aspect were selected for statistical significance.

These training sets were then evaluated using a $\chi^2$ value of the Hosmer–Lemeshow test, Cox and Snell $R^2$, and Nagelkerke $R^2$. The test showed that the goodness of fit of the equation could be accepted because the significance of $\chi^2$ is larger than 0.05. The values of Cox and Snell $R^2$ and Nagelkerke $R^2$ showed that the independent variables could explain the dependent variables in a way. The relative operating characteristic (ROC) value of 0.834 was obtained in this study, which reflects the sign of good correlation between the independent and dependent variables. Table 2 shows the overall statistics of the regression model.

### Table 1
Multicollinearity diagnosis indexes for independent variables.

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>TOL</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect</td>
<td>0.896</td>
<td>1.117</td>
</tr>
<tr>
<td>Elevation</td>
<td>0.600</td>
<td>1.666</td>
</tr>
<tr>
<td>Distance to settlement</td>
<td>0.930</td>
<td>1.075</td>
</tr>
<tr>
<td>Land cover</td>
<td>0.845</td>
<td>1.184</td>
</tr>
<tr>
<td>Local relief</td>
<td>0.129</td>
<td>7.778</td>
</tr>
<tr>
<td>Plan curvature</td>
<td>0.748</td>
<td>1.337</td>
</tr>
<tr>
<td>Profile curvature</td>
<td>0.663</td>
<td>1.508</td>
</tr>
<tr>
<td>Distance to drainage lines</td>
<td>0.715</td>
<td>1.399</td>
</tr>
<tr>
<td>Distance to road network</td>
<td>0.808</td>
<td>1.238</td>
</tr>
<tr>
<td>Slope angle</td>
<td>0.901</td>
<td>1.110</td>
</tr>
<tr>
<td>Terrain roughness</td>
<td>0.133</td>
<td>7.536</td>
</tr>
<tr>
<td>Shape of slope</td>
<td>0.651</td>
<td>1.535</td>
</tr>
<tr>
<td>Soil type</td>
<td>0.784</td>
<td>1.275</td>
</tr>
<tr>
<td>Petrofabrics type</td>
<td>0.942</td>
<td>1.062</td>
</tr>
</tbody>
</table>

Table 2 Overall statistics of the logistic regression model with eight independent variables.

<table>
<thead>
<tr>
<th>Hosmer and Lemeshow test $\chi^2$</th>
<th>df</th>
<th>Sig.</th>
<th>–2 Log likelihood</th>
<th>Cox and Snell $R^2$</th>
<th>Nagelkerke $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>666.428</td>
<td>8</td>
<td>0.281</td>
<td>12,406.168</td>
<td>0.317</td>
<td>0.425</td>
</tr>
</tbody>
</table>

Table 3 Regression coefficients obtained for the eight independent variables.

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect</td>
<td>0.00091</td>
</tr>
<tr>
<td>Elevation</td>
<td>-0.00413</td>
</tr>
<tr>
<td>Slope angle</td>
<td>0.03662</td>
</tr>
<tr>
<td>Land cover</td>
<td>0.00069</td>
</tr>
<tr>
<td>Petrofabrics type</td>
<td>0.00097</td>
</tr>
<tr>
<td>Plan curvature</td>
<td>0.22566</td>
</tr>
<tr>
<td>Soil type</td>
<td>0.00075</td>
</tr>
<tr>
<td>Distance to drainage lines</td>
<td>-0.00130</td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.09592</td>
</tr>
</tbody>
</table>

In addition, the accuracy of the classification for all training sets was calculated by the SPSS software package. The predicted accuracy of calibration dataset was 82.6% for landslides and 78.1% for non-landslides; the overall predicted accuracy was 84%. The overall success of the classification was 80.4%. After logistic regression analysis with the eight significant independent variables, the regression coefficients were obtained (Table 3).

### 4.5. Validation

Using the logistic regression coefficients, the landslide occurrence probability was computed (Fig. 6). The probability ranges between 0 and 1, where 0 represents no probability of a landslide occurrence and 1 represents certainty of a landslide occurrence. The probability ranges were divided into several classes (Fig. 6). The classification results were then evaluated using the receiver operating characteristic (ROC) curve, which is a graphical representation of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

Fig. 5. Distribution of randomly selected 6793 landslide-free grids, 6793 seed cell grids and 8499 landslides grids.
Fig. 6. Probability map obtained by logistic regression analyses.

