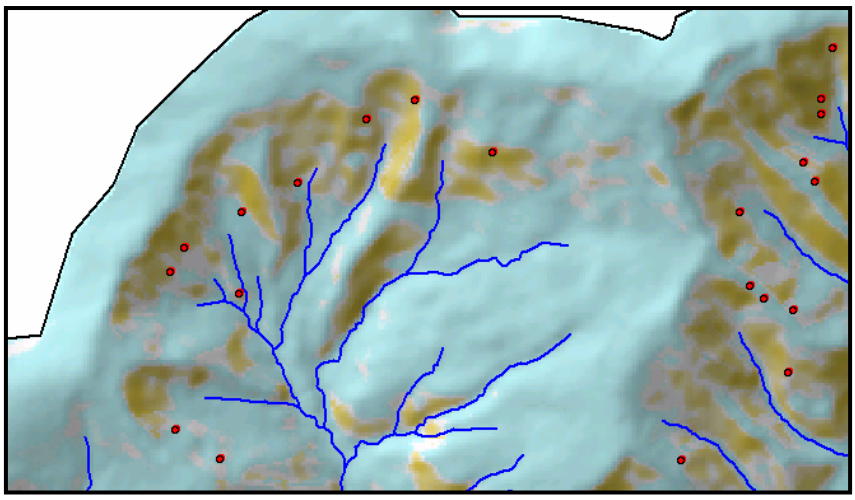
DAMOCLES

DEBRISFALL ASSESSMENT IN MOUNTAIN CATCHMENTS FOR LOCAL END-USERS

Contract No EVG1 - CT-1999-00007

LANDSLIDE HAZARD MAPPING BY MULTIVARIATE STATISTICS: COMPARISON OF METHODS AND CASE STUDY IN THE SPANISH PYRENEES



Santiago Beguería & Adrián Lorente

Instituto Pirenaico de Ecología, CSIC, Campus de Aula Dei, Apartado 202, 50080-Zaragoza, Spain.

This paper, written as a deliverable of the DAMOCLES project, is a review of the different existing methodologies to landslide hazard mapping by multivariate statistics. Within the DAMOCLES project, multivariate statistical models have been applied to different study regions in Italy and Spain. The experience gained has allowed to write this revision, addressing to the differences and advantages of the different tested procedures.

I. Comparison of methods

Introduction. Natural hazards modelling

Natural hazard has been defined as the probability of occurrence of a potentially damaging phenomenon in a given area and in a given period of time (Varnes et al., 1984). Thus, any natural hazards mapping project might address and give answers to the three key questions: the magnitude, the location and the time recurrence of the dangerous process.

The type and magnitude of landslides can be assessed by traditional geomorphological survey, normally based on field work and aerial photo interpretation. Geomorphological work usually includes the mapping of the observed landslides, so it also constitutes the principal input to natural hazard modelling. In the early stages of the research, however, geomorphological work should be addressed to the identification and classification of the landslides present in the study area. A correct discrimination between different types of landslides is very important, as every type is governed by different physical processes, and so should require a different modelling approach. Also, the magnitude of the process can help or even influence important methodological decisions, like the spatial design of the model.

The temporal framework of landsliding, as it has been seen before, is a difficult task most of the time. In some places, historical records exist on the occurrence of landslides, especially if personal or economical damage is involved. If such a record exists, assessing the recurrence of landslides is relatively easy. When no written records exist, geomorphological techniques can be used to fix the chronology of landslides. These techniques can include dendrochronology, lichenometry, isotope-dating, etc. For relatively frequent processes, like shallow debris-flows or soil-slips, diachronic

mapping using sequential aerial photos can perfectly resolve the question of the time recurrence of the process.

Once defined the type and magnitude of the landslides and their temporal recurrence, hazard mapping can be addressed. Below the different procedures for hazard mapping are the same basic assumptions:

- Slope failures leave a discernible morphological signature in the landscape, what permits the identification and mapping all the landslides occurred in a certain time period.
- Landsliding will occur under the same conditions as in the past.
- The basic mechanisms that have leaded to the observed landslides can be determined. The instability factors can be assessed, and hazard can be evaluated.

Hazard modelling procedures

The development of GIS has greatly improved the possibilities of hazard modelling, and many different approaches have been described since the end of the seventies. The different landslide hazard evaluation procedures can be classified in two main groups: qualitative and quantitative methods (figure 1). Qualitative methods include geomorphologic mapping and heuristic or index based (weighting of different thematic layers) approaches. They are very flexible and permit a complete inclusion of expert knowledge. The main pitfall is that they involve a great level of subjectivity, so the maps produced by different researchers can be very different. Quantitative methods include statistical modelling as well as process based or geotechnical modelling, and recent approaches based in neural networks. Although a completely objective procedure does not exist, quantitative methods assure that the same results can be achieved provided the same basic assumptions. A detailed analysis of the two groups can be found in Guzzetti et al. (1999) and King & Zeng (2001).

