# MULTIVARIATE REGRESSION ANALYSIS FOR LANDSLIDE HAZARD ZONATION

CHANG-JO F. CHUNG
*Geological Survey of Canada, Ottawa*

ANDREA G. FABBRI and CEES J. VAN WESTEN
*International Institute for Aerospace Survey and Earth Sciences, Enschede*

## ABSTRACT

Based on several layers of spatial map patterns, multivariate regression methods have been developed for the construction of landslide hazard maps. The method proposed in this paper assumes that future landslides can be predicted by the statistical relationships established between the past landslides and the spatial data set of map patterns. The application of multivariate regression techniques for delineating landslide hazard areas runs into two critical problems using GIS (geographic information systems): (i) the need to handle thematic data; and (ii) the sample unit for the observations. To overcome the first problem related to the thematic data, favourability function approaches or dummy variable techniques can be used.

This paper deals with the second problem related to the sample units. In this situation, the unique condition subareas are defined where each subarea contains a unique combination of the map patterns. Weighted least squares techniques are proposed for the zonation of landslide hazard using those unique condition subareas. The traditional pixel-based multivariate regression model becomes a special case of the proposed weighted regression model based on the unique condition subareas. This model can be directly applied to vector-based GIS data without the need of rasterization.

A case study from a region in central Colombia is used to illustrate the methodologies discussed in this paper. To evaluate the results adequately, it was pretended that the time of the study was the year 1960 and that all the spatial data available in 1960 were compiled including the distribution of the past landslides occurred prior to that year. The statistical analyses performed are based on these pre-1960 data about rapid debris avalanches. The prediction was then compared with the distribution of the landslides which occurred during the period 1960-1980.

*A. Carrara and F. Guzzetti (eds.), Geographical Information Systems in Assessing Natural Hazards,* 107–133.

## 1. Introduction

Landslides are natural geologic processes that contribute significantly to shape the landscape of the Earth. Landslides become hazardous processes when they interfere with human activities. Often human activities such as deforestation and urban expansion, in fact, accelerate the process of landslides. This problem is especially serious in developing countries where environmental protection and management are harder to sustain. For example, over 95% of all disaster and fatalities related to landslides occur in developing countries (Hansen, 1984), and up to 0.5% of gross national product of these countries have been lost by landslides (Fournier D'Albe, 1976). In 1990, the General Assembly of the United Nations declared the decade 1990-2000 as the International Decade for Natural Disaster Reduction. Annual economic losses due to landslides are estimated to be in the order of two to five billions US dollars (Schuster, 1994).

Landslides are usually triggered by events such as extreme rainfall, earthquakes, volcanic eruptions, and land-use-changes. The prediction of future landslides in landslides-prone areas is an important aspect for future land use planning. Landslide hazard zonation aims at delineating potential areas for the occurrences of future landslides by using geoscience data such as soil types, slope angle and other geomorphologic features in the area of study. In this paper we discuss a regression model for landslide hazard zonation under the following two assumptions: (i) that the characteristics of the past landslides in the study area can be described by the input spatial geoscience data; and (ii) that the future landslides will occur under similar conditions in which the past landslides took place.

The use of multivariate regression models for landslide hazard zonation and prediction has mainly been developed in Italy by Carrara (1983, 1988) and his colleagues (Carrara *et al.*, 1992). In their earlier applications (Carrara 1983; 1988), a square grid is first overlaid on a study area, and the size of grid was determined so that the total number of grid cells was reasonably small, generally not more than a few thousands. The square cell becomes the sample unit for the regression analysis. Therefore, for each square cell, one observation is made for each layer. Similar cell-based multivariate regression models were developed for the prediction of mineral potential areas based on several layers of map data (Chung, 1978; Chung and Agterberg, 1980; and Chung, 1983).

If the sample unit covers a large area (e.g., large rectangular cell of 100 m x 100 m or larger), it becomes a difficult task to properly represent a large area by one observation for each layer, because the observation of the cell may not properly represent geomorphologic meaning of the cell. A way to avoid the difficulty is to make the size of the grid spacing small (e.g., sample units of 12.5 m x 12.5 m was used in the Colombian example) so that the each grid cell covers a small area on the ground. As a size of the grid spacing gets smaller, it gets easier for one observation to represent the cell, but the number of grid cells becomes larger. Often, the number of the sample units becomes too large for regression analysis.

To shun the difficulty, Carrara and his collaborators (Carrara *et al.*, 1991) have proposed the use of morphometric units as sample units. The advantage of using morphometric units is that they can be delineated automatically on the basis of a digital elevation model (DEM). The mean size of those morphometric units can also be adjusted to the average size of the landslides occurring in a study area. The main disadvantage of the morphometric units is that the overall conditions may be very heterogeneous within such units. The morphometric units are also irregularly shaped and sized. With the statistical techniques used by Carrara *et al.* (1992) it is possible to say whether a unit as such is stable or not. Within a unit which may still be in the order of hectars, however, no differentiation in hazard can be made.

We deal with this problem in this contribution by proposing a new weighted regression model based unique condition subareas. As in Appendix A, it can be shown that the traditional cell-based multivariate regression model becomes a special case of the proposed weighted regression model based on the unique condition subareas. This unique condition subarea-based model can be directly applied to vector-based GIS data without the need of rasterization.

Another difficulty in the application of regression analysis to that kind of spatial input is the handling of thematic data such as digitized maps representing bedrock lithologies and land uses. Traditionally, this problem was unraveled by generating a series of dummy binary variables to represent the presence or absence of the various map units in cells (Chung, 1983; Chung and Agterberg, 1980; and Carrara, 1983, 1988). Instead of the dummy variable approach, we can make use of the favourability functions proposed by Chung and Fabbri (1993). The use of the favourability function approaches for the zonation of the landslide hazard are shown in Chung and Leclerc (1994).

Wang and Unwin (1992) have proposed the log-linear model for the prediction of landslide hazard. To apply the log-linear models (Christensen, 1990) proposed by Wang and Unwin (1992), it must be assumed that the cell (or something similar to that) is a sample unit. Let us consider the m+1 dimensional contingency table containing all the possible combinations of the thematic classifications of the m input layers and the landslide layer. At each slot in the table containing a combination, we count the number of cells which have that combination of thematic classes. In the example used in Wang and Unwin (1992), only three layers were used. Two layers consisted of three classes each and one layer consisted of only two classes. Their three-dimensional 3 x 3 x 2 contingency table contained 18 slots (two slots contained no observations (see Tab. 3 of Wang and Unwin, 1992). The data from Colombia used in this paper contain eight input layers including the distribution map of the past landslides. The eight-dimensional table contains 388,800 slots and each slot represents one of 388,800 possible combinations of thematic classes. In the table, however, only 4728 (the same as the number of unique condition subareas assuming that 12.5 m x 12.5 m cell is used) contain one or more cells, and the remaining 384,072 slots contain no cells. Because of the large number of empty slots, the log-linear model based on the m+1 dimensional contingency table can not be utilized for analysis of these types of spatial data.

## 2. Study area

The catchment of the Rio Chinchina, located on the western slope of the central Andean mountain range (Cordillera Central) in Colombia, near the Nevado del Ruiz Volcano was used as a test area for various landslide hazard zonation techniques (van Westen *et al.*, 1993). A part of this catchment with an area of 68 km$^2$ was used as the training area for the application of different quantitative techniques. Fig. 1 illustrates the distribution of landslides occurred prior to 1960, whereas the distribution of landslides occurred between 1960 and 1980 is shown in Fig. 2.

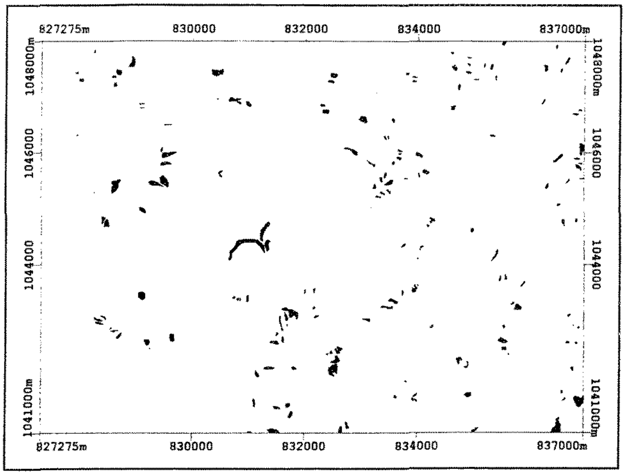


Figure 1. Distribution of pre-1960 rapid debris avalanches (termed *derrumbes* in Spanish).

