Cross-Verification of Spatial Logistic Regression for Landslide Susceptibility Analysis: A Case Study of Korea

S. Lee [a,*]

[a]Geoscience Information Center, Korea Institute of Geoscience and Mineral Resources (KIGAM) 30, Gajeong-Dong, Yuseong-Gu, Daejeon, 305-350, Korea - leesaro@kigam.re.kr

Abstract - The aim of this study is to cross-verify of multiple logistic model at Korea using a Geographic Information System (GIS). Landslide locations were identified in the 3 study areas from interpretation of aerial photographs and satellite image, field surveys, and maps of the topography, soil type, forest cover and land cover were constructed to spatial data-sets. 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 cover was classified from the Landsat TM image satellite image. Landslide susceptibility was analyzed using the landslide-occurrence factors by multiple logistic regression models. For the verification and cross-verification, the result of the analysis was applied to study areas.

Keywords: Susceptibility; GIS; Multiple logistic regression; Verification; Korea

1. INTRODUCTION

In Korea, frequent landslides often result in significant damage to people and property, the most recent having occurred in 1991, 1996, 1998, 1999 and 2002. In the study area, Boun, Janghung and Youngin in Korea, 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 thus decrease landslide damage through proper preparation. In order to achieve this, landslide hazard analysis techniques were verified in the study area using multiple logistic regression models.

A key assumption using this approach is that the potential (occurrence possibility) of landslides will be comparable to the actual frequency of landslides. First, the study area was selected. Then landslide occurrence areas were detected in the Boun area, Korea by interpretation of aerial photographs and field surveys. A map of recent landslides was developed from aerial photographs, in combination with the GIS, and this was used to evaluate the frequency and distribution of shallow landslides in the area. The factors such as altitude, slope, aspect and curvature 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, and land cover data from Landsat TM image were used. Using the detected landslide locations and the constructed spatial data sets, a landslide analysis method were applied and verified. For this, the calculated and extracted factors were converted to a 10m × 10m grid (ARC/INFO GRID type). Using the detected landslide locations and the constructed spatial data-sets, a multiple logistic regression model was applied and landslide susceptibility map was made. Then, the susceptibility map was verified using existing landslide location.

The first study area, Boun, 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. The landslides occurred where the maximum daily rainfall is 407 mm. The second study area, Janghung lies between latitudes 37°43' N and 37°46' N, and longitudes 126°56' E and 127°01' E, and covers an area of 40.74 km^2. The study area is in the northwestern part of the Kyonggi gneiss complex, which is composed mainly of gneisses. In the study area, the landslides occurred where the maximum daily rainfall is 208.5 mm. The third study area, the Youngin, 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. The bedrock geology of the study area consists mainly of granite and gneiss. The landslides occurred where the maximum daily rainfall exceeded 114 mm, with a maximum hourly rainfall of 40 mm.

In this study, GIS (Geographic Information System) software, ArcView 3.2 and ARC/INFO 8.1 NT version, and statistical software, SPSS 10.0 were used as the basic analysis tool for spatial management and data manipulation.

2. SPATIAL DATABASE

Identification and mapping of a suitable set of instability factors (thematic mapping) bearing a relationship with slope failures requires an a priori knowledge of the main causes of landslides (Guzzetti and others 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 thematic data varies largely, depending on the type, scale, and method of data acquisition. A digitized map of landslide boundaries was produced, and these digital data were input to the GIS. A vector-to-raster conversion was undertaken to provide a raster data of landslide areas. Maps relevant to landslide occurrence were constructed in vector-type spatial data sets using the ARC/INFO GIS software package. These included 1:5000-scale topographic maps, 1:25000 or 1:50,000-scale soil maps, and 1:25000-scale forest maps. In the Janghung, 1:50,000-scale soil map was used because there is no published

1:25:000-scale soil map. A land-use map was extracted from Landsat TM satellite images having a resolution of 30 m. Contour and survey base points that had an elevation value read from the topographic map were extracted, and a Digital Elevation Model (DEM) was constructed. Using the DEM, the slope, aspect and curvature were calculated. The topographic type, texture, drainage, material, and thickness were acquired from a soil map. The type, diameter, age and density were obtained from forest maps, and land cover data was classified according to LANDSAT TM satellite images. In the study areas, the data sets were was divided into a grid with 10 m × 10 m cells. The Boun data set was composed of 555 rows by 734 columns, so the total cell number is 407,370 and the cell number where landslides occurred is 107. The Janghung data set was composed of 555 rows by 734 columns, so the total cell number is 407,370 and the cell number where landslides occurred is 107. The Youngin data set was composed of 555 rows by 734 columns, so the total cell number is 407,370 and the cell number where landslides occurred is 107.

