4th Japan-Taiwan Joint Worlshop on
Geotechnical Hazards from Large Earthquakes and Heavy Rainfalls
25-28, October, 2010, Sendai, Japan

USING LOGISTIC REGRESSION TO PREDICT THE SUSCEPTIBILITY OF DEBRIS FLOW HAZARD IN HUALIEN, TAIWAN

Yeng Li, Chen[1] Wen-Chun Lo[2] Chen-Yu Chen[3] Che-Wei Shen[4*] Ting-Chi Tsao[5]

ABSTRACT

Taiwan was recently ravaged by frequent natural disasters. The typhoons and heavy rainfall caused landslides, often causing uncalculatable damage. Because of this, this study focuses on the 160 potential Debris flow creeks in Hualien County announced by the Soil and Water Conservation Bureau. It is based on landslide inventory and uses the topographic factors raised by many in the previous studies. These topographic factors are divided into five categories: factors of form, slope grade factors, height factors, slope aspect factors, and landslide factors in catchment. Through principle component analysis and correlation analysis, 10 significant topographic factors are chosen. The study used logistic regression to analyze each Debris flow susceptibility value. Through Classification Error Matrix, Receiver Operator Characteristic curve, and historical disasters, it attests to the accuracy of pattern classification. After using statistical tests to prove model significance, based on susceptibility classification criteria, the study creates a debris flow susceptibility map, establishes a susceptibility analysis flow chart with statistical theoretical foundation, providing reference materials for drafting slope disaster prevention policies and risk management.

Keywords: debris flow, logistic regression, topographic factors, susceptibility map.

INTRODUCTION

Soil and Water Conservation Bureau of Council of Agriculture (SWBC) conducted the "The investigation of vulnerability factors of debris flows torrents and it's counter measurements", targeting the 1,420 (currently 1,552) potential Debris flow creeks to undertake field investigation and GIS data collection. This study uses the 160 potential Debris flow creeks in Hualien County announced by SWBC as analysis sample, using catchment as analytical unit and calculating the factors of form, slope aspect factors, height factors, slope grade factors, and landslide factors in catchment, totaling 20 topographic factors. Through principle component analysis (PCA) and correction

[1] Deputy Director General, Soil and Water Conservation Bureau, Council of Agriculture, Nantou 540, Taiwan, R.O.C.
2 Assistant Engineer, Debris Flow Disaster Prevention Center, Soil and Water Conservation Bureau, Council of Agriculture, Nantou 540, Taiwan, R.O.C.
3 Director, Debris Flow Disaster Prevention Center, Soil and Water Conservation Bureau, Council of Agriculture, Nantou 540, Taiwan, R.O.C.
4 Research Engineer, Geotechnical Engineering Research Center, Sinotech Engineering Consultants, INC., Taipei 110, Taiwan, R.O.C.
5 Engineer, Geotechnical Engineering Research Center, Sinotech Engineering Consultants, INC., Taipei 110, Taiwan, R.O.C.

STUDY AREA

1. Terrain and Geology

Hualien County is a mountainous region, with plains making up only a small part. In addition to the plains distributed around the Meilun River alluvial fan, majority is distributed along the two sides of the East Rift Valley in strips. Mountains account for majority of Hualien County's total area (approximately 87%), with 40 mountains exceeding 3,000 m. Of those belong to the Central Range, Hsiukuluan Peak is the tallest (3,833m). Of those belong to the Coastal Range, Xingang Peak is the tallest (1,628m). The topographic map is shown in figure 1.

The county's terrain can be divided into Central Range area, Coastal Range area, and Rift Valley plain area. Hualien's geology (as shown in figure 2) can be roughly divided into three eras and igneous rock: (1) the First Tertiary and Paleogene Periods metamorphic rock region, with Tananao Schist, Xicun Formation (Eocene slate and phyllite in Snow Mountain Range), Xingao Formation (Eocene slate and phyllite in Backbone Range) distributed along the east side of the Central Range; (2) Miocene Lushan Formation is distributed in the mountains to the west of Dawu and Chihpen. Tuluanshan Formation is distributed along the Coastal Range. (3) Dagangkou Formation and Chimei Formation from Pliocene and Pleistocene Periods, Lichi Melange, Puyuma Hill Conglomerate, accumulation of red earth platform, uplifted coral reef, and alluvial layer were distributed along Coastal Range, Taitung Longitudinal Valley, and the east coast (There is no Puyuma Hill Conglomerate in the Hualien area.). (4) Igneous rock: Serpentinite and mafic igneous rock from the First Tertiary Period are distributed around the Central Range. Andesite from the Miocene Period as well as gabbro, peridotite, basalt, serpentinite, and agglomerate from an unknown period are distributed in the Coastal Range.

