Asian Journal of Environment and Disaster Management
Vol. 6, No. 2 (2014) 117–129

doi:10.3850/S179392401400323X

Landslide Susceptibility Analysis Using GIS and Logistic Regression Model A Case Study In Malang, Indonesia

Shahroz Hina[1,a], Akiyuki Kawasaki[1,2] and Muhammad Qasim[1,b]
[1]*Asian Institute of Technology, P.O. Box 4, KlongLuang, Pathumthani 12120, Thailand*
E-mail: [a]shahrozhina@gmail.com, [b]qasim.ait@gmail.com
[2]*Department of Civil Engineering, The University of Tokyo*
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
E-mail: kawasaki@hydra.t.u-tokyo.ac.jp

Landslide susceptibility mapping is one of the most important counter measures in landslide risk reduction, as this paper will show. A method for determining landslide-prone areas by combining multivariate statistical analysis and GIS was demonstrated, with Malang, Indonesia, as the study area. Seven spatial parameters – elevation, slope, aspect, flow accumulation, land use/land cover, geology and soil – were used in the analysis. Three of these parameters were identified as being more likely to cause landslides. These particular parameters were used to produce a landslide susceptibility map, divided into five classes. Gain statistics were then applied to assess the accuracy of the model; 77% accuracy was the result. The output was overlaid with a land use/land cover dataset to investigate which areas were prone to landslides. The result showed that in the study area, forest and upland food crops are most vulnerable to landslide, followed by mixed tree crops and settlements.

Keywords: GIS, Landslide, Susceptibility, Logistic regression model.

1. Introduction

Indonesia is located at the meeting point of three tectonic plates; these plates are dynamic and constantly moving. The country is located in a tropical region where the rainy season lasts six months. These factors, coupled with the nature of the soil, the slope of the land, and how the land is used, make some parts of Indonesia prone to landslides. Statistical data reveals that approximately 11% of disaster events in Indonesia in 2008 were landslides, which caused 73 casualties and affected 1,599 people.[1]

Landslide susceptibility has been analyzed using various methods. Comparative evaluation of landslide susceptibility analysis has been performed using the information value model and the logistic regression model, based on the Geographic Information System (GIS).[2,3] The logistic regression model, which makes use of multivariate statistical analysis, can be fitted with spatially-dependent data.[3] The generalized additive model[4,5] extends logistic and linear regression models by introducing nonlinear transformations of the predictor variables, while maintaining an interpretable additive structure, that is the effects of different predictors are still added together.

A multivariate logistic regression model within a GIS environment was used to map landslide hazards in a mountainous environment, with the aim of determining which factors cause landslides.[6,7] Likelihood ratio models, logistic regression models, and artificial neural networks models were applied and verified for analysis of landslide susceptibility in Youngin, Korea.[8]Brenning(2008) introduced the generalized additive model (GAM) with terrain attributes, combined with logistic and linear regression models.[9] Pradhan and Lee (2010) developed multi modal procedures for delineating landslide hazard areas by using frequency ratio models, logistic regression models, and artificial neural network models. Using statistical analysis in a GIS environment, the probability of landslide occurrence was then calculated.[10] A landslide susceptibility map has been generated to analyze spatial distribution of slope failure in a specific area. This could be an effective landslide hazard mitigation measure.

The nature of dependent and independent variables guide the selection of the most appropriate model. Binary logistic regression was used to choose between categorical and continuous predictor variables in a regression analysis.[11] Bivariate and multivariate statistical techniques were used to analyze data collected on landslides from a particular section of the study area (referred to as the training area). The statistical techniques were then extended to the whole study area. Nandi and Shakoor (2009) concluded that sampling from a small representative area can produce realistic results, and with less time and effort than is required to analyze the entire study area.[12]Cascini (2002) and Cascini *et al.* (2005) coupled GIS with statistical analysis in two different models to determine the impact of landslide-influencing factors (with and without sub-class respectively).[13,14]

The objective of this paper is to demonstrate an integrated approach for assessing landslide susceptibility using GIS and statistical analysis. We captured the distribution of landslide occurrences through satellite remote-sensing, then applied GIS and the statistical method. Thus, we produced a susceptibility map for potential landslides in the study area and assessed land-use types that are prone to landslide.

