GIS-aided Statistical Landslide Susceptibility Modeling And Mapping Of Antipolo Rizal (Philippines)

A J Dumlao, J A Victor
Institute of Civil Engineering
University of the Philippines Diliman, Quezon City, Philippines
E-mail: jaimeangelovictor@yahoo.com

Abstract. Slope instability associated with heavy rainfall or earthquake is a familiar geotechnical problem in the Philippines. The main objective of this study is to perform a detailed landslide susceptibility assessment of Antipolo City. The statistical method of assessment used was logistic regression. Landslide inventory was done through interpretation of aerial photographs and satellite images with corresponding field verification. In this study, morphologic and non-morphologic factors contributing to landslide occurrence and their corresponding spatial relationships were considered. The analysis of landslide susceptibility was implemented in a Geographic Information System (GIS). The 17320 randomly selected datasets were divided into training and test data sets. K- cross fold validation is done with k= 5. The subsamples are then fitted five times with k-1 training data set and the remaining fold as the validation data set. The AUROC of each model is validated using each corresponding data set. The AUROC of the five models are; 0.978, 0.977, 0.977, 0.974, and 0.979 respectively, implying that the models are effective in correctly predicting the occurrence and non-occurrence of landslide activity. Field verification was also done. The landslide susceptibility map was then generated from the model. It is classified into four categories; low, moderate, high and very high susceptibility. The study also shows that almost 40% of Antipolo City has been assessed to be potentially dangerous areas in terms of landslide occurrence.

1. Background

Landslide is one of the natural hazards that is unexpected and with high magnitude that threatens humans and properties. It is a downward movement of rocks and soil, which includes rock, falls, and deep failure of slopes, shallow debris flows and avalanches. The movement is triggered by gravity due to other hazards like earthquake and high amount of rainfall. Other factors such as geology, morphology, elevation and human activities affect the slope stability of an area.

The study will be conducted in Antipolo City due to its increase in population and establishments on its steep slope covering its 306 km² area. As of 2009, government records show that 545,157 are residing in the city but an unofficial count estimates the population to reach 800,000 due to informal settlers. The number of business establishments also continues to rise with 6,903. The subdivisions in the city are also increasing with a total of 405 and 10 are still under development.

Although Antipolo City has experienced several landslide events in the past years, preparedness and mitigation in the area are lacking. It is because there are not enough data, and hazard experts available to guide the concerned authorities.
2. Research Objectives

The object of this study is to investigate and apply multivariate statistical analysis (specifically logistic regression) within a GIS environment to determine which independent intrinsic variables significantly influence the occurrence of contemporary landslide activity within the study area. Results will be utilized to create a susceptibility map and analyzed to assess the viability of utilizing GIS modeling as part of a broader decision support system pertaining to slope instability monitoring, mitigation and indeed planning for natural hazards.

3. Logistic Regression

Landslide susceptibility analysis may be approached with the use logistic regression as method of mathematical modeling. This method relies on the estimation of future landslides based on the current presence and absence of landslides associated with the values of a set of predictor variables. (Lee, 2005).

In logistic regression, the normal distribution of the dependent variable is not required. When compare with multiple linear regression, one of the main advantages of logistic regression is that it does not assume that there is a linear relationship between the dependent and independent variables. It assumes a linear relationship between the outcome and the logit of the independent variables. Also, there are no assumptions required with regards to the homogeneity of the variance and normally distributed error terms. (Nandi et. al, 2008). The dependent variable can also have 2 values, the event occurring and the event not occurring. (Dai and Lee, 2002)

In landslide susceptibility assessment, the outcome demands to be predicted is the occurrence/ non-occurrence of landslide in a given area. The independent variables are the physical controlling factors that affect the instability of an area and the dependent variable is the presence or absence of landslides. Because of the binary response, two alternative groups are established; mapping units free of landslides, and mapping units having landslides.