2. River System

The drainage system within Hualien County is composed of primarily three river basins: Heping River, Hualien River, and Hsiukuluan River. Heping River is located in the northeastern part of Taiwan, on the border of Ilan and Hualien. To its east is the Pacific Ocean, and to the west is Tachia River. To its south lies the Liwu River, and to the north, Nanau river and Lanyang River.

3. Fault Distribution

The faults within Hualien County include the Chimei Fault, Yuli Fault,Yuemei Fault, Meilun Fault, and Chihshang Fault(as shown in figure 3).

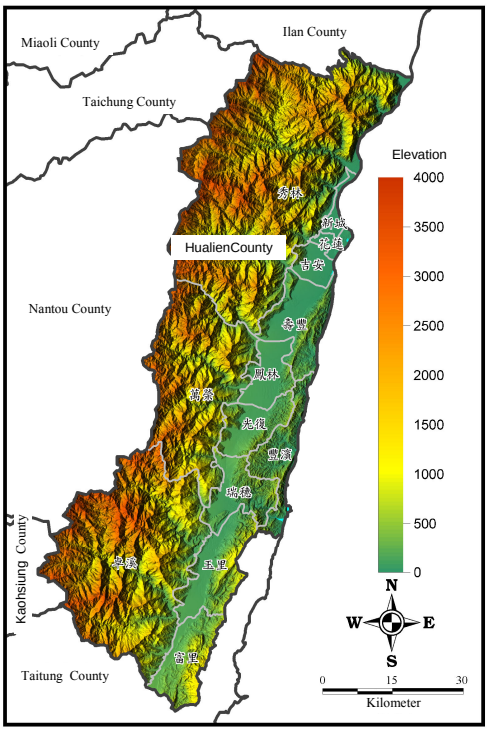


Fig.1 topographic map (redraw from Sinotech,2008)

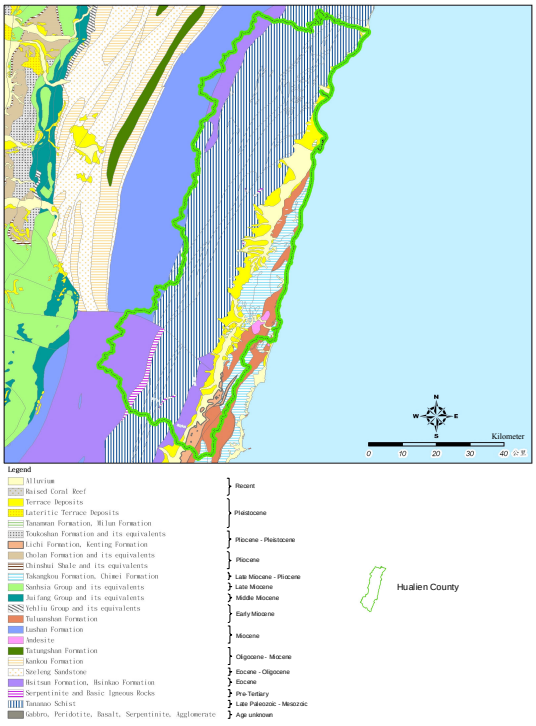


Fig.2 1/250,000 geologic map (redraw from CGS-MOEA geologic map)

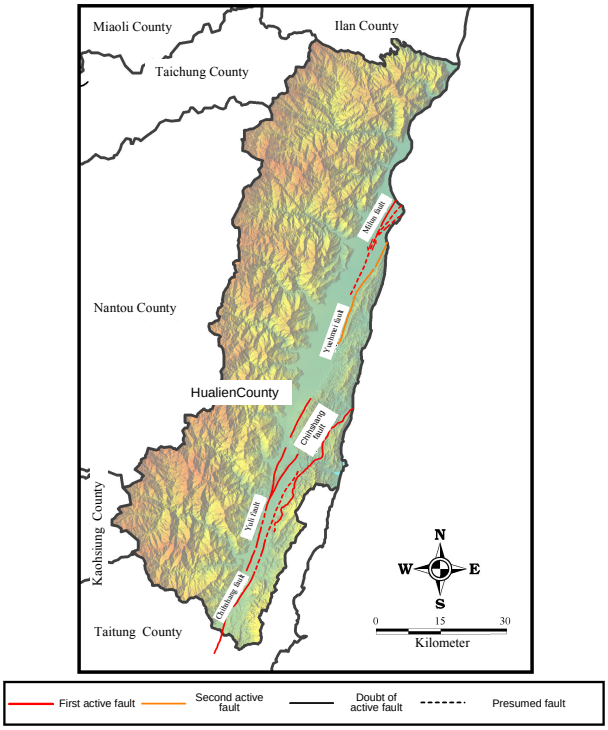


Fig.3 Fault Distribution map (redraw from CGS-MOEA activity fault map)