2. Study Area

The island of Java is one of numerous regions in Indonesia that experiences landslides on a yearly basis. These landslides have a major impact on social and economic life. The study area is located between the western flank of the

Mt Semeru volcano and KarangKates dam, and covers approximately six districts of the Malang regency in East Java (112° 22′ 49″ ∼ 112° 54′ 59″ east longitude and 8° 09′ 52″ ∼ 8° 20′ 52″ south latitude) (Fig. 1). The area measures 1736 km^2 and is upland. Vegetables are produced in this horticultural region. The relief and topography of the study area vary widely, from mountainous active volcanoes, to hills of sediment material, to gently-sloping fluvial sediments. Mean annual rainfall in the study area ranges from 1,000 to 2,550 mm. The highest monthly rainfall (450–530 mm) is in January or February, while from June to September/October monthly rainfall is less than 100 mm.

While annual landslide frequency varies significantly, the average is 49 per year.[1] The number of landslides increases from June to November and decreases significantly from January to July. More landslides occur in January and November than other months: the average figures for these months are 9 and 7 respectively. This distribution is closely related to rainfall levels. The number of landslides in the area and the number of people killed by these events tend to increase year by year.[15] Human activities help to induce landslides: deforestation, reservoir construction, road construction, the expansion of settlement areas and land use changes all contribute to landslides.[16]

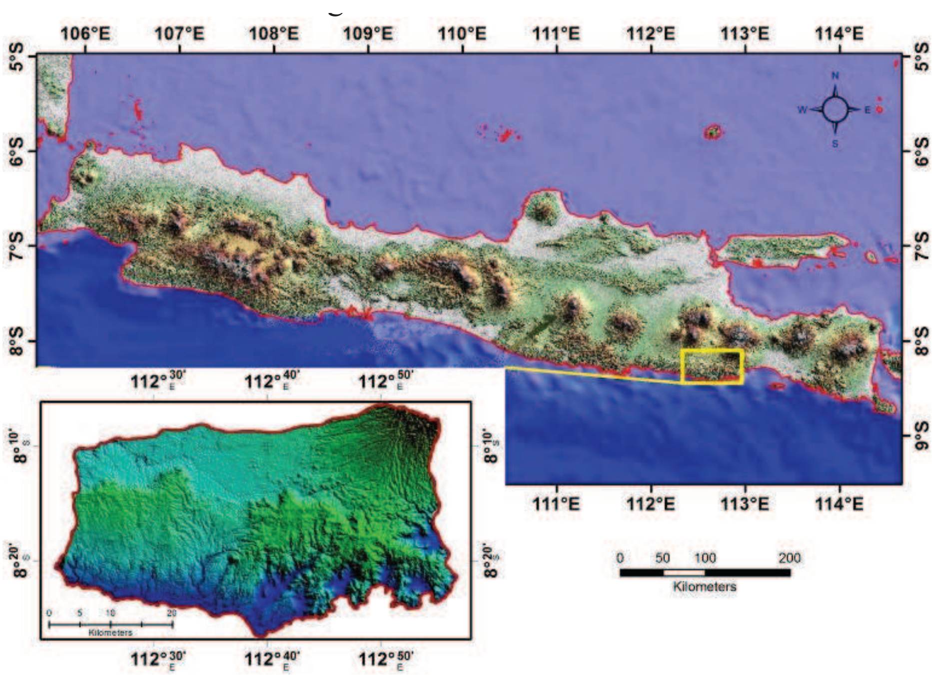

Figure 1 Location of study area.

3. Methodology

3.1. *Outline*

Geospatial technology and spatial data analysis have been employed widely for landslide hazard zonation and mapping.[17–19] Advancements in remote-sensing technology have allowed us to identify features on the surface of earth in finer scale and for the creation of the high resolution Digital Elevation Model (DEM). In this paper we demonstrate an integrated approach to landslide susceptibility analysis using the logistic regression model and spatial analysis in a GIS environment. A diagram of the analysis process is shown at Fig. 2.

3.2. *Data Processing*

Database preparation was a preliminary and crucial step in conducting our landslide susceptibility assessment. Our data set included information on landslide locations, land use/land cover, soil, geological features, flow accumulation, and DEM generation. Some data was gathered through a satellite remote-sensing process. This was coupled with data from a field survey in the form of a landslide inventory provided by the Indonesian Soil Research Institute (ISRI). Data sets used for analysis are summarized at Table 1.