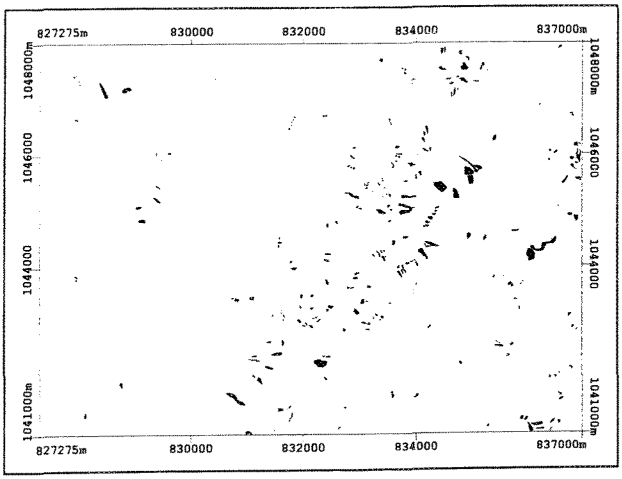


Figure 2. Distribution of 1960-1980 rapid debris avalanches (termed *derrumbes* in Spanish).

The severity of natural hazard in the study area combined with extensive industrial and agricultural activity and a high population density, has caused considerable damage and loss of lifes in the past. The area is susceptible to mass movements, earthquakes, and volcanic hazards. The geology of the study area has been reviewed by van Westen (1993).

Most of the lithological units have a marked north-south directionality, which is related to the Romeral fault pattern. Paleozoic schists, quartzites, and marbles were uplifted during the Upper Palaeozoicum as result of Late Hercynian orogenesis. These rock are intruded by a number of intermediate and acidic batholites, stocks and dikes of Jurassic and Cretaceous age. Tectonic uplift of the area started near the end of the Cretaceous and continued throughout the Tertiary into the Quaternary. The major tectonic uplift, in which the two fault systems of Romeral and Palestina played an important role, took place during the Late Cretaceous and Early Tertiary.

Most of the rocks experienced intensive metamorphism, and intrusives related to the Romeral fault zone occurred locally. Practically all rocks have faulted contacts. The later stages of tectonic uplift were accompanied by important volcanic activity so that flows unconformably overlie the Paleozoic rocks. During Late Miocene and Early Pliocene large volume of sediments related to volcanic activity were deposited throughout the area. Most of these materials were later removed by subsequent erosion. Volcanic activity continued during Pleistocene, with the formation of lava flows, now mostly restricted to presently existing valleys. Below the maximum limits of the lava flows the valleys were filled with debris flows and pyroclastic flows. During periods of glaciation the increased ice volume in higher parts of the area generated large debris flows. Another very important effect of Pleistocene and Olocene volcanism is the deposition of a thick blanket of ash over the terrain. The ash sequences vary in thickness and composition, depending on the distance from the volcanoes and the amount of erosion since deposition. These ash deposits are of great importance in the occurrence of mass movements. Contacts between the relatively permeable ashes and the underlying, less-permeable, weathering soils often serve as the failure surface for landslides.

The study area is located in a zone of important seismic activity, in which earthquakes with magnitude 6 or larger on the Richter scale have occurred with an approximate return period of 15 years. At various locations displacement in ash sequences was observed, indicating Quaternary fault activity. The relationship between the faults and mass movements is, however, more due to their width and amount of milonitization and deformation than due to their seismic activity, In the main Romeral fault an area up to 500 m wide is affected.

Geomorphologically, the region represents a typical Andean environment: an active mountain chain in the wet equatorial zone, characterized by deep weathering, strong Pliocene uplift and associated deep fluvial incision, mass movement problems, and active volcanism at higher elevations interfering with Pleistocene glaciation. With the exception of the terraces, the geomorphological zones in the area are oriented north-south as a consequence of the tectonic framework, lithology and altitude. Five general geomorpho-logical zones can be distinguished:

1. Rounded hills between the Cauca and the Romeral fault zone. Mass movements are not abundant in this region. Fossilized landslides are found along terrace edges. Active landslides occur on the terrace slopes.
2. Romeral fault zone. The area between Chinchina and Manizales is characterized by a number of fault-related noth/south-oriented valleys and ridges with steep slopes. The relationship between the faults and the drainage pattern is very clear: the Rio Chinchina changes its course four times between Manizales and La Manuela, making turns of 90 degrees. Landslide problems can be severe, expecially in fault zones, where the bedrock material is highly deformed. The most common type of mass movement is soil slip or soil avalanche.
3. Dissected Tertiary planation surface. The area between the Romeral zone and west of the Tertiary lava deposits is characterized by remnants of a Tertiary planation surface of Late Eocene-Early Oligocene age. The area has an almost continuous cover of ash, with uniform sequences of silty sand and lapilli, and is characterized by the occurrence of large flow slides, which have shown little differentiation in size or activity since the 1940s. The steep slopes of the major valleys, and the fault scarps, have by far the highest frequency of active surficial landslides.
4. Volcanic complex. The highest part of the study area consists of a series of lava flow levels, among which the upper part has been shaped by glacial erosion. From an altitude of 2300 m to the maximum glacial limit, the terrain is characterized by very steep slopes, covered by original Andean forest, with a high density of surficial debris avalanches. On these slopes gulley erosion and solifluction are the most common denudational processes.
5. Terraces. A large number of different terrace levels can be observed throughout the area. They are quite homogeneous in composition and may occur at different levels due to differential uplift. Most terraces are composed of debris flow material and alluvial material. They differ with respect to the degree of weathering and ash cover.

After the 1985 eruption of the Nevado del Ruiz a great amount of research was done to determine the volcanic history of the area, as a basis for a better volcanic hazard map (van Westen, 1993, p. 41). On the basis of studies of stratigraphic columns and radiometric dating a total of 24 important eruptions have been identified for the last 6247 years. The following hazards are associated with volcanic eruptions: lava flows, pyroclastic flows, lateral blasts, pyroclastic falls and lahars (debris flow of pyroclastic material, incorporating rock fragments, eroded alluvial deposits, trees, ice, and water, triggered by a volcanic eruption).

The geological map and the slope map of the study area are shown in Fig. 3 and Fig. 4, respectively.

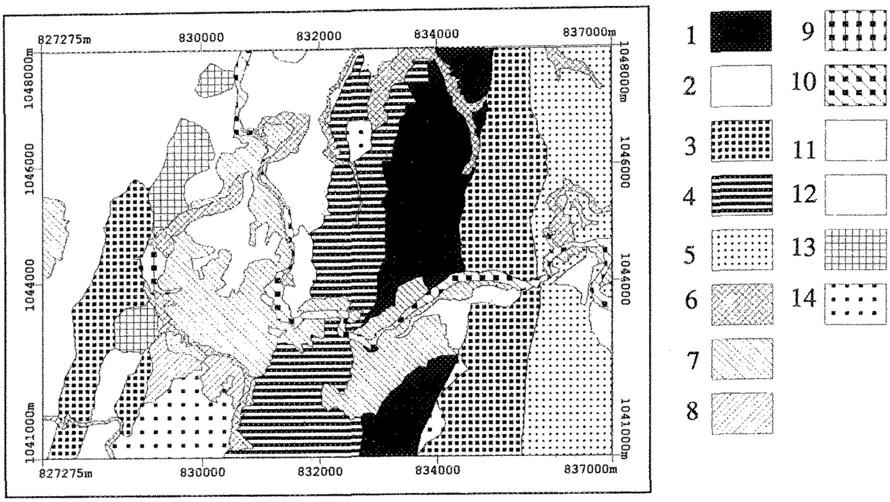


Figure 3. Geological map of the Colombian landslide hazard study area. Legend: 1) Gneiss. 2) Schists. 3) Volcanic. 4) Gabbro. 5) Metasediment. 6) Alluvial. 7) Mix Debris. 8) Weathered Debris. 9) Lake Deposit. 10) Lahar Deposit. 11) Flow Deposit. 12) Pyroclastic Flow. 13) Andesitic. 14) Tertiary Sediments.

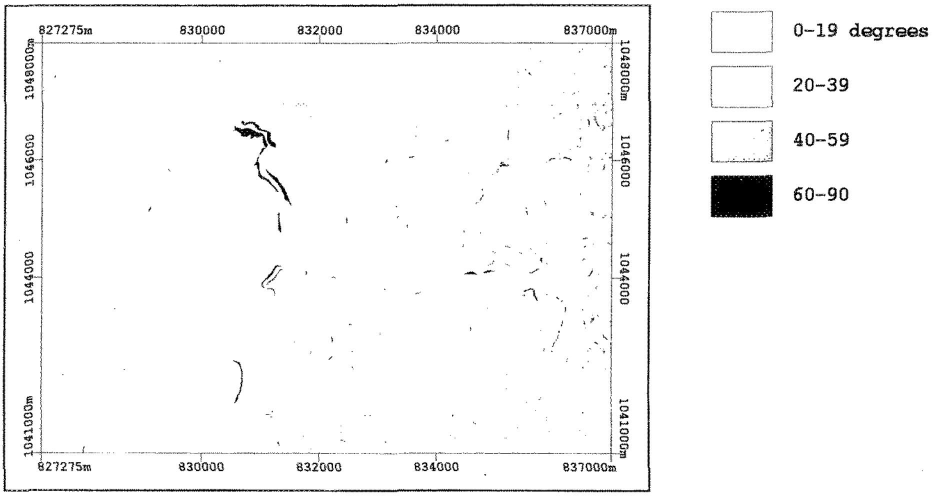


Figure 4. Slope map of the Colombian landslide hazard study area.