3. LOGISTIC MULTIPLE REGRESSION

Logistic multiple regression allows one to form a multivariate regression relation between a dependent variable and several independent variables. A limitation of ordinary linear models is the requirement that the dependent variable is numerical rather than categorical. But many interesting variables are categorical in landslide analysis. The logistic multiple regression is easier to use than discriminant analysis when we have a mixture of numerical and categorical regressors, because it includes procedures for generating the necessary dummy variables automatically. Just like linear regression, logistic multiple regression gives each regressor a coefficient b_1 that measures the regressor's independent contribution to variations in the dependent variable. But there are technical problems with dependent variables that can only take values of 0 and 1.

The advantage of logistic multiple regression over simple multiple regression is that, through the addition of an appropriate link function to the usual linear regression model, the variables may be either continuous or categorical or any combination of both types. Moreover, when the dependent variable has only two groups, logistic multiple regression may be preferred over discriminant analysis that is also can use categorical data for several reasons. First, discriminant analysis relies on strictly meeting the assumptions of multivariate normality and equal variance-covariance matrices across groups-assumptions that are no met in many situations. Logistic multiple regression does not face these strict assumptions and is much more robust when these assumptions are not met, making its application appropriate in many more situations. Second, even if the assumptions are met, many researchers prefer logistic multiple regression because it is similar to regression. Both have straightforward statistical tests, the ability to incorporate nonlinear effects, and a wide range of diagnostics. For these and more technical reasons, logistic multiple regression is equivalent to two-group discriminant analysis and may be more suitable in many situations (Hair et al., 1998).

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:

$$p = 1 / (1 + e^{-z}) \quad (1)$$

where p is the probability of an event occurring. In the present situation, the p is the estimated probability of landslide occurrence. The p is estimated probabilities of landsliding based on the intrinsic properties only, and this we term susceptibility to landsliding. The probability varies from 0 to 1 on an S-shaped curve and z is the linear combination. It follows that logistic multiple regression involves fitting to the data an equation of the form.

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

where b_0 is the intercept of the model, the b_i (i = 0, 1, 2, …, n) are the slope coefficients of the logistic multiple regression model and the x_i (i = 0, 1, 2, …, n) are the independent variables (Dai and Lee, 2002). The linear model formed is then a logistic multiple regression of presence or absence of landslides (present conditions) on the independent variables (pre-failure conditions).

Although logistic multiple regressions finds a "best fitting" equation just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, which maximizes the probability of getting the observed results given the fitted regression coefficients. A consequence of this is that the goodness of fit and overall significance statistics used in logistic multiple regression is different from those used in linear regression.

The decision process for logistic multiple regression is, as with all multivariate applications, setting the objectives is the first step in the analysis. Then the researcher must address specific design issues and make sure the underlying assumptions are met. The analysis proceeds with the derivation of the logistic function and the determination of whether a statistically significant function can be derived to separate the tow groups. The logistic multiple regression results are then assessed for predictive accuracy by developing a classification matrix. Next, interpretation of the discriminant function determines which of the independent variables contributes the most to discriminating between the groups. Finally, the logistic function should be verified with a holdout sample (Hair et al., 1998).

4. APPLICATION 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, though, overall significance is tested using Model Chi-square, 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: greatest reduction in the –2LL values, 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 compared 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. The Cox and Snell R^2 and Nagelkerker R^2 measure operates which higher values indicating greater model fit. In our instance, the Cox and Snell value is .000 and the Negelkerke value is .096.

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 and calculated the correlation of landslide to each factor. First all of the factors that were constructed in the database were considered. Then logistic multiple regression coefficients of the factors calculated. The coefficients of the logistic multiple regression model are estimated using the maximum-likelihood method. In other words, the 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 association such as slope and negative association such as curvature. After interpretation, formulae (3) and (4), which predict the landslide-occurrence possibility, were created.