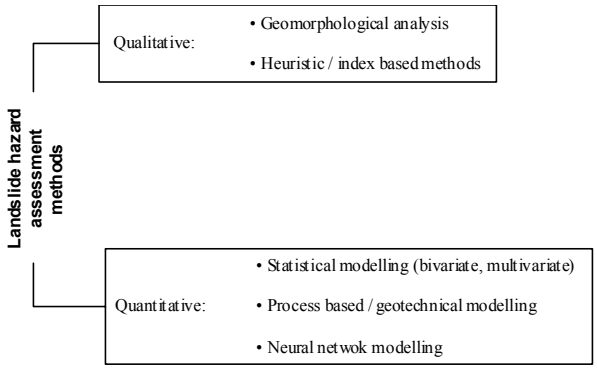


Figure 1. Different landslide hazard assessment methodologies.

This report is specially devoted to the statistical modelling of landslides. Statistical models are based on contiguity analysis of the observed landslides and a set of variables that can potentially be considered instability factors. Statistical approaches are data-based, 'black box' models, and their conclusions do not imply cause-effect relationships (but they can give certainty to well posed hypothesis). Provided sufficiently good input variables, statistical modelling can successfully shape the hazard of landsliding in a given area; however, its conclusions can hardly be applied to different places, or used to test simulation scenarios. This constitutes, probably, the main drawback of statistical procedures.

On the other side, geotechnical approaches are based on the basic physical principles that govern landsliding; i.e., they are process-based. This means that their findings can be applied to many different situations, because the same processes must be in action. Geotechnical models, however, usually require intensive parameterisation. Many of the physical variables that are necessary for running these models are not usually available, and their acquisition is often very costly. For this reason, geotechnical models are normally used at detailed scales, and for testing different scenarios. The statistical approach is, thus, specially useful in regional hazard assessment.

The literature presents different multivariate statistical approaches with potential use for landslide hazard assessment, including linear regression, discriminant analysis and logistic regression. The nature of the dependent and independent variables must suggest the selection of the most appropriate model. Discriminant analysis and logistic regression should be used if the predicted variable is the presence or absence of landslides within a given mapping unit (binary, or dummy, variable). Among the two methods, logistic regression handles better with categorical or binary variables, and is

robust to the violation of the multinormality assumption. Multiple linear regression is designed for the case when the predicted variable is continuous, like for example the density of debris flows in a given land unit.

Spatial design

In the spatial domain the researcher need to be sure that all the geo-hydrological heterogeneity in the area is adequately represented in the model. Basically, two main approaches have been used in hazard modelling: distributed and lumped models (figure 2). In a distributed model the variables are considered to present a continuous variation in space, whereas in lumped models space is subdivided into regions based on certain hydrological or morphological criteria. This refers directly to the question of the of the mapping unit, a topic that has been discussed by several authors (Carrara *et* al, 1992; Guzzetti *et al*, 1999; Aleotti & Chowdbury, 1999). Mapping units are portions of the land surface that are considered homogeneous, and are assigned a unique hazard value, so they are the minimum meaningful spatial units in the analysis. The selection of the mapping unit is not trivial and has important conceptual and practical implications.

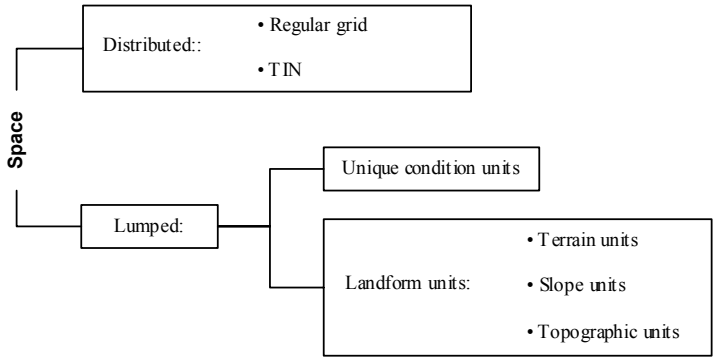


Figure 2. Consideration of space in statistical hazard modelling.

A grid (also known as raster format) consists in a regular orthogonal pattern that divides the space in small units called cells. If the grid resolution is adequate to the scale of the analysis, it can be considered a finite approximation to a continuous field. A single cell (the mapping unit) can be thus considered an integrated approximation to the exact value of the variable at this location. For this reason, grids are specially adapted to the representation of continuous distributed variables, like elevation or slope (figure 3). The grid format offers many advantages due to the simplicity of operation through

matrix algebra, and has been used by many researchers in heuristic or statistical analysis.

