EVALUATING THE USE OF TRAINING AREAS IN BIVARIATE STATISTICAL LANDSLIDE HAZARD ANALYSIS

A case study from Chinchiná - Santa Rosa De Cabal, Colombia

J.L. Naranjo[1], C.J. van Westen[2] and R. Soeters[2]

[1] Universidad de Caldas, Facultad de Geologia y Minas, Apartado Aereo 1729, Manizales, Colombia.
[2] International Institute for Aerospace Survey and Earth Sciences (ITC), P.O.Box 6, 7500 AA, Enschede, the Netherlands.

ABSTRACT

The assessment of landslide hazard is a costly and time-consuming activity which requires the collection, manipulation and analysis of a large number of factors related with landslide occurrence. To reduce these costs and time investments a study was conducted to test the use of training areas in statistical landslide hazard assessment. For this purpose two areas with the same terrain conditions in the Colombian Andes were selected, one in Santa Rosa de Cabal, Risaralda, and the other one in Chinchiná, Caldas. The testing was performed using a statistical bivariate method for landslide hazard analysis in both areas, followed by a comparison of the results. An examination of the obtained weight values and densities of mass movements in the hazard classes for both areas revealed that the landslide densities within the various parameter maps from the two areas showed large differences. Although the calculated weight values revealed more or less the same trend, they were quite different in magnitude. Care must be taken in the selection of training areas. Expert driven combination of weights in the two areas resulted in rather similar hazard maps. These could, however, only be obtained by carefully selecting hazard classes with equal percentages of landslides. When a landslide map for the prediction area is not available the method will give erroneous results.

INTRODUCTION

Mass movements in mountainous terrain are natural degradational processes. Under the influence of a variety of causal factors, and triggered by events such as earthquakes or extreme rainfall, most of the terrain in mountainous areas has been subjected to slope failure at least once. The zonation of landslide hazard must be the basis for any landslide mitigation project and should supply planners and decision makers with adequate and understandable information. Landslide hazard is defined by Varnes [6] as "the probability of occurrence of a potentially damaging phenomenon within a specified period of time and within a given area".
Because many factors can play a role in the occurrence of mass movements, the analysis of landslide hazard is a complex task. Not only it requires a large number of input variables, but techniques of analysis may be very costly and time-consuming. During the last decades, the increasing availability of computers has created opportunities for more detailed and rapid analyses of landslide hazard.
In a study by Van Westen on the application of geographic information systems in landslide hazard zonation [7], the recommendations were given for suitable data collection and analysis techniques at various working scales. Several techniques for landslide hazard zonation can be

applied, which can be subdivided into heuristic, statistical and deterministic techniques. Before starting a hazard study an earth scientist should be aware of the desired degree of detail of the hazard map, given the requirements of the study. When a degree of detail and a working scale have been defined, the cost-effectiveness of obtaining input data must be considered. In the above mentioned study recommendations were given as to which kind of data can be suitably collected at each working scale (regional, medium and large scale). The availability of data determines the type of analysis that can be performed.

Statistical techniques are considered the most appropriate approach for landslide hazard zonation at scales of 1:25.000 to 1:50.000. On this scale it is possible to map out in detail the occurrences of past landslides, and to collect sufficient information on the relevant factors that are considered to be related with the occurrence of landslides.

Two different statistical approaches are used in landslide hazard analysis: multivariate and bivariate statistical analysis.

Multivariate statistical analyses of important factors related to landslide occurrence may give the relative contribution of each of these factors to the total hazard within a defined land unit or pixel. Several multivariate methods have been proposed in the literature. Most of these, such as discriminant analysis or multiple regression, require the use of external statistical packages. GIS is used to sample variables for each land unit. Recent examples of multivariate statistical analysis in landslide studies using GIS have been presented mainly by Carrara [3,4]. His work has developed from the use of large rectangular grid cells as the basis for analysis [3] towards the use of morphometric units, which can be automatically extracted from a digital elevation model [4]. The method itself has not undergone major changes. For each homogeneous unit the presence/absence or the percentage cover of landslides and of the landslide controlling factors are sampled into a data matrix. Then multivariate statistical models are applied to the matrix and prediction formulae are obtained. In the application of multivariate models two major difficulties exist:

(1) The numerical representation of classified data. Most of the data-layers presented in table 1, such as geology, landuse, geomorphology, do not consist of continous data types, but of classified discrete patterns. To use this kind of data in multivariate statistical analysis the data should be changed. This is mostly done through the creation of so-called dummy variables by converting each variable class (e.g. each geological unit) into a binary patterns (presence/absence). This will increase the number of variables enormously, causing problems in the analysis.

(2) The large number of observations. In GIS-based landslide hazard analysis the ideal case would be to do the statistical analysis based on individual pixels. However, this results in such voluminous matrices that the calculation using standard statistical packages becomes a real problem.

In bivariate statistical analysis, overlaying of variable maps and calculation of landslide densities form the core of the analysis. The importance of each variable, or specific combinations of variables, can be analyzed individually [2]. For each variable class a weight value can be calculated by relating the landslide density within the class with the overall landslide density in the map. These weight values are defined by Chung and Fabbri as favourability functions [5], indicating the "sureness, probability, certainty, belief, plausibility, possibility or compatibility that the pixels within the variable class will have a landslide". Specific combination rules are applied to derive at a final hazard map. The main steps in the procedure are shown in figure 1.