The relationship of the occurrence of landslide and the preparatory factors affecting it is expressed as (Hosmer and Lemeshow, 2000):

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

\[
Z_n = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n
\]

Where \( P_r \) is the estimated probability of landslide event with values varying from 0 to 1, \( b_0 \) is the intercept of the model, \( n \) is the number of independent variables or predictors, \( b_{(i-n)} \) is the \( i^{th} \) coefficient of the model, \( x_{(i-n)} \) is the \( x^{th} \) independent variable or predictor.

The data analysis was utilized using Statistical Package for Social Sciences (SPSS). The raster map from the ArcGIS was used to obtain the values that will be needed for the data analysis. Each raster map was divided into cells, which comprised the whole study area. The DEM was used to define the location of the cells using the sample tool in ArcGIS 10.0. A total of 3527765 pixels were extracted to each raster map. The values obtained in ArcGIS 10.0 were saved as a dbf file for compatibility in the SPSS. A total of 41 independent variables were used including six continuous and 35 categorical variables. The independent set consists of the Aspect \([\text{Flat, North}(0-22.5), \text{Northeast, East, Southeast, South, Southwest, West, Northwest, North}(337.5-360)]\), Slope angle, Elevation, Distance to River, Distance to Road, TWI, SPI, Planform Curvature \([\text{Convex, Flat, Concave}]\), Profile Curvature \([\text{Convex,} \]


Flat, Concave], Soil Type [Mountain soil, Antipolo soil, Antipolo clay, Binangonan clay, Marikina silt loam, Antipolo clay loam, Binangonan lowland clay, Marikina clay loam], and Geology [Cretaceous, Oligocene-miocene, Paleocene-eocene, Oligocene-miocene (sedimentary and metamorphic), Recent, Oligocene, Neogene, Upper miocene-pliocene, Pliocene-quaternary, Pliocene-pleistocene, Cretaceous-paleocene]

The dependent variable, which is the observed landslide occurrence (1580 landslide events), is coded as 1 and 0 for presence or absence of landslide respectively. The statistical analysis was performed using randomly selected training data sets with equal number of presence and absence of landslide value. A total of 17320 cases were chosen as subsamples for determining the relationship of the dependent to the independent variables.

Before performing the regression analysis of the datasets, dummy variables for the categorical variables were created. Since some of the variables will be insignificant in the model it is much easier to remove it by means of the generation of dummy variables. Dummy variables were also used as binary representation of categorical data (e.g. soil type, lithology) based on the presence (1) and absence (0).

The logistic regression analysis was performed using the binary logistic tool under the analyze tab in SPSS. The output in the SPSS produced different summary of the model. It includes Homer and Lemeshow Test, -2 Log likelihood, Cox and Snell R Square, and Nagelkerke R Square. The analysis was then again performed as the insignificant variables were removed. Insignificant variables were determined as variables exceeding the threshold value of 0.05 (level of significance). Variables exceeding the tolerance value accept the null hypothesis thus affecting the model insignificantly. The coefficients of the variables in the model were identified after all the variables portray significant influence in the occurrence of the event.

4. Validation

To have scientific significance, LSI prediction models should have validation measures (Chung and Fabbri, 2003). The models generated were validated using k-fold cross-validation. The randomly selected dataset were randomly divided into five folds in the SPSS. The performance of each model was measured using the corresponding test data set by creating a confusion matrix. The coefficients of the model were used to manually compute the predicted probability of each case of the test data set. A cut-off value of 0.5 was used for the predicted probability. The predicted probability of each case is then compared to the observed outcome of the landslide occurrence. The correctly predicted and wrongly predicted cases of occurrence and non-occurrence of landslide were tabulated in a table to compute for the overall correct percentage of each model.

The AUROC are also performed using the a separate independent validation data set of each model (approximately 20% of the total input data).