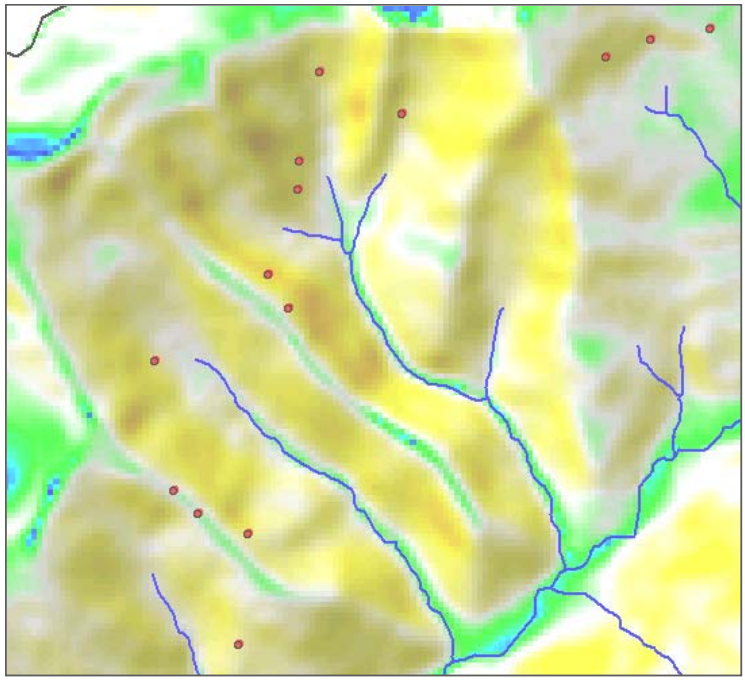


Figure 3. Grid or raster representation of a continuous variable (slope).
The red dots show the location of debris flows scars.

Other researches, however, have preferred the aggregation of the land surface in discrete units, what is often called a spatially lumped model. Different ways to define the land units have been proposed. Unique condition units are constructed by the overlay of different categorical maps, so each map unit is defined by a unique combination of attributes. The procedure permits to construct perfectly homogeneous units, but unique conditions lack the main advantage of landform units, which is a conceptual correspondence between mapping and process units. Landform units, however, are constructed based only upon morphological information, and so they can have a physical meaning that the other mapping units lack. The delineation of terrain units is done manually by the geomorphologist using aerial photos or detailed topographic maps. This approach permits to introduce the expert knowledge on the process, but introduces a high level of subjectivity, as demonstrates the work of Van Westen *et al.* (1999). More objective methods have been outlined, like the slope units approach (Carrara *et al*, 1995). Slope units are automatically constructed by the intersection between drainage lines and divides. Topographical units, proposed by O'Loughlin (1986), are based on a similar, but more detailed, approach (intersection

between contours and flow lines), and have been used basically in surface runoff generation and surface erosion models.

An example of a slope-units aggregation procedure is shown in figure 4, where every slope unit is given a different colour. It can be seen that slope units integrate a variable portion of the landscape, and so can greatly diverge in size. The physical sounding of the procedure is evident, with every land unit consisting on the right or left slope draining to a given flow line. The use of land units forces to integrate the spatial information of distributed variables. For example, if the distribution of debris flow scars is considered, only the number of them observed in each slope unit can be reported, with the loss of information on the exact spatial location.

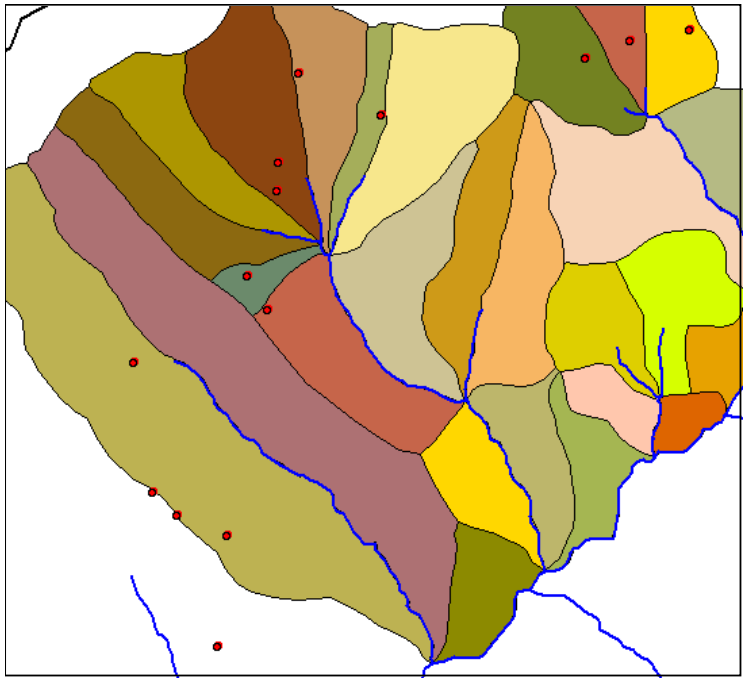


Figure 4. Space aggregation in slope units. The red dots show the location of observed debris flows.