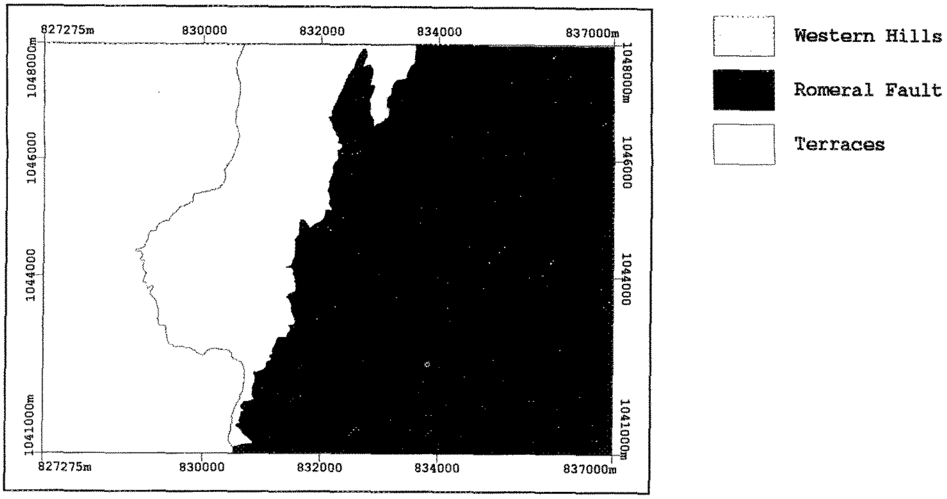


Figure 5. Geomorphological map of the Colombian landslide hazard study area.

From West to East three main terrain complexes as shown in Fig. 5, can be differentiated in this test area: (i) *the western hills*, with gentle slope, mainly underlain by schists, with a relatively thick cover of volcanic ashes, and with rotational slides as the most predominant mass movement feature; (ii) *debris flow terraces*, located in a graben structure in the central part of the area; and (iii) *Romeral fault zone*, with steep slopes in metamorphic rocks and shallow ash cover, located within one of the major active fault-zones of Colombia, characterized by the frequent occurrences of rapid debris avalanches. In addition to these three maps, four layers were used for the analysis. All seven input layers and their corresponding classes are contained in Tab. 1.

## 3. Representation of data

The input data for landslide hazard zonation usually consists of several layers of spatial information digitized from maps. The types of the spatial input layers considered here are illustrated in Tab. 1. Some layers represent continuous measurements such as slope angles and distances, while other layers represent non-scaled thematic data such as bedrock lithology units and land use classes. We will refer non-scaled thematic data as "thematic classification" data where an observed categorical pixel value does not have numeric meaning, but it only represents a thematic class.

Table 1. Data layers and classes used for analysis.

| *MAP* | *CLASSES* |
|---|---|
| Geology | • gneissic intrusive • schists • volcanic and metasedimentary<br>• gabbro and diorite • alluvial sediments<br>• flow materials and alluvial and ashes<br>• weathered debris flow materials • lake deposits • lahar deposits<br>• pyroclastic flow deposits<br>• mix of pyroclastic and debris flow<br>• andesitic intrusive • tertiary sediments |
| Geomorphology Complexes | • Western hills • terrace<br>• Romeral fault zone |
| Slope Intervals | • 0-9° • 10-19° • 20-29° • 30-39° • 40-49° • 50-59° • 60-69° • 70-79° • 80-90° |
| Landuse | • traditional farming system • technified farming system<br>• modern intermediate farming system • other crops • construction<br>• bare • grass • shrubs • forest |
| Distance to Roads | • < 25 meters • 25-50 m • > 50 m |
| Distance to Valley Heads | • < 25 meters • 25-50 m • > 50 m |
| Distance to Faults | • < 50 meters • 50-99 m • 100-149 m • 150-199 m • 200-249 m<br>• > 250 m |

The map information captured in digital form are stored in either raster or vector format (Aronoff, 1989; Chung and Fabbri, 1993). We are here assuming that the digital data are stored according to the raster model, because it is easier (i) to illustrate the methodology proposed; and (ii) to compare the techniques with the traditional cell-based regression technique. As in traditional cell-based regression procedure, raster data are obtained by overlaying a square grid over each map, although the spacing of the grid tends to be small such as 10 m or 30 m. Each cell is now called a pixel (picture element) and each map is represented by a rectangular matrix of numbers in which each number indicates the class membership of a pixel that is in one-to-one correspondence with a small area on the map. In the study area in Colombia, each map consists of 779 x 561 number matrix, each pixel representing a 12.5 m x 12.5 m area. The value of the pixel in a layer (e.g., the slope image) represents the slope angle and the pixel value in another layer (e.g., the geology image) for the corresponding small area on the ground, indicates the class of the bedrock lithology.

Even for a continuous measurement such as slope angle, it is necessary, in practice, to quantize the data, usually dividing the angles into a number of classes instead of the actually observed slope angle for the pixel. For example, we divided the slope angle in the Colombian study into the following 10 classes: class #0: no observation is available or unmapped area; class #1: $0 \leq$ slope $< 10$; class #2: $10 \leq$ slope $< 20$; $\cdots$ ; class #9: $80 \leq$ slope $\leq 90$. Figs. 1, 2 and 3 show three of the seven maps used as input in this study. The value k, ranging from 0 to 9, is used to represent the pixels which belongs to class #k as shown in Fig. 4 in a simplified version. In this example, we convert a map layer containing a continuous measurement into the ten-class thematic map. As it was done for the Colombian study, it can be assumed that every layer represents a thematic map which contains a discrete number of thematic classes.

## 4. Preparation of the data for statistical analysis

When the data are stored in the raster format, a pixel is a natural choice of the sample unit for the statistical analysis. The study area in Colombia consists of 779 x 561 (= 437,019) pixels and each pixel represents 12.5m x 12.5m square area in the ground. However, if a smaller pixel of the size 6.25m x 6.25m was used for the Colombia study area, 1,748,076 (= 437,019 x 4) pixels were needed to cover the area. Let us suppose that we have a digitized map containing the distribution of the past landslides, in addition to m input thematic maps in a study area. The data base consists of m input layers and n pixels to cover the entire study area. At the i-th pixel, we have, for $i = 1, \cdots, n$,

$$(Y_i ; X_{i1}, \cdots, X_{im}), \tag{4.1}$$

where $Y_i$ denotes the presence or the absence of the landslide at the i-th pixel, and $X_{i1}, \ldots, X_{im}$ represent m input layers at the i-th pixel.

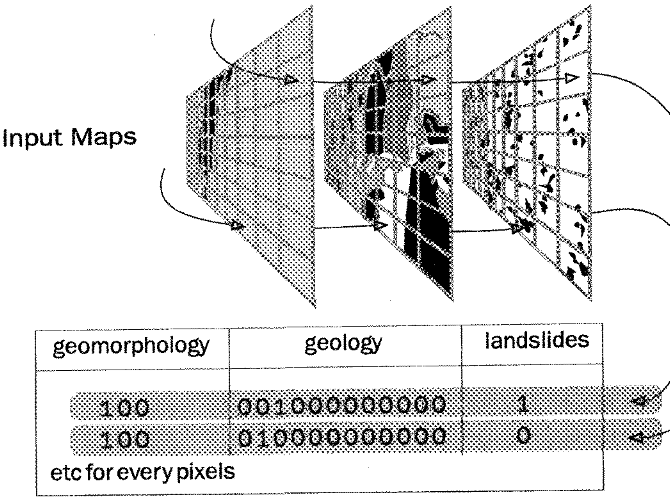
Because $X_{ij}$ represents a thematic class of the i-th pixel in the j-th layer, the numeric value $X_{ij}$ can not be used directly in regression analysis. To avoid the difficulty of the thematic data, a commonly used technique is to generate a binary variable (called dummy variable) for each thematic class (Chung and Agterberg, 1980; Carrara, 1988) to indicate the presence or absence of that class at each pixel. The use of the dummy variable model here is identical to the linear model approach in the Analysis of Variances (Searle, 1971). Suppose that we have $h_j$ thematic classes in the j-th layer ($j = 1, \cdots, m$). Then, at the i-th pixel, instead of one observation $X_{ij}$, we generate $h_j$ binary variables, $B_{ij1}, \cdots, B_{ijh_j}$, where one of the $B_{ij1}, \cdots, B_{ijh_j}$ is equal to 1 and all the others are equal to 0. Fig. 6 illustrates graphically how the dummy binary variables were generated for a regression model using two layers, geomorphological map (Fig. 5) and geological map (Fig. 3) and the distribution map (Fig. 1) of the pre-1960 landslides in the Colombian data.