TOPOGRAPHIC FACTORS

This debris flow susceptibility assessment model in this study considers both topographic and landslide characteristic factors. Using 5m-resolution digital elevation model (DEM) and extration from the post-Typhoon-Toraji landslide inventory, 20 factors under 5 categories are obtained. The following summarizes the topographic factor categories considered (Sinotech, 2008):

1. Catchment factors of form: including catchment area, catchment length, catchment circumference, form factor, elongation ratio, circularity ratio. River basin form factor was raised by Horton in 1932 and is defined as $F=A/L^2$, where L is the total river length, and A is the catchment area. Form factor represents river basin area per main river length, primarily reflecting the length of time of concentration in catchment in order to accurately reflect the effect on catchment by river width.
2. Catchment slope aspect factors: including the average slope aspect X vector, average slope aspect Y vector, slope aspect standard deviation, average slope aspect. Catchment slope aspect factors mainly reflect strike of slope.
3. Catchment height factors: including Hypsometric Integral (or residual soil rate), average elevation, elevation standard deviation, elevation coefficient of variation. Height factors reflect the degree of ups and downs in the terrain of the study area. Greater ups and downs signify greater erosion by rain or wind.
4. Catchment slope grade factors: including average slope grade, first 5% average slope grade, first 10% average slope grade, first 15% average slope grade. Slope grade factors reflect the fact the steeper the slope, the easier it is to failure.
5. Catchment landslide characteristic factors: based on the satellite images after Typhoon Aere (2004), the total landslide area and the rate of collapse within 50 m of the two sides are obtained through digitization. The distribution of the landslides in catchment is considered, mainly reflecting the possibility of existing landslides and exposed areas to collapse again.

ESTABLISHING A FLOW CHART THROUGH SUSCEPTIBILITY MODEL

Based on the landslide inventory established by the "Hazard Factors Investigation in Areas of Potential Debris flow" conducted between 2006 and 2008, potential Debris flow creek catchments are divided into Debris flow creeks and non-Debris flow creeks. Debris flow creeks are those that had landslides and had actual damages and losses, and the opposite is true for non-Debris flow creeks. The following points illustrate the steps in establishing susceptibility model (Sinotech, 2008). Analysis is shown in figure 4.

1. Establishing overflow points and calculating the topographic factors

First, establishing overflow points is needed. This study used the definition by SWCB: " First, the affected range of 105° sector area as calculated by Hiroshi Ikeya Formula is used as base map. Possible overflow points (such as valley entrance, obstructions, or where terrain suddenly becomes smooth) are repositioned through site inspection. Afterward, they are corrected based on the current terrain, and the areas where debris flow would be impossible are eliminated." Based on this rule, the distribution of the 160 potential Debris flow creeks within Hualien and the overflow points are completed, as shown in figure 5. Through 5 m digital elevation model (DEM) and the landslide inventory made after Typhoon Toraji (2001), extraction of the topographic factors in catchments above overflow points is made.

2. Principle Component Analysis

Principle Component Analysis is a technique used to simplify data sets. his study chooses a representativeness (also known as cumulative proportion) of 85% and 5 principle components (Stone and Brooks, 1990). Through eigenvalues and cumulative variance percentages, representativeness for the principle components is obtained (as

shown in table 1), and the first five principle components are selected.

3. Correlation Analysis

This study uses Pearson product-moment correlation coefficient to examine the independence of the various topographic factors. The range is between –1 and 1. When the value approaches 1, the two variables are positively correlated. When the value approaches –1, the two variables are negatively correlated. When the value equals 0, the two variables are completely independent (Fisher, 1928). A correlation matrix is obtained through the analysis, as shown in figure 6. Based on the outcome of the analysis, this study eliminates the dependent factors selected in Principle Component Analysis, and the combination of significant topographic factors in the studied area is obtained.