The adoption of a particular mapping unit implies important conceptual and practical questions. Probably, the most important question is the consideration of hazard. In a grid based approach, landslide hazard, or the probability of occurrence of a landslide in a point within a given time period, is treated as a continuous distributed variable. This implies that, in theory, the final user of the map can know exactly the probability of landsliding in a given place. If the user is not interested in a single point, but in a wider area, it is possible to integrate the probabilities over all the surface, obtaining the expected number of land movements in a given time period. In a lumped

model, however, hazard is considered a continuous, spatially aggregated, variable. The same probability of experiencing landslides is given to the entire land unit, and abrupt changes may occur between adjacent units. The final map is a zonation of the entire area into homogeneous landslide hazard units.

This links directly to the spatial resolution of the final maps. Generally speaking, grids offer the maximum resolution over the rest of the procedures. Although lumped procedures allow to estimate the probability of finding a slope failure within a given slope (or a number of them, or a given proportion of terrain affected by them), they provide no information about which part of the slope is more likely to be affected. The more sharp look of grid maps, however, does not necessarily imply that they are more accurate. The minimization of mapping errors in the inventory stage is particularly critical in grid based approaches, and the final map would be as accurate as the worst of the input layers. In spatially integrated models, however, a certain degree of uncertainty is tolerated, if it does not imply the assignation of the slope movement to the wrong land unit.

The adopted procedure not only affects the consideration of the response variable (the probability or hazard), but also involves the treatment of the independent variables. As it has been said before, grids are specially adequate to the modelling of continuous variables. Categorical variables can be easily adapted to a grid format, as well. Land units do not work well with continuous variables, on the other hand. In the unique conditions approach, where the land units are defined by the intersection between categorical layers, continuous variables must be categorised. This must be done, preferably, according to a previous exploratory analysis. In landform units approaches, continuous variables can only be treated by statistics describing the distribution of the variable within the land unit (mean slope, maximum slope, etc.). Reducing a continuous variable to a discrete or categorical scale, or to a statistic, implies a great loss of information. Even categorical variables must be transformed in a landform units approach. As landform units are not defined upon the independent variables, but are based on morphological considerations, they are not usually homogeneous. Landform units usually contain different configurations of the landscape, like different lithologies or vegetation types, and this introduces new problems. Normally, the statistic used is the fraction of the surface affected by each situation. A scheme of all this considerations is shown in table 1.

Map unit	Variable type		
	Categorical	Ordinal	Cuantitative (interval or ratio)
Grid based	Discrete *Distributed*	Discrete *Distributed*	Continuous *Distributed*
Unique conditions	Discrete *Lumped*	Discrete *Lumped*	Discrete *Lumped*
Landform units	Statistic *Lumped*	Statistic *Lumped*	Statistic *Lumped*

Table 1. Treatment (scale, normal, and spatial representation, italic) of the different variable types according to the selected mapping unit.

Despite this methodological considerations, the final decision about the mapping unit should be conditioned by the characteristics of the process being modelled. It should be a concordance between the analysis and the process unit. For example, small movements like shallow debris flows or soil slips can be adequately represented as points in a grid, provided that the resolution of the cells is similar or greater than the mean size of the movements. Modelling bigger landslides, however, is more problematic within a grid, as it implies depicting a single movement into many different cells. In the statistical package those cells would be considered random independent variables, what is not the case; evidently, the cells that belong to the same mass movement must be considered the same thing. As big landslides normally affect an entire slope, a landform lumping procedure seems more appropriate to them.

Summarizing, the selection of the mapping unit depends on the type and size of the process being modelled, the final scale of the model, the scale of the available information, and several methodological considerations.

II. Case study: debris flow hazard modelling in the Garcipollera valley

The precedent considerations are applied in this section to a case study in the Garcipollera valley (central Spanish Pyrenees).

The Garcipollera valley (54.6 km^2) corresponds to the catchment of the Ijuez river, a small tributary of the Aragón river. It is a mountain catchment, with altitudes ranging between 800 and 2200 m a.s.l. The entire catchment lies in the Eocene flysch sector of the Pyrenees. The mean annual precipitation is about 1070 mm at the Bescós de Garcipollera climatic station (905 m), and more than 1300 mm can be estimated in

the highest parts. The catchment presents actually a relatively dense vegetation cover, represented by pines (65.5% of the total surface), oaks and beeches (6.5%), dense scrubland (8.2%) and mountain pastures in the highest parts (5.3%). The 13.5% of the territory is occupied by bare soils or rock outcrops.

Magnitude and recurrence of landsliding in the Garcipollera valley

The flysch sector of the Spanish Pyrenees has been identified as prone for debris flows (Lorente *et al.*, 2002). The most part of the observed debris flows are of the hillslope type, what are one of the most common geomorphic phenomena in mountain areas (Innes, 1983; Johnson & Rodine, 1984; Blijenberg, 1998). They consist on a rupture area or scar, very similar to a shallow landslide; a tongue with lateral levees; and a frontal deposit.

