APPLICATION AND CROSS-VALIDATION OF SPATIAL LOGISTIC MULTIPLE REGRESSION FOR LANDSLIDE SUSCEPTIBILITY ANALYSIS

SARO. LEE
Geoscience Information Center, Korea Institute of Geology & Mineral Resources (KIGAM)
30, Gajung-dong, Yusung-gu, Daejeon, 305-350, Korea
leesaro@kigam.re.kr

Abstract: The aim of this study is to apply and cross-validate a spatial logistic multiple-regression model at Boun, Korea, using a Geographic Information System (GIS). Landslide locations in the Boun area were identified by interpretation of aerial photographs and field surveys. Maps of the topography, soil type, forest cover, geology, and land-use were constructed from a spatial database. The factors that influence landslide occurrence, such as slope, aspect, and curvature of topography, were calculated from the topographic database. Texture, material, drainage, and effective soil thickness were extracted from the soil database, and type, diameter, and density of forest were extracted from the forest database. Lithology was extracted from the geological database and land-use was classified from the Landsat TM image satellite image. Landslide susceptibility was analyzed using landslide-occurrence factors by logistic multiple-regression methods. For validation and cross-validation, the result of the analysis was applied both to the study area, Boun, and another area, Youngin, Korea. The validation and cross-validation results showed satisfactory agreement between the susceptibility map and the existing data with respect to landslide locations. The GIS was used to analyze the vast amount of data efficiently, and statistical programs were used to maintain specificity and accuracy.

Keywords: Landslide; Susceptibility; GIS; Logistic regression; Validation

1. Introduction

There are frequent landslides in Korea that often result in significant damage to people and property. The most recent occurred in 1996, 1998, and 1999. In the study area, Boun, much damage was caused on these occasions. The reason for the landslides was heavy rainfall, and, as there was little effort to assess or predict the event, damage was extensive. Through scientific analysis of landslides, we can assess and predict landslide-susceptible areas and so landslide damage can be reduced through proper preparation. In order to achieve this, landslide hazard analysis techniques were developed, applied, and validated in the study area using spatial logistic multiple-regression.

In logistic multiple-regression, a corresponding comparison is to assess the logistic relationship between each independent variable, such as each factor, and each dependent variable, such as landslide occurrence. Logistic multiple-regression allows one to form a multivariate regression relationship between a dependent variable and several independent variables. The advantage of logistic multiple-regression over simple multiple regressions is that, through the addition of an appropriate link function to the usual linear regression model, the variables may be continuous, categorical, or any combination of these. In the present situation, the dependent variable is a binary variable representing the presence or absence of landslides. Where the dependent variable is binary, use of the logistic link function is appropriate (Atkinson and Massari, 1998).

The Boun area of study had much landslide damage following heavy rain in 1998 and was selected as a suitable case to evaluate the frequency and distribution of landslides. The site lies between the latitudes 36 °25' 21" N and 36° 30' 00" N, and longitudes 127° 39' 36" E and 127° 45' 00" E, and covers an area of 68.43km^2. The bedrock geology of the study area consists mainly of biotite granite. In the study area, the landslides were mainly soil slide and the landslides occurred when the maximum daily rainfall is 407 mm. For cross-validation, another study area, Youngin, Korea was used. The Youngin study area had high landslide damage after heavy rain in 1991 and was selected as a suitable case to evaluate the frequency and distribution of landslides. The site lies between the latitudes 37.14° N and 37.19° N and longitudes 127.11° E and 127.23° E, and covers an area of 66 km^2. In the study area, the landslides were mainly debris flows and shallow soil slips that occurred during 3–4 hours of high intensity rainfall, or shortly afterwards. The landslides occurred when the maximum daily rainfall exceeded 114 mm, with a maximum hourly rainfall of 40 mm. The bedrock geology of the study area consists mainly of granite and gneiss.

A key assumption in using this approach is that the potential (occurrence possibility) of landslides will be comparable to the actual frequency of landslides. Following selection of the study area, the places where landslides had occurred in the Boun area were identified by interpretation of aerial photographs and field surveys. A map of recent landslides was developed from 1:20,000 scale aerial photographs, in combination with the GIS, and this was used to evaluate the frequency and distribution of shallow landslides in the area. Topography, soil, forest, geology, and land-use databases were constructed as part of the analysis. Topographic factors such as altitude, slope, aspect, and curvature were extracted from the topographic database; soil texture, material, drainage, effective thickness, and topography from the soil database; forest type, forest diameter, and forest density from the forest map; lithology from the geological database; and land-use data from Landsat TM images. Using the detected landslide locations and the constructed spatial database, a landslide analysis method was applied and validated. For this, the calculated and extracted factors were converted to a 5 m x 5 m grid (ARC/INFO GRID type). The grid data was converted into ASCII file and then imported to the statistical program used. Then, using a logistic multiple-regression model, the spatial relationships between the landslide location and each landslide-related factor, such as topography, soil, forest, geology, and land-use, were analyzed, and a formula of landslide occurrence possibility was extracted using the relationships in the statistical program. This formula was used to calculate the landslide susceptibility index and was mapped using the grid. The susceptibility map was validated using existing landslide locations. For cross-validation, the formula was also applied to another study area, Youngin in Korea, to calculate and map the landslide susceptibility index there. The same spatial database and factors for the Youngin area as for Boun were available to us. Finally, the susceptibility map was again validated using the existing landslide location.