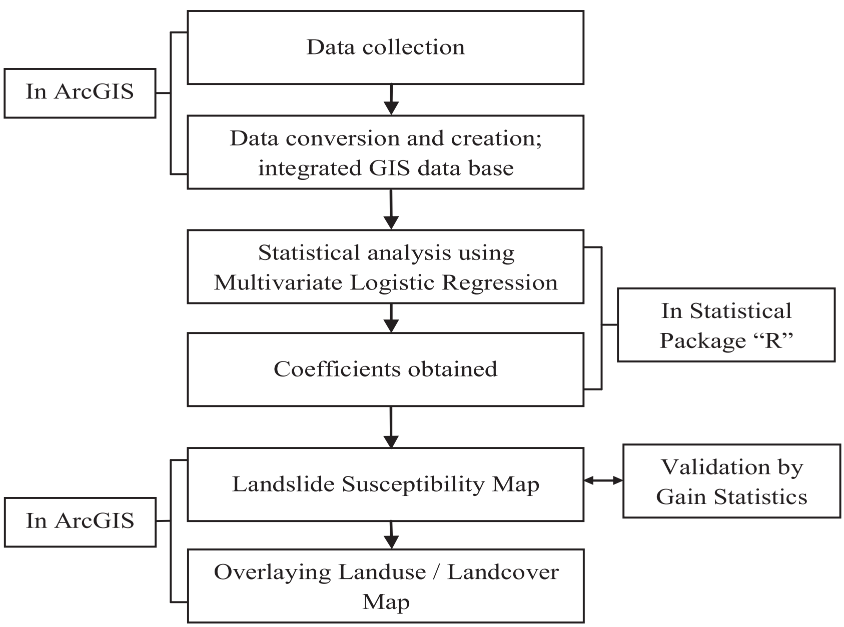

Figure 2 Analysis process

Table 1 Description of data set for analysis

Classification	Sub-classification		Data type	Source
Geological	Landslide		Point feature	ISRI
hazard	1. Silty sand	4. Sandy clay loam		
Soil map	2. Silty clay	5. Silty loam		
	3. Clay loam			ISRI
Geological	1. Ext. intermediate lava	Polygon feature		
maps	2. Ext. intermediate polymic			
	3. Ext. intermediate pyroclastic			
	4. Intr. intermediate lava			
	5. Sedimentary clastic alluvium			
	6. Sedimentary clastic breccia			ISRI
	7. Sedimentary clastic limestone			
	8. Sedimentary clastic sand			
	9. Sedimentary reef limestone			
	10. Sedimentary medium sand			
Topography	1. Elevation		GRID	Aster GDEM
map	2. Slope			DEM
	3. Aspect			DEM
Land use/	1. Rice fields	7. Golf course	Polygon feature	ISRI
land cover	2. Mixed tree crops	8. Airport		
	3. Settlement	9. Water body		
	4. Upland food crops	10. Plantation		
	5. Grass	11. Shrub		
	6. Forest	12. Coastal fish pond		
Drainage map			Line feature	DEM

Mostly the locations of landslide sites were extracted using Quick Bird satellite images. 121 landslide points were identified in the study area (Fig. 3(a)). While we endeavored to digitize the actual area of landslides as polygons in the study area, it was difficult to digitize all landslide boundaries given that parts of landslides were covered by forest and the resolution of available images is limited. Nonetheless, we were able to determine that landslide boundaries often fit within a 30 m buffer from landslide points. Accordingly, this figure was taken as the assumed landslide area in this study. The total area of these 30 m buffered zones(from 121 landslide points) was converted to 563 cells as a GIS raster data format.

In order to generate a land use/land cover data set, we used an ALOS/AVNIR-2 image acquired from JAXA on 21 November 2009. The spatial resolution of AVNIR-2 ETM+ is 30 m. Prior to categorization, the satellite image was systematically investigated using spatial and spectral profiles to determine digital numbers for land use/land cover. Image examination was then performed to obtain a good difference between land use/land cover surface appearances. This was achieved by employing the supervised image classification technique and the ERDAS IMAGINE application. Finally, the classified image was evaluated and units from both the visual interpretation and classified image were integrated to prepare the final land use/land cover map (Fig. 3(b)).

ASTER GDEM of 30 m resolution was used for elevation data (Fig. 3(c)). Geological and soil maps were obtained through ISRI (Figs. 3(d) & 3(e)).