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

$$p = 1 / (1 + e^{-z}) \text{ or } p = e^{z} / (1 + e^{z}) \quad (4)$$

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

Using these formulae, a landslide susceptibility map was made. The logistic multiple regression analysis is performed by dividing the study area into a 5 m × 5 m size grid, and the factors were divided into a 5 m × 5 m, and converted to an ASCII file to use the statistical package. In the study area, the total cell number is 2,729,160 and the cell number where landslides occurred is 483. The distribution of the calculated the possibility is made to landslide susceptibility map. The value is 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), very high (0.00030 <). Also, using the formulae (3) and (4), the other study area, Youngin, was analyzed for cross-verification of landslide susceptibility. The logistical multiple regression analysis is performed by dividing the study area and the factors. In the study area, the total cell number is 2,633,346 and the cell number where landslides occurred is 1,149. The distribution of the calculated possibility is shown as map. The 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-VERIFICATION OF LANDSLIDE SUSCEPTIBILITY MAPPING

The landslide susceptibility analysis result verified using the landslide locations for the same study areas and cross-verified using the landslide locations of the others study areas. The verification method was performed by comparison of existing landslide data and landslide susceptibility analysis results for the Boun of the study area. The comparison results are shown as a line graph, with logistic multiple regression method at the case 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.

To obtain the relative ranks for each prediction pattern, the calculated index values of all cells in the study area were sorted in descending order. The above procedure also was adapted for the Janghung and Youngin of the study area by comparing the classes obtained with the distribution on the Janghung and Youngin of the study area.

The success rate verification is from the landslide susceptibility analysis result verified in the Boun area using the landslide occurrence locations, for the logistic multiple regression methods. Therefore, strictly speaking, the success rate is not a suitable verification method. However, the success rate verification method needs information about the properties of analysis method, and checks the landslide susceptibility analysis calculation for major errors. It also needs to be tested against the prediction rate verification method.

The success rate verification results are divided into classes of accumulated area ratio % according to the landslide susceptibility index value. In the case of Boun, the 90 to 100% (10%) class that highest possibility of landslide contains 51.9% of the Boun area in success rate. A 0-20% class (20%) contain 71.4% and 0-30% class (30%) contain 86.3% of the Boun area. In the case of Janghung, the 90 to 100% (10%)

class that highest possibility of landslide contains 55.5% of the Boun area in success rate. A 0-20% class (20%) contain 77.8% and 0-30% class (30%) contain 92.5% of the study area. In the case of Youngin, the 90 to 100% (10%) class that highest possibility of landslide contains 43.5% of the Boun area in success rate. A 0-20% class (20%) contain 67.2% and 0-30% class (30%) contain 83.7% of the study area.

The prediction rate verification is from the landslide susceptibility analysis result verified in the Youngin area using the landslide occurrence locations that were unused in the calculation. Therefore, strictly speaking, the prediction rate is a true verification method. The prediction rate verification results are divided into classes with accumulated area % according to landslide susceptibility index value. In the case of Youngin ratio for Janghung and Boun, the 90 to 100% (10%) class that highest possibility of landslide contains 36.5% of the Janghung area and 36.7% of the Youngin area in prediction rate. In the case of Janghung ratio for Youngin and Boun, the 90 to 100% (10%) class that highest possibility of landslide contains 29.3% of the Boun area and 38.0% of the Youngin area in prediction rate. In the case of Boun for Youngin and Janghung, the 90 to 100% (10%) class that highest possibility of landslide contains 29.0% of the Janghung area in prediction rate.

6. CONCLUSION AND DISCUSSION

Landslides are among the most hazardous natural disasters. Government and research institutions worldwide have attempted for years to assess the landslide hazard and risk and to show its spatial distribution. In this study, a verification of probabilistic approach to estimating the susceptible area of landslides using 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, Janghung and Youngin, Korea.

For the landslide susceptibility analysis, multiple logistic regression model, was applied and verified for the study area of Youngin, Korea, using the spatial data-sets. Using the 13 factors, likelihood relation model was applied to analyze the landslide hazard. Then, the results were verified by calculating the correlation observed between landslide occurrence location and the predicted occurrences. Generally, the verification results showed satisfactory agreement between the susceptibility map and the existing data on landslide location.

In comparison between success rate and prediction rate, success rate showed the better accuracy than prediction rate for all cases. In the Janghung case for success rate show the best accuracy among the all cases in success rate. Among the all cases, Janghung rate for Boun showed the best accuracy.

In this study, only the susceptibility analysis was performed, because the small area studied did not allow us to determine the distribution of rainfall. However, if data on factors causing the landslides, such as rainfall, earthquake shaking, or slope cutting, exist, then the possibility analysis could also be done. In particular, if the data could be combined with a hydrological model, a more accurate analysis could be done. If the factors relevant to vulnerability of buildings and other property were available, risk analysis could also be done. 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.

7. REFERENCES

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

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

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