GIS-based logistic regression model for landslide susceptibility mapping: a case study along the E-W highway, Malaysia

Tareq H. Mezughi & Yones A. Abulghasem

Keywords: Logistic regression (LR), landslide causative factors, landslide susceptibility, East-West highway (Malaysia), Area under a Curve (AUC)

Abstract

In this paper, a GIS-based methodology has been used to produce a landslide susceptibility map. The area selected is along the E-W highway in Malaysia where frequent landslides occur. The susceptibility mapping was based on a multivariate statistical method namely the logistic regression. The spatial database for factors that influence landslide occurrence were prepared from different sources including topographical maps, geological maps, satellite data, hydrological data, soil data and field data. Ten prepared thematic maps of factors were: slope gradient, slope aspect, elevation, curvature, and distance from road, drainage density, lithology, lineament density, soil, and rainfall. All maps were subdivided into different classes by its value or feature and then were converted to raster format in the ArcGIS 9.3, each representing an independent layer of causative factor in the constructed spatial database. the contribution of each factor towards landslide susceptibility was evaluated using the logistic regression model .The Wald test in logistic regression analysis suggests that slope gradient, lineament density, rainfall, distance from road and lithology play a positive important role in the landslide susceptibility. However, the curvature and drainage density factors play a negative important role in the landslide susceptibility in the study area. The results of the analysis have been validated by calculating the AUC of the prediction rate curve which shows an accuracy of 80.97%, indicating a high quality susceptibility map obtained from the logistic regression model. The map could be used by decision makers as basic information for slope management and land use planning

INTRODUCTION

Landslides are among the most costly and damaging natural hazards in the mountainous terrains of tropical and subtropical environments, which cause frequently extensive damage to property and occasionally result in loss of life. Over the last two decades, many governments and international research institutes in the world have investigated considerable resources in assessing landslide hazards and construct maps portraying their spatial distribution (Guzetti et al., 1999). These maps describe areas where landslides are likely to occur in the future and classify those areas into different susceptibility zones from very low to very high susceptible zones according to their susceptibility to landslides. Such as landslide susceptibility maps

are useful for planners and developers to choose favorable locations for future developments.

The Geographical Information Systems (GIS), enables data acquisition, storage, retrieval, modeling and manipulation. The GIS systems have the capability to incorporate various geographical technologies including remote sensing and global positioning systems hence they have become very vital for landslide susceptibility mapping. The analytical and combinational capacity of GIS has enabled the production of techniques used in landslide assessment for generating more precise maps, detailing the probable landslide hazard prone areas.

Landslide susceptibility mapping can vary from simple methods that use a minimum data to sophisticated mathematical methods that use practical mathematical methods using complex databases in computer-based geographic information system (GIS). For assessing landslide hazard different methodologies are proposed which are mainly grouped as: qualitative and quantitative methods. Both qualitative and quantitative approaches are based on the principle that future landslides are more likely to occur under the same conditions that led to past slope instability. In qualitative methods, the factors leading to landslides are ranked and weighted according to their expected importance in causing slope failure based on an earth scientist's experience. These methods are often useful for regional assessments (Aleotti and Chowdhury, 1999; van Westen et al., 2003).To overcome the subjectivity of qualitative methods, many statistical approaches have been developed and employed to become a major topic of research in landslides susceptibility studies during the last decade. In statistical analysis methods weighting values are computed based on the mathematical relationship between existing landslide distribution and their controlling factors.

In this study the logistic regression model, which is a multivariate statistical model was used to produce a landslide susceptibility map for an area located at the central northern part of Peninsular Malaysia along the E-W highway (Gerik – Jeli). The study area is frequently subjected to landslides following heavy rains, especially alongside the highway since it was constructed

Study Area

The study area lies in the central northern part of Peninsular Malaysia along the E-W highway between 5°:24":6' N to 5°:45":56.5' N latitude and 101°:7":53.6' E to 101°:50":26' E longitude, with a total area of 1205 km² (Fig. 1). It is characterized by rugged hills and mountain terrains covered by forest. The study area is frequently subjected to landslides following heavy rains, especially alongside the highway since it was constructed. The common types of landslides identified in the area were soil slides, soil slumps, rock falls, rock plane failure, wedge failure, toppling, and erosion failure.