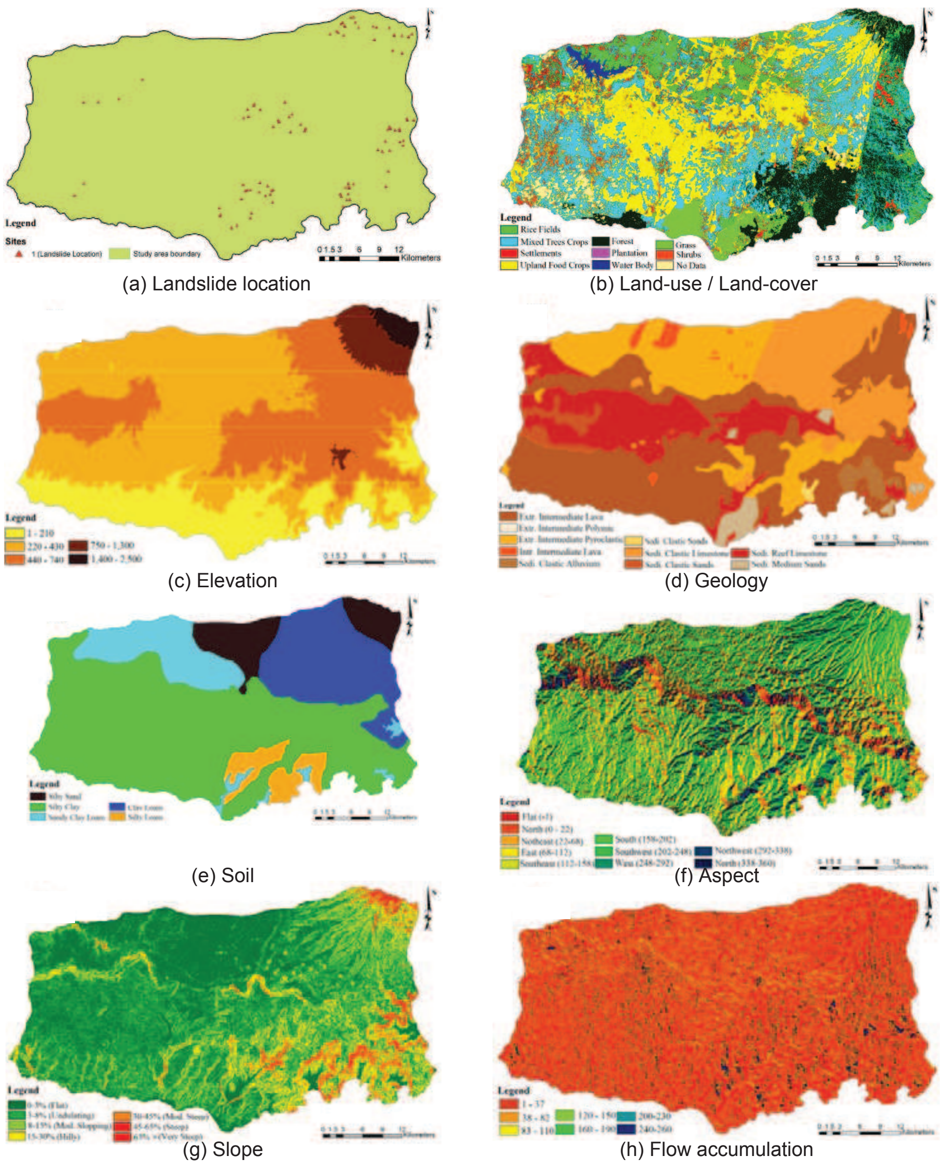

Figure 3 Thematic maps of data sets used for analysis.

3.3. *Data set Setting for Logistic Regression Analysis*

For logistic regression analysis, the dependent variable was landslide cells ($n = 563$) coded as 1, with no landslide cells ($n = 563$) coded as 0 (Fig. 3(a) "Landslide Location Map"). To enable further statistical analysis, equal numbers of landside cells and no landslide cells were selected randomly.The independent variables consist of the following environmental data sets:

- Land use/land cover: a categorical variable measuring the occurrence between 0 and 1 dependent variable for each class (range 0–13) (Fig. 3(b) "Landuse/Landcover Map")
- Elevation: a continuous variable measuring the elevation in meters above sea level (range = 0-1,979 m) (Fig. 3(c) "Elevation Map")
- Geology: a categorical variable measuring the occurrence between 0 and 1 dependent variable for each class (range 0–10) (Fig. 3(d) "Geology Map")
- Soil: a categorical variable measuring the occurrence between 0 and 1 dependent variable for each class (range 0–5) (Fig. 3(e) "Soil Map")
- Aspect: a continuous variable measuring slope direction (range 0–360 degrees) (Fig. 3(f) "Aspect Map")
- Slope: a continuous variable measuring slope steepness (range 0–69 degrees) (Fig. 3(g) "Slope Map")
- Flow accumulation: a continuous variable measuring flow from drainage areas (range 0–255) (Fig. 3(h) "Flow Accumulation Map").