Assuming that all m layers are representing thematic classes, instead of expression (4.1), we have, at each pixel i, for $i = 1, \cdots, n$,

$$(Y_i ; B_{i11}, \cdots, B_{i1h_1}, \cdots, B_{ij1}, \cdots, B_{ijh_j}, \ldots, B_{im1}, \cdots, B_{imh_m}). \tag{4.2}$$

In the Colombian data set, 48 binary "dummy" variables were obtained, although only eight layers, including the distribution of the past landslides, were introduced. For 12.5

m x 12.5 m pixel size, the expression (4.2) generates the matrix of the size, 437,019 x 48 which was subjected to multivariate regression analysis. For 6.25m x 6.25m pixels, 1,748,076 x 48 matrix was required for the statistical analysis. In practice, however, it is difficult to use such a large matrix for the analysis, i.e., it is impractical to use the pixel as a sample unit for multivariate analysis because the number of pixels is usually too large.



| geomorphology | geology | landslides |
|---|---|---|
| 100 | 001000000000 | 1 |
| 100 | 010000000000 | 0 |
| etc for every pixels | | |

Figure 6. Graphical illustration of generating binary dummy variables.

Consider all the pixels for which ($B_{i11}$, $\cdots$ , $B_{i1h_1}$, $\cdots$ , $B_{ij1}$, $\cdots$ , $B_{ijh_j}$, $\cdots$ , $B_{im1}$, $\cdots$ , $B_{imh_m}$) and/or ($X_{i1}$, $\cdots$ , $X_{im}$) are identical. Such sets of all the pixels where the observed values of the m input layers are identical are termed the "unique condition subareas". The whole map area can be subdivided into a small number (small relatively to the total number of pixels) of unique condition subareas which has a unique combination of ($X_{i1}$, $\cdots$ , $X_{im}$) and/or ($B_{i11}$, $\cdots$ , $B_{i1h_1}$, $\cdots$ , $B_{ij1}$, $\cdots$ , $B_{ijh_j}$, $\cdots$ , $B_{im1}$, $\cdots$ , $B_{imh_m}$).

Given an input map, the number of unique condition subareas is relatively independent of the pixel size. While the number of pixels increases exponentially with the increase in resolution, given the same input map, the increase in the number of unique condition subareas will simply reflect the presence of the finer details not representable at a given resolution. In fact, the pixel-based or vector-based representaion is completely irrelevant to divide the area into a set of unique condition subareas.

In the Colombian data set, only 4728 unique condition subareas were obtained, in contrast to the 437,019 pixels covering the study area. Some of the 4728 unique condition subareas were very small and consisted of only 1 pixel, whereas some of the large subareas contained more than 8000 pixels. In particular, 2280 only among the 4728 unique condition subareas contained more than 10 pixels.

Suppose that we have q unique condition subareas, and consider a t-th unique condition subarea. Let $n_t$ be the total number of the pixels in the t-th subarea and let $k_t$ ($\leq n_t$) be the number of pixels effected by the past landslides. Then at the t-th unique condition subarea, we have, for t = 1, $\cdots$, q,

$$(k_t, n_t, B_{t11}, \cdots, B_{t1h_1}, \cdots, B_{tj1}, \cdots, B_{tjh_j}, \cdots, B_{tm1}, \cdots, B_{tmh_m}). \tag{4.3}$$

In the Colombian data, it was this matrix of size 4728 x 49 that was subjected to a new procedure of weighted regression models.

## 5. Quantitative prediction model

The first step toward the construction of a quantitative prediction model is to determine the sample unit where the observations of the s input layers and of the landslides are made. The estimation of the unknown parameters depends on the sample unit. Two sample units, the pixel and the unique condition subarea were introduced in the previous section.

### 5.1 ANALYTIC (NOT STATISTICAL) MODEL

The next step for a quantitative prediction model is to postulate the landslides as a function of the s input layers, $(L; Z_1, \cdots, Z_s)$:

$$L = f(d_1, \cdots, d_k; Z_1, \cdots, Z_s), \tag{5.1}$$

where L represents the occurrence of the landslides, the function $f(.)$ specifies a quantitative equation as a prediction function of L. $f(.)$ usually contains several unknown parameters, $d_1, \cdots, d_k$, and these parameters are estimated from the input data such that the estimated parameters satisfy certain analytically "optimal" properties.

For example, suppose that we have the observations for n sample units,

$$(L_i; Z_{i1}, \cdots, Z_{is}) \text{ for } i = 1, \cdots, n, \tag{5.2}$$

where $L_i$ represents the occurrences of the landslides in the i-th unit, and $Z_{ij}$ indicates the observation of the i-th unit in the j-th layer. Then the equation (5.1) is rewritten, for each i-th unit, as,

$$L_i = f(d_1, \cdots, d_k; Z_{i1}, \cdots, Z_{is}) \text{ for } i = 1, \cdots, n. \tag{5.3}$$

We may obtain the estimators, $\bar{d}_1, \cdots, \bar{d}_k$ for $d_1, \cdots, d_k$ such that $\sum_{i=1}^{n} | \bar{L}_i - L_i |$ is the *minimum* where $\bar{L}_i = f(\bar{d}_1, \cdots, \bar{d}_k ; Z_{i1}, \cdots, Z_{is})$ for i = 1, ⋯, n, and the estimators, $\bar{d}_1, \cdots, \bar{d}_k$ are termed $L_1$-estimators (Rao, 1977). Or we may obtain the estimators, $\tilde{d}_1, \cdots, \tilde{d}_k$, for $d_1, \cdots, d_k$ such that $\sum_{i=1}^{n} (\tilde{L}_i - L_i)^2$ is the minimum where $\tilde{L}_i = f(\tilde{d}_1, \cdots, \tilde{d}_k ; Z_{i1}, \cdots, Z_{is})$ for i = 1, ⋯ , n, and the $\tilde{d}_1, \cdots, \tilde{d}_k$ are termed as $L_2$-estimators.

The idea is to obtain the estimators, $\bar{d}_1, \cdots, \bar{d}_k$ ; or $\tilde{d}_1, \cdots, \tilde{d}_k$ for $d_1, \cdots, d_k$ such that they minimize the differences between the estimator $\bar{L}_i$ or $\tilde{L}_i$ and the unknown true model $L_i$ for all i = 1, ⋯ , n.

The advantages of the analytic model are that: (1) no statistical assumption related to the occurrences of the landslides as random variables with the respective distribution functions, is needed; and (2) the interpretation of the estimators is simple (all we are trying to do is to find $\bar{d}_1, \cdots, \bar{d}_k$ or $\tilde{d}_1, \cdots, \tilde{d}_k$ such that the differences between $\bar{L}_i$ or $\tilde{L}_i$ and $L_i$ are as small as possible for all i = 1, ⋯ , n). The disadvantages of the model are, however, that: (3) there is no way to test whether the estimators are "good"; and (4) no inference is possible. The estimators can not be used outside the input layers ($L_i$ ; $Z_{i1}, \cdots, Z_{is}$) for i =1, ⋯ , n.

## 5.2. STATISTICAL MODEL

The first step to construct a statistical prediction model is to assume that the occurrences of landslides are the random variables and the "expected" landslides (Roussas, 1973) can be postulated as a function of the s input layers, (L; $Z_1, \cdots, Z_s$). The model then can be written as:

$$L = f(d_1, \cdots, d_k; Z_1, \cdots, Z_s) + \varepsilon, \tag{5.4}$$

where L represents a random variable for the occurrences of the landslides, $f(.)$ specifies a quantitative equation, $d_1, \cdots, d_k$ are the unknown parameters, and ε is an error random variable with the "expected" value E(ε) = 0. The unknown parameters, $d_1, \cdots, d_k$, are estimated from the input data such that the estimated parameters satisfy certain statistical "optimal" properties.

For example, suppose that we have the observations, ($L_i$; $Z_{i1}$, ... , $Z_{is}$) for each of the n samples for i = 1, ⋯ , n where $L_i$ represents the random variable for the occurrences of the landslides in the i-th unit, and $Z_{ij}$ indicates the observation of the i-th unit in the j-th layer. Then the equation (3.3) is rewritten, for each i-th unit, as,

$$L_i = f(d_1, \cdots, d_k; Z_{i1}, \cdots, Z_{is}) + \varepsilon_i \text{ for } i = 1, \cdots, n. \qquad (5.5)$$

In this example as it was done in the analytic model, we may obtain the estimators $\overline{d}_1$, ⋯, $\overline{d}_k$ for $d_1$, ⋯ , $d_k$ such that $\sum_{i=1}^{n} E(\overline{L}_i - L_i)^2$ is the minimum and $E(\overline{L}_i) = L_i$ for i = 1, ⋯ , n, where "E(.)" is the "expected value" (Roussas, 1973) and $\overline{L}_i = f(\overline{d}_1, \cdots, \overline{d}_k : Z_{i1}, \cdots, Z_{is})$ for i = 1, ⋯ , n, and the estimators, $\overline{d}_1$, ⋯ , $\overline{d}_k$ are termed the mean squares estimators (Rao, 1973).

The idea is to obtain the estimators, $\overline{d}_1$, ⋯ , $\overline{d}_k$ for $d_1$, ⋯ , $d_k$ such that they minimize the "expected" differences between the estimator $\overline{L}_i$ and the unknown true model $L_i$ for all i = 1, ⋯ , n.