Bivariate statistical analysis

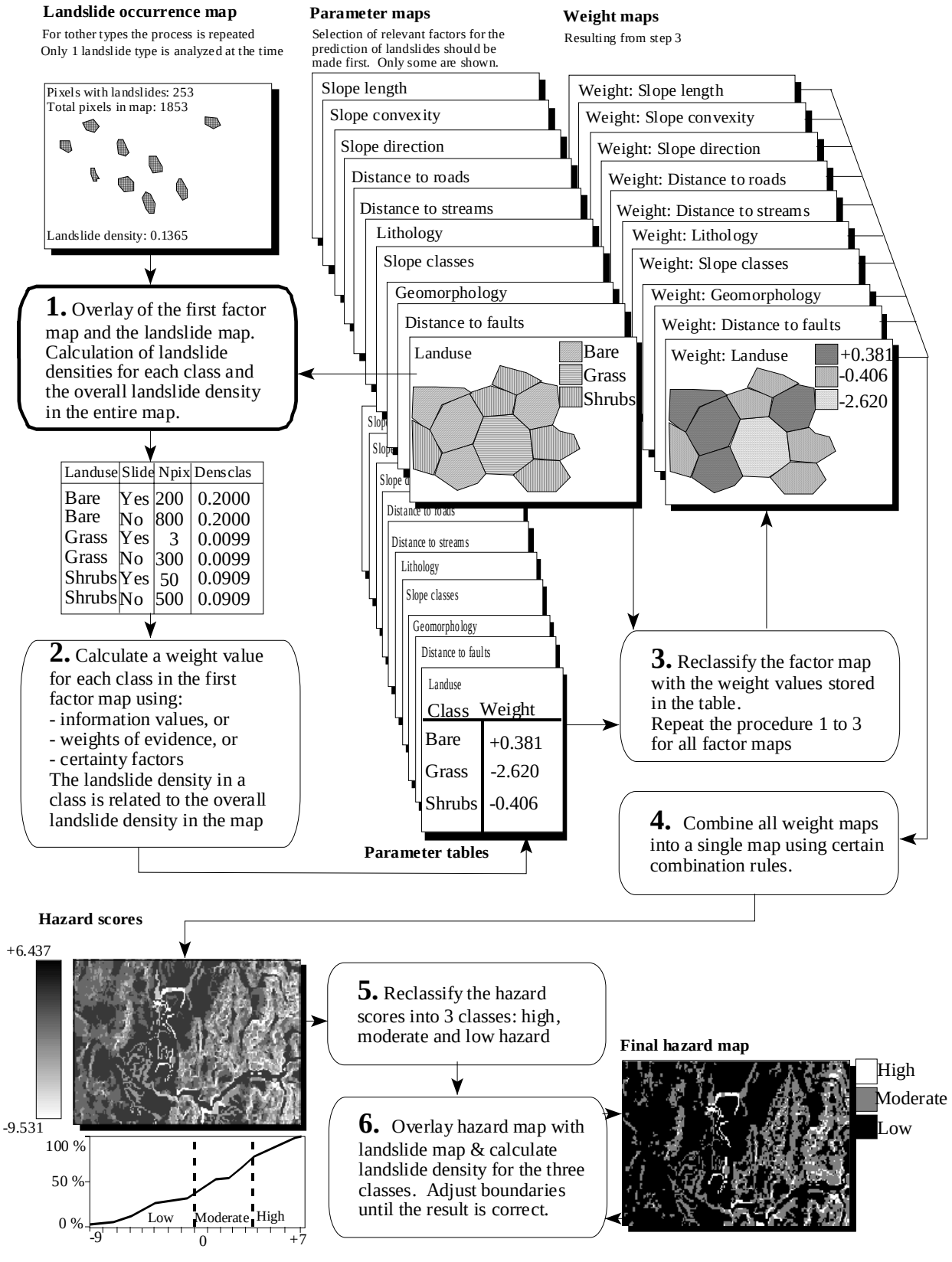

Figure 1: Simplified procedure for bivariate statistical landslide hazard analysis

Several statistical methods can be applied to determine the favourability functions and their combination, such as landslide susceptibility [2,7], the information value method [8], weights of evidence modelling, Bayesian combination rules, certainty factors, Dempster-Shafer belief function and fuzzy sets [5]. The use of bivariate statistical analysis has the following advantages:

- The results are reproducible, because the mathematical operations are fixed,
- The results are easy to interpret as each factor map can be evaluated separately,
- It is possible to include expert opinion, as specific combinations of factors can be selected and tested for their importance in generating landslides.
- The accuracy of the resulting prediction maps can be checked by overlaying it with the landslide distribution, especially when the prediction is based on the landslide distribution in the past, and compared with the present landslide distribution.

However, the use of bivariate statistical analysis also has serious drawbacks:

- The assumption of conditional independence is mostly not valid. Two variables (A and B) may be considered conditionally independent with respect to the occurrence of landslides, when the probability that a pixel is in the intersection of two variables (A_B), given it is a landslide pixel, is equal to the multiplication of the individual probabilities that a pixel is in A or in B, given that it is a landslide pixel. Violation of this assumption may lead to favourability functions which are deviating from reality.
- The collection of landslide distribution and a large series of parameters over a large area is very time consuming. Furthermore, during the analysis it may be so that a factor map, collected with much effort, turns out to be of no importance, and is not used in the final prediction.

To overcome this last problem the concept of training areas is often applied. A training area or sample area is defined as a small area within the overall research area where the spatial distribution of landslides and the various factors is (or should be) well known. Statistical analysis is performed on this subset, and a prediction formula is derived. When this evaluation is completed, the results are extrapolated to larger areas called target, study or prediction areas. The extrapolation is based on the assumption that the factors which created the phenomenon in the sample area are the same in the target area. This concept has been used in mineral exploration [1], and in multivariate statistical landslide hazard zonation [3,4]. When the most important variables (factor maps) are known, from the analysis in the training area then they are collected for the target area. This is important because it decreases the costs not only in the acquisition of some maps which are unimportant in the analysis, but also in the fact that no time is spent on digitizing and processing irrelevent data.
In this study the applicability of using training and target areas in bivariate statistical analysis is evaluated using two data sets from neighbouring areas.