4. Sample Selection

The ratio between Debris flow creek samples and non-Debris flow samples is often disproportional. Considering the importance of unbiased estimate of the samples, analysis should have a sample ratio of 1:1 between Debris flow creeks and non-Debris flow creeks. This study uses Simple Random Sampling to select the training sample for analysis (Lee et al., 2008). The procedure includes all of the Debris flow creek samples and uses Random Sampling to select non-Debris flow creek samples so that there are equal number of Debris flow creek samples and non-Debris flow-creek samples. The two types of samples are combined to form the training sample used in the analysis.

5. Susceptibility Analysis

Susceptibility Analysis uses Logistic Regression in Data Mining Prediction Algorithm. It was executed through data mining software Polyanalyst6.0. Through multivariate statistics, a set of linear combination functions composed of topographic factors and regression coefficients that represent the degree of contribution of each factor are obtained. Through induction, we obtained a discerning Debris flow susceptibility analytical empirical formula. The following summarizes the basis for Logistic Regression. Logistic Regression is a special form of Log-Linear Model. It is a function that can be applied to a binary variable as dependent variable (ie. Debris flow and non-Debris flow) and defines a series of independent variables (the Debris flow topographic factors in this study). Logistic Regression can be expressed as following (Hosmer and Lemeshow, 1989) :

$$P(y_i = 1 \mid x_i) = \frac{1}{1 + e^{-Z}} \tag{1}$$

$$Z = \alpha + \beta x_i \tag{2}$$

Z is a linear polynomial of probability factors that affect the occurrence of an event. Its range lies between negative infinity and positive infinity. Substituting the value of Z into equation (2), we can obtain the value for P, which lies between 0 and 1. This represents the possibility for an event to occur. x_i is an independent variable, α and β are regression intercept (constant) and regression coefficient respectively. In the analysis of the probability for debris flow to occur, x_i is the value for each debris flow impact factor. β is the weight of each factor. Substituting Z into (1), we can obtain the value for P, the Debris flow susceptibility value. Next, we follow the suggestion by Lillesand and Kiefer (2000) to use Classification Error Matrix to indicate the accuracy of pattern analysis. This study recommends classification accuracy greater than 70% as the outcome of the pattern analysis is better.

6. Statistical Tests

Individual parameter testing may follow recommendation from Hair (1998), using Cox & Snell R square value in model summary testing. The primary goal is to determine the significance of model variables. In other words, the higher the Cox & Snell R square value, the more the independent variables used in model can discern Debris flow set from non-Debris flow set, and the more accurate the model is.

In terms of overall model testing, indices are used to determine the Goodness of Fit of the overall model, and they are, respectively, $\chi 2$ value and Hosmer-Lemeshow test (Hosmer and Lemeshow, 1989). When $\chi 2$ is significant, it indicates that there needs at least one independent variable that can predict samples' probability value in the dependent variables and that value for P is not significant (P>0.05) for the Goodness of Fit of the overall model to be good.

Receiver Operator Characteristic curve (ROC curve) was raised by Swets in 1988. ROC curve primarily shows the accuracy of analysis model. The proportion of incorrect interpretations from Classification Error Matrix makes up the x-axis, and the proportion of correct interpretations makes up the y-axis. A curve is thus drawn. The higher the proportion of correct interpretations, the greater the area is under the curve (AUC), this indicates the analysis model has better classification accuracy, and AUC value falls between 0 and 1. This study recommends setting AUC to 0.7 as benchmark for judging analysis model. If AUC value is greater than 0.7, the requirement for model accuracy is met.

Table.1 Results of principal component analysis.

Principle Component	**Eigenvalue**	**Variance Percentage, %**	**Cumulative Variance,%**
First Principle Component	10.079	50.393	50.393
Second Principle Component	2.863	14.313	64.706
Third Principle Component	1.648	8.241	72.947
Fourth Principle Component	1.337	6.687	79.634
Fifth Principle Component	1.164	5.821	**85.455%>85%**

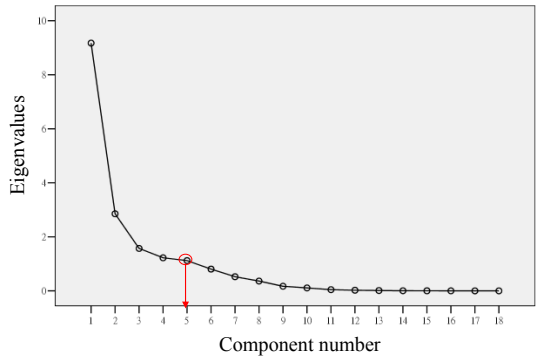


Table.2 Combination of topographic factors significantly List.