The advantages of the statistical model are that: (1) the estimators can be tested; and (2) statistical inference is possible. The estimators can be used outside the sample units ($L_i$ ; $Z_{i1}$, ⋯ , $Z_{im}$) for i = 1, ⋯ , n . The disadvantages of the model are, however, that: (3) statistical assumptions related to the occurrences of the landslides as random variables with the relative distribution functions are needed; and (4) statistical and physical interpretations of the estimators and optimality properties are difficult to obtain.

## 5.3 PREDICTION MODEL

When we find such "optimal" estimators, $\overline{d}_1$, ⋯ , $\overline{d}_k$ , regardless of whether we deal with an analytic or statistical model, for any given values of the s input layers at a unit, ($Z_{o1}$, ⋯ , $Z_{os}$), the prediction for $L_o$ is given by:

$$\overline{L}_o = f(\overline{d}_1, \cdots, \overline{d}_k ; Z_{o1}, \cdots, Z_{os}). \qquad (5.6)$$

It is also important to note that even under identical models and from identical input data, the estimators for the parameters can be drastically different depending upon how the "optimal" properties are defined. Under certain conditions, the computational procedures and numerical results may be identical for the analytic and statistical models, but the interpretations of the results will be very different.

## 6. Multivariate regression analysis

Consider the landslides and the s input parameters,($L$ ; $Z_1, \cdots , Z_s$) as discussed in (5.1) and (5.3). As argumented in the previous Section 5, regression models can be interpreted either as analytic or statistical models, but we will deal with regression analysis as a statistical model only. We specify a linear function:

$$L = d_0 + d_1 Z_1 + \cdots + d_s Z_s + \varepsilon , \tag{6.1}$$

where ( $d_0$ , $d_1$ , $\cdots$ , $d_s$ ) are s+1 unknown parameters to be estimated from the input data (Draper and Smith, 1981). Several techniques related to the linear model in (6.1) were applied to the Colombian data set.

### 6.1 MODEL 1 - STANDARD MODEL

Consider a 437,019 x 48 matrix as in (4.2). At each pixel i, we have,

$$(Y_i ; B_{i11}, \cdots , B_{i1h_1}, \cdots , B_{ij1}, \cdots , B_{ijh_j}, \cdots , B_{im1}, \cdots , B_{imh_m}), \tag{6.2}$$

where $Y_i$ represents the presence (1) or the absence (0) of the past landslides at the i-th pixel and $B_{ijk}$ represents the presence (1) or the absence (0) of the k-th thematic class of the j-th layer at the i-th pixel. All seven layers and all the corresponding thematic classes are listed in Tab. 1.

In each layer, one variable which had the least correlation with the past landslides was excluded from the model to avoid the perfect collinearity among the dummy binary variables. The excluded seven classes were: (i) "mixed old debris" class for the Lithology layer; (ii) class "2" representing 10-20 degrees for the Slope-Angle layer; (iii) "Romeral Fault" class for the Geomorphology layer; (iv) "technical coffee growing" class for the Landuses layer; (v) class "3" representing "Distance farther than 50 m from the Roads" for the Road layer; (vi) class "3" representing "Distance farther than 50 m from Valley Heads" for the Valley Heads layer; and (vii) class "5" representing "Distance farther than 250 m from the Faults" for the Faults layer. In addition "unmapped area" in the Geomorphology layer is excluded because the identical class is also shown in the Lithology layer. The exclusion of the classes did not mean that the evidence that they provided was not being used in the model: it simply meant that collinearity of data was avoided.

After excluding these eight variables, we had a 437,019 x 40 matrix,

$$(Y_i\,;B_{i1},\cdots,B_{ir}), \tag{6.3}$$

where we assumed that the $B_{ik}$'s are reindexed for notational simplicity. Among the 40 variables, the first variable for the occurrences of the landslides was used as the "dependent" or "response" variable and the remaining 39 variables were "independent" or "predictor" variables in the linear regression model in (6.1).

The linear model postulated for each i-th pixel, was

$$Y_i = d_0 + d_1 B_{i1} + \cdots + d_r B_{ir} + \varepsilon_i\,. \tag{6.4}$$

Using the data in (6.3) we obtained the LS estimators, $\bar{d}_1, \cdots, \bar{d}_k$ shown in Tab. 2 for the unknown parameters $d_0, \cdots, d_r$. Under certain conditions (Draper and Smith, 1981) including that the variances of $\varepsilon_i$ are all equal:

$$\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(Y_i) = C \text{ for } i = 1, 2, \cdots, n, \tag{6.5}$$

where C is a constant, and the LS estimators have statistical "optimal" properties. The condition of the equal variances of $\varepsilon_i$'s (the equal variances of $Y_i$'s) implies that the "reliabilities" of $Y_i$'s are identical or each sample unit provides an identical amount of information to estimate the unknown parameters. The predicted values for the probability of the occurrences of the landslides at each pixel using the estimators shown in Tab. 2 are shown in Fig. 7a.

Note that it was pretended that the time of the study was the year 1960 and all the spatial data used here were pre-1960. The estimators in Tab. 2 are based on these pre-1960 data. Because all the variables in the model are binary variables representing presences or absences of the corresponding variables, the interpretation of the regression coefficients in Tab. 2 is relatively simple. For example, the negative coefficients indicate that the presence of the corresponding variables is related to safe areas, while the presence of the variables with the positive coefficients implies the possible areas for landslide hazards.

All the variables with the estimated coefficients near zero have very little effect on the prediction of possible areas of landslide hazards. Among all the variables, the binary variable representing the class of Slope: 70-79 (estimated coefficient is 0.07945) is the most effected single indicator for the landslide hazards, while the variable representing the lithological unit, volcanic and metasedimentary (estimated coefficient is -0.02018), is the best single indicator for the least dangerous area for the landslide hazards. Contrary to the usual notion, the areas with Slope: 80-90 (estimated coefficient is -0.00344) appear safe areas for landslide hazards, because such areas perhaps do not contain any debris or soils except for bedrocks. On other hand, the flat areas with Slope: 0-9 (estimated coefficient is 0.00831) have a positive coefficient indicating possible areas for landslide hazards. Obviously such estimated coefficients are not properly interpretable and it is one of the deficiencies of the regression models.

Table 2. Regression coefficients estimated for the models.

| *ESTIMATOR* | *COEFFICENT* | *ESTIMATOR* | *COEFFICENT* |
|---|---|---|---|
| Constant term | 1.55155E-02 | Slope: 60-69 ° | 1.56512E-02 |
| gneissic intrusive | 4.00725E-03 | Slope: 70-79 ° | 7.94524E-02 |
| schists | -6.44134E-03 | Slope: 80-90 ° | -3.43960E-03 |
| volcanic and metasedimentary | -2.01823E-02 | traditional farming | -1.43843E-02 |
| gabbro and diorite | -1.60107E-02 | modern intermediate farming | -9.42330E-04 |
| alluvial sediments | -4.54183E-03 | other crops | -2.65025E-04 |
| flow materials, alluvial and ashes | -8.95750E-03 | construction | 5.72468E-03 |
| weathered debris flow materials | -6.89581E-03 | bare | -1.26819E-02 |
| lake deposits | -3.37607E-03 | grass | 3.12856E-03 |
| lahar deposits | -9.03179E-03 | shrubs | -3.36146E-03 |
| pyroclastic flow deposits | -8.10535E-03 | forest | 9.35692E-05 |
| andesitic intrusive | -5.77758E-03 | Roads: < 25 m | 2.96225E-03 |
| tertiary sediments | 7.61305E-03 | Roads: 25-50 m | 2.20617E-02 |
| Geomorp.: Western hills | 2.22365E-02 | Valley: < 25 m | 1.04442E-02 |
| Geomorp.: terrace | -1.81741E-02 | Valley: 25-50 m | 3.82563E-04 |
| Slope: 0-9 ° | 8.30965E-03 | Fault: <50 m | -4.16950E-04 |
| Slope: 20-29 ° | -7.28955E-03 | Fault: 50-99 m | 2.89060E-03 |
| Slope: 30-39 ° | -7.06489E-03 | Fault: 100-149 m | 1.83681E-03 |
| Slope: 40-49 ° | 9.68498E-03 | Fault: 150-199 m | 8.44423E-04 |
| Slope: 50-59 ° | 1.61332E-02 | Fault: 200-249 m | 4.30893E-03 |

Although the model contains such flaws, the prediction map not only covers the past landslides before 1960 well, but also it adequately predicts the 1960-1980 landslides, as shown in the last column "Pixel-Based & 1/n weight" of Tab. 3. In Tab. 3, the whole study area is divided into five classes depending on the regression scores. Five percents the pixels of the highest scores is classified as "very high", and the 5, 10 and 25 percents of the pixels of the next highest scores are classified as "high", "medium" and "low", repectively. For example, from Tab. 3, the areas predicted as "medium" to "very high" occupies 25% of the whole area, but the predicted areas contain 51.5% of all the landslides occurred during the next 20 years, 1960-1980.