5. Thematic Data Layers

There are several factors that can cause a landslide event. The selection and preparation of these factors are one of the crucial components for modeling a landslide susceptibility map. The dataset that will be selected must be available all over the investigated area and categorical data must be defined at a comparable spatial accuracy. The factors selected were namely: elevation, slope, curvature, aspect, topographic wetness index (TWI), stream power index (SPI), lithology, distance from roads and rivers. Most of the data were obtained from Department of Science and Technology (DOST) specifically from Project Nationwide
Operational Assessment of Hazards (NOAH) and Disaster Risk and Exposure Assessment for Mitigation (DREAM) Program while others were available from online resource databases. The DEM used was obtained from Synthetic Aperture Radar (SAR), which were post-processed by the UP Training Center for Applied Geodesy and Photogrammetry (UP-TCAGP). It has a resolution of 10 m, tile size of 10 x 10 km, and an absolute vertical map accuracy of LE90

5.1 Morphometric Preparatory Factors
5.2 Non-morphometric Preparatory Factors

Figure 1. Thematic factor maps of morphometric variables related to landslide occurrence: Slope, Aspect, Elevation, Flow direction, Planform curvature, Profile Curvature, SPI, and TWI raster maps of Antipolo City.

Figure 2. Non-morphometric factors considered in this study: Distance to river, Distance to road, Soil and Geology raster map.

6. Tests for Multicollinearity
The results of the multicollinearity tests done on the set of independent variables are summarized in table 1. An evidence of presence of multicollinearity is tolerance value of less than 0.2, and consequently, a variance inflation factor of greater than 5 (Menard, 1995). Logistic regression can proceed with the chosen set because multicollinearity among the
variable is not evident in the set presented by tolerance values greater than 0.2 and consequently, variance inflation factor values less than 5.

### Table 1: Results of Multicollinearity Tests.

| Model Coefficient | Tolerance | VIF  |
|-------------------|-----------|------|
| (Constant)        |           |      |
| Distance to river | .807      | 1.239|
| Aspect            | .983      | 1.018|
| Distance to road  | .555      | 1.801|
| Elevation         | .461      | 2.169|
| Geology           | .795      | 1.258|
| Plan curvature    | .800      | 1.250|
| Profile curvature | .939      | 1.065|
| Slope             | .873      | 1.145|
| SPI               | .752      | 1.329|
| TWI               | .552      | 1.811|
| Soil              | .483      | 2.071|

Several runs of logistic regression were performed until all significant values of the variables were under the threshold value. Despite the good fit statistics show that each model accurately predict the outcome of the dependent variable, it is still not clear which variables have an active influence to the landslide occurrence. So the usual threshold of significance value of 0.05 is set to identify those variables that will be effective in susceptibility assessment. It shows in the table that removal of those variables that deemed to be insignificant mostly gives positive effect on the significance value of the remaining variables.

The coefficients derived from the logistic regression indicate the relationship of each independent variable to the probability of occurrence of the event. The exponential value of the coefficient represent the log-odds. Thus positive coefficients tend to increase the probability of a landslide activity and decrease it otherwise. The constant or the intercept of the models represent the odds of an outcome happening with all the variables removed.

From all the models, the geology 3 yields the highest value followed by aspect 9. The aspect 9, are slopes facing northwest, and geol 3, Paleocene-eocene rock, implies that these variables have the most significant value in explaining the occurrence of landslide. Other factors such as the slope, TWI, SPI, and curvature always yield lower values than the other aspect and geology factors hence does not explained well the occurrence of landslides.

