Statistic GIS Modeling

Lee, S. and Pradhan, P. 2007.

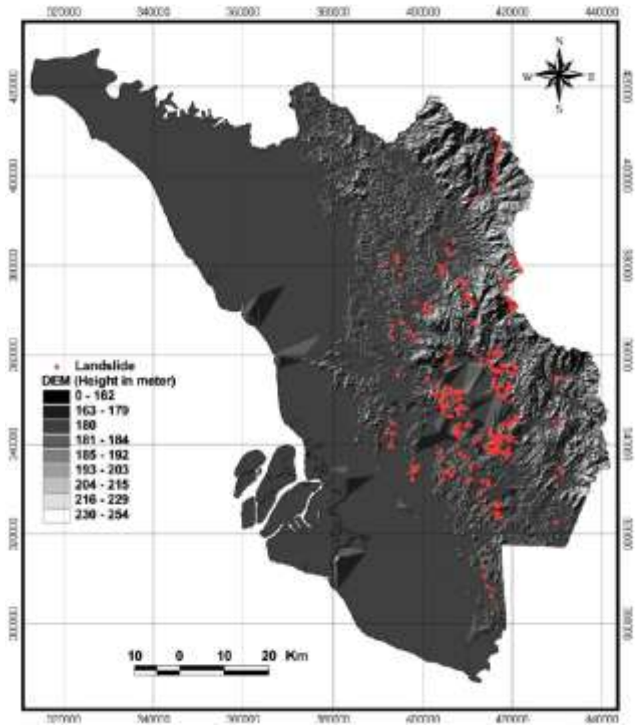
Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models.

Landslides, 4: 33–41.

Introduction

- Evaluate landslide hazards (landslide hazard mapping)
- Landslide risk factors:
  - Slope, aspect, curvature
  - Distance from drainage
  - Geology
  - Distance from lineament (fault lines)
  - Soil
  - Land cover
  - Precipitation

Selangor, Malaysia



Data

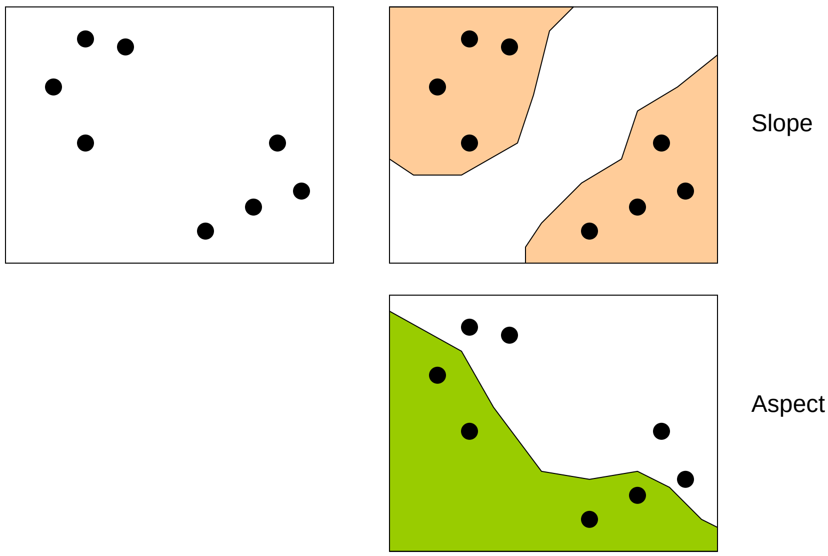
- Landslide locations (polygons)
  - Aerial photographs interpretation + field survey
- 10 m DEM (grid)
  - Topo map contours + survey point data
- 30 m Landsat Thematic Mapper (TM) image
  - Geological lineament, land cover, vegetation index (NDVI)
- 1 m distance surfaces (grid)
  - Distance from drainage and lineament
- Soil map (polygons)
- 100 m precipitation data (grid)

- All data layers were resampled to 10 m resolution

Methods

- Frequency Ratio
- Logistic regression
- Assumption
  - Observed actual landslides represent the potential for slope stability