Table 3. Success rate of 1960-1980 landslides predictions using the model proposed with three different weights. Fig. 8 shows this table in graphic form.

| *HAZARD* | *CUMMULATIVE AREA (%)* | *CUMMULATIVE MAPPED 1960-80 LANDSLIDES (%)* | |
|---|---|---|---|
| | | No Weight | Pixel-Based & 1/n weight |
| very high | 5 | 9.2 | 13.2 |
| high | 10 | 21.5 | 25.8 |
| medium | 25 | 45.5 | 51.5 |
| low | 50 | 76.0 | 82.5 |
| very low | 100 | 100.0 | 100.0 |

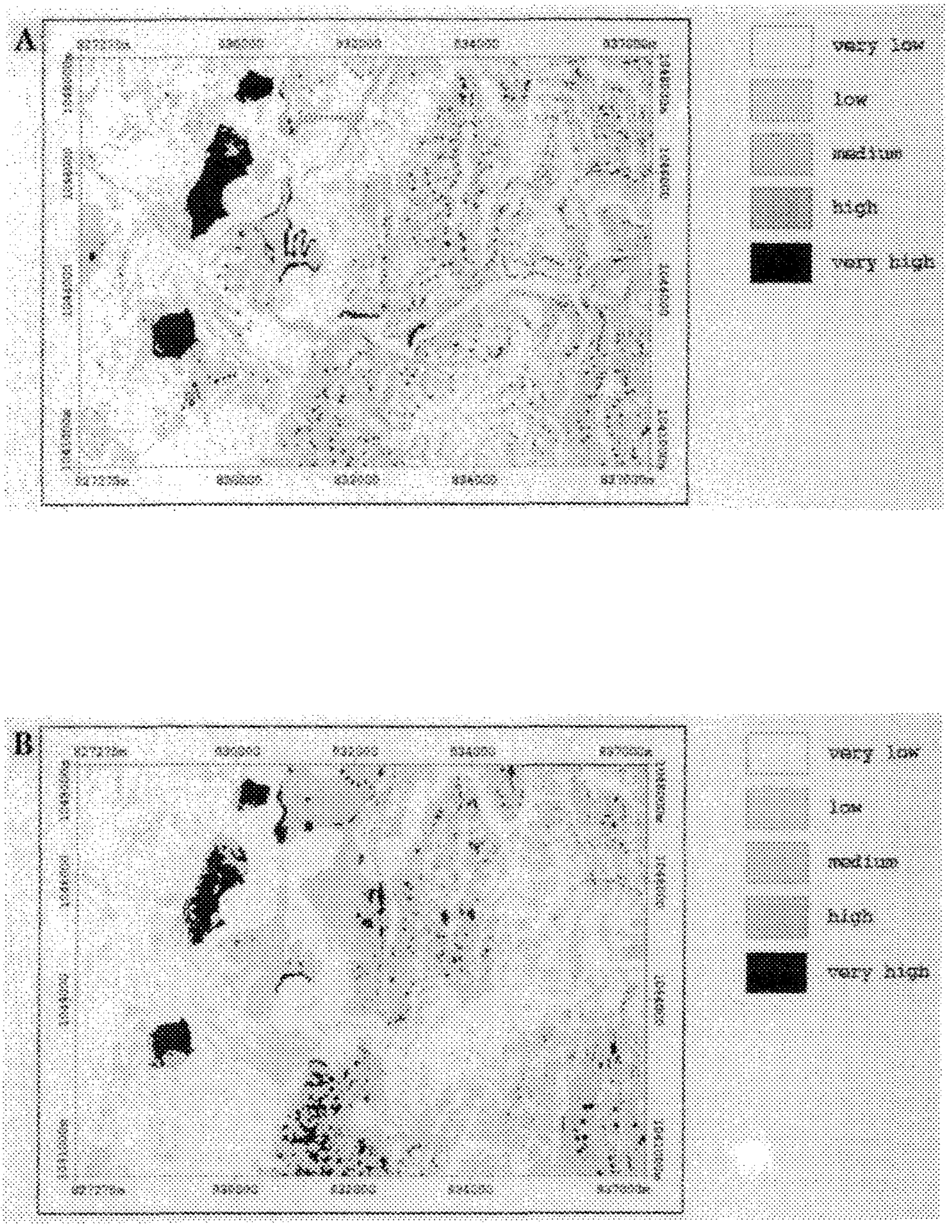


Figure 7. Predition map of weighted regression model based on 4728x40 matrix in (6.7) using (a) $1/n_u$ as a weight; (b) no weight. The prediction map shown in (a) is identical to the output from the pixel-based regression model in (6.4) using 437,019x40 matrix in (6.3).

The main flaw of the technique, however, is the requirement of the rasterization of the data before the analysis. This method is also computationally heavy, in addition to the requirement of a large storage disk space (over 40 MB for 12.5m x 12.5m pixels). If the pixels of the size 6.25m x 6.25m were used instead, 1,748,076 x 48 matrix were subjected to the analysis. In order to overcome this difficulty, we are proposing the following weighted least squares model based on the unique condition subareas which was discussed in (4.2) and the results from the proposed model are identical to this model.

## 6.2 MODEL 2 - UNIQUE CONDITION SUBAREA MODEL

### *Model 2.1 - OLS (Ordinary Least Squares) model*

Instead of considering a 437,019 x 48 matrix as it was done in (6.2), let us consider the 4728 x 49 matrix shown in (4.2) based on the unique condition subareas,

$$( k_u, n_u, B_{u11}, \cdots, B_{u1h_1}, \cdots, B_{uj1}, \cdots, B_{ujh_j}, \cdots, B_{um1}, \cdots, B_{umh_m}). \qquad (6.6)$$

As carried out in Model 1, we exclude the eight collineated variables from (6.6) and reindex the remaining variables. We have the 4728 x 41 matrix,

$$( k_u, n_u, B_{u1}, \cdots, B_{ur}). \qquad (6.7)$$

Among the 41 variables at each unique condition subarea, the ratio, denoted by R, of the first two variables (the number of pixels affected by the past landslides divided by the total number of pixels in the unique condition subarea) is used as the "dependent" or "response" variable. As in Model 1, the remaining 39 variables are used as "independent" or "predictor" variables in the regression analysis.

The linear model postulated is: for each u-th unique condition subarea,

$$R_u = \beta_0 + \beta_1 B_{u1} + \cdots + \beta_r B_{ur} + \varepsilon_u . \qquad (6.8)$$

Using the data in (6.7) we obtain the LS estimators, $\bar{\beta}_0, \cdots, \bar{\beta}_r$ for the unknown parameters $\beta_0, \cdots, \beta_r$. As discussed in (6.5), the LS estimators have an "optimal" property under the assumption that

$$\mathrm{Var}(\varepsilon_u) = C \text{ for } u=1,2, \cdots, q, \qquad (6.9)$$

where C is a constant and the predicted values for the probability of the occurrences of the landslides at each pixel are shown in Fig. 7. Although it predicts the past landslides before 1960 well as shown Tab. 3, it does poorly for the 1960-1980 landslides. The prediction pattern shown in Fig. 7b is different from the pattern in Fig. 7a obtained from Model 1. Fig. 8 illustrates the prediction performances between these two models.

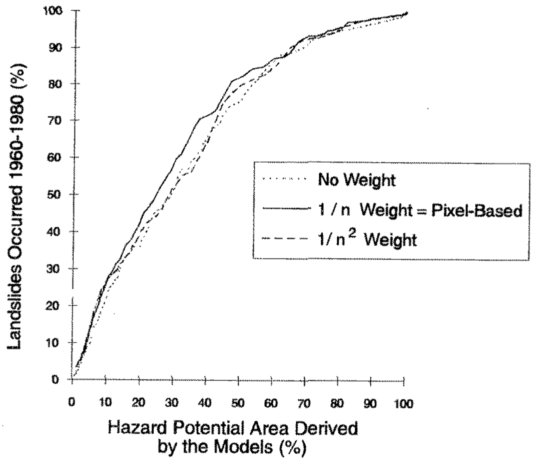


Figure 8. Prediction performance of weighted regression models. The prediction was built using pre-1960 data and it was compared with the landslides between 1960-1980. Tab. 3 contains the numbers used for this graph.

*Model 2.2 - WLS (Weighted Least Squares) model*

For the LS estimators of the Model in (6.8) under the assumption of (6.9), we have that the condition (6.9) implies that every unique condition subarea has the same weight or significance regardless of the sizes of the subareas. In other words, a unique condition subarea provides an equal significance to determine the LS estimators based on (6.9) whether it contains one pixel only or 8000 pixels. It is one of the reasons why the LS estimators based on (6.9) are not as effective as the estimators from (6.5).