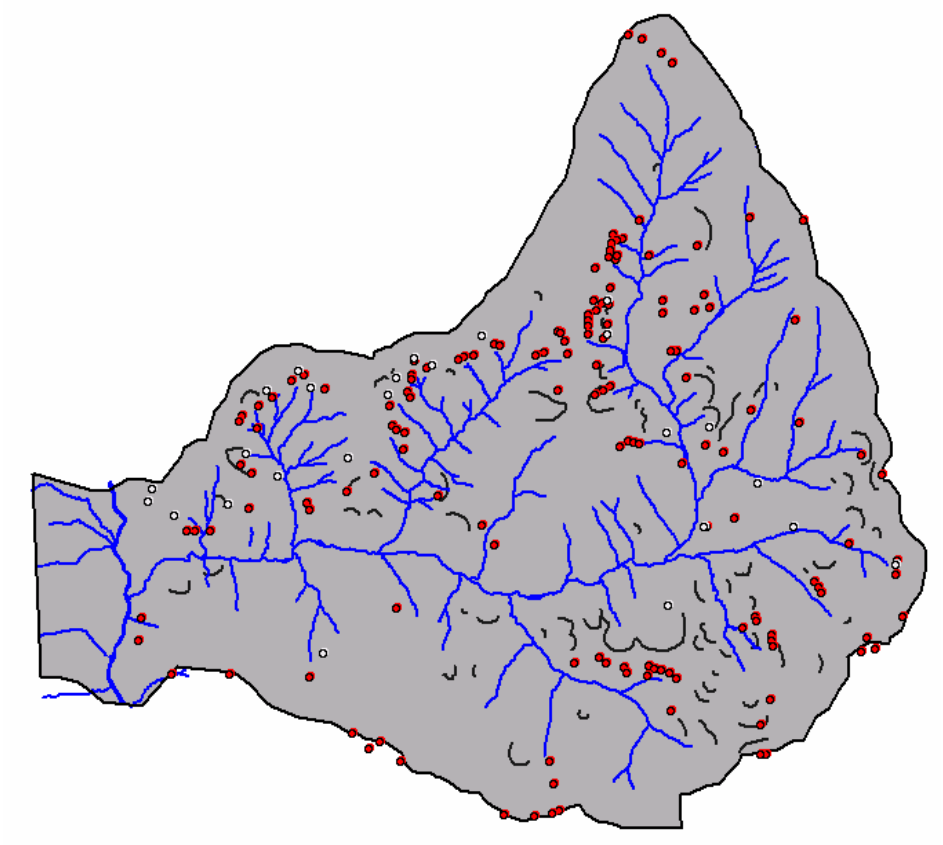


Figure 5. Distribution of debris flows (red dots: before 1990; white dots: 1990-2001) and rotational landslide scars (black lines) in the Garcipollera valley

The debris flows observed in the area are the typical small movements affecting mainly the soil and the regolith. The scar is usually small (15.9 m in average), and the

average total length is about 55.7 m. The figure 5 shows the incidence of debris flows in the area, from the aerial photos of 1956, 1977 and 1990, completed by field survey on 2001. The debris flows scars have been represented as points.

The sequence of aerial photos and the field campaign has allowed to assess the timing or recurrence of debris flows in the area. A wider area that comprises the Garcipollera valley has been used, in order to obtain more general results about the flysch sector. Fig. 6 shows (black dots) the cumulative number of debris flows observed in different moments. A linear disposition of the dots can be clearly seen, and demonstrated by the high coefficient of determination of the adjusted line ($r^2 = 0.997$). The high linear trend on the occurrence of debris flows demonstrates that, far from being a rare phenomena, shallow landsliding is a relatively common and constant process in the Ijuez catchment (and, generally speaking, in the flysch sector of the Spanish Pyrenees). The slightly lower than expected number of debris flows mapped in 2001 can be attributed to the change in the methodology, as field recognition mapping is less exhaustive than aerial photo analysis. The mean rate of occurrence is 3.417 debris flows per year.

