Landslide Susceptibility Mapping using Probability and Statistics Models in Baguio City, Philippines

S. Lee [a,*], Digna G. Evangelista[b]

[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
[b]Mines and Geosciences Bureau, Department of Environment and Natural Resources, North Avenue, Diliman, Quezon City, Philippines

Abstract - For landslide susceptibility mapping, this study applied, verified and compared a probability model, a frequency ratio and statistical model, a logistic regression to Baguio city, Philippines, using a Geographic Information System (GIS). Landslide locations were identified in the study area from interpretation of aerial photographs and field surveys; and a spatial database was constructed from topographic maps, geology and land cover. The factors that influence landslide occurrence, such as slope, aspect, curvature, distance from drainage and terrain mapping unit were calculated from the topographic database. Lithology and distance from fault were extracted and calculated from the geology database. Land cover was classified from Landsat TM satellite imagery. The relationship between the factors and the landslides were calculated using frequency ratio and logistic regression models. The relationships, frequency ratio and logistic regression coefficient, were overlaid to determine each factor's rating for landslide susceptibility mapping. Then the landslide susceptibility map was compared with known landslide locations and verified. The logistic regression model had higher prediction accuracy (80.01%) than the frequency ratio model (78.42%).

Keywords: Frequency ratio; Logistic regression; GIS; Philippines

1. INTRODUCTION

Landslides cause extensive damage to property, and occasionally result in the loss of life. Specifically, recent landslides occurred in Philippines. It is therefore necessary to assess and manage areas that are susceptible to landslides in order to mitigate any damage associated with them. Among the many causes, landslides triggered by heavy rainfall are the most common throughout Philippines. The resultant need to predict landslide occurrences has led to the development of numerous stochastic and process-based models, with increasing emphasis on the use of a GIS. To remedy this, it is necessary to assess scientifically the area susceptible to landslide. This assessment may be carried out by applying the frequency ratio and logistic regression models, with verification of the results. Therefore, the objective of this study was to apply and verify the models for landslide hazard zonation in the Baguio city of Philippines using GIS. The study area is bounded by 16º23'00"–16º 29'00" latitude and 120º34'00"–120º37'00" longitude.

Landslides may occur as a consequence of a number of determining and triggering factors. In order to assess susceptibility from landslide it is therefore necessary to identify and analyze the factors leading to landslide. The following parameters were used: slope, aspect, curvature, proximity to drainage, lihology, proximity to major structures, land cover, geomorphologic/Terrain Units. The July 16, 1990 earthquake-induced landslide inventory from the report of Arboleda and Regalado (1990) was used as bases for landslide susceptibility mapping.

Using GIS as the basic analysis tool for landslide hazard mapping can be effective for spatial and data management and manipulation, together with some reasonable models for the analysis. In this regard, there have been many studies of landslide hazard mapping using GIS and many of these studies have applied probabilistic methods. One of the statistical methods available, the logistic regression method, has also been applied to landslide hazard mapping as has the geotechnical method and the safety factor method. As a new approach to landslide hazard evaluation using GIS and data mining such as fuzzy logic, and artificial neural network methods have been applied. The difference in this study is the application and comparison of GIS-based methods to landslide susceptibility mapping in Philippine situation.

2. THEORY: FREQUENCY RATIO AND LOGISTIC REGRESSION

The frequency ratio is the ratio of occurrence probability to non-occurrence probability, for specific attributes. In the case of landslides, if we set the landslide occurrence event to B and the specific factor's attribute to D, the frequency ratio for D is a ratio of conditional probability. So if the ratio is greater than 1, the greater the relationship between a landslide and the specific factor's attribute; and if the ratio is less than 1, the lower the relationship between a landslide and the specific factor's attribute.

Logistic regression, which is a multivariate analysis model, 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 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:

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

[*] Corresponding author

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 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).

3. DATA AND METHODOLOGY

Data preparation involved the digitization or creation of GIS database, which include the topographical, geomorphologic, geological and land cover data. A digitized map of landslide location which detected from satellite imagery and field surveys was produced, and these digital data were input to the GIS. A vector-to-raster conversion was undertaken to provide raster data of landslide areas with 10 m by 10 m pixels. Factor maps related to landslide occurrence were constructed in a vector-type spatial database. These included topographic maps and geological maps. A land cover map was extracted from Landsat TM satellite imagery with 30 m resolution. The factors such as slope, aspect, curvature, proximity to drainage, lithology, proximity to major structures, land cover and geomorphologic/terrain units were used. The study area was divided into a grid with 10 m × 10 m cells, occupying 560 rows and 541 columns: totaling 295,637 grid-cells and landslides fell in 61 of these.