The LS estimators from (6.5) for the linear model in (6.4) are obtained using 437,019 pixel data in (6.3). There each pixel had the same significance to compute the LS estimators. When we use the unique condition subarea data in (6.7) to estimate the unknown parameters, $\beta_0$, $\cdots$, $\beta_r$ in (6.8), each sample unit, which is a unique condition subarea, should not be treated as having an equal significance, because it does not provide an equal information or “reliabiliity”. The larger is the size of the subarea, the "more likely reliable" it is or/and “more information" it provides than the smaller ones. For example, let us compare two unique condition subareas where one contains only one pixel and the other contains 8000 pixels. Suppose that both subareas are completely covered by the past landslides. Of course, both unique conditions appear to be hazardous environments for the landslides. We should consider the unique condition of the subarea containing 8000 pixels, however, more serious than that of the one pixel subarea. It implies that the data from the 8000 pixel unique condition subarea is "more reliable" or provides "more information" than that from the one pixel subarea.

It also suggests that the condition in (6.9) should be modified to:

$$\mathrm{Var}(\varepsilon_u) = g(n_u) \text{ for } u=1,2, \ldots , q, \tag{6.10}$$

where $n_u$ is the total number of pixels in the unique condition subarea u discussed in (6.7) and g(.) is a function. Under the condition in (6.10) which states that the variance is a function of the total number of pixels in the unique condition subarea, the LS estimators do not have the "optimal" properties.

Under (6.10), the "optimal" estimators for $\beta_0, \cdots , \beta_r$ in (6.7) are the weighted least squares (WLS) estimators. The WLS estimators depend on the function g(.) in (6.10). When $g(n_u) = C$ for all u's, then the WLS estimators are identical to the LS estimators under the condition in (6.9).

As discussed in Appendix A, it can be shown that if we set the weighting function, $g(n_u) = {}^{1}\!/_{n_u}$ for all u's, then the WLS estimators are identical to the LS estimators from (6.7) based on all the n pixels. In other words, the LS estimators from the linear model in (6.4) based on the 437,019 x 40 matrix in (6.3) are identical to the WLS estimators from the linear model in (6.8) based on the 4728 x 41 matrix in (6.7) using the weights of the subareas, $g(n_u) = {}^{1}\!/_{n_u}$. Therefore the corresponding predicted values for the probabilities of the occurrences of the landslides at each pixel are identical for both these two models. Figure 3.a shows the prediction pattern. If we were to use the vector format, rather than the raster format, to capture the digital data for the spatial information and we obtained the unique condition subareas through vector operators, we would use the weighting function $g(n_u) = {}^{1}\!/_{w_u}$ where $w_u$ is the size of the u-th unique condition subarea, instead of $g(n_u) = {}^{1}\!/_{n_u}$.

One interpretation of the weight function, $g(n_u) = {}^{1}\!/_{n_u}$ is that the "reliability" of the ratio between the observed number of occurrences and the total number of pixels in the u-th unique condition subarea is inversely proportional to the size of the subarea, because the weighting function $g(n_u)$ is the variance of the ratio.

## 7. Discussion and concluding remarks

Several considerations can now be made about the multivariate regression approaches presented in this contribution. Some of the immediately beneficial aspects of regression analysis of integrated multilayered spatial data are:

- to adequately test the regression model, we have constructed the model based on pre-1960 data and the prediction were empirically compared with the distribution of the landslides which occurred during 1960-1980. This illustrates that regression is indeed useful for landslide hazard zonation as demonstrated for mineral exploration (Chung, 1983);

- the traditional pixel-based regression which requires vector-to-raster conversion is superseded by regressing unique condition subareas or sub-polygons, thereby minimizing processing time and data storage. Pixel-based regression is a special case of the proposed weighted regression;

- Regardless the pixel-size, the number of sample units for the regression model is relatively constant and small, and hence it is simple to implement the technique in a personal computer;

- the regresssion model proposed can easily be adapted to other types of tessellation of space, such as the morphometric units proposed by Carrara *et al.* (1991).

In this study, all experiments and applications dealt with one particular type of landslides: the rapid debris avalanches (derrumbes). Beside the application of regression to other types of landlsides, it would be beneficial to develop query and visualization tools to bring the geologist/geomorphologist or the hazard assessor into a more familiar interactive environment which could include 3-D representations (see color Plate 2) and animation.

The following is a list of possible developments out of the regression approach:

- the addition of the capability to query the spatial data base for interpreting the prediction of each subarea or of each pixel;

- to allow to reproject the results of regresssion in the space of scanned aerial photographs so that the photo-interpreter might start seeing features that might be left unnoticed (also enhanced images of aerial photos, digital elevation models, spaceborne TM, SPOT, ERS-1, etc., should be used);

- to allow to transform risk into hazard by considering the presence of man and of settlements in the spatial database;

to represent landslide processes both in space and in time. Although we have not discussed the "time concept" in this contribution, the concept is a necessary condition for predicting "future" landslides such as the landslides occurring within a predefined period of time. To study such landslides, however, the distribution map of the past landslides that was used in the Colombian data set is not adequate. This is due to the fact that the distribution of the past landslides did not identify the time of all the occurrence of lanslide phenomena. To study such "time" related landslides we should at least have the distribution of the past landslides within predefined time periods. Only in such a situation we may be able to develop new techniques to deal with such time-dependent sliding phenomena.

This approach overcomes the problems encountered in pixel-based regresssion, therefore, it can be implemented on any personal computer. Regression, however, is a data-driven approach which cannot incorporate expert opinion in the analysis. For applications in which expertees and subjectivity are also inputs to predictive processes, methods such as Bayesian approaches are more appropriate (Chung and Fabbri, 1993).

## Acknowledgements

We wish to thank Yvon Leclerc of the Geological Survey of Canada who has produced the graphic representation of the results for this manuscript and Dr. Ping An of the Geological Survey of Canada who has read an earlier draft of the manuscript and provided many useful comments. The study was partially funded by Japan Science Technology Fund granted to the Geological Survey of Canada from the Canadian Department of Foreign Affairs. The Geological Survey of Canada contribution number for this paper is 38694.

## References

Aronoff S., 1989. *Geographic Information Systems: A management perspective.* WDL Pub., Ottawa. 294 pp.

Carrara A., 1983. *Multivariate Models for landslide hazard evaluation.* Mathematical Geology, v. 15:3, 403-427

Carrara A., 1988. *Landslide hazard mapping by statistical methods. A "black box" approach.* The Proceedings of the Workshop on Natural disasters in European Meditteranean Countries, Italy, 205-224

Carrara A., Cardinali M., Detti R., Guzzetti F., Pasqui V., and Reichenbach P., 1991. *GIS techniques and statistical models in evaluating landslide hazard.* Earth Surface Processes and Landforms, v. 16:5, 427-445

Carrara A., Cardinali M., and Guzzetti F., 1992.. *Uncertainty in assessing landslide hazard and risk.* ITC Journal, v. 2, 172-183

Christensen R., 1990. *Log-Linear Models. Springer-Verlag.* New York, 408 pp.

Chung C.F., 1978. *Computer program for the logistic model to estimate the probability of occurrence of discrete events.* Geological Survey of Canada Paper 78-11, 23 pp.

Chung C.F., 1983, *SIMSAG: Integrated computer system for use in Evaluation of mineral and energy resources.* Math. Geology, v. 15:1, 47-58

Chung C.F., and Agterberg F.P., 1980. *Regression models for estimating mineral resources from geological map data.* Math. Geology, v. 12:5, 473-488

Chung C.F., and Fabbri A.G., 1993. *The representation of geoscience information for data integration.* Nonrenewable Resources, v. 2:2, 122-139

Chung C.F., and Leclerc Y., (in preparation). *Quantitative data integration techniques for landslide hazard mapping*

Draper N.R., and Smith H., 1981. *Applied Regression Analysis.* 2nd ed., Wiley, New York, 709 pp.

Fournier D'Albe E.M., 1976. *Natural disasters.* Bulletin Int. Assoc. Engin. Geol., v. 14, 187

Hansen A., 1984. *Landslide hazard.* In: Brunsden D., and Prior D.B., (Editors), Slope Instability, Wiley, New York, 523-602

Rao C.R., 1973. *Linear Statistical Inference and its Applications.* 2nd ed., Wiley, New York, 366-374

Roussas G., 1973. *A First Course in Mathematical Statistics.* Addison-Wesley, Reading, Mass. 506 pp.

Schuster R.L., 1994. *Socioeconomic significance of landslides.* In: Turner A.K., and Schuster R.L., (Editors), Landslides, investigation and mitigation, Transport Research Board Manual. (in press).

Searle S.R., 1971. *Linear Models.* Wiley, New York, 532 pp.

van Westen C.J., 1993. *Application of Geographic Information Systems to Landslide Hazard Zonation.* Ph.D. Thesis, Technical University of Delft, International Institute for Aerospace Surveys and Earth Sciences, Enschede, The Netherlands, ITC Pubblication 15, v. 1, 245 pp.

van Westen C.J., van Duren H.M.G., Kruse I., and Terlien M.T.J., 1993. *GISSIZ: Training Package for Geographic Information Systems in Slope Instability Zonation.* ITC Publication 15, ITC, Enschede, The Netherlands. Volume 1 - Theory, 245 pp., v. 2 - Exercises, 359 pp. with 10 diskettes

Wang S.-Q., and Unwin D.J., 1992. *Modeling landslide distribution on loess soils in China: an investigation.* International Journal of Geographic Information Systems, v. 6:5, 391-405

## Appendix A

We illustrate that the pixel-based regression model is a special case the weighted regression model based on the unique condition subarea by looking at a simple case, because a general case is a notational nightmare.