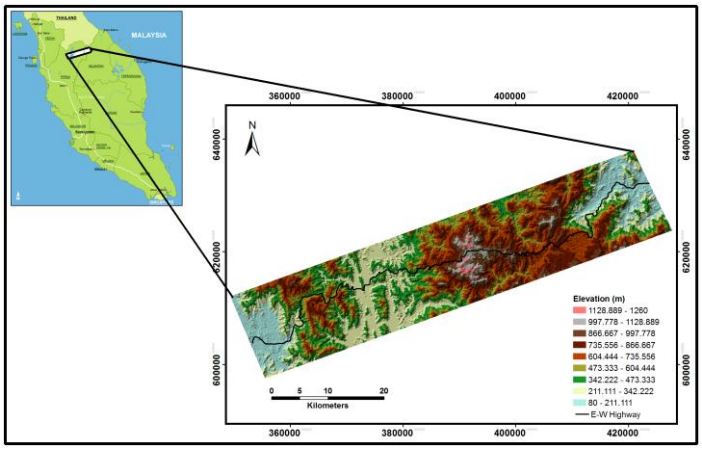

Fig. 1: Location of the study area shown with the TIN map

MATERIAL & METHOD

Data collection and database construction

The study method was applied in four main steps: data collection, construction of geospatial database, logistic regression model analysis, and validation of the results. In this study a landslide location map and thematic maps of the various causative factors were prepared in ArcGIS 9.3 software. The landslide location map was prepared based on interpretation of the aerial photographs, literature review and field work. Data layers of considered factors were obtained from different sources such as topographical maps, geological maps, satellite data, hydrological data, soil data and field data. Ten prepared thematic layers of factors were: slope gradient, slope aspect, elevation, Curvature, distance from road, drainage density, lithology, lineament density, soil, and rainfall.

Digital Elevation Model and Its derivatives

Digitized contour and survey base points from the 1:50,000-scale topographic maps were extracted and processed to generate a 20 m DEM using the TIN module of 3D Analyst tool extension on ArcGIS 9.3. Using this DEM, the slope angle map, slope aspect map, curvature map and elevation map were automatically derived. Slopes in the area were found to be vary from 0° to 87° and they were classified into 5 classes, slope aspect map was classified to 9 classes (fig. 3), the elevation in the studied area was found to range from 0 to 1268 m and it was classified into six classes.

Lineaments map

In this study, the band 4 (0.75 -0.90µm) of Landsat 7 ETM+ image was used to delineate lineaments because it has the ability to reflect the best contrast between lineaments and the surroundings in tropic areas where most of the rocks are covered

by vegetation. Lineaments were traced from visual interpretation of band 4 of Landsat 7 ETM+ image, and from filtered images obtained from four directional sobel filters (Fig 2) which applied on the band 4 of Landsat 7 ETM+ image. The produced lineaments map was then used to compute lineaments density map.

N-S			NE-SW			E-W			NW-SE		
-1	0	1	-2	-1	0	-1	-2	-1	0	1	2
-2	0	2	-1	0	1	0	0	0	-1	0	1
-1	0	1	0	1	2	1	2	1	-2	-1	0

Figure 2. 3 by 3 Sobel kernel directional filters in four principle directions

The lineaments density for each 20- by 20-m cell was computed using the line density analyst extension on ArcGIS 9.3, and classified into five equal classes: very low (<0.5 km/km²), low (0.5 to 1 km/km²), moderate (1 to 1.5 km/km²), high (1.5 to 2 km/km²), and very high (> 2 km/km²), density. Distance to lineaments map was also generated from the lineaments map using the straight line distance of spatial analyst extension tool on ArcGIS 9.3 and classified to five classes.

Lithology map

From the lithological point of view, 10 units were digitized from six geological maps (scale 1:63,360) covering the area. These units are described in Table 1.