### Table 2. Summary of the coefficients of each training model.

| Parameters/ Coefficients | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|--------------------------|---------|---------|---------|---------|---------|
| Slope                    | 0.37    | 0.40    | 0.32    | 0.33    | 0.34    |
| SPI                      | 2.33    | 2.10    | 2.72    | 2.67    | 2.54    |
| TWI                      | -2.31   | -2.06   | -2.69   | -2.63   | -2.52   |
| Flat                     | -2.91   | -2.61   | -4.64   | -2.86   | -2.69   |
The summary shown on table 2 presents that not all variables were effective in determining the outcome of the landslide occurrence. It can be observed that for all the models; the slope, SPI, Pliocene-quaternary, and Cretaceous-paleocene showed positive values of coefficients. Thus, regions with high slope and SPI are more probable to landslide as well as areas lying in bedrocks composed of Pliocene-quaternary, and Cretaceous-paleocene rocks. For models 1 and 5 convex planform curvature also have positive coefficient hence also increase the instability of the area.

| Model | Hosmer and Lemeshow test | Cox & Snell R Square | Nagelkerke R Square | Accuracy |
|-------|-------------------------|---------------------|---------------------|----------|
| Model 1 | Chi-square: 130.926 | 0.635 | 0.847 | 90.2 % |
| Model 2 | Chi-square: 105.452 | 0.638 | 0.851 | 90.7 % |
| Model 3 | Chi-square: 91.033 | 0.638 | 0.850 | 90.8 % |
| Model 4 | Chi-square: 84.556 | 0.640 | 0.853 | 91.0 % |
| Model 5 | Chi-square: 100.437 | 0.638 | 0.850 | 90.7 % |

The summary shown on table 2 presents that not all variables were effective in determining the outcome of the landslide occurrence. It can be observed that for all the models; the slope, SPI, Pliocene-quaternary, and Cretaceous-paleocene showed positive values of coefficients. Thus, regions with high slope and SPI are more probable to landslide as well as areas lying in bedrocks composed of Pliocene-quaternary, and Cretaceous-paleocene rocks. For models 1 and 5 convex planform curvature also have positive coefficient hence also increase the instability of the area.

| Table 3. Statistics and accuracy of the each model. |
|-------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Hosmer and Lemeshow test | Chi - square: 130.926 | 0.635 | 0.847 | 90.2 % |
| df | 8 | 8 | 8 | 8 | 8 |
| Sig | 0 | 0 | 0 | 0 | 0 |
| -2 Log likelihood | 5224.28 | 5143.24 | 5134.65 | 5029.28 | 5151.01 |
| Cox & Snell R Square | 0.635 | 0.638 | 0.638 | 0.640 | 0.638 |
| Nagelkerke R Square | 0.847 | 0.851 | 0.850 | 0.853 | 0.850 |
| Accuracy | 90.2 % | 90.7 % | 90.8 % | 91.0 % | 90.7 % |
The summary of the statistics in the table 3 shows goodness of fit of the model to the data and its accuracy. The Hosmer and Lemeshow test yields very small significance values implying that the models were statistically significant. The Cox & Snell and Nagelkerke R squares show a reliable value which indicates the effectiveness of the model. The accuracy confirms the value of the R squares yielding more than 90% of the training data sets were correctly predicted.

7. Results Validation
The AUROC of each model is validated using each corresponding data set. The AUROC of the five models are; 0.978, 0.977, 0.977, 0.974, and 0.979 respectively. The value of the AUROC implies that the models are effective in correctly predicting the occurrence and non-occurrence of landslide activity. The data used in evaluating the AUROC is the independent validation set that is not used in the susceptibility model calibration. Confusion matrices were also done to determine the predictive capability of the models, as show in tables 4 to 8.

**Table 4.** Confusion matrix of validation data set 1 evaluated using model 1.

|           | landslide | percentage correct |
|-----------|-----------|---------------------|
|           | absence   | presence            |
| landslide | 1568      | 165                 | 90.48               |
| presence  | 153       | 1581                | 91.18               |
| Overall Percentage | 90.83     |                     |

**Table 5.** Confusion matrix of validation data set 2 evaluated using model 2.

|           | landslide | percentage correct |
|-----------|-----------|---------------------|
|           | absence   | presence            |
| landslide | 1635      | 102                 | 94.13               |
| presence  | 228       | 1455                | 86.45               |
| Overall Percentage | 90.29     |                     |