In this study, Geographic Information System (GIS) software, ArcView 3.2 and ARC/INFO 8.1 NT software, and the statistical software SPSS 12.0 were used as the basic analysis tools for spatial management and data manipulation.

2. SPATIAL DATABASE

Identification and mapping of a suitable set of instability factors bearing a relationship to slope failures requires an *a priori* knowledge of the main causes of landslides (Guzzetti et al., 1999). These instability factors include surface and bedrock lithology and structure, bedding altitude, seismicity, slope steepness and morphology, stream evolution, groundwater conditions, climate, vegetation cover, land-use, and human activity. The availability of such thematic data varies greatly, depending on the type, scale, and method of data acquisition. Geomorphologic, lithological, structural geologic, soil, forest, and land-use data should be available for the entire area. To apply the logistic multiple-regression method, maps relevant to landslide occurrence were constructed to a vector-type spatial database using the GIS software ARC/INFO. These included 1:5,000 scale topographic maps, 1:25,000 scale soil maps, 1:25,000 scale forest maps, and 1:50,000 scale geological maps. These data are available in Korea either as a paper map or as a digital map. The land-use was classified from satellite images such as those from Landsat TM.

There are 13 factors considered in calculating the landslide probability. These factors were extracted from the constructed spatial database. Contour and survey base points had their elevation value read from the topographic map and were used to build a Digital Elevation Model (DEM). The DEM has 5 m resolution. Using the DEM, the slope angle, slope aspect, and slope curvature were calculated. The topography, texture, drainage, material, and thickness of soil were acquired from the soil map, and the type, diameter, and density of forest were obtained from the forest maps. The lithology map was obtained from the geologic map. Finally, land-use data were classified from a LANDSAT TM image using the unsupervised classification method. The five classes (urban, water, forest, agricultural area, and barren area) were extracted for land-use mapping.

3. THEORY OF LOGISTIC MULTIPLE REGRESSION

Logistic regression allows one to form a multivariate regression relation between a dependent variable and several independent variables. Logistic regression, which is one of the multivariate analysis methods, is useful for predicting the presence or absence of a characteristic or outcome based on values of a set of predictor variables. 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. In the case of multi-regression analysis, the factors must be numerical, and in the case of a similar statistical method, determinant analysis, the variables must have a normal distribution. In the present situation, the dependent variable is a binary variable representing presence or absence. Where the dependent variable is binary, the logistic link function is applicable (Atkinson and Massari, 1998). For this study, the dependent variable must be input as either 0 or 1, so the method applies well to landslide occurrence possibility analysis. Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model.

In the present situation, the dependent variable is a binary variable representing the presence or absence of landslides. Quantitatively, the relationship between the occurrence and its dependency on several variables can be expressed as:

$$p = \exp(z) / (1 + \exp(z)) \quad (1)$$

where p is the probability of an event occurring. In the present situation, the value p is the estimated probability of landslide occurrence. The probability varies from 0 to 1 on an S-shaped curve and z is the linear combination. It follows that logistic regression involves fitting an equation of the following form to the data:

$$z = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n \qquad (2)$$

where z is parameter, b_0 is the intercept of the model, the b_i ($i = 0, 1, 2, \ldots, n$) are the slope coefficients of the logistic regression model, and the x_i ($i = 0, 1, 2, \ldots, n$) are the independent variables. The linear model formed is then a logistic regression of presence or absence of landslides (present conditions) on the independent variables (pre-failure conditions).

4. APPLYING AND INTERPRETING LOGISTIC MULTIPLE-REGRESSION FOR

LANDSLIDE SUSCEPTIBILITY MAPPING

A key concept for understanding the tests used in logistic multiple-regression is that of log likelihood. Usually, however, the overall significance is tested using the chi-squared test, which is derived from the likelihood of observing the actual data under the assumption that the model that has been fitted is accurate. It is convenient to use -2 times the log (base e) of this likelihood (-2LL). The log likelihood value (-2LL) here is 8418.480. Several criteria can be used to guide entry: these include the greatest reduction in the –2LL values, or the greatest Wald coefficient.