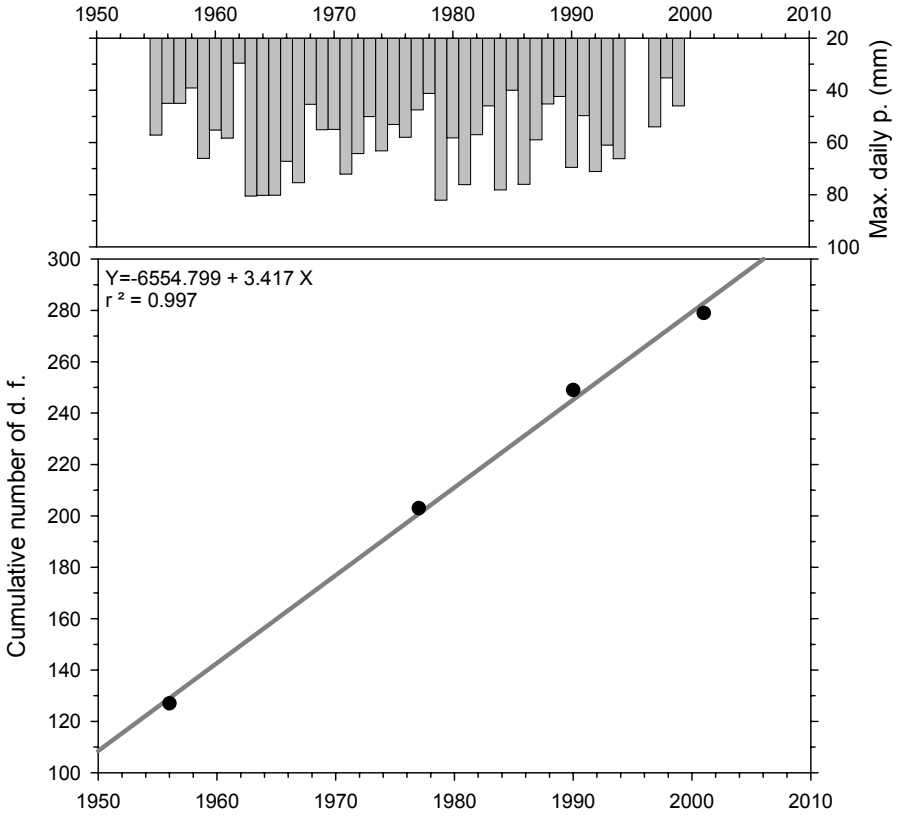


Figure 4. Cumulative number of debris flows in the Garcipollera valley, and annual maximum precipitation series at Bescós station

The analysis of the sequential aerial photos has also confirmed the important changes in land cover that occurred after farmland abandonment of the Ijuez catchment and the human-induced reforestation during the fifties. It is noticeable that, despite the great changes in land cover, the timing of debris flow occurrence does not show any change. This lessens the effectiveness of reforestation as a debris flow mitigation practice in the area, and enhances the importance of other factors like topography or soils.

Variables in the model and spatial design

The variables entered in the model came from very different sources. A detailed digital elevation model (DEM) was constructed at a resolution of 10 m, from topographic maps at 1:10 000 scale. Several morphological variables were derived from the DEM: slope, aspect, planform and profile curvature, upstream slope length, contributing area and the topographic index. Several variables were log or power transformed to better adjust to a normal distribution. Several variables, referring mainly to the vegetation cover, were obtained from a Landsat TM summer image: the normalized difference vegetation index (NDVI), and the three first components of the tasselled cap transformation (namely brightness, greenness and soil humidity). Several thematic variables, like plant cover or the past land use, where obtained from thematic digital maps. Finally, the distance to a rotational landslide scar was calculated within the GIS. The complete list of the variables entered in the model is shown in table 2.

A 10 m grid format was selected for the model. This resolution allowed to fully exploit the detailed morphological information, as topographic variables like slope are known to play a very important role in debris flow triggering. The rest of the variables were adapted to that resolution. The grid format was considered optimum for this kind of process, as the size of the debris flows were similar to the grid cells. For that reason, each debris flow scar was recorded as a single pixel.

Group	Name	Variable	Type
Response			
	dfmodel	Debris flow	Dummy
Morphology			
	elev	Elevation	Quantitative
	slope	Slope	Quantitative
	slopex5	Slope x5	Quantitative
	cosasp	Cos (aspect)	Quantitative
	plancurv	Planform curvature	Quantitative
	procurv	Profile curvature	Quantitative
	lengthup	Upstream slope length	Quantitative
	lgcontr	Log (contributing area)	Quantitative
	topondx	Topographical index	Quantitative
	sqrndx	Sqrt (topind)	Quantitative
Satellite			
	ndvi	NDVI	Quantitative
	tcap1	Tasseled Cap 1	Quantitative
	tcap2	TC 2	Quantitative
	tcap3	TC 3	Quantitative
Thematic			
	vege	Vegetation	Categoric (8 cats.)
	fields	Abandoned fields	Categoric (4 cats.)
	south	South exposition	Dummy
	north	North exposition	Dummy
	west	West exposition	Dummy
Other			
	scardist	Scar distance	Quantitative

Table 2. Variables entered in the model

Logistic regression approach to rare events

Due to the binary character of the response and some predictor variables, and the dubious normality of some of the variables, a logistic regression procedure was selected. Logistic regression states that the natural logarithm of odds (logit) is linearly related to the independent variables:

$$\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_n \cdot X_n \qquad \text{(eq. 1)}$$

where π_1 is the probability of occurrence of a debris flow, X_n is a set of n independent variables, and β_n is a set of n+1 parameters. Developing expression 1:

$$\pi_1 = \frac{\exp\left(\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_n \cdot X_n\right)}{1 + \exp\left(\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_n \cdot X_n\right)} \qquad \text{(eq. 2)}$$

As landsliding is normally a rare event, the population can have hundreds or even thousands of times fewer events (ones) than non-events (zeros). This is specially true in grid or raster based models, but is also frequent in spatially lumped ones (based on

unique condition or landform units). It is well known that common statistical multivariate procedures, such as discriminant analysis and logistic regression, are designed to work with groups that are more or less equal in size. When dealing with rare events, like landslides, the groups tend to be very unequal, and the models tend to sharply underpredict the probability of rare events. This was the case for the Garcipollera valley model, where the pixels with debris flow were around 2.5 in 10 000 cases.