The training and prediction areas

For testing the validity of using training and prediction areas two adjacent areas on the western slope of the Central Cordillera in Colombia were selected. One within the municipality of Santa Rosa de Cabal in Risaralda and the other one in Chinchiná, located in the department of Caldas. Both areas are located within the same geological zone: the western side of the Romeral fault zone, one of the most important fault zones in Colombia.
Igneous, metamorphic and metasedimentary rocks are the most common rock types in the areas. The region where the two areas are located was affected by volcanic activity at least three times during the Quaternary period. Volcanic activity of the nearby volcanos of Nevado del Ruiz, Santa Isabel and Paramo de Santa Rosa resulted in large emissions of acid pyroclastics, accumulation of lavaflows, and generation of volcanic debris flows as a consequence of thawing of the icecaps, covering the volcanoes. An overview of the training area (Santa Rosa) is given in figure 2.

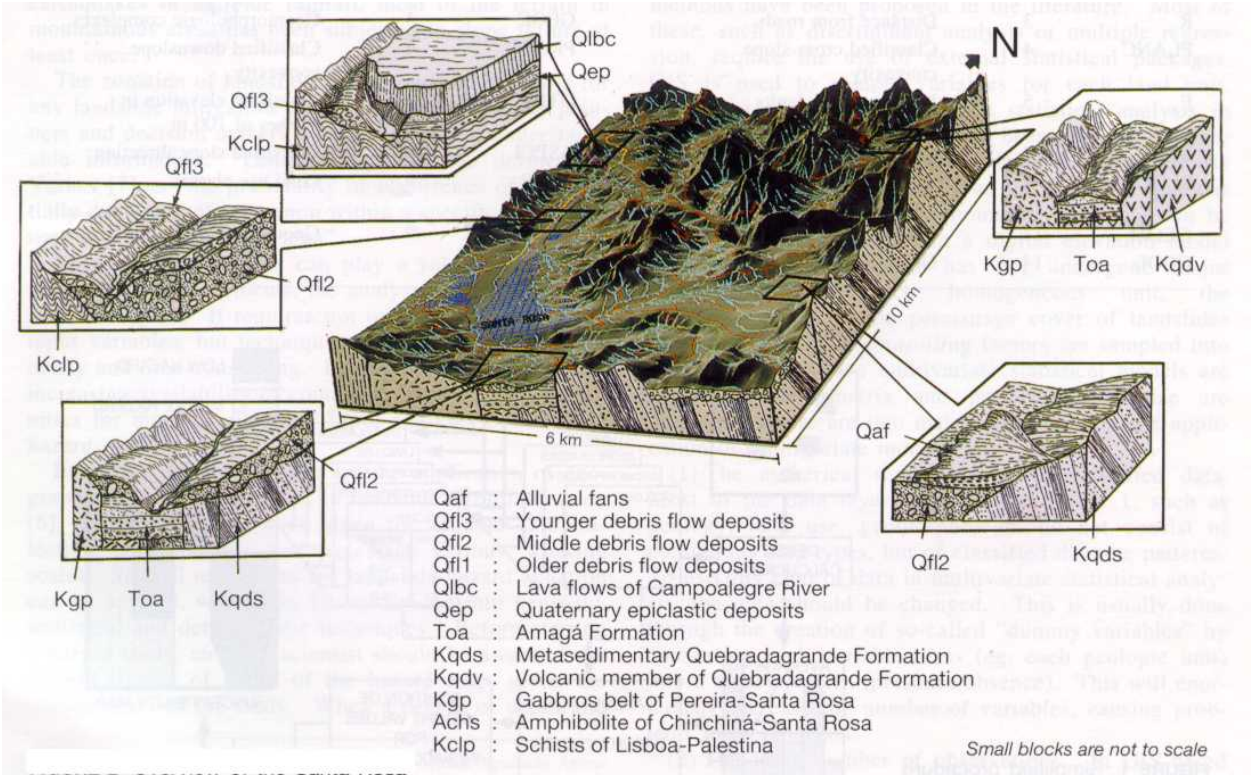

Figure 2: Overview of the Santa Rosa area with block-diagrams showing the main geological formations.

The geomorphological conditions in the two areas are quite similar. Generally three geomorphological units can be differentiated:

- Fault related valleys with steep slopes in the eastern part,
- Large terrace levels in the central part,
- Rounded hills in metamorphic rocks in the western part.

The western zone is more extensive in the Chinchina area, whereas the eastern part is larger in the Santa Rosa area. The zone is representative of the Andean environment, where there is a high frequency of landslides due to topographic conditions, deeply weathered profiles, high rainfall rates, and very intense landuse and earthquakes. The land use within the two areas is also comparable. Both areas are located within the major coffee producing region of Colombia.
A large series of factor maps were obtained for both areas (Table 1). The maps were rasterized with a pixel size of 12.5 m. Landslide distribution maps (Figure 3) were obtained using multi-temporal airphoto-interpretation and detailed fieldwork, making use of check-lists described in Van Westen [7]. A classification based on six types of mass-movements and three subtypes was used to identify the processes in the area.