**Table 6.** Confusion matrix of validation data set 3 evaluated using model 3.

|           | landslide | percentage correct |
|-----------|-----------|---------------------|
|           | absence   | presence            |
| landslide | 1547      | 185                 | 89.32               |
| presence  | 168       | 1570                | 90.33               |
| Overall Percentage | 89.83     |                     |
Table 7. Confusion matrix of validation data set 4 evaluated using model 4.

|          | landslide absence | landslide presence | percentage correct |
|----------|-------------------|--------------------|---------------------|
| landslide absence | 1533              | 210                | 87.95               |
| landslide presence  | 148               | 1629               | 91.67               |
| Overall Percentage |                   |                    | **89.81**           |

Table 8. Confusion matrix of validation data set 5 evaluated using model 5.

|          | landslide absence | landslide presence | percentage correct |
|----------|-------------------|--------------------|---------------------|
| landslide absence | 1537              | 178                | 89.62               |
| landslide presence  | 147               | 1581               | 91.49               |
| Overall Percentage |                   |                    | **90.56**           |

8. Final Landslide Susceptibility Map

The landslide susceptibility map generated from model 1 is shown in figure 3. It is classified into four categories; low, moderate, high and very high susceptibility. Most of the areas with low susceptibility to landslide are lying in the low altitude areas with gentle slopes. The susceptibility increases as it traverses eastward because of the mountain ranges that covers the terrain. It can also be observed that small portions with low susceptibility are river networks.

Figure 3: Landslide susceptibility map of Antipolo City.
9. Conclusions
Based on the results of the logistic regression analysis of the landslide susceptibility of Antipolo City, the objectives of the study had been attained. The study shows the ability of the model to delineate areas of instability.

From the study, it shows that area with more than 15° of steepness and facing the southwest direction have high probability of failure. The lithology of the area also contributes to the occurrence of landslide. It was found out that area lying in bedrocks composed of Cretaceous rocks and topsoil cover of Antipolo-origin soil was more prone to landslide compared to other bedrocks and top soil covers.

The study also shows that almost 40% of Antipolo City has been assessed to be potentially dangerous areas in terms of landslide occurrence. The landslide susceptibility of each barangays was also examined. Brgy. San Juan yields the highest percentage of highly susceptible land area with more than 65%. Other barangays such as Bagong Nayon, Calawis, and Inarawan should also be aware of instability with more than 40% of their areas were situated in high landslide susceptibility.

The model generated also shows high reliability based on the several statistical tests applied using test data set. It can be concluded that the model can accurately predict landslide occurrence given the physical factors that are present in the area. Thus it can be used for further studies such as landslide hazard and risk assessments. The generated map can also be used by concerned authorities to properly mitigate areas that were found to be unstable. Residences that were lying in highly susceptible areas should also be force evacuated to avoid further property damages and loss. As well as the establishments developed should also be removed or converted to other kind of recreational area. Based on the map, an extensive zoning should also be conducted since the population, business establishments, and subdivisions are increasing rapidly in the city.

10. References

1. Lee S (2005) Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data journals. Int J Remote Sens 26(7):1477–1491.

2. Nandi A, Shakoor A (2008) Application of logistic regression model for slope instability prediction in Cuyahoga River Watershed, Ohio, USA. Georisk 1:12

3. F.C. Dai, and S. Lee, "Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong", Geomorphology, Vol. 42, 2002, PP. 213-227.

4. Hosmer D, Lemeshow S (2000). Applied Logistic Regression, Second Edition, John Wiley & Sons, Inc., Print ISBN: 9780471356325

5. C. J. F. Chung and A. G. Fabbri, Validation of spatial prediction models for landslide hazard mapping, Natural Hazards, vol. 30, pp. 451-472, 2003.

6. Menard SW (1995) Applied logistic regression analysis. SAGE, Thousand Oaks