Contour (5-meter interval) 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 gradient, slope aspect and curvature were calculated. The slope gradient of a surface refers to the maximum rate of change in z values across a region of the surface and the slope aspect of a surface is the compass direction maximum rate of change in z in the downward direction. The curvature represents the morphology of the topography. A positive curvature indicates that the surface is upwardly convex at that cell, and a negative curvature indicates that the surface is upwardly concave at that cell. A value of zero indicates that the surface is flat. The drainage buffer was calculated in 100 m intervals.

In the geology, four (4) formations cover the study area these are: (1) *Zigzag Formation* –consist of conglomerates, sandstones, and some limestone lenses; (2) *Kennon Formation* – consists principally of massive biohermal limestone, calcarenites and calcirudites. The basal portion consists of wackes, conglomeratic calcarenite, with clasts of volcanic rocks, diorite pebbles and cobbles (David, 1994); (3) *Klondyke Formation* – lithologies are clastic sedimentary rocks consisting mainly of polymictic conglomerates with interbedded sandstones, siltstones, shales and in places intercalated with flow breccias and pyroclastic rocks. It rests unconformably over the Kennon Limestone and underlies wide areas of the western sides of Baguio City Quadrangle at higher elevation. (4) *Baguio Formation* consists of tuff, volcanic conglomerate, volcanic breccia, glassy andesite, and porphyritic andesite, with minor sandstones. The type of geology of an area plays an important factor in the development of landslide. Correlating the landslide inventory map with the geological map of the area resulted to the tabulated values below: Land cover data were classified from Landsat TM satellite imagery. For calculating proximity to major structures, a distance buffer on both sides of a major structure was generated to see the occurrence of landslides with respect to fault lines. Visually, one could say that most of the landslides that occurred after the 1990 earthquake were within the 250 and 500 meters distance buffer.

Landslides occurring at different terrain units such as floodplain, deep valley, plateau, karst, wide valley, limestone hills, basin and shallow valley. Floodplain is cultivated flat terrain, pronounced meanders and deep valley is narrow and deep valleys; wide drainage divider. Plateau is well-drained round ridges and peaks dominated by ridges and narrow plateaus rather than valleys. Karst is rugged terrain with poorly defined drainage lines characterized by sinkholes. Wide valley is wide valleys with narrow floodplain dominated by active erosion processes. Limestone hills is poorly drained, rounded contours and absent of sinkholes. Basin is shallow depression with rounded contours and poor development of drainage line. Shallow valley is characterized by narrow and shallow valleys with steep slopes; generally rugged terrain.

Using the detected landslide locations and the constructed spatial database, landslide analysis models were applied and verified. To represent the distinction quantitatively, frequency ratio and logistic regression models were used. For this analysis, the calculated and extracted factors were mapped to a 10 m resolution grid. The raster data were converted for the statistical program used. Then, using the frequency ratio and logistic regression models, the spatial relationships between the landslide location and each landslide related factor, such as topography, soil, forest and land cover, were analyzed in the statistical program, and a formula of landslide occurrence possibility was extracted using the relationships. The formula was used for calculating the landslide susceptibility index, which was mapped to each grid cell. Finally, the susceptibility map was verified using known landslide locations and success rates were calculated for quantitative verification. In this study, GIS software, ArcView 3.3 and ARC/INFO 8.1 NT version, and statistical software, SPSS 12.0, were used as the basic analysis tools for spatial management and data manipulation.

4. LANDSLIDE PROPERTIES BY ANALYSIS OF THE RELATIONSHIP BETWEEN LANDSLIDES AND FACTORS

The relationship between areas where a landslide has occurred and landslide-related factors can be distinguished from the relationship between areas without past landslides and landslide-related factors. To represent this distinction quantitatively, the frequency ratio was used. The factors chosen, such as the slope, aspect, curvature, distance from drainage, lithology, distance from lineament, landuse, and vegetation index were evaluated using the frequency ratio method to determine the level of correlation between the location of the landslides in the study area and these factors. Probabilistic approaches are based on the observed relationships between each factor and the distribution of landslides.