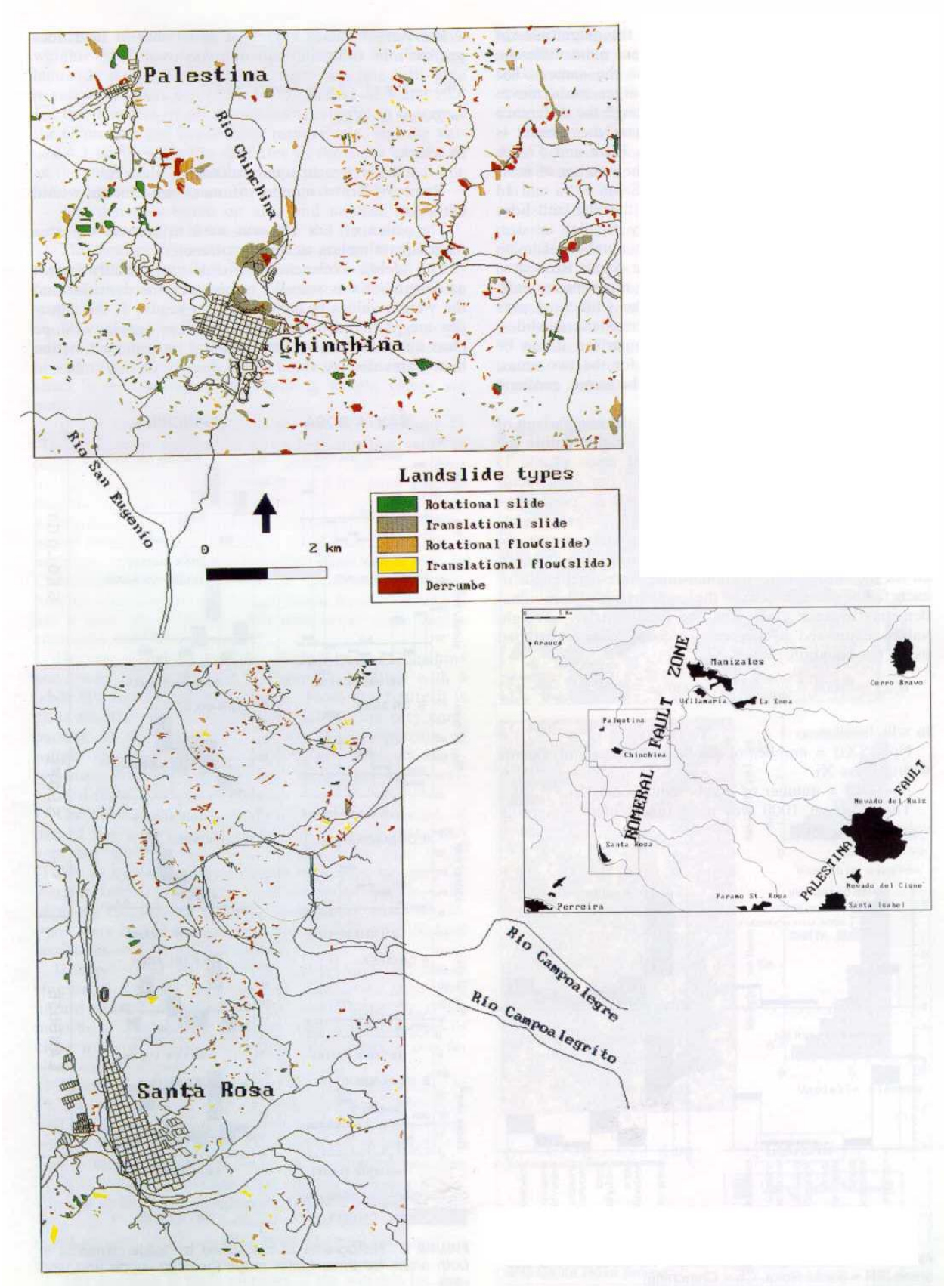

Figure 3: Landslide distribution maps. A: Santa Rosa, B: Chinchina

DATA ANALYSIS

The first step in the analysis was a comparison between the landslide distribution maps of the two areas. In the Santa Rosa area, with an extension of 60 km^2, 2% was covered by landslides, and in the Chinchina area (68 km^2) 5.7%. For both areas the percentages of specific combinations of landslide type, subtype and activity were calculated (Figure 4), expressed as percentage cover or as number of landslides per km^2.

Map code	Number of classes	Description	Map code	Number of classes	Description
DRS	3	Distance from valley heads	DRT	3	Distance from streams
R	3	Distance from roads	GEOC	3	Geomorphological complexes
PLANC	4	Classified cross-slope convexity	PROFC	4	Classified downslope convexity
F	5	Distance from faults	DTMC	9	Classified altitude in classes of 100 m
SLLC	7	Classified distance from ridges	ASPCL	8	Classified slope direction
SLOC2	9	Classified slope map in groups with of 10°	LUSE	9	Land-use classes
GEOM	11	Geomorphological main units	GEOS	9	Geomorphological subunits
GEOL	14	Geological units			

Table 1: Summary of the input maps used in the statistical analysis at the medium scale. The number of classes, the code of the file, and a description are given.

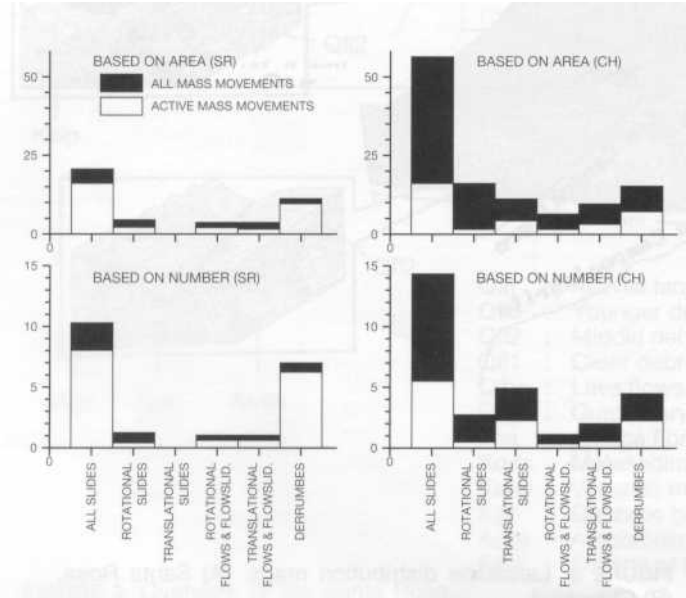

FIGURE 4 Comparison of landslide densities of the two areas (SR = Santa Rosa, Ch = Chinchiná)

Figure 4: Comparison of landslide densities of the two areas. SR= Santa Rosa, Ch= Chinchina.