Accumulated rainfall levels for 2007 were taken from the TRMM (Tropical Rainfall Measuring Mission), but this data was not included in the analysis due to its almost uniform effect of the rainfall in the study area by the TRMM data.

All thematic maps relating to elevation, slope, aspect, flow accumulation, soil, geology and land use/land cover were loaded in the ArcGIS environment and then, using the "intersect point tool" (Hawths tools of ArcGIS Desktop), "landslide" and "no landslide" features were selected as a target source to extract the information from each thematic layer. This function exports analysis results in CSV format for further analysis in a statistical analysis environment, such as R.

This study followed the workflow for a logistic regression model, beginning with:

- exploratory evaluation of possible relationships between the dependent variable and each separate covariate
- selection of one or more covariates as promising predictors and their further analysis by preliminary logistic modeling and significance testing
- creation of a final multivariate regression model which suggests the different levels of influence of the predictor (independent) variables
- prediction onto a raster surface of the likelihood of a site being present at each cell location, typically expressed as a probability ranging between 0 and 1.

Logistic regression analysis was used to build the landslide susceptibility index. The first step involved extracting intercept and class parameter values. The linear combination then was calculated by summarizing the intercept and parameters. Finally, the probability value was computed. A multivariate logistic regression was created using all of the predictors initially detected by univariate methods.

4. Results

4.1. *Parameters and Independent Variables Test*

A combination of covariates and different scenarios were tested (by removing or adding covariates) to determine the best model, using the AIC (Akaike Information Criteria) method. Calculating the multivariate logistic regression provided the regression coefficients for the independent variables (Table 2). Together with intercept values, these coefficients were used to produce the landslide susceptibility map.

In the creation of "Model-1" , initially 7 independent variables – "elevation", "slope", "aspect", "flow accumulation", "land use/land cover", "soil" and "geology" – were prepared, but the model selected only 3 independent variables, with sub-classes in each.

The "elevation" parameter was divided into 3 sub-classes: "low elevation" (0-500m), "mid-elevation" (500–1000m), and "high elevation" (1000-1979m). The model then excluded "high elevation". "Mid-elevation", and 'low elevation" in particular, exert great influence on landslide susceptibility.

Table 2 Regression coefficients obtained for Model-1.

Independent variables	Class	Coefficients		
	Dummy variables	Symbol	Value	Symbol
Elevation	Low elevation (0–500)	x_1	1.2085	b_1
	Mid-elevation (500–1000)	x_2	0.7494	b_2
Slope	Low slope (0–10^0)	x_3	−1.3999	b_3
	Mid-slope (10^0–20^0)	x_4	2.9814	b_4
	High slope (20^0–60^0)	x_5	4.1851	b_5
Land use/land cover	Rice fields	x_6	−16.3092	b_6
	Mixed tree crops	x_7	0.8285	b_7
	Upland food crops	x_8	−1.8834	b_8
	Grass	x_9	1.6739	b_9
	Forest	x_{10}	2.0934	b_{10}
Intercept			−1.8939	b_0

The "slope" parameter was also divided into 3 sub-classes: "low slope" (under 10 degrees), "mid-slope" (10–20 degrees), and "high slope" (over 20 degrees). The model selected all classes of slope, with better coefficient values shown for "mid-slope" and "high slope". This indicates that "mid-slope" and "high slope" are good predictors for landslide susceptibility.

The "land use" parameter was divided into 13 sub-classes, of which the model selected the 5 most effective in terms of landslide prediction. These 5 sub-classes are: "land use1 (rice fields)", "land use2 (mixed tree crops)", "land use4 (upland food crops)", "land use5 (grass)", and "land use8 (forest)". Of these,"mixed tree crops", "grass" and "forest" are most susceptible to landslides.