In addition to the distribution map of the occurrences of landslides in a given study area, consider two layers in which the first layer consists of three classes, A, B and C and the second layer contains two classes, D and E only. Suppose that the study area consists of 100 pixels. At each pixel i, the following four dummy binary variables, $B_{i1}$, $B_{i2}$, $B_{i3}$ and $B_{i4}$, for the five classes in two layers, were generated. Let:

$Y_i$ represents the presence (1) or the absence (0) of the occurrence of the landslide,
$B_{i1}$ represents the presence (1) or the absence (0) of the class A in the first layer,
$B_{i2}$ represents the presence (1) or the absence (0) of the class B in the first layer,
$B_{i3}$ represents the presence (1) or the absence (0) of the class D in the second layer,
$B_{i4}$ represents the presence (1) or the absence (0) of the both class C in the first layer and E in the second layer.

As discussed in the text, two separate binary variables for the class C and class E were not used to avoid the collinearlity (Draper and Smith, 1981). In this appendix, one more binary variable $B_{i4}$ was added into the model, although the similar operations were not carried out in the Colombian study because the large numbers of classes and the layers.

As in (6.3), we have the following 100 x 6 matrix: For

$$(Y_i; 1, B_{i1}, B_{i2}, B_{i3}, B_{i4}), \qquad i = 1, 2, \cdots, n\ (=100), \tag{A.1}$$

where "1" is for the constant term in the regression model.

The linear model in (6.4) for each i-th pixel is written

$$Y_i = d_0 + d_1 B_{i1} + d_2 B_{i2} + d_3 B_{i3} + d_4 B_{i4} + \varepsilon_i \tag{A.2}$$

where the five unknown parameters $(d_0, d_1, d_2, d_3, d_4)$ will be estimated from the 100 x 6 matrix in (A.1). It can be shown (Draper and Smith, 1981) that the LS estiamtors $(\bar{d}_0, \bar{d}_1, \bar{d}_2, \bar{d}_3, \bar{d}_4)$ of the $(d_0, d_1, d_2, d_3, d_4)$ are:

$$\begin{pmatrix} \bar{d}_0 \\ \bar{d}_1 \\ \bar{d}_2 \\ \bar{d}_3 \\ \bar{d}_4 \end{pmatrix} = \begin{pmatrix} n & n_A & n_B & n_D & n_{CE} \\ n_A & n_A & 0 & n_{AD} & 0 \\ n_B & 0 & n_B & n_{BD} & 0 \\ n_D & n_{AD} & n_{BD} & n_D & 0 \\ n_{CE} & 0 & 0 & 0 & n_{CE} \end{pmatrix}^{-1} \bullet \begin{pmatrix} m \\ m_A \\ m_B \\ m_D \\ m_{CE} \end{pmatrix} \tag{A.3}$$

where:

$n = 100$: the total number of pixels,
$n_A$: the number of pixels in the class A in the first layer,
$n_B$: the number of pixels in the class B in the first layer,
$n_D$: the number of pixels in the class D in the second layer,
$n_{AD}$: the number of pixels in the both class A and D,
$n_{BD}$: the number of pixels in the both class B and D,
$n_{CE}$: the number of pixels in the both class C and E,
m: the total number of the occurrences of the landslides,
$m_A$: the number of the occurrences in the class A in the first layer,
$m_B$: the number of the occurrences in the class B in the first layer,
$m_D$: the number of the occurrences in the class D in the second layer,
$m_{CE}$: the number of the occurrences in the both class C and E.

Now consider the unique condition subareas in the study area. we have at most following six unique condition subareas:

AD: the subarea overlapped by both A in the first layer and D in the second layer,
AE: the subarea overlapped by both A in the first layer and E in the second layer,
BD: the subarea overlapped by both B in the first layer and D in the second layer,
BE: the subarea overlapped by both B in the first layer and E in the second layer,
CD: the subarea overlapped by both C in the first layer and D in the second layer,
CE: the subarea overlapped by both C in the first layer and E in the second layer.

Similar to (6.7), we have the following 6 x 7 matrix from the above six unique condition subareas:

$$\begin{array}{lccccc}
( Y_i \quad ; & 1, & B_{i1}, & B_{i2}, & B_{i3}, & B_{i4} ), \\
\hline
( m_{AD} , n_{AD} ; & 1, & 1, & 0, & 1, & 0 ), \\
( m_{AE} , n_{AE} ; & 1, & 1, & 0, & 0, & 0 ), \\
( m_{BD} , n_{BD} ; & 1, & 0, & 1, & 1, & 0 ), \\
( m_{BE} , n_{BE} ; & 1, & 0, & 1, & 0, & 0 ), \\
( m_{CD} , n_{CD} ; & 1, & 0, & 0, & 1, & 0 ), \\
( m_{CE} , n_{CE} ; & 1, & 0, & 0, & 0, & 1 ),
\end{array} \qquad (A.4)$$

where $m_u$ represents the number of the occurrences and $n_u$ represents the number of pixels in the u-th unique condition subarea.

For each u-th unique condition subarea, let $R_u$ be the ratio between $m_u$ and $n_u$. i.e., the ratio between the number of occurrences and the number of pixel in each subarea. The regression model at each u-th subarea is:

$$R_u = a_0 + a_1 B_{u1} + a_2 B_{u2} + a_3 B_{u3} + a_4 B_{u4} + \varepsilon_u \qquad (A.5)$$

where the five unknown parameters ($a_0$, $a_1$, $a_2$, $a_3$, $a_4$) will be estimated from the 6 x 7 matrix in (A.4). Let us assume that

$$\mathrm{Var}\,(\varepsilon_u) = g(n_u) = {}^{1}\!/_{n_u} \qquad (A.6)$$

where $n_u$ is the total number of pixels in the unique condition subarea. The weighted LS estimators $(\bar{a}_0 , \bar{a}_1 , \bar{a}_2 , \bar{a}_3 , \bar{a}_4 )$ which have many "optimal" statistical properties for $( a_0 , a_1 , a_2 , a_3 , a_4 )$ in (A.5) are obtained (Draper and Smith, 1981):

$$\begin{pmatrix} \bar{a}_0 \\ \bar{a}_1 \\ \bar{a}_2 \\ \bar{a}_3 \\ \bar{a}_4 \end{pmatrix} = \left( \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} n_{AD} & 0 & 0 & 0 & 0 & 0 \\ 0 & n_{AE} & 0 & 0 & 0 & 0 \\ 0 & 0 & n_{BD} & 0 & 0 & 0 \\ 0 & 0 & 0 & n_{BE} & 0 & 0 \\ 0 & 0 & 0 & 0 & n_{CD} & 0 \\ 0 & 0 & 0 & 0 & 0 & n_{CE} \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \right)^{-1} \bullet$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} n_{AD} & 0 & 0 & 0 & 0 & 0 \\ 0 & n_{AE} & 0 & 0 & 0 & 0 \\ 0 & 0 & n_{BD} & 0 & 0 & 0 \\ 0 & 0 & 0 & n_{BE} & 0 & 0 \\ 0 & 0 & 0 & 0 & n_{CD} & 0 \\ 0 & 0 & 0 & 0 & 0 & n_{CE} \end{pmatrix} \begin{pmatrix} m_{AD}/n_{AD} \\ m_{AE}/n_{AE} \\ m_{BD}/n_{BD} \\ m_{BE}/n_{BE} \\ m_{CD}/n_{CD} \\ m_{CE}/n_{CE} \end{pmatrix}$$

and hence, the WLS estimators, ( $\bar{a}_0$ , $\bar{a}_1$ , $\bar{a}_2$ , $\bar{a}_3$ , $\bar{a}_4$ ) from the unique condition subarea-based data in (A.4) are:

$$\begin{pmatrix} \bar{a}_0 \\ \bar{a}_1 \\ \bar{a}_2 \\ \bar{a}_3 \\ \bar{a}_4 \end{pmatrix} = \begin{pmatrix} n & n_A & n_B & n_D & n_{CE} \\ n_A & n_A & 0 & n_{AD} & 0 \\ n_B & 0 & n_B & n_{BD} & 0 \\ n_D & n_{AD} & n_{BD} & n_D & 0 \\ n_{CE} & 0 & 0 & 0 & n_{CE} \end{pmatrix}^{-1} \bullet \begin{pmatrix} m \\ m_A \\ m_B \\ m_D \\ m_{CE} \end{pmatrix} . \qquad (A.7)$$

They are the exactly identical to the LS estimators, ( $\bar{d}_0$ , $\bar{d}_1$ , $\bar{d}_2$ , $\bar{d}_3$ , $\bar{d}_4$ ) in (A.3) from the pixel-based data in (A.1).