Topographic factors, such as slope, aspect, curvature, and distance from drainage were used. In the case of the relationship between landslide occurrence and slope, below a slope of 12°, the ratio was <1, which indicates a low probability of landslide occurrence. For slopes above 13°, the ratio was >1, which indicates a high probability of landslide occurrence. This means that the landslide probability increases according to slope angle. As the slope angle increases, then the shear stress in the soil or other unconsolidated material generally increases. Gentle slopes are expected to have a low frequency of landslides because of the generally lower shear stresses associated with low gradients. Steep natural slopes resulting from outcropping bedrock, however, may not be susceptible to shallow landslides. In the case of the relationship between landslide occurrence and aspect, landslides were most abundant on south-facing and northeast-facing slopes. The frequency of landslides was lowest on northwest-facing, north-facing, and west-facing slopes, except in flat areas and highest on southeast-facing, east-facing, and south-facing slopes. In the case of the relationship between landslide occurrence and curvature, the more positive or negative a value is, then the higher is the probability of a landslide occurrence. Flat areas had a low curvature value of 0.71. The curvature values represent the morphology of the topography. A positive curvature indicates that the surface was upwardly convex at that grid. A negative curvature indicates that the surface was upwardly concave at that grid. A value of zero indicates that the surface was flat. The reason for this is that following heavy rainfall, a convex or concave slope contains more water and retains this water for a longer period. Analysis was carried out to assess the influence of drainage lines on landslide occurrence. For this purpose, the proximity to a drainage line was identified by buffering. In the case of the relationship between landslide occurrence and distance from drainage, as the distance from a drainage line increases, the landslide frequency generally decreases. At a distance of < 259 m, the ratio was > 1, indicating a high probability of landslide occurrence, and at distances > 260 m, the ratio was < 1, indicating a low probability. This can be attributed to the fact that terrain modification caused by gully erosion and undercutting may influence the initiation of landslides.

In the case of the relationship between landslide occurrence and lithology, the frequency ratio was higher in Kennon formation where consist principally of massive biohermal limestone, calcarenites and calcirudites, at 1.55, and was lower in klondyke formation where consisting mainly of polymictic conglomerates with interbedded sandstones, siltstones, shales and in places intercalated with flow breccias and pyroclastic rocks, at 0.00. In the case of the relationship between landslide occurrence and distance from a lineament, the closer the distance was to a lineament, then the greater was the landslide-occurrence probability. For distances to a lineament of < 478 m, the ratio was > 1, indicating a high probability of landslide occurrence, and for distances to a lineament of > 479 m, the ratio was < 1, indicating a low probability landslide occurrence. This means that the landslide probability decreases with increasing distance from a lineament. As the distance from a lineament decreases, the fracture of the rock increases, and in addition, the degree of weathering increases.

In the case of the relationship between landslide occurrence and land cover, landslide-occurrence values were higher in agricultural lands, bushes and grass areas, and lower in broadleaf area. The reason for this is that landslides occurred mainly in none or a few tree areas. In the case of the relationship between landslide occurrence and terrain mapping unit, landslide-occurrence values were higher in karst and wide valleys areas, and lower in flood plain, deep valleys, plateau, basin and shallow valley areas. The reason for this is that the geomorphology of terrain have effected to landslide occurrences.

5. APPLICATION OF THE FREQUENCY RATIO MODEL AND LOGISTIC REGRESSION MODEL

For calculation of the frequency ratio, the area ratio for landslide occurrence and non-occurrence was calculated for the class or type of each factor, and an area ratio for the class or type of each factor to total area was calculated. So, frequency ratios for the class or type of each factor were calculated by dividing the landslide occurrence ratio by the area ratio. The frequency ratios of each factor's type or class were summed to calculate the landslide susceptibility index (LSI), as shown in Equation (3)

$$\mathrm{LSI} = \Sigma \mathrm{LR} \tag{3}$$

where LR = Frequency ratio of each factor's type or class

If the LSI value is high, it means a higher susceptibility to landslide; a lower value means a lower susceptibility to landslides. The index was classified into equal areas and grouped into five classes for visual and easy interpretation. The minimum value is 3.35 and maximum value is 12.83, the mean value is 8.17 and the standard deviation value is 1.51.