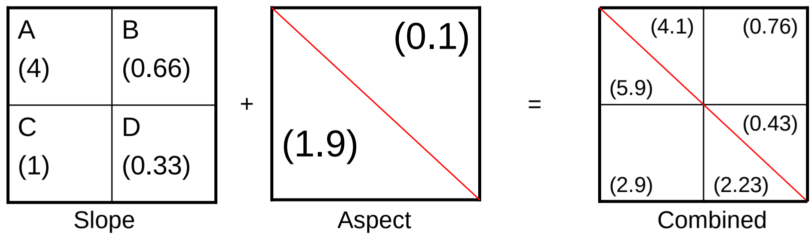
Environmental Determinants of Spatial Phenomenon



Frequency Ratio Model

Land Type	% Area (PA)	% Occur Freq (POF)	Freq Ratio (= POF / PA)	Remarks
TYPE A	10	40	4	Association
TYPE B	30	20	0.66	Avoidance
TYPE C	30	30	1	Average
TYPE D	30	10	0.33	Avoidance
Sum	100%	100%		

Combining FR of multiple factors



Regression Analysis

- Tells us the directions and magnitudes of the effects of independent variables on dependent variables and the interactions between independent variables.

- Example: used car value model

Value($) = 16000 – 1000 x age(year) – 0.15 x miles

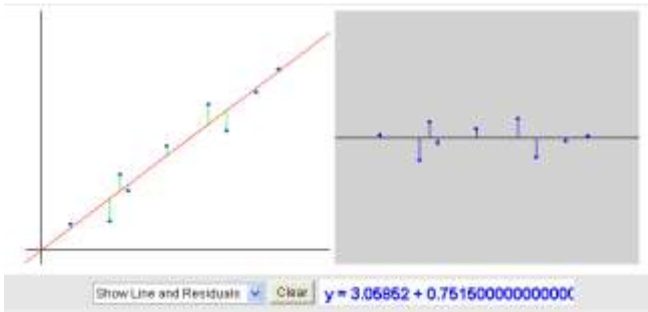
	Age	**Mileage**	**$**
Car 1	5	10000 miles	
Car 2	10	20000 miles	

Regression Analysis

- A power tool for discerning the relations between **dependent** (response) variables (Y) and **independent** (explanatory) variables (Xs). ε is the error term (or **residual**).

$Y = a + bX + \varepsilon$ (bivariate regression)

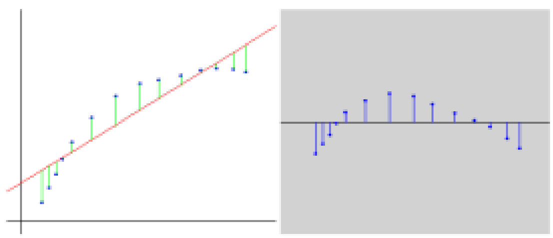
$Y = a + b_1X_1 + b_2X_2 + \ldots + \varepsilon$ (multivariate regression)



http://www.math.csusb.edu/faculty/stanton/m262/regress/regress.html

Assumptions of OLS Regression Analysis

- Lack of measurement error
- Linearity (if linear regression is used)
- Normality
- No multicollinearity
- Homoscedasticity (zero means and equal variances in residuals)
- No autocorrelation of the residuals



Logistic Regression Model

- A multivariate regression model with a binary dependent and/or independent variables:

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

$$z = a + b_1X_1 + b_2X_2 + \ldots + \varepsilon$$

$$e^{-z}: 1 / 2.71828182845904^{z}$$

$$\begin{aligned} z_p &= (0.0780 \times Slope) + Aspect_c + (-0.0032 \times Curvature) \\ &\quad + (-0.0048 \times Drainage) + Lithology_c \\ &\quad + (0.0001 \times Lineament) + (-1.3633 \times NDVI) \\ &\quad + Landcover_c + (0.0043 \times Precipitation) - 16.4726 \end{aligned}$$

Examples: $p = 1/(1+e^{-z})$

$$z = a + b_1X_1 + b_2X_2 + \ldots + \varepsilon$$

$$e^{-z}: 1 / 2.71828182845904^{z}$$

Z	P
10	0.999955
3	0.952574
1	0.731059
0	0.5
-1	0.268941
-3	0.047426
-10	0.000045

Assumptions of Logistic Regression Analysis

- Lack of measurement error
- Linearity
- Normality
- No multicollinearity
- Homoscedasticity (zero means and equal variances in residuals)
- No autocorrelation of the residuals