As can be observed from figure 4 the magnitudes of landslide densities for the two areas are quite different, although the trends are more or less the same. The Chinchiná area contains more mass movements, with larger sizes, than the Santa Rosa area. Although the difference in percentage of area covered by mass movements is quite different in both areas (20✦ in Santa Rosa and 57✦ in Chinchiná), the difference in the number of mass movements is less (10 slides/km^2 in Santa Rosa and 14 slides/km^2 in Chinchiná). This means that the landslides in the Santa Rosa area are generally smaller in size. Another noticeable difference is the number of mass movements identified as *active*. The Santa Rosa area contains a larger percentage of active massmovements, mainly of the derrumbe type (local term for debris avalanches), whereas the Chinchina area contains more inactive rotational and translational slides, flows and flowslides. From this comparison it can be concluded that the landslide patterns for the two areas, although adjacently located within the same geological zone, is quite different.
The second step in the analysis was the comparison of the calculated landslide densities and weights within the two areas. For both areas the parameter maps (Table 1) were overlain with the landslide map. For each parameter class the landslide density, expressed as percentage cover or as number of landslides per km^2, was calculated for different combinations of types and subtypes of mass movements. Also landslide densities were calculated for the entire map. Weight values were calculated for each parameter class, based on the comparison of individual densities in each class with the total density. Weight values expressed as perentage cover were calculated using formula:

$$W_{area} = 1000 \times \frac{Npix(SXi)}{NPix(Xi)} / 1000 \times \frac{\sum Npix(SXi)}{\sum NPix(Xi)}$$

in which:
Npix(SXi) = Number of pixels with mass-movements within class **Xi.**
Npix(Xi) = Number of pixels within class **Xi**.
The value of 1000 is used to calculate weighting values in permillage.
For weight values expressed as number of landslides per km^2 the following equation was used:

$$W_{number} = \frac{1x10^6}{Area(Xi)} \times Number(SXi) / \frac{1x10^6}{\sum Area(Xi)} \times \sum Number(SXi)$$

in which:
Area(Xi) = Area in square meters of class **Xi**.
Number (SXi) = Number of mass-movements within the class **Xi**.

The value of 1x10^6 is used to convert the area from square meters to square kilometres.
The landslide type *derrumbe* (debris avalanche) was selected to compare the densities and the weight values in both areas. In figure 5 the densities are plotted for three selected maps **(GEOL, SLOC2, LUSE)**. The horizontal lines in each one of the histograms display the average density of derrumbes in each of the areas. The bars located above the line will result in positive weights (landslide density in class is higher than average landslide density), and those lower than the line will give negative weights. For example the values **15.4** and **10.8** are the densities of all derrumbes in permillage in the areas for Chinchiná and Santa Rosa respectively, and the values **4.3** and **6.9** are the densities in the areas expressed as the number of derrumbes per square kilometer in both areas respectively.

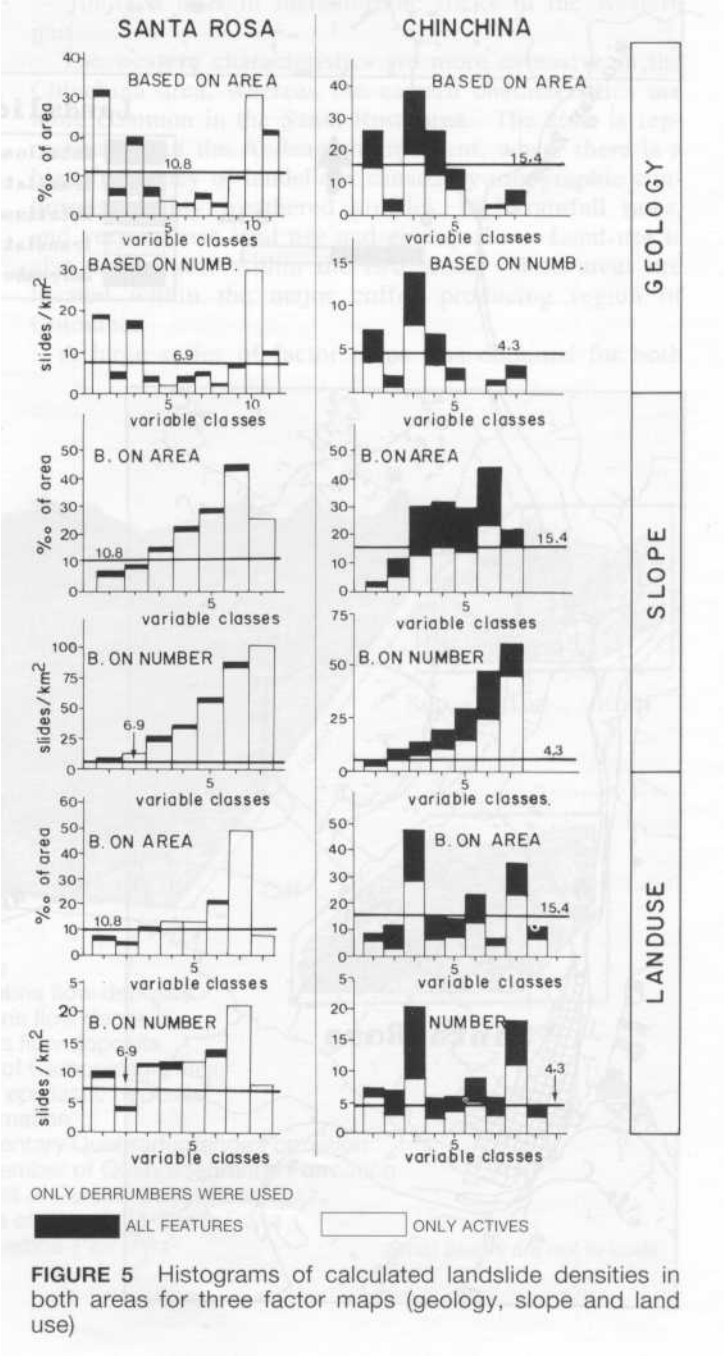

FIGURE 5 Histograms of calculated landslide densities in both areas for three factor maps (geology, slope and land use)