Table 1. Lithological units

Lithological unit	Description
LU1	Granite
LU2	Metagreywacke and metasandstone
LU3	Quartz-chlorite schist, sericite schist, graphitic schist and phyllite
LU4	Quartz-mica schist, quartz-graphite schist, and minor amphibole
LU5	Metatuff of rhyolitic composition
LU6	Chert, shale, slate and metasiltstone
LU7	Metarenite
LU8	Phyllite and slate
LU9	Marble with calcereous matesediments
LU10	Granite, granodiorite and syenite

Drainage map

The drainage map was digitized from the topographic maps of scale 1:50.000. Then the drainage density map was computed considering a 20- by 20-m cell and classified

into five equal intervals classes (fig. 3): very low (<0.876 km/km²), low (<0.876 - 1.752 km/km²), moderate (1.752 to 2.629 km/km²), high (2.629 to 3.505 km/km²),and very high (>3.505 km/km²).

Rainfall map

Annual rainfall data for the years 2000, 2005 and 2009 were collected from three meteorological stations and then they were used to produce a rainfall map using an interpolation method. The area was classified to four rainfall zones which were (fig. 3): (<2000 mm/yr, 2000-2500 mm/yr, 2500-3000 mm/yr, and >3000 mm/yr).The maximum rain fall in the area is 3970 mm/yr and mostly occurs in the east. On the other hand, the lowest rainfall is 1590 mm/yr and occurs to the western part of the area.

Road distance map

The road distance map was digitized from the topographic map and then classified to seven distance buffer classes (fig. 3) (0 -50 m ,50 - 100 m ,100 - 150 m,150 - 200m, 200m - 250m, 250-300 m, and >300m) calculated on both sides of the roads.

Soil map

A soil map was prepared using 32 soil samples collected in the field from residual soils formed by weathering processes on the rocks. In this study the soil were classified according to the unified soil classification system (USCS). The grain size distribution of gravel and sand particles were measured using the sieve analysis. Fine-grained soils, which could be silts or clays, cannot be measured using the sieves therefore they were classified according to their Atterberg limits. The liquid limit (LL), plastic limit (PL) and plasticity index (PI) values of fine-grained soils were determined from laboratory analysis. The values of plasticity index and liquid limit are plotted on a plasticity chart, and the fine-grained soils were classified according to its plotting region on the chart. The two types of soil were identified were SILT-sandy and SAND-silty (fig. 3).

Landslides location map

Landslides location map was prepared based on interpretation of the aerial photographs, the previous studies conducted on the area (Abdul Ghani Rafek et al., 1989) and the collected historical information on landslides occurrences. In addition, field work has been carried out to map the recent landslides. A total of 143 landslides were mapped in the area. The common types of landslides identified in the area were rock slumps, rock falls, wedge slides, topplings, soil slides and soil slumps.