The same problem has been analysed by King and Zeng (2001). They propose a design based in endogenous stratified sampling, or sampling within categories of the dependent variable. The strategy is to select all the cases for which (Y=1) and a random selection of cases for which (Y=0). This sampling procedure is specially useful when, as is the case of landslide inventory, the researcher knows the exact proportion of ones in the population (prior knowledge). The number of zeros to collect is a decision of the researcher. A number of zeros ten times higher than ones has been used for the Garcipollera valley model.

The endogenous stratified sampling procedure requires correcting the estimated probabilities based on the prior information about the proportion of ones in the population. Derived from the work of King and Zeng (2001), the following correction has been used:

$$\pi' = \pi \cdot \exp\left(\frac{1-\tau}{\tau} \cdot \frac{y}{1-y}\right) \qquad \text{(eq. 3)}$$

π' being the corrected or posterior probability, π the estimated probability, τ the proportion of ones in the population or prior probability, and y the proportion of ones in the sample or sampling probability.

Exogenous sampling prior correction was probably first used by Prentice and Pyke (1979). Other correction procedure available for exogenous sampling is the weighting maximum-likelihood estimator formulated by Manski and Lerman (1977). For the Garcipollera valley model the former procedure has been selected due to its ease of use.

A forward stepwise procedure has been used to introduce the variables, with a probability to enter of 0.05. This procedure selects only the variables that significantly contribute to improve the model.

Results

The model results are shown in table 3. Apart from the intercept, four variables where selected by the stepwise procedure: slope (with a 5x5 filter), alpine pastures, south exposition and north exposition. Only the slope is a continuous variable, the rest being categorical. In the table are shown the β coefficients with their standard errors (s.e.), the Wald statistic (the squared ratio of β to s.e.) and its significance level, and the exp of β, or the change in odds for a unit increase in the independent variable.

Variable	β	s.e.	Wald	Sig.	Exp(β)
intercept	-11.833	0.528	117.722	0.000	1.629E-08
X_1: $slope_f$	0.142	0.018	62.340	0.000	1.153
X_2: alpine pastures	-1.363	0.414	10.822	0.001	0.256
X_3: south	0.472	0.220	4.581	0.032	1.603
X_4: north	-0.730	0.328	4.952	0.026	0.482

Table 3. Model results: coefficients

This last statistic shows the relative importance of the variables. The prior or mean probabilities are represented by the intercept, β_0. Its low value reflects the fact of the scarcity of positive cases in the data set (the value has been corrected using eq. 3). Slope and south exposition have a positive effect in the triggering of debris flows, whereas alpine pastures and north exposition have a negative effect. Slope is the most important variable in the model, as slopes in the Garcipollera valley range typically from 0 to 45 degrees. This means that $\beta_1 X_1$ can yield values in the range (0-6.39). The other variables, due to their binary character, can only yield the values of their β parameters.

A graphical interpretation of the model can be seen in figure 5. The horizontal line (a) represents the mean or prior probability of debris flow triggering in the area. The effect of slope is incorporated in the bold line (e); the change in the probability of debris flow triggering due to changes in the slope comprise three orders of magnitude. The rest of the lines in the figure represent the cases for the different categorical variables in the model. As they are binary variables, their only effect is changing the intercept of the model, but they do not affect the slope of the line. The most favourable case for debris flow triggering (south exposition) is represented by line f, whereas the safest case (north exposition with pastures) is represented by line b. The difference between the two cases is about one order of magnitude.