Results

Table 2 Coefficient values for frequency ratio and logistic regression in the case of each factor

Factor	Class	Number of pixels showing landslide occurrence	Percentage of pixels showing Landslide occurrence	Pixels in domain	Pixel %	Frequency ratio	Coefficients of logistic regression
Slope	0~15 degree	67,777,334	82.86	115	35.17	0.42	0.0780
	16~25 degree	8,426,979	10.30	87	26.61	2.58	
	26~35 degree	4,648,328	5.68	68	20.80	3.66	
	35~85 degree	940,212	1.15	57	17.43	15.16	
Aspect	Flat	32,746,440	40.04	0	0.00	0.00	−9.1703
	North	5,541,254	6.77	55	16.82	2.48	−1.9392
	Northeast	6,678,670	8.17	45	13.76	1.69	−1.3081
	East	6,077,711	7.43	27	8.26	1.11	−1.2635
	Southeast	6,181,783	7.56	28	8.56	1.13	−1.9959
	South	5,564,596	6.80	29	8.87	1.30	−1.1984
	Southwest	6,721,149	8.22	35	10.70	1.30	−2.0135
	West	6,105,258	7.46	46	14.07	1.88	−2.0177
	Northwest	6,175,992	7.55	62	18.96	2.51	0.0000
Curvature	Concave	13,288,765	16.25	66	20.18	1.24	−0.0032
	Flat	55,283,859	67.59	15	4.59	0.07	
	Convex	13,220,229	16.16	246	75.23	4.65	

Model Comparison

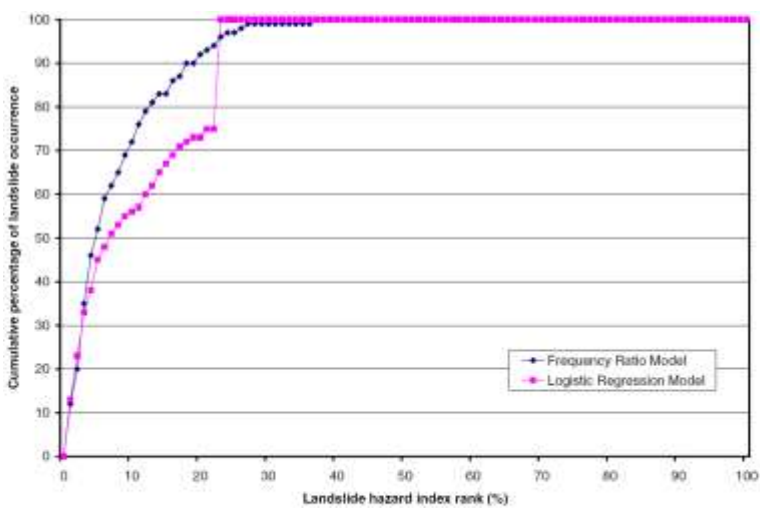
- Risk Scores of observed landslide sites

Model	1	2	3	4	5	6	7	8	9	10
A	100	100	90	80	80	70	60	40	20	10
B	100	100	100	100	100	50	40	20	20	10

- Which model is better? A or B?

Model	>= 90	>= 80	>= 70	>= 60	>= 50	>= 40	>= 30	>= 20	>= 10	Sum
A	30%	50%	60%	70%	70%	80%	80%	90%	100%	630
B	50%	50%	50%	50%	60%	70%	70%	90%	100%	590

Verification



/ Landslide probability

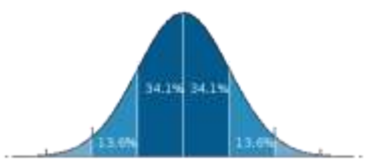
Comments

- Factors weights of Frequency Ratio Model?
- Spatial autocorrelation & statistic models
  - Select uncorrelated samples
  - Spatial Regression
    - Spatial expansion method (location as an additional independent variable)
    - Geographically weighted regression (location as weights of independent variables)

Normalization of data (statistical normalization)

- Converting to Z scores (i.e., expressing data in the unit of standard deviation)

$$z = \frac{x - u}{\sigma}$$



x is a raw score to be standardized

μ is the mean of the population

σ is the standard deviation of the population.