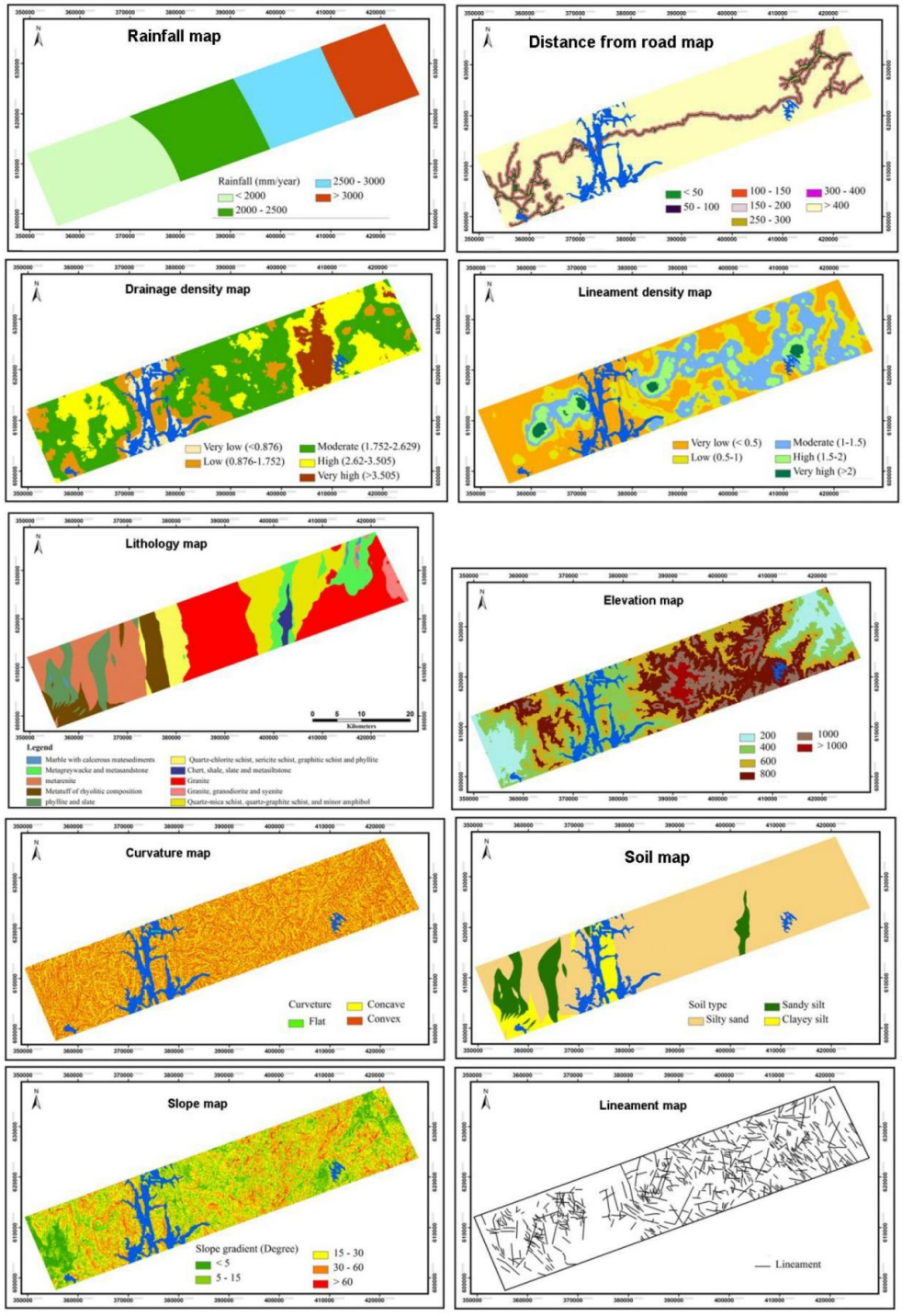

Figure 3 Landslide factors maps

Landslide susceptibility mapping using the logistic regression model

a. Sampling

The spatial relationship between landside and landslide influencing parameters was evaluated using the logistic regression method. Three different datasets were selected and the logistic regression model was run with each dataset by SPSS (2010) software package. These training sets were then evaluated using a chi-square of Hosmer-Lemeshow test, Cox and Snell R^2 and Nagelkerke R^2. In addition, the accuracy percentages of classification for the three training sets were calculated. Table 2 shows the statistics used to evaluate the three datasets. As stated by Ayalew & Yamagishi (2005), the chi-square value is considered as the key element in standard analysis of the test as it shows the significant test for logistic regression. The chi-square value in sample one is fairly higher than that of the others and it can be concluded that the causal factors have a sufficient influence on the landslide occurrences. Moreover, sample one shows higher values for the R^2 value of Cox & Snell and Nagelkerke. The higher R^2 indicates the extent to which the model fits the data. R^2 value prior to 1 means that the model fits the data perfectly, whereas 0 indicates there is no relationship with the data (Ayalew & Yamagishi 2005). However, when the R^2 is greater than 0.2, this is an evidence of relatively goodness of fit (Clark & Hosking 1986). The final step was concerned with comparing the accuracy percentages of classification for the three training sets. Based on this comparison, it was revealed that the first sampling dataset gained the highest overall accuracy, so it was selected to be used in the logistic regression analysis.

Table 2 Summary statistics for the three samples datasets

Training set no.	-2ln Log likelihood	Cox and Snell pseudo R^2	Nagelkerke Pseudo R^2	Chi-square	Overall accuracy %
1	131.379	0.605	0.807	267.874	88.9
2	184.108	0.526	0.702	215.145	85.8
3	251.611	0.401	0.535	147.641	78.5

b. Forward stepwise and checking the fitness of the models

The logistic regression model of dataset number one was constructed to analyze the data representing the ten independent factors by using the forward stepwise logistic regression method. The null hypothesis used to test is that the coefficient of the independent variable (b) is 0. The statistical test used is the Wald chi-square value (χ^2) at 5% significance level interval for the corresponding degree of freedom (df):

$\chi^2 = (b/SE)^2$

where, S.E. is the standard error and is given as SE = s/n, where s is the standard deviation of the input data samples and n is the number of pixels in the input data.