Figure 5: Histograms of calculated landslide densities in both areas for three parameter maps (Geology, Slope and Landuse)

The following conclusions can be obtained from the figure:

- The densities based on area and number show the same pattern in the histograms, for the same area.
- The densities for the two areas are quite different, especially when all derrumbes are used. The large difference in active landslides, discussed in the previous sections, is also very clear in this figure.
- The values for the slope classes and the geological units more or less show the same trends for both areas.
- Although there are generally large differences in densities in both areas, the resulting weight values are quite similar.
- If we look at the geology we can see that the unit 11 (Toa= Tertiary sediments) has a high positive value in Santa Rosa and a low negative one in Chinchiná. This unit is covered by recent deposits in Chinchiná and therefore the mapped size is smaller than in Santa Rosa

where the cover is absent, and due to a larger number of landslides in Santa Rosa. The unit four (Kqds) has a negative value in Santa Rosa and a positive one in Chinchiná, which is due to the small amount of derrumbes identified in this unit in Santa Rosa, although it has a relatively large area, but lower slope angles as compared to Chinchiná.

The two classified weighted slope maps for the Chinchina and Santa Rosa area are given in figure 6, together with a table of calculated weight values. From this figure it is clear that the results for the slope classes are very comparable between the two areas.

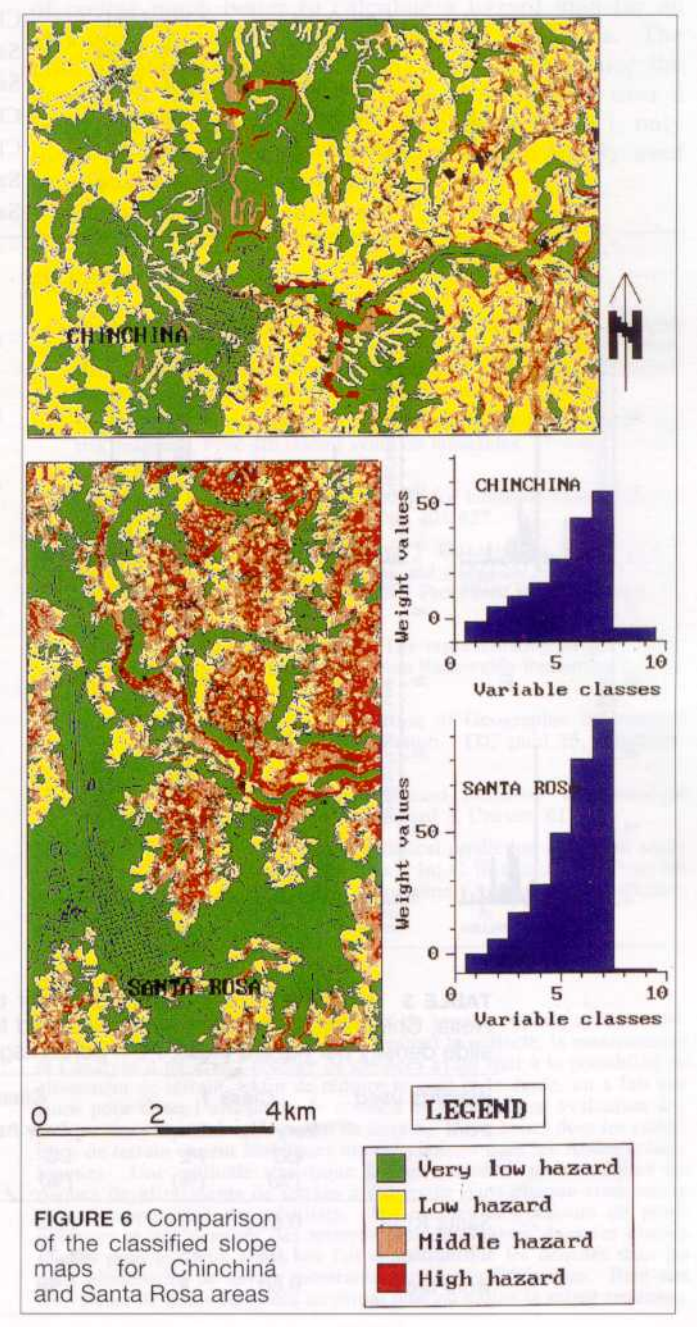

Figure 6: Comparison of the classified slope maps for Chinchiná and Santa Rosa areas

From the comparison of weight values it can be concluded that, although density values may be quite different, the weight values calculated in both areas are similar.

The third step in the analysis was the calculation of hazard maps, using the weight values of one area, and applying it to the other. Eight maps were made (Table 2) by subtracting a hazard map made by applying weights from another area, from a hazard map made by applying the weights from the same area. All calculations were based on numbers/km^2 and calculated for all

landslides, and for all derrumbes.

	Hazard map 1		Hazard map 2		
	Area	Weights	Area	Weights	Landslides
A	Chinchiná	Chinchiná	Chinchiná	Santa Rosa	all
B	Chinchiná	Chinchiná	Chinchiná	Santa Rosa	derrumbes
C	Santa Rosa	Santa Rosa	Santa Rosa	Chinchiná	all
D	Santa Rosa	Santa Rosa	Santa Rosa	Chinchiná	derrumbes
E	Chinchiná	Chinchiná	Chinchiná	Both	all
F	Chinchiná	Chinchiná	Chinchiná	Both	derrumbes
G	Santa Rosa	Santa Rosa	Santa Rosa	Both	all
H	Santa Rosa	Santa Rosa	Santa Rosa	Both	derrumbes

Table 2: Maps created to evaluate the use of weight values from another area. For each of the maps A to H Hazard Map 2 was subtracted from Hazard Map 1.