There are Wald statistics for each regressor in each model, together with a corresponding significance level. The Wald statistic has a chi-squared distribution, but apart from that, it is used in just the same way as the *t* values for individual regressors in linear regression.

In assessing model fit, several measures are available. Smaller values of the –2LL measure indicate better model fit. The goodness-of-fit measure compares the predicted probabilities to the observed probabilities, with higher values indicating better fit. The value for the single variable model is 8418.480. Next, three measures comparable to the R^2 measure in multiple regression are available.

Using the logistic multiple-regression method, the spatial relationship between landslide-occurrence location and landslide-related factors was calculated. The statistical method used was logistic multiple-regression analysis. A statistical program was used to calculate the correlation of a landslide event to each factor. Firstly, all of the factors that were constructed in the database were considered, and then logistic multiple-regression coefficients of the factors were calculated. The coefficients of the logistic multiple-regression model were estimated using the maximum-likelihood method. In other words, coefficients that make the observed results most likely are selected. Since the relationship between the independent variables and the probability is nonlinear in the logistic multiple-regression model, an iterative algorithm is necessary for parameter estimation (Dai and Lee, 2002). There are positive associations, such as slope, and negative associations, such as curvature. After interpretation, formulas (1) and (3), which predict the landslide-occurrence possibility, were created.

$$z = (0.0262 \times \mathrm{SLOPE}) + (-0.0245 \times \mathrm{CURVA}) + \mathrm{TOPOw} + \mathrm{TEXTUREw} + \mathrm{MATERIALw} + \mathrm{DRAINw} + \mathrm{THICKw} + \mathrm{TYPEw} + \mathrm{DIAMETERw} + \mathrm{DENSITYw} + \mathrm{GEOLw} + \mathrm{LANDUSEw} - 33.173 \quad (3)$$

where Slope is slope value; Curva is Curvature value; TOPOw, TEXTUREw, MATERIALw, DRAINw, THICKw, TYPEw, DIAMETERw, DENSITYw, GEOLw, and LANDUSEw are logistic multiple-regression coefficients; z is a parameter; and p is the landslide-occurrence possibility.

Using these formulae, a landslide susceptibility map was made. Logistic multiple-regression analysis was performed by dividing the study area into a 5 m × 5 m sized grid, and the factors were divided into a 5 m × 5 m array and converted to an ASCII file to use in the statistical package. In the study area, the total cell number was 2,729,160 and the cell number where landslides occurred was 483. The calculated possibility value was classified by equal areas and grouped into five classes for easy interpretation: very low (0.00000), low (0.00000–0.00003), medium (0.00003–0.00010), high (0.00010–0.00030), and very high (>0.00030). Using the coefficients and formulas (1) and (3), the other study area, Youngin, was analyzed for cross-validation of landslide susceptibility. Logistical multiple regression analysis is performed for the Youngin area. In this study area, the total cell number is 2,633,346 and the cell number where landslides occurred is 1,149. The calculated possibility value is classified by equal areas and grouped into five classes: very low (0.0000), low (0.0000–0.0009), medium (0.0009–0.0033), high (0.0033 –0.0083), very high (0.0083 <).

5. CROSS-VALIDATION OF LANDSLIDE SUSCEPTIBILITY

For validation of these landslide susceptibility calculation methods, two basic assumptions are needed. One is that landslides are related to spatial information such as topography, soil, forest, geology, and land-use, and the other is that future landslides will be precipitated by a specific impact factor, such as rainfall or an earthquake. In this study, the two assumptions are satisfied because the landslides are related to the spatial information, and the landslides were precipitated by one cause, heavy rainfall in the Boun and Youngin areas.