Thus, the variables with estimated coefficients having a significance value (Sig.) of less than 0.05 are found to be significant, or in other words, these are accepted as influential independent variables. Application of the forward stepwise method usually begins with a model which does not involve any independent variables, and then, the following steps are concerned with determining the variables with a significance value less than 0.05 as significant and adding it to the model while at the same time rejecting all other variables with a significance value greater than 0.05 because such variables do not significantly affect the outcome of the dependent variable. The model summary of the stepwise is shown in Table 3.

Table 3 Model Summary statistics of the forward stepwise for sample no.1

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square	Chi-square	Accuracy (%)
1	235.643	0.433	0.578	163.610	83
2	209.750	0.482	0.643	189.503	81.9
3	191.451	0.514	0.685	207.801	86.8
4	175.055	0.541	0.721	224.198	85.1
5	160.797	0.563	0.751	238.456	85.8
6	153.218	0.574	0.766	246.034	86.1
7	143.702	0.588	0.784	255.551	87.5
8	138.422	0.596	0.794	260.830	88.5
9	132.627	0.605	0.807	267.874	88.9

As shown in Table 3, the extent to which the final model improves over the null model is measured by the difference in the -2 log likelihood (-2LL) values in the nine steps. Generally, the lower the value of -2LL of the model is, the better is the fitness of the model to the data. In the case of the present study, there was a decrease in the -2LL value from 235.643 at the first step to 132.627 at the final step. Moreover, the usefulness of the model was measured by using the Cox & Snell's and Nagelkerke's R^2. A higher R^2 value of Cox & Snell and Nagelkerke is an indicative of a better model. Thus, it was found that the most significant model was achieved in the final iteration step (9) with an overall calibration accuracy of 88.9 % as shown in Table 4.

*Table **Error! No text of specified style in document.** Classification table of training dataset for Logistic regression base model*

	Observed		Predicted		
			LANDSLID		Percentage
			0.00	1.00	Correct
Step 1	LANDSLID	0.00	111	32	77.6
		1.00	17	126	88.1
	Overall Percentage				82.9
Step 2	LANDSLID	.00	116	27	81.1
		1.00	25	118	82.5
	Overall Percentage				81.8
Step 3	LANDSLID	.00	114	29	79.7
		1.00	10	133	93.0
	Overall Percentage				86.4
Step 4	LANDSLID	0.00	115	28	80.4
		1.00	15	128	89.5
	Overall Percentage				85.0
Step 5	LANDSLID	0.00	119	24	83.2
		1.00	17	126	88.1
	Overall Percentage				85.7
Step 6	LANDSLID	0.00	121	22	84.6
		1.00	18	125	87.4
	Overall Percentage				86.0
Step 7	LANDSLID	0.00	122	21	85.3
		1.00	15	128	89.5
	Overall Percentage				87.4
Step 8	LANDSLID	0.00	124	19	86.7
		1.00	16	127	88.8
	Overall Percentage				87.8
Step 9	LANDSLID	0.00	125	18	87.4
		1.00	14	129	90.2
	Overall Percentage				88.9

c. Test of single variables

For estimating the significance of the coefficients for each single variable, it was tested with the Wald test. To obtain this, the maximum likelihood estimate of every variable was compared with its estimated standard error (Hosmer & Lemeshow 1989; Van Den Eeckhaut et al. 2006). Thus, a coefficient is considered significant if the "null hypothesis estimating that coefficient is 0" and can be rejected at a 0.05 significance level. According to their different levels of influence on the landslide occurrence, different variables have different coefficient values. The regression coefficients for the variables retained from the final model are given in Table 5.