4.2. *Predicting Areas Prone to Landslides*

Using the coefficients from the logistic regression analysis, we generated a prediction surface on ArcGIS environment using the raster calculator. This computed the prediction for each cell by inserting the values in Eq. (1). This is shown as follows:

$$\begin{aligned}&-1.893 + ([low\ elevation] * 1.208) + ([mid - elevation] * 0.749) + ([low\ slope] * \\ &-1.399) + ([mid - slope] * 2.981) + ([high\ slope] * 4.185) + ([land\ use1] * \\ &-16.309) + ([land\ use2] * 0.828) + ([land\ use4] * -1.883) + (land\ use5) * 1.673) \\ &+([land\ use8] * 2.093)\end{aligned} \tag{1}$$

The output of this formula was named "logodds". We then computed the following Eq. (2) to retrieve the relative probabilities, leading to the final landslide susceptibility map of "Model-1" with sub-classes of parameters (Fig. 5).

$$(\mathrm{Exp}([\mathrm{logodds}]))/(1 + (\mathrm{Exp}([\mathrm{logodds}])) \tag{2}$$

The same procedure was repeated to create another landslide susceptibility map in which parameters are not divided into sub-classes. This was done to check the impact of major classes of parameters on susceptibility map production. The coefficient values obtained from the model and inserted in Eq. (3) are as follows:

$$-4.684 + ([\mathrm{slope}] * 0.379) + ([\mathrm{soil}] * 0.194) + ([\mathrm{geology}] * 0.060)(3) \tag{3}$$

By applying formula (2), we generated the second susceptibility map – "Model-2" without sub-classes of parameters (Fig. 4).

The "Model-1" and "Model-2" landslide susceptibility maps were divided into 5 categories: "extremely low susceptibility" (0–0.20), "very low susceptibility" (0.20–0.40), "low susceptibility" (0.40–0.60), "medium susceptibility" (0.60–0.80),

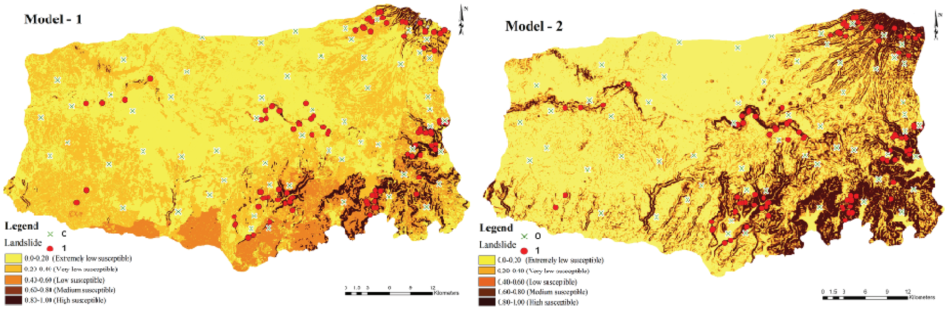

Figure 4 Landslide susceptibility maps (Model-1, with sub-classes of parameters) and (Model-2, without sub-classes of parameters).

and "high susceptibility" (0.80–1.0). Of the total study area, 4.9% was classed as "high susceptibility" (land mostly covered by mixed tree crops and forest); 0.2% as "medium susceptibility" (mountainous environment with steep slopes); 3.8% as "low susceptibility" (land covered by mixed tree crops); 17.8% as "very low susceptibility" (moderate to gently sloping land covered by mixed tree crops, upland food crops, rice fields and containing some settlements); and 73.2% as "extremely low susceptibility" (land covered by plantations, mixed tree crops, rice fields, forest and containing settlements).

4.3. *Gain Statistics*

A series of gain statistics was applied to both models to evaluate the ultimate effectiveness of the predictive model. The gain statistic is calculated using the following equation:

$$G_p = 1 - S_p/O \tag{4}$$

"S_p" is the percentage of the total surface area for which sites are predicted, given a specific p (probability) threshold, and "O" is the percentage of observed sites that fall withinS_p. The resulting "G_p''value is close to "1" for a perfect model and close to "0" for a poor one. Gain statistics have been computed for the two output surfaces.

Gain statistics for "Model-1" were obtained by using equation (4): the p-value > 0.90 and p-value > 0.80 are 0.76 and 0.77 respectively (Table 3). For "Model-2", p-value > 0.90 is 0.72 and p-value > 0.80 is 0.67. These verifications show that greater accuracy resulted when sub-classes of independent variables were applied ("Model-1"), compared to when no sub-classes of parameters were applied ("Model-2"). The results are listed in Table 3.

Table 3 Gain statistics for the two models' surfaces.