Fig. 7. Reclassified landslide susceptibility map.
0.018 and 0.957, and if values are closer to 1, landslides are more likely found. Can et al. (2005) mentioned two rules for a spatially effective landslide susceptibility mapping: i) observed landslide areas should coincide with the areas having high-susceptibility values, and ii) high-susceptibility values should cover only small areas. When the susceptibility value of 0.5 was used as the cut-off value (Dai and Lee, 2002), 81.97% of the observed landslides are located on the areas having high-susceptibility values.

The correct classification percentage and the root mean square error (RMSE) were calculated to validate the dataset and evaluate the performance of the models constructed during the training stages. The predicted accuracy of the validation dataset for landslides is 84.4% and that for non-landslides is 78.3%; the overall predicted accuracy is 81.4%, and the RMSE value is 0.392.

In obtaining a landslide susceptibility map, the method adopted in literature is to divide the histogram of the probability map into different categories (Guzzetti et al., 1999; Dai and Lee, 2002; Ohrmacher and Davis, 2003; Ayalew et al., 2004; Süzen and Doyuran, 2004a,b; Ayalew and Yamagishi, 2005). In this study, the standard deviation method was used to generate class breaks, and the probability map was divided into five susceptibility classes: very low, low, medium, high, and very high (Fig. 7). Among the five susceptibility zones, 18.15% of the study area was designated as very low susceptible zone (Table 4), and low, medium, high, and very high susceptible zones made up 36.24%, 26.72%, 16.09%, and 2.80%, respectively. As noted, the area with landslides covered 2.04% of the study area. The approximately equal percentages for the landslide area and the very highly susceptible area indicate that the system adopted to divide the probability map was appropriate. Then, the validity of the result was checked using the data for the whole study area. Our working hypothesis was that an active landslide must be triggered in areas with at least medium values of susceptibility (class 3 in Table 4), and will be more likely for higher susceptibility values (classes 4 and 5 in Table 4). Most of the landslides in the whole study area (89.50%) and 81.69% of seed cells occur in areas with susceptibility classes 3 to 5.

5. Conclusions

Landslides in the Three Gorges Reservoir region of China have received considerable attention. The construction of the dam and other human engineering activities severely affect the geological environment in the region, and thus landslides and collapses occur frequently. There are many GIS-based qualitative and quantitative techniques useful to analyze the relationship between landslides and environmental factors. This paper applied a logistic regression model within the framework of a GIS for landslide susceptibility mapping in a 260.9 km² study area in the landslide-prone Three Gorges Reservoir region.

In the study area, the seed cells were used to be the dependent variable in order to construct a reliable logistic regression model. The seed cells were derived from landslide polygon boundaries and represented the pre-failure undisturbed morphological conditions. The method differed from using all landslide area with depletion and accumulation zones as the dependent variable. The difference between the proposed logistic regression model and the existing models in statistical landslide susceptibility mapping is that the latter uses all continuous variables as they are, and the landslide density is used to transform nominal variables to numeric variables.

A susceptibility map shows where landslides may occur. The landslide susceptibility map ranks the slope stability of the study area in categories that range from stable to unstable. The susceptibility mapping was validated and this prediction capability shows that the landslide susceptibility map produced in this research paper can be used for planning protective and mitigation measures by the local and regional authorities.

The main types of maps related to landslide mitigation are landslide inventory, landslide susceptibility, landslide hazard and landslide risk maps. Landslide susceptibility maps are fundamental to quantify landslide risk. The methods used in this study may also be applicable to other landslide-prone areas with similar conditions.

Acknowledgements

The authors would like to thank Professors Cees van Westen, Mehmet Lütfi Süzen and Lulseged Ayalew for their help. This study was supported by the National Natural Science Foundation of China (Project Nos. 40801212 and 49971064, and Key Project No. 2007CB416602), the National Key Basic Research Program of the early special issues (No. 2007CB416602), and the Opening Fund of State Key Laboratory of Geohazard Prevention and Geoenvironment Protection of Chengdu University of Technology, China (No. GZ2007–11).

References