Furthermore, determination of the effect of that variable on the probability of landslide occurrence can be done by examining the sign of a dependent variable's coefficient estimate. The positive coefficient means those classes are positively responsible for the landslide occurrence and the negative coefficient means that they negatively influence the landslide. Table 5 indicates that the coefficients of slope, lineament, rainfall, road factors and the lithology units LU1, LU2, LU6, LU7 and LU8 which are composed of granite, metagreywacke and metasandstone, chert, shale, slate and metasiltstone, metarenite and phyllite and slate respectively are very significant and positive in the logistic regression; hence, they play a positive important role in the landslide susceptibility. However, the coefficients of the curvature, drainage factors are significant but negative, so this will reduce the landslide susceptibility. In other words, these factors play a negative important role in landsliding in the study area. Lithology units LU3, LU4, LU5, LU9 which are composed of quartz-chlorite schist, sericite schist, graphitic schist and phyllite, quartz-mica schist, quartz-graphite schist, and minor amphibole, metatuff of rhyolitic composition respectively and soil units S1and S2 which are composed of silty sand and clayey silt are not significant for landsliding in the study area based on the model, and, hence, are less important for the prediction of landsliding.

Table 5 The regression coefficients estimated for retained independent variables in the logistic regression model

Variable	B	S.E.	Wald	Sig.
Slope	0.053	0.026	4.050	0.044
Curvature	-1.006	0.480	4.395	0.036
Lineament	1.020	0.504	4.098	0.043
Drainage	-1.037	0.357	8.410	0.004
Rainfall	0.003	0.001	9.968	0.002
Road	0.0003	0.000	7.747	0.005
Soil			5.203	0.074
Soil (1)	-16.507	4214.780	0.000	0.997
Soil (2)	-17.879	4214.780	0.000	0.997
Lithology			26.531	0.002
LU(1)	3.670	1.285	8.156	0.004
LU(2)	0.387	1.112	0.121	0.728
LU(3)	-31.001	7451.488	0.000	0.997
LU(4)	-6.062	3.340	3.295	0.070
LU(5)	-32.615	6907.712	0.000	0.996
LU(6)	2.685	1.627	2.722	0.099
LU(7)	3.668	1.544	5.647	0.017
LU(8)	5.369	1.792	8.980	0.003
LU(9)	-20.244	25797.138	0.000	0.999
Constant	9.006	4214.782	0.000	0.998

B: Coefficients for each class S.E.: Standard Error of estimate values
Wald: Wald chi-square value Sig.: Significance of the value

d. Landslide probability and susceptibility map

The logistic regression yields the intercept of the model and the coefficients values for all retained variables from the final step of the logistic regression analysis model. These coefficients values were then transferred into Arc Map in ArcGIS 9.3 to be assigned as weights for the individual independent variables. Using all these coefficients and intercept value, the predicted probability (landslide susceptibility index LSI) for the area was calculated using the following equation:

P (landslide susceptibility index) = 1/1 + exp [9.006 + (0.053* Slope) + (-1.006* Curvature) + (1.020* Lineament) + (-1.037* Drainage) + (0.003* Rainfall) + (-0.0003* Road) + ($Soil_c$) + (0.098* Foliation) + $Lithology_c$

where, P (landslide susceptibility index) is the landslide-occurrence possibility, Slope is the slope value, Curvature is curvature value, Lineament is lineament density value, Drainage is the drainage density value, Rainfall is the rainfall precipitation value, Road is distance from road value, $Lithology_C$ and $Soil_c$ are logistic multiple regression coefficients for the lithologic units and soil units as listed in Table 5. The calculated P (landslide susceptibility index LSI) , probability values of the entire study area lie in the range of 0 to 0.99 which were considered as predicted probabilities of landslide in a pixel in the presence of the independent variables considered in the model. The pixels which have a value of probability being higher than (0.5) are more subject or susceptible to slope failure. Finally, the predicted probability values were categorized into five landslide-susceptible zones based on the success rate curve method. The LSZ map, thus, produced is given in Figure 4. The statistical calculation of the areas and percentages of landslide susceptibility zones in the study area are listed in Table 6. The percentages of landslide susceptibility classes and the landslide occurrence in each class are shown in Fig.5.