	Sp			*O*			
Output Surface *Output Surface*	# of cell with "*p*"	Total Cell	% %	# of LS with "*p*"	Total LS points	%	Gain Statistics *(Gp)*
Model-1 - p > 0.90	88136	1903123	4.63	108	563	19.18	0.76
Model-1 - p > 0.80	92895	1897190	4.90	121	563	21.49	0.77
Model-1 - p > 0.70	94069	1897190	4.96	123	563	21.85	0.77
Model-2 - p > 0.90	185639	1900710	9.77	199	563	35.35	0.72
Model-2 - p > 0.80	252556	1900710	13.29	228	563	40.50	0.67
Model-2 - p > 0.70	308974	1900710	16.26	243	563	43.16	0.62

4.4. *Overlay Analysis with Land Use/Land Cover*

Having obtained gain statistics for the two models, "Model-1" was selected for further overlay analysis using GIS. The final landslide susceptibility map was overlaid with a land use/land cover map. This made it possible to see the nature of the land (land use/land cover) in areas that have a medium to high susceptibility to landslides.

The land cover/land use sub-classes which overlapped with regions of medium susceptibility (Fig. 5) were "upland food crops", "shrub", "plantation", "settlement" and "mixed tree crops". In regions highly susceptible to landslides, the main land cover/land use sub-classes were "forest" and "mixed tree crops" (Fig. 6). The overlay analysis revealed that "forest", "upland food crops", "grass" and "mixed tree crops" are the sub-classes more susceptible to landslides, but that overall there is no significant relationship between land use/land cover and susceptibility to landslides.

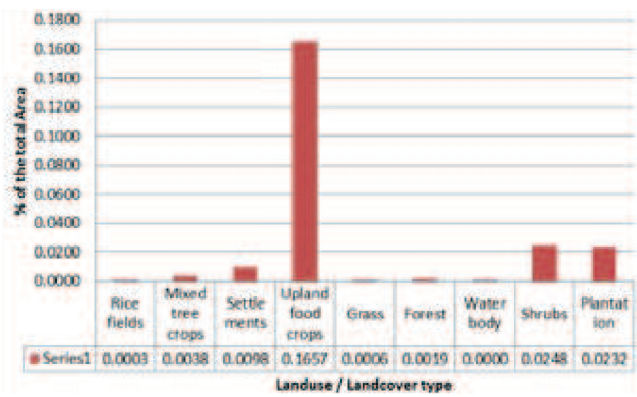

Figure 5 Land use types in areas of medium landslide susceptibility.

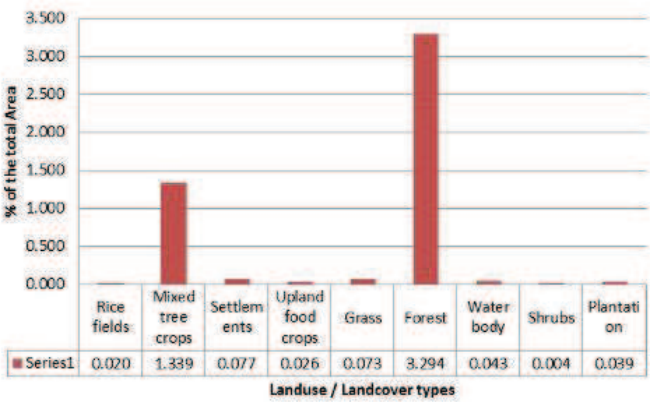

Figure 6 Land use types in areas of high landslide susceptibility.

4.5. *Conclusion*

This paper demonstrated an integrated approach for assessing landslide susceptibility. We captured the distribution of landslide occurrences using satellite remote sensing, and then applied statistical methods and GIS to assess landslide susceptibility. As a result, we produced a susceptibility map for potential landslides in the study area and assessed which land-use types are susceptible to landslide.

It was found that in areas where slopes are greater than 10 degrees, elevation is under 500m, and the land use/land cover types are "mixed tree crops", "grass" or "forest", slope failure is more likely.

Two output maps were prepared to predict the impact of parameters on landslide susceptibility. These maps were verified by gain statistics. From the results it was observed that "Model-1", in which parameters were divided into sub-classes, was more accurate (77%) than "Model-2" (67%), in which parameters were not divided into sub-classes.

The overlay analysis showed that "forest", "upland food crops", "mixed tree crops", "shrub" and "settlement" are the land use/land cover types most susceptible to landslide.