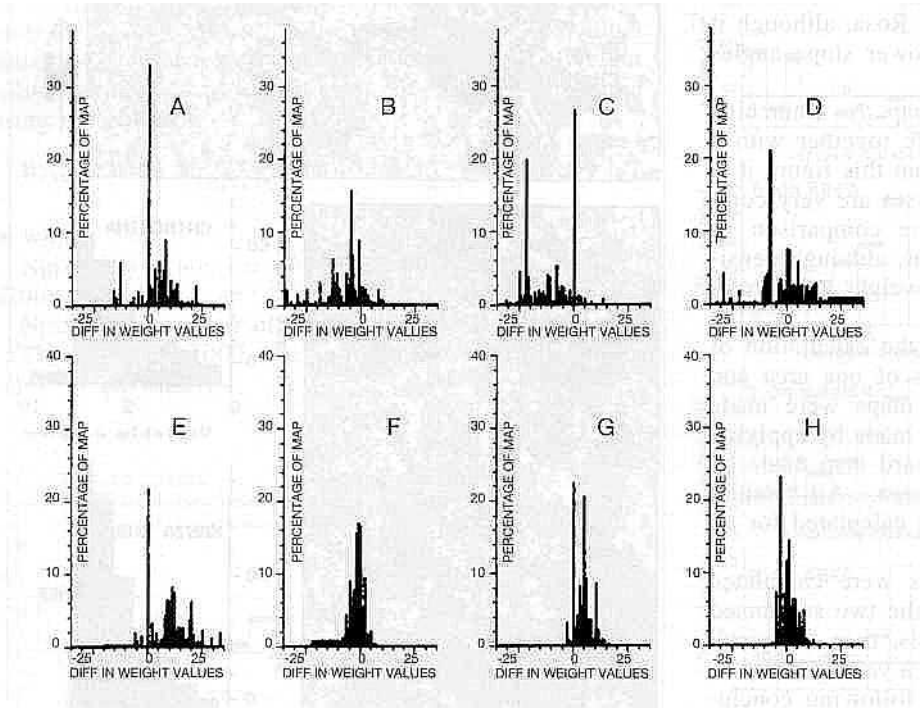

Figure 7: Histograms displaying the differences in summated weight values between a hazard map calculated with weights derived from the same area and a hazard map made with weights from another area. See table 2 for explanation of the letters.

For each map A to H histograms of the resulting values were calculated (Figure 7). If the two subtracted hazard maps have similar weight values, then the result must be a large number of pixels with value zero or close to zero. From this figure, the following conclusions can be drawn:

- Since the weight values for all landslides in Chinchiná are higher than in Santa Rosa, the difference will result in an underestimation (positive values) when weight values from Santa Rosa are used to calculate a hazard map of Chinchiná (A). In the reverse situation - weight values from Chinchina applied to Santa Rosa- they result in an overestimation (negative values) as seen in graph C. Only 30 percent of the area will obtain equal weights if we use data from one area and apply it to the other.
- The situation is more complex if the weights for derrumbes are used. In this case only very few pixel will actually obtain the same weight values. Weight values for derrumbes are higher in Santa Rosa, resulting in an overestimation if data from Santa Rosa is used in

Chinchiná, and underestimation in the reverse case.

- The use of weight values derived from both areas does not influence the results significantly when derrumbes are used as can be seen in histograms F and H, but the influence is strong when all mass movements are used (histograms E and G). This is related to the fact that the Santa Rosa area only contains a limited number of other landslides.

The last step in evaluating the effect of using data from other areas was the construction of final hazard maps, by classiying the resulting weights in 4 classes (very low, low, moderate and high). As can be deduced from figure 7 the use of equal classification tables would result in serious over- or underestimations of the hazard. The boundary values of the hazard classes in the classification tables were adjusted repetitively, untill the pattern of the final hazard map was in accordance with our knowledge from the terrain. Each time the produced hazard map was overlain with the landslide distribution map to evaluate the landslide density in each of the hazard classes. Table 3 gives the results for the Santa Rosa area, calculated for all derrumbes based on nr/km^2, using weight values from Santa Rosa, Chinchiná and both areas. As can be seen from this table, the use of different classification boundaries resulted in maps which are quite comparable. Figure 8 gives an example of a resulting hazard map calculated for derrumbes, with weights expressed in number/km^2, using data from Santa Rosa.

Weights used from	CLASS 1 Very low hazard		CLASS 2 Low hazard		CLASS 3 Moderate Hazard		CLASS 4 High hazard	
	LC %	PL %	LC %	PL %	LC %	PL %	LC %	PL %
Santa Rosa	0.69	10	1.32	18	2.18	30	3.18	42
Chinchiná	1.03	13	1.51	19	2.49	30	3.06	38
Both areas	0.60	8	1.80	20	2.70	33	3.10	39

Table 3: Resulting hazard classification for the Santa Rosa area, using weights from Santa Rosa, Chinchiná, and both areas, calculated for all derrumbes, using nr/km^2. LC: landslide density per hazard class, PL = percentage of all landslides within the hazard class.