No.	**Debris Flow Susceptibility Factor Categories**	**Susceptibility Factor Types**	**Principle Component Type**	**Unit**
1	Catchment Factors of form	Catchment Area	Second Principle Component	km^2
		Form Factors	Second Principle Component	(Dimensionless)
		Circularity Ratio	Fifth Principle Component	(Dimensionless)
2	Catchment Slope Aspect Factors	Slope Aspect Standard Deviation	Third Principle Component	(Dimensionless)
		Average Slope Aspect	Fourth Principle Component	Degree
3	Catchment Height Factors	Hypsometric Integral	Fourth Principle Component	(Dimensionless)
		Height Coefficient of Variation	Fifth Principle Component	(Dimensionless)
4	Catchment Slope Grade Factors	Average Slope	Third Principle Component	Degree
		Average Slope for the First 5%	First Principle Component	Degree
5	Catchment Landslide characteristic factors	Total landslide Area Within 50 m of Each Side	First Principle Component	m^2

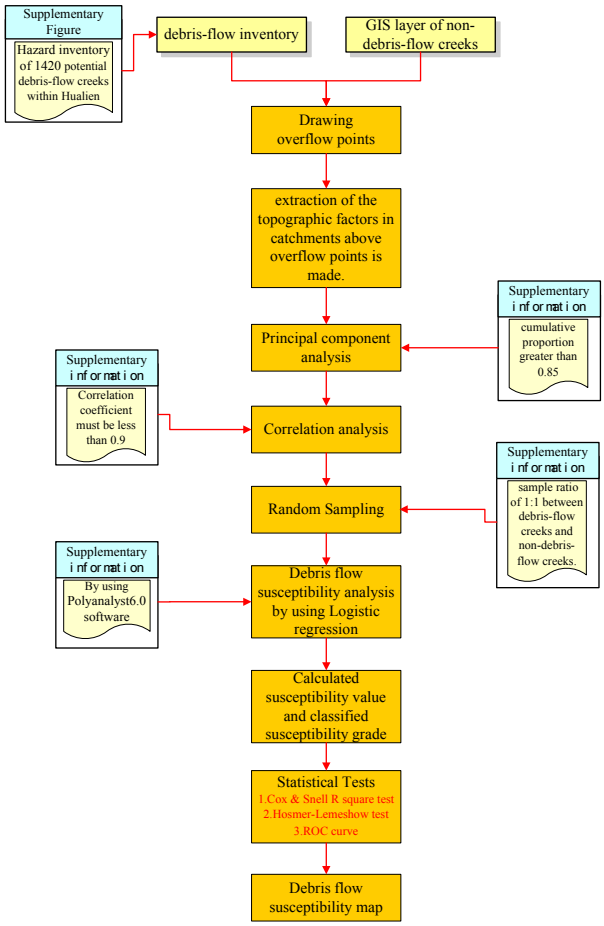
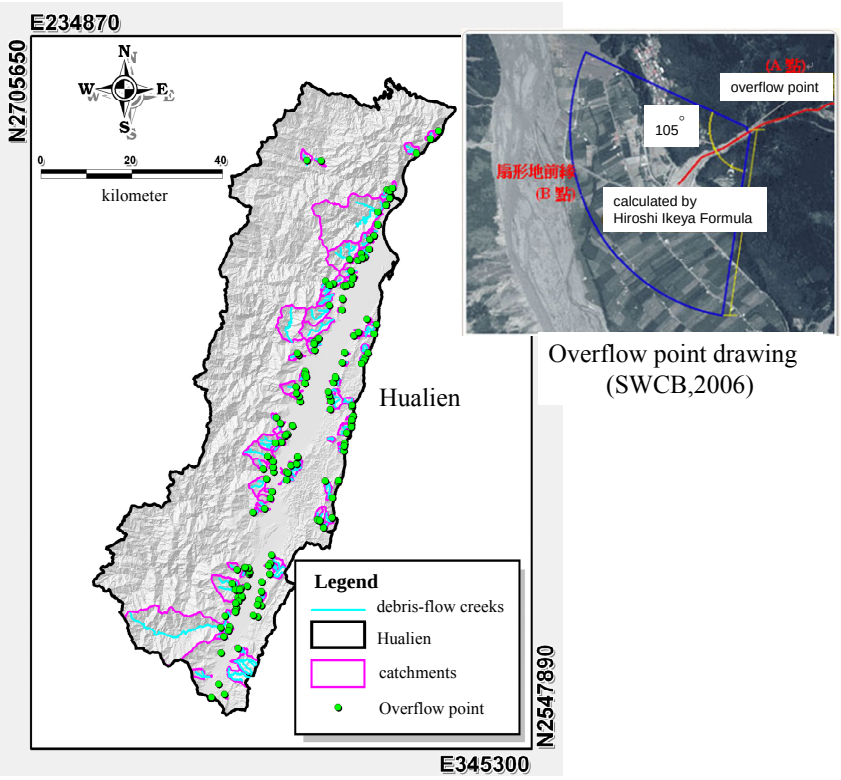
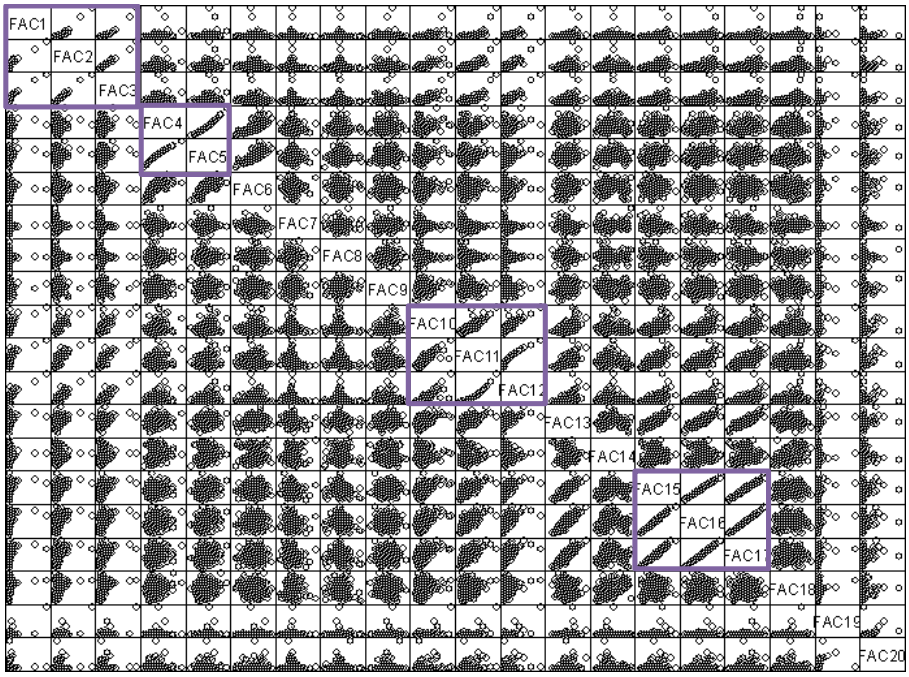