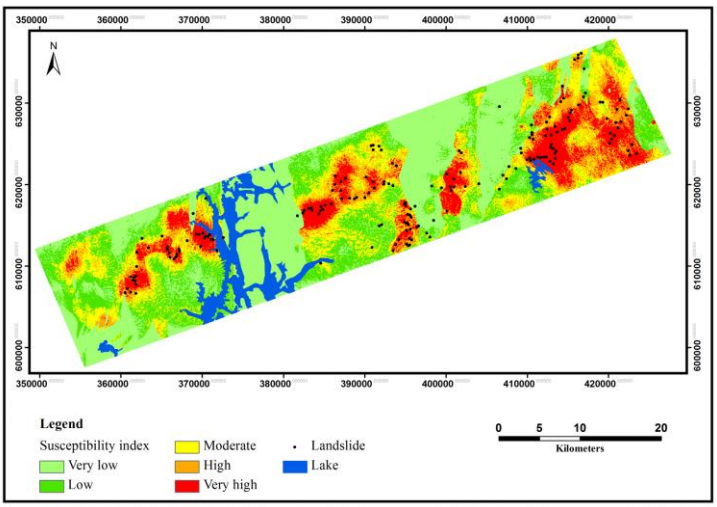

Figure 4 The landslide susceptibility map of the study area based on logistic regression method

Table 6 Area of different landslide susceptibility zones obtained with the logistic regression method

Landslide susceptibility zones	Area	
	(Km^2)	(%)
Very low landslide susceptibility	447.83	37.25
Low landslide susceptibility	263.62	21.93
Moderate landslide susceptibility	213.76	17.78
High landslide susceptibility	135.64	11.28
Very high landslide susceptibility	141.59	11.78

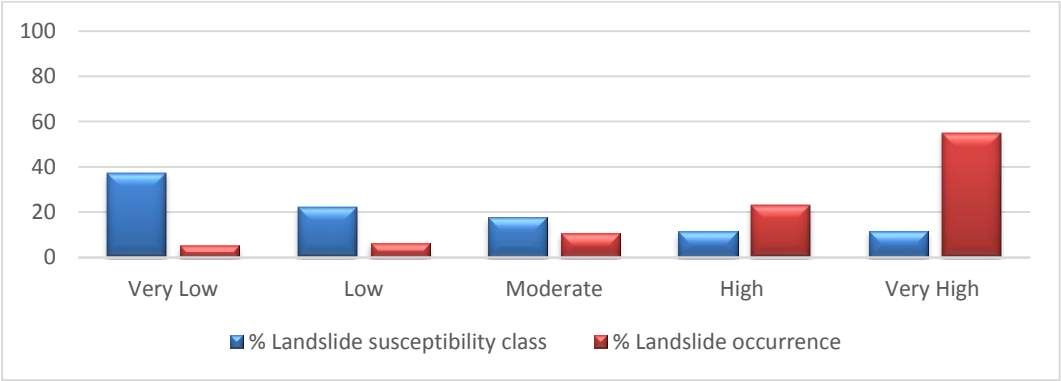

Figure 5 Histograms showing the percentages distribution of landslide susceptibility classes with the percentages landslide occurrence in each class

Verification of the model

To verify the results, the landslide susceptibility index, which indicates calculated probability values of the entire study area was compared with known landslides. The LSI was assessed in terms of its predictive power validity by calculating the prediction rate curve. To produce the prediction rate curve, the computed index values of all cells in the targeted area were arranged in descending order, and divided into 100 equal classes ranging from very highly susceptible classes to non-susceptible classes. Then the 100 classes were overlaid and intersected with known landslides to establish the percentage of landslide incidences in each susceptible class. Fig. 6 illustrates the prediction rate curve as a line graph. The Fig. 6 also indicates the satisfactory results, highest susceptibility pixels that envelop 10% of the study area includes 50% of known landslides , while the 20% high susceptible area covers more than 65% of landslides. Later, the prediction of the map was validated more precisely using the area under the curve (AUC) by ascertaining that the ideal prediction will have highest AUC of 1. In our study, the AUC was found to be 0.8097. Consequently, it indicates that the prediction precision of the acquired

map is 80.97% with respect to the ideal value of 100%, which is comparatively satisfied.