The landslide susceptibility information presented in this paper was compiled using the minimum level of required data. For better results and precision, the compilation of more specific data, particularly on soil properties, is recommended.

References

1. Indonesian Disaster Data and Information retrieved 21^{st} May-2009, (National Disaster Management Agency, 2009).
2. Wang, H. B. and Sassa, K., Comparative evaluation of landslide susceptibility in Minamata area Japan. *Journal of Environmental Geology*, 47, pp. 956–966 (2005).

3. Brenning, A., Spatial prediction models for landslide hazards: review, comparison and evaluation. *Journal of Natural Hazards and Earth System Sciences*, 5(6), pp. 853-862 (2005).
4. Hastie, WJ., Landslip hazard modeling, University of Otago, Dunedin. *Conf. Proc. of the Inaugural Colloquium of the Spatial Information Research Center*, pp. 96–105 (1990).
5. Wood, S.N., Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of American Statistical Association*, 99, pp. 673–686 (2006).
6. Zhaohua, C. and Jinfei, W., Landslide hazard mapping using logistic regression model in Mackenzie Valley, Canada. *Journal of Natural Hazards*, 42(1), pp. 75–89 (2007).
7. Hosmer, D.W. and Lemeshow, S., *Applied logistic regression*. John Wiley and Sons, New York. 2nd edition, pp. 375 (2000).
8. Lee, S., Ryu, J. H. and Kim, I. S., Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: case study of Youngin, Korea. *Journal of Landslide*, 4(4), pp. 327–338 (2007).
9. Brenning, A., Statistical Geocomputing combining R and Saga: The example of landslide susceptibility analysis with generalized additive models. In: J. Böhner, T. Blaschke, L. Montanarella, (Eds.), *SAGA–Seconds Out*, (Hamburg Contributions to Physical Geography and Landscape Ecology. 19, pp. 23–32 2008).
10. Pradhan, B. and Lee, S., Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models. *Journal of Environmental Earth Sciences, 60*, pp. 1037–1054 (2010).
11. Mathew, J., Jha, V.K. and Rawat, G.S., Landslide susceptibility zonation mapping and its validation in part of Garhwal Lesser Himalaya, India, using binary logistic regression analysis and receiver operating characteristic curve method. *Journal of Landslide*, 6, pp. 17–26 (2009).
12. Nandi, A. and Shakoor, A., A GIS-based landslide susceptibility evaluation using Bivariate and multivariate statistical analyses. *Journal of Engineering Geology*, 110 (1–2), pp. 11–20 (2009).
13. Cascini, L., The landslide risk in urban areas-southern Apennines Center, *Proc. XXI Italian Geotechnical Congress*, September, 11–13. PàtronEditore, pp. 127–134 (2002).
14. Cascini, L., Bonnard, Ch., Corominas, J., Jibson, R. and Montero-Olarte, J., *Landslide hazard and risk zoning for urban planning and development*. In: O. Hungr, R. Fell, R. Couture and E. Eberthardt, (Eds.), *Landslide Risk Management*. Taylor and Francis, London, pp. 199–235 (2005).
15. Christanto, N., Hadmoko, D.S., Westen, C.J., Lavigne, F., Sartohadi, J. and Setiawan, M.A., Characteristic and Behavior of Rainfall Induced Landslides in Java Island, Indonesia: an Overview. *Geophysical Research Abstracts,* Vol. 11, EGU2009-4069-7, (EGU General Assembly, 2009).
16. Highland, L.M. and Bobrowsky, P., *The landslide handbook - A guide to understanding landslides*. (Reston, Virginia, U.S. Geological Survey Circular 1325, p. 129, 2008).
17. Dhakal, A.S., Amada, T. and Aniya, M., Landslide hazard mapping and its evaluation using G1S: An investigation of sampling scheme for grid-cell based quantitative method. *Journal of Photogrammetric Engineering & Remote Sensing*, 66(8), pp. 981–989 (2000).
18. Saha, A.K., Gupta, R.P. and Arora, M.K. GIS-based landslide hazard zonation in the Bhagirathi (Ganga) valley, Himalayas. *International Journal of Remote Sensing*, 23(2), pp. 357–369 (2002).
19. Sarkar, S. and Kanungo, D.P., An Integrated Approach for Landslide Susceptibility Mapping Using Remote Sensing and GIS. *Journal of Photogrammetric Engineering and Remote Sensing*, 70(5), pp. 617–625 (2004).