Fig.4 Flow chart of debris flow susceptibility analysis.



Overflow point drawing (SWCB,2006)

Fig.5 Debris flow catchment area and the overflow point distribution in Hualien County. (Sinotech，2008)



Note: catchment area (numbered FAC1), catchment length (numbered FAC2), catchment circumference (numbered FAC3), form factor (numbered FAC4), elongation ratio (numbered FAC5), circularity ratio (numbered FAC6), the average slope aspect X vector (numbered FAC7), average slope aspect Y vector (numbered FAC8), slope aspect standard deviation (numbered FAC9), average slope aspect (numbered FAC10), Hypsometric Integral (numbered FAC11), average elevation (numbered FAC12), elevation standard deviation (numbered FAC13), elevation coefficient of variation (numbered FAC14), average slope grade (numbered FAC15), first 5% average slope grade (numbered FAC16), first 10% average slope grade (numbered FAC 17), first 15% average slope grade (numbered FAC18), the total landslide area (numbered FAC19) and the rate of landslide (numbered FAC20) within 50 m of the two sides.

Fig.6 Correlation matrix of topographic factors in Hualien County.

SUSCEPTIBILITY ANALYSIS RESULTS

1. Logistic Regression Equation

The outcome of debris flow susceptibility analysis needs to be verified by Classification Error Matrix and ROC curve for classification accuracy. After the model significance is verified by statistical tests, a Debris flow susceptibility map can be drawn based on susceptibility classification criteria.