Using the logistic regression model, the spatial relationship between landslide-occurrence and factors influencing landslides was assessed. The spatial databases of each factor were converted to ASCII format files for use in the statistical package, and the correlations between landslide and each factor were calculated. In addition, logistic regression formulae were created as shown in equations (4) and (1). Finally, the probability that predicts the possibility of landslide-occurrence was calculated using the spatial database, data from the coefficients and equations (4) and (1):

$$z = (0.013 \times \mathrm{SLOPE}) + (\ 0.000 \times \mathrm{CURVA}) + (-0.001 \times \mathrm{DRAIN}) + (-0.002 \times \mathrm{FAULT_b}) + \mathrm{ASPECT_b} + \mathrm{GEOL_b} + \mathrm{LANDCOVER_b} + \mathrm{TERRAIN_b} - 2.578 \tag{4}$$

where SLOPE is slope value; CURVA is curvature value; DRAIN is distance from drainge value; FAULT is distance from fault value; $\mathrm{ASPECT_b}$, $\mathrm{GEOL_b}$, $\mathrm{LANDCOVER_b}$, $\mathrm{TERRAIN_b}$ are logistic regression coefficient values listed; z is a parameter

The possibility was classified by equal areas and grouped into five classes for visual interpretation. The minimum value is 0.00 and maximum value is 0.374. The mean value is 0.040 and the standard deviation value is 0.048.

6. VERIFICATION OF THE LANDSLIDE SUSCEPTIBILITY MAPS

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

The landslide susceptibility analysis result was validated using known landslide locations. Validation was performed by comparing the known landslide location data with the landslide susceptibility map. Each factor used and frequency ratio was compared. The rate curves were created and its areas of the under curve were calculated for all cases. The rate explains how well the model and factor predict the landslide. So, the area under curve in can assess the prediction accuracy qualitatively. 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. Then the ordered cell values were divided into 100 classes, with accumulated 1% intervals. For example, in the case of frequency ratio model, 90 to 100% (10%) class of the study area where the landslide susceptibility index had a higher rank could explain 36% of all the landslides. In addition, the 80 to 100% (20%) class of the study area where the landslide susceptibility index had a higher rank could explain 55% of the landslides. In the case of logistic regression model, 90 to 100% (10%) class of the study area where the landslide susceptibility index had a higher rank could explain 39% of all the landslides. In addition, the 80 to 100% (20%) class of the study area where the landslide susceptibility index had a higher rank could explain 57% of the landslides.

To compare the result quantitative, the areas under the curve were recalculated as the total area is 1 which means perfect prediction accuracy. So, the area under a curve can be used to assess the prediction accuracy qualitatively. In the case of all factor and logistic regression model used, the area ratio was 0.8001 and we could say the prediction accuracy is 80.01%. In the case of all factor and frequency ratio model used, the area ratio was 0.7842 and we could say the prediction accuracy is 78.42%. Overall the case of all factor and logistic regression model used showed a higher accuracy than cases of each factor and logistic regression used and all factor and frequency ratio model used.

7. CONCLUSION AND DISCUSSION

Landslides are among the most hazardous of natural disasters. Government and research institutions worldwide have attempted for years to assess landslide hazards and risks and to show their spatial distribution. In this study, a statistical approach for identifying the susceptible area of landslides using GIS shows considerable promise.

The result of validation of logistic regression and frequency ratio model, the logistic regression model showed the better prediction accuracy more than 1.59%.

The frequency ratio model is simple, the process of input, calculation and output can be readily understood. The large amount of data can be processed in the GIS environment quickly and easily. The logistic regression model requires conversion of the data to ASCII or other formats for use in the statistical package, and later reconversion to incorporate it into the GIS database. Moreover, it is hard to process the large amount of data in the statistical package. However, correlation of landslide and other factors can be analyzed qualitatively. The logistic regression model showed better accuracy than frequency ratio model in this study. In the case of a similar statistical model (determinant analysis), the factors must have a normal distribution, and in the case of multi-regression analysis, the factors must be numerical. However, for logistical regression, the dependent variable must be input as 0 or 1, therefore the model applies well to landslide occurrence analysis.

Statistical packages can allow analysis of landslide susceptibility, but they are inconvenient for management of spatial data. A GIS has few if any functions for statistical analyses, but has many functions for database construction, display, printing, management and spatial analysis. Therefore it is necessary to integrate the GIS and the statistics to reduce the restrictions of using the two 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.