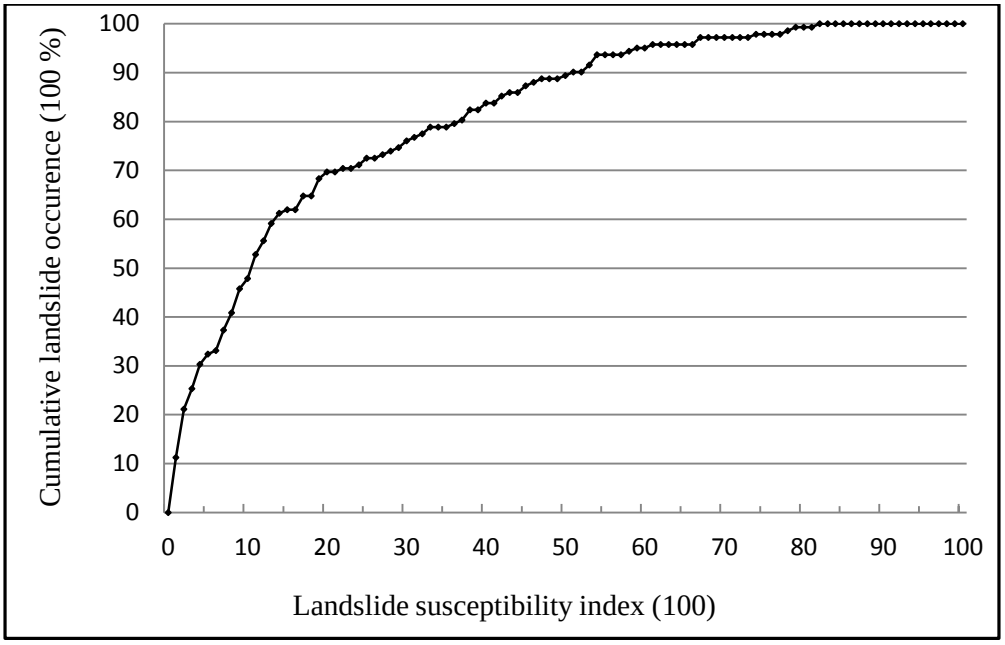

Figure 6 Cumulative frequency diagram showing percentage of study area classified as susceptible (x-axis) in cumulative percent of landslide occurrence (y-axis).

Conclusion

This work has provided a landslide susceptibility assessment using logistic regression model with the aid of GIS for a landslide prone area located in central northern Malaysia. This model is cost effective and capable of quickly contributing to the landslide assessment by manipulating data and performing the essential analysis. In order to accomplish this purpose ten landslide control factors were employed in the analysis which includes: slope gradient, slope aspect, elevation, Curvature, distance from road, drainage density, lithology, lineament density, soil, and rainfall. A logistic regression analysis was implemented in order to obtain the weights for every factor and class using direct pairwise comparison, later based on these weights, thematic maps of factors were combined by weighted overly techniques and the landslide susceptibility map of the study area was created. The obtained map was classified into five susceptibility classes which specified that the high and very high susceptible zones include about 23.06% of the total area, while about 29.18 % were classified as low and very low susceptible zones and 17.78 % is moderately susceptible zone. At the end, the map was validated with known landslides data based on the area under curve (AUC) method, by which the prediction precision of 80.97% was established.

References

Guzzetti, F., Carrara, A., Cardinali, M., &Reichenbach, P. 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. Geomorphology 31 (1–4):181-216.

Aleotti, P., & Chowdhury, R. 1999. Landslide hazard assessment: summary review and new perspectives. Bulletin of Engineering Geology and the Environment 58 (1):21-44.

Van Westen, C. J., Rengers, N., &Soeters, R. 2003. Use of Geomorphological Information in Indirect Landslide Susceptibility Assessment. Natural Hazards 30 (3):399-419.

Abdul Ghani, R., Komo, I., &Tan, T. H. 1989. Influence of geological factors on slope stability along the East-West Highway, Malaysia. Paper read at Proceeding of the international conference on engineering geology in tropical terrains, 26-28 June, at Bangi, Malaysia.

Ayalew, L., & Yamagishi, H. 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology 65 (1–2):15-31.

Clark, W. A. V., & Hosking, P. L. 1986. Statistical methods for geographers: Wiley.Clerici, A., Perego, S., Tellini, C., &Vescovi, P. 2002. A procedure for landslide susceptibility zonation by the conditional analysis method. Geomorphology 48 (4):349-364.

Hosmer, D. W., & Lemeshow, S. 1989. Applied Regression Analysis. New York, USA: Wiley.

Van Den Eeckhaut, M., Vanwalleghem, T., Poesen, J., Govers, G., Verstraeten, G., &Vandekerckhove, L. 2006. Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium). Geomorphology 76 (3–4):392-410.