Through Logistic Regression, debris flow susceptibility model in Hualien County is obtained (the Logistic Regression Equation), as shown in equation (3). Substituting equation (3) in Logistic Regression Equation (2), each of the debris flow susceptibility value can be obtained. The values are between 0 and 1. Greater the susceptibility values means greater probability of debris flow occurring. They are relative values.

$$\lambda = 0.002X_1 - 5.829X_2 + 8.941X_3 + 12.286X_4 + 0.000X_5 - 0.357X_6 - 0.016X_7 + 0.114X_8 - 0.020X_9 + 0.000X_{10} - 0.406 \quad (3)$$

X_1 is the catchment area; X_2 is the form factor; X_3 is the circularity ratio; X_4 is Hypsometric Integral; X_5 is height coefficient of variation; X_6 is the average slope gradient; X_7 is the average slope aspect; X_8 is the average slope aspect for the first 5 percent; X_9 is the standard deviation for slope aspect; X_{10} is the landslide area within 50 m of the two sides.

2. Susceptibility Classification Method

Using 0.5 as Logistic Regression index *P* to divide the two groups, this study recommends classifying those susceptibility values greater than 1.5 times that of Logistic index (in other words, susceptibility value≧ 0.75) as high susceptibility class, classifying susceptibility values between Logistic index and 1.5 times that of Logistic index as intermediate-high susceptibility class (in other words, 0.5≦susceptibility value<0.75), classifying susceptibility values between Logistic index and half of Logistic index as intermediate susceptibility class (in other words, 0.25≦susceptibility value<0.5), and classifying susceptibility values less than half of Logistic index as low susceptibility class (susceptibility value<0.25).

3. Pattern Analysis Accuracy

Classification Error Matrix is made from the result of susceptibility analysis, indicating the accuracy of pattern analysis. The result is shown in table 3, and the Receiver Operator Characteristic curve is shown in figure 7. After comparison with landslide inventory, the result indicates that the overall accuracy of the analysis result for Debris flow susceptibility in Hualien County is close to 80%. Receiver Operator Characteristic curve also meets the significance requirement (AUC>0.7).

4. Results of Statistical Tests

The result from the above analysis then undergoes Homer and Lemeshow test as well as Cox&Snell R square test. Hosmer and Lemeshow test indicates that p value is greater than the significance level of 0.05. This test assumes if p value is greater than the significance level 0.05, then the analysis model is significant. The greater the Cox&Snell R square value is, the better the "Goodness of fit" will be. In this model, the correlation is the medium relationship of Cox&Snell R square test (Hair, 1998). The statistical testing results are compiled in table 4. The results indicate that the statistical tests passed the requirement, and a susceptibility map can be drawn based on the susceptibility classification, as shown in figure 8.

Table.3 Classification error matrix results.

Accuracy Types	Hualien County
Non-Debris flow Set Accuracy (%)	81.5
Debris flow Set Accuracy (%)	77.8
Overall Accuracy (%)	79.6

Table.4 Logistic regression model statistical test results.

Test Types / Area	Hosmer and Lemeshow Test			Model Summary
Hualien County	Chi-square test	Degree of Freedom	p Value	Cox & Snell R square
	9.22	8	0.102＞0.05:significant	0.359:medium relationship

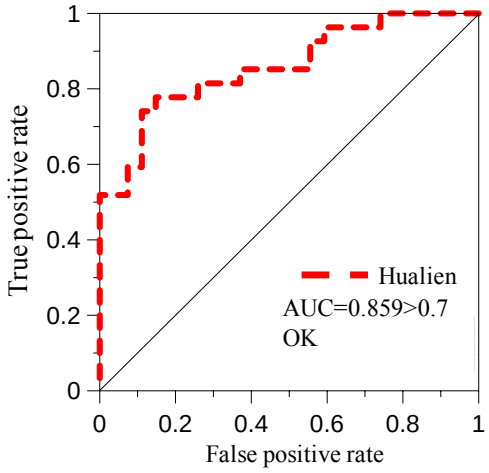


Fig.7 Receiver Operator Characteristic curve of debris flow susceptibility in Hualien County.