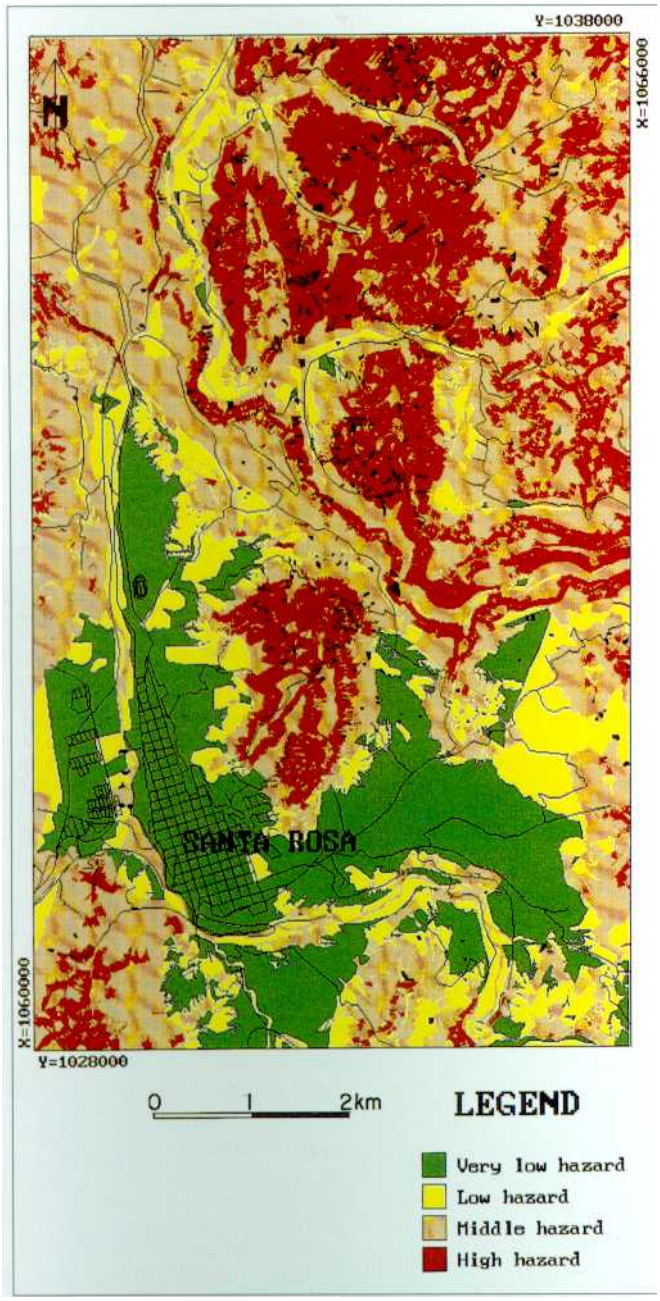

Figure 8: Landslide hazard map for the Santa Rosa area resulting from the combination of the weights from the geological, slope and landuse maps. The calculation was made for all derrumbes, using weight values from Santa Rosa, expressed in number/km^2.

DISCUSSION AND CONCLUSIONS

The results from this study show that there are serious limitations to the use of training and prediction areas in bivariate statistical analysis. Although two adjacent areas were selected with the same geological, topographical and landuse characteristics, the resulting landslide density values, calculated for the various factor maps, were quite different from one area to the other. The differences between the two areas were however partly reduced when the weights were calculated. Nevertheless, the weights derived from the Chinchiná area were generally higher than for Santa Rosa if all landslides were taken into account, and lower if only derrumbes were used. Also the hazard maps calculated for the same area showed a large deviation when different data sets were used.

The crucial part in the analysis turned out to be the classification of the summated weight values into a final hazard map. When a process of trial and error was used, in which the result of each classification was checked against the expected map pattern, similar maps could be made with the data set from the two areas.
This leads to the conclusion that the application of training areas in bivariate statistical analysis is possible, be it with a number of serious limitations. The most difficult problem in this respect is the selection of a training area whith the same characteristics as the entire prediction area. In this case, the initial assumption that the two areas were very similar was not valid of the landslide patterns were compared. In this respect it is important to specify that the data sets of the two areas were collected by two different persons. This will have inluenced the result in a negative sense, especially regarding the mapping and classification of landslides. The elaboration of a landslide map is a highly subjective and error prone procedure (Dunoyer and van Westen, this volume). However, if the effects of misclassifications were ruled out by taking all landslides together, the resulting landslide densities still were quite different.
Another issue is related to the practical use of the concept of training and prediction areas. From this research it turned out to be of crucial importance to have a landslide map for the prediction area, in order to be able to make a correct classification of the summated weight values. Furthermore, also the relevant parameter maps should be gathered for the prediction area, in order to apply the weights. Therefore, when one disposes of these maps, it is of course much better to calculate a hazard map for an area using the weights derived from the same area. The only advantage of using training areas is to reduce the number of factor maps to be made and digitized over a large area. From the factor maps given in table 1, only those of Geology, Slope and Landuse were finally used in this study.

REFERENCES

1 **Bonham-Carter, G.F., Agterberg, F.P., Wright, D.F. 1990.** Weights of evidence modelling: a new approach to mapping mineral potential. In: *Geological Survey of Canada Paper 8-9.* Agterberg, F.P. and Bonham-Carter, G.F. (eds). Ottawa, Canada, pp.171-183.
2 **Brabb, E.E. 1984.** Innovative approaches to landslide hazard and risk mapping. Proceedings 4th International Symposium on Landslides, Toronto, Canada, Vol. 1, pp 307-324.
3 **Carrara, A. 1983.** Multivariate models for landslide hazard evaluation. Mathematical Geology, 15, No. 3, pp. 403-427.
4 **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, Vol. 16, No. 5, pp. 427-445.
5 **Chung, C.J. and Fabbri, A. 1993.** The representation of Geoscience information for data integration. Non renewable resources, Vol. 2, No. 3, pp 122-139.
6 **Varnes, D.J. 1984**. Landslide hazard zonation: A review of principles and practices. UNESCO, Natural Hazard No. 3, 61 pp.
7 **Westen, C.J. van 1993.** Application of geographic information systems to landslide hazard zonation. ITC-Publication Nr 15, ITC, Enschede, The Netherlands, 245 pp.
8 **Yin, K.L. and Yan, T.Z. 1988.** Statistical prediction model for slope instability of metamorphosed rocks. In: *Proceedings 5th International Symposium on Landslides, 10-15 July 1988, Lausanne, Switzerland*. C. Bonnard (ed). Balkema, Rotterdam, Vol. 2, pp 1269-1272.