The landslide susceptibility analysis result was validated using the landslide locations for the Boun study area and cross-validated using the landslide locations for the Youngin study area. The validation method was performed by comparison of existing landslide data and landslide susceptibility analysis results for the Boun study area. The comparisons are performed using the logistic multiple-regression method at the cases of success rate and prediction rate. The success rates illustrate how well the estimators perform with respect to the left-side landslides used in constructing those estimators. The prediction rates, on the other hand, are used as measurements of how well the probability model and its estimators predict the distribution of future landslides (Chung and Fabbri, 1999). To obtain the relative ranks for sucess pattern, the calculated index values of all cells in the study area were sorted in descending order. The success rate validation results were divided into classes of accumulated area ratio percentage, according to the landslide susceptibility index value. The above procedure also was adapted for the Youngin area by comparing the classes obtained with the distribution in Youngin to obtain the prediction rate. The success rate validation results have obtained by comparing the susceptibility calculation results and landslide occurrence location using the logistic multiple-regression method. The success rate validation results are divided into classes of accumulated area ratio percentage according to the landslide susceptibility index value. For example, the 90–100% (10%) class that has highest possibility of landslide contains 63.8% of the Boun area in its success rate using the logistic multiple-regression method. A 0–20% class (20%) contain 74.9%, and the 0–30% class (30%) contains 80.0% of the study area. The values of 63.8%, 74.9%, and 80.0% are very high, and very accurately reproduce the existing result (Lee et al., 2002; Pistocchi et al., 2002). The success rate validation is from the landslide susceptibility analysis result validated in the Boun area using the landslide occurrence locations and logistic multiple-regression methods. Therefore, strictly speaking, the success rate is not a perfect validation method. However, the success rate validation method needs information about the properties of the analysis method, and checks the landslide susceptibility analysis calculation for major errors. It also needs to be tested against the prediction rate validation method.

The prediction rate validation results have found by comparing the susceptibility calculation results and landslide occurrence locations using the logistic multiple-regression method. The prediction rate validation results are divided into classes with accumulated area percentage according to landslide susceptibility index value. For example, the 90–100% (10%) class, with the highest possibility of landslide contains 45.4% of the Boun area in its success rate using the logistic multiple-regression method. The 0–20% class (20%) contains 57.0%, and the 0–30% class (30%) contains 68.4% of the study area. In addition, the success rate is better than the prediction rate for all classes. The prediction rate validation is from the landslide susceptibility analysis result validated in the Youngin area using some landslide occurrence locations that were unused in the calculation. Therefore, strictly speaking, the prediction rate is a true validation method.

6. CONCLUSIONS AND DISCUSSION

Landslides are among the most hazardous natural disasters. Government and research institutions worldwide have attempted for years to assess the hazard of landslides, estimate their risk, and show their spatial distribution. In this study, a statistical approach to estimating the susceptibility of an area to landslides using aerial photography and the GIS is presented. For the landslide susceptibility analysis, landslide location was detected using aerial photographs, and a landslide-related database was constructed for the study area of Boun and Youngin, Korea. For the landslide susceptibility analysis, logistic multiple-regression methods were applied and validated for the study area of Youngin, Korea, using the spatial database. Generally, the validation results showed satisfactory agreement between the susceptibility map and the existing data on landslide locations. With respect to the Boun study area, the success rates of the logistic multiple-regression method showed more accurate result than prediction rates of the Youngin study area. Generally, the success rate is higher than the prediction rate for all classes.

The statistical program can allow analysis of landslide susceptibility but it is inconvenient for the management of spatial data, and modification of its input data is difficult. A GIS has none or few functions for statistical and artificial neural network analyses but has many functions for database construction, display, printing, management, and analysis. Therefore, it is necessary to integrate the GIS and statistics to reduce the restrictions of using the three applications separately. The benefits of integrating GIS and statistical programs are efficiency and ease of management, input, display, and analysis of spatial data for landslide susceptibility.

Landslide susceptibility maps are of great help to planners and engineers for choosing suitable locations to implement developments. These results can be used as basic data to assist slope management and land-use planning.

REFERRENCE

P.M.Atkinson and R.Massari, 1998, Generalized linear modeling of susceptibility to landsliding in the central Apennines, Italy. *Computer & Geosciences*, 24(4), 373-385.

C.F. Chung, A.G.Fabbri, 1999, Probabilistic prediction models for landslide hazard mapping, *Photogrammetric Engineering & Remote Sensing,* 65:1389-1399.

F.C.Dai and C.F.Lee, 2002, Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. *Geomorphology,* 42:213– 228.

F.Guzzetti, A.Carrarra, M.Cardinali and P.Reichenbach, 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy, *Geomorphology,* 31:181-216.

J.F.Hair, R.E.Anderson, R.L.Tatham and W.C.Black, 1998, Multivariate data analysis. 5ed., Prentice-Hall, London.

S.Lee, and U.C.Choi, 2003, Development of GIS-based geological hazard information system and its application for landslide analysis in Korea, *Geosciences Journal,* 7(3): 243-252.

S.Lee S, J.Choi and K.Min, 2002, Landslide susceptibility analysis and verfication using the Bayesian probability model, *Environmental Geology.* 43:120-131.

A.Pistocchi, L.Luzi and P.Napolitano, 2002, The use of predictive modeling techniques for optimal exploitation of spatial databases: a case study in landslide hazard mapping with expert system-like methods. *Environmental Geology*, 41: 765–775.