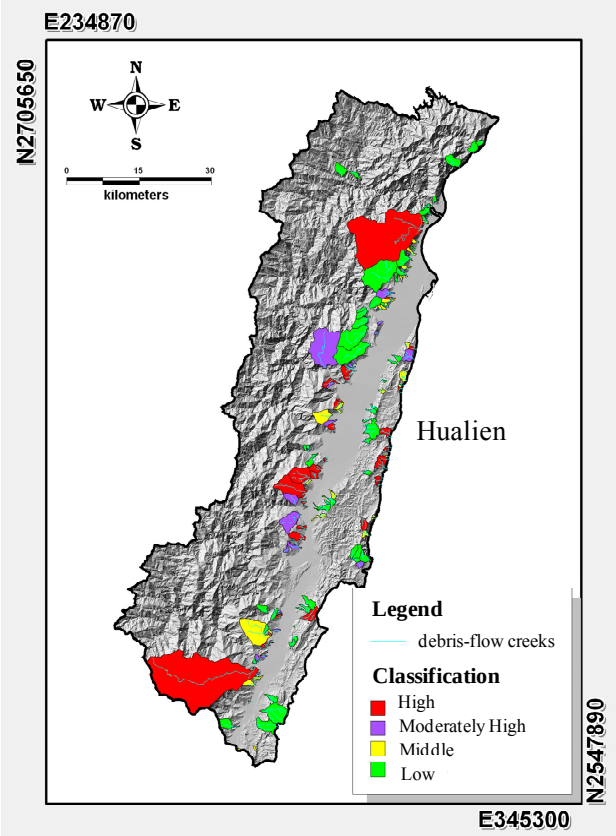


Fig.8 Debris flow susceptibility map of Hualien County.

CONCLUSIONS

This study raises a susceptibility analysis flow chart that is based on statistical theories and is available for future debris flow related studies.

The overall accuracy for debris flow susceptibility analysis in Hualien County exceeds 80%. Moreover, the statistical tests meet the requirement for a significant pattern. These indicate that the analysis model is fairly reliable.

From the debris flow susceptibility map, the probability of landslides happening in the potential debris flow creek catchments is obtained. The result is available for

classification management and disaster prevention.

REFERENCES

Chen, C.Y., Chen, T.C., Yu, F.C., and Hung, F.Y. (2004): A Landslide Dam Breach Induced Debris Flow – A Case Study on Downstream Hazard Areas Delineation, Environmental Geology, Vol.47: pp.91-101.

Fisher, R. A. (1928): The general sampling distribution of the multiple correlation coefficient, *Proceedings of the Royal Society of London*, Ser. A, 121, pp.654-673.

Hair, J.F. et al. (1998): Multivariate Data Analysis, 5th edition, Prentice Hall.

Hosmer, D. W and Lemeshow, S. (1989): Applied Logistic Regression, John Wiley & Sons, Inc.

Horton, R.E. (1932): Drainage Basin Characteristics, Trans. Amer. Geophys. Union, 13, pp.350-361.

Lee, C. T., Huang, C. C., Lee, J. F., Pan, K. L., Lin, M. L., Dong, J. J. (2008): Statistical approach to earthquake-induced landslide susceptibility, *Engineering Geology*, Vol.100 (1-2), pp.43-58.

Lillesand, T. M. and Kiefer R. W. (2000): Remote Sensing and Image Interpretation, Fourth Edition, John Wiley & Sons, Inc., pp.750.

Lin, M.L., Wang, K.L., and Huang, J.J. (2005): Debris Flow Run Off Simulation and Verification – Case Study of Chen-You-Lan Watershed, Taiwan, Natural Hazards and Earth System Sciences, Vol.5, pp.439-445.

Swets, J.A. (1988): Measuring the Accuracy of Diagonstic Systems", Science, Vol. 204, No. 4857, pp.1285-1293.

Sinotech Engineering Consultants, INC.(2006): The investigation of vulnerability factors in debris flow areas and it's counter measurements", Report to the Soil and Water Conservation Bureau, Council of Agriculture, Executive Yuan. (in Chinese)

Sinotech Engineering Consultants, INC. (2007): The investigation of vulnerability factors of debris flows torrents and it's counter measurements", Report to the Soil and Water Conservation Bureau, Council of Agriculture, Executive Yuan. (in Chinese)

Sinotech Engineering Consultants, INC. (2008): The Investigation of Vulnerability Factors and Risk Analysis, Risk Management of Debris Flows, Report to the Soil and Water Conservation Bureau, Council of Agriculture, Executive Yuan. (in Chinese)

Sinotech Engineering Consultants, INC. (2009): The Investigation of Vulnerability Factors and Risk Analysis, Risk Management of Debris Flows, Report to the Soil and Water Conservation Bureau, Council of Agriculture, Executive Yuan. (in Chinese)

Stone, M. and Brooks, R. J.(1990): Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares, and Principal Components Regression, *Journal of Royal Statistical Society*, Vol.52, No.2, pp.237-269.