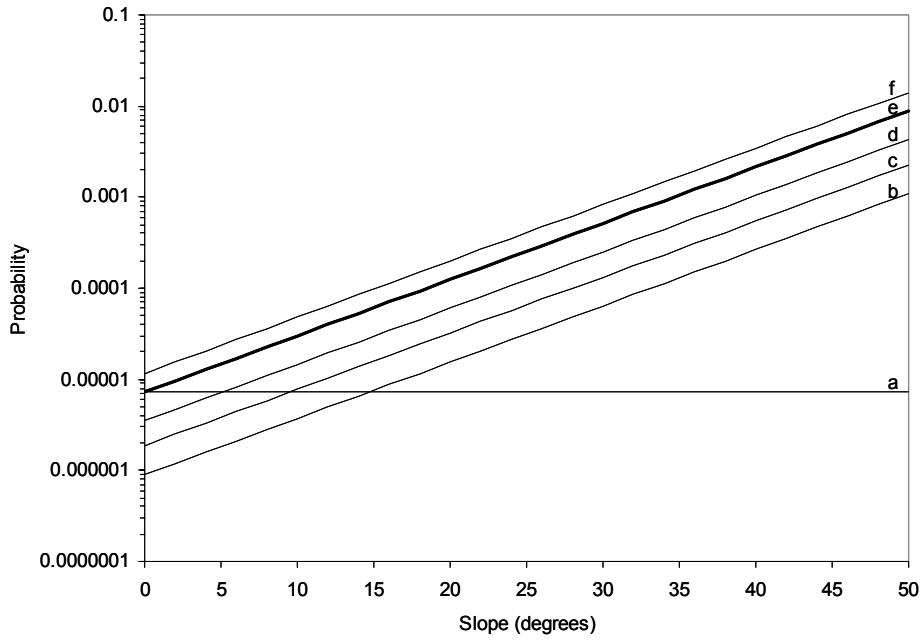


Figure 5. Graphical interpretation of the model. a: intercept (prior probability); b: most favourable case (slope * north * pastures); c: slope * pastures; d: slope * north exposition; e: slope only; f: least favourable case (slope * south exposition).

The temporal framework of the model equals that of the original sample, that is 33 years (1955-1990). So, the estimated probabilities are referred to this time period. It is easy to calculate a probability for a different time period of T years by multiplying the value yielded by the model by the correction factor T / 33.

Thanks to the implementation of the model in a GIS environment, a hazard map can be displayed (figure 6). In this map, the probability of debris flow triggering is shown by a colour ramp, and the exact probability of experiencing a debris flow at an exact point can be known.

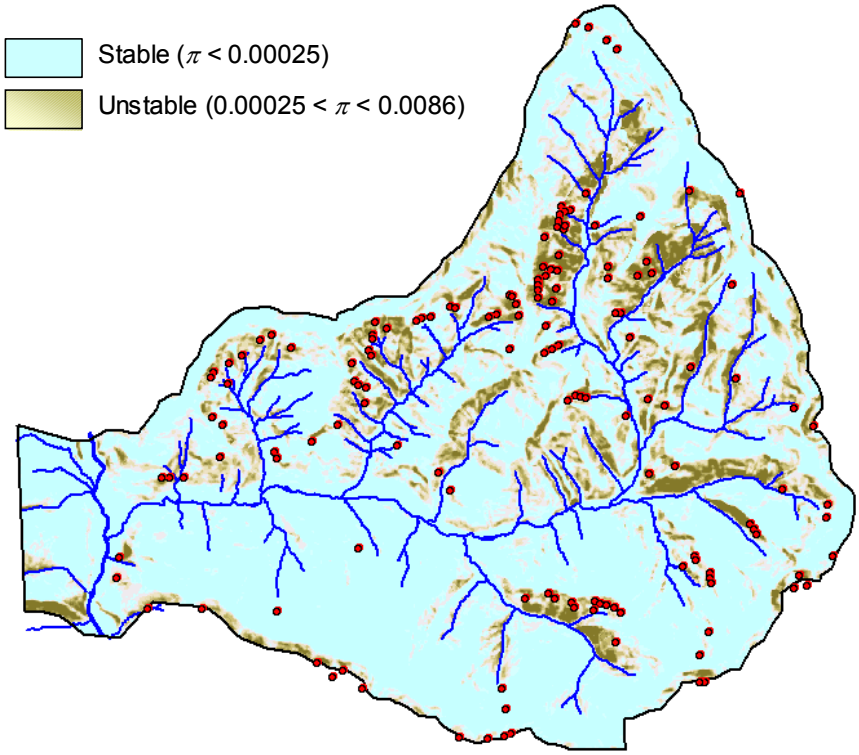


Figure 6. Debris flow hazard map of the Garcipollera valley. The red dots show the places where debris flows have been observed.

The low probabilities of the model are related to the resolution of the grid, that makes the proportion of ones very low. There is a dependency of the probabilities predicted by the model on the resolution of the grid, in such a manner that, the smaller the cells are, the lower are the values. This fact should not be considered a fault of the model, since is a very common mathematical property of any discretization procedure, as is a grid. A very well known example is the construction of a frequencies histogram, where the frequencies depend on the size of the sampling intervals used. Even if the histogram looks different when the sampling intervals are changed, the total number of cases is always the same. Analogously, in the hazard map is possible to integrate (add) the probabilities of a group of cells that form a spatial unit that interests the researcher. As normally the land planner is not interested in a single point, but in a given area (the place to build a house, or a mountain road reach, for example), this operation should be very normal in the use of grid based hazard maps. The operation is exemplified in figure 7. The at-point (cell) probabilities of three different locations are shown, and the integrated probability of a given sub-catchment. This last value can be considered as the

expected number of debris flows produced in the area within the reference period (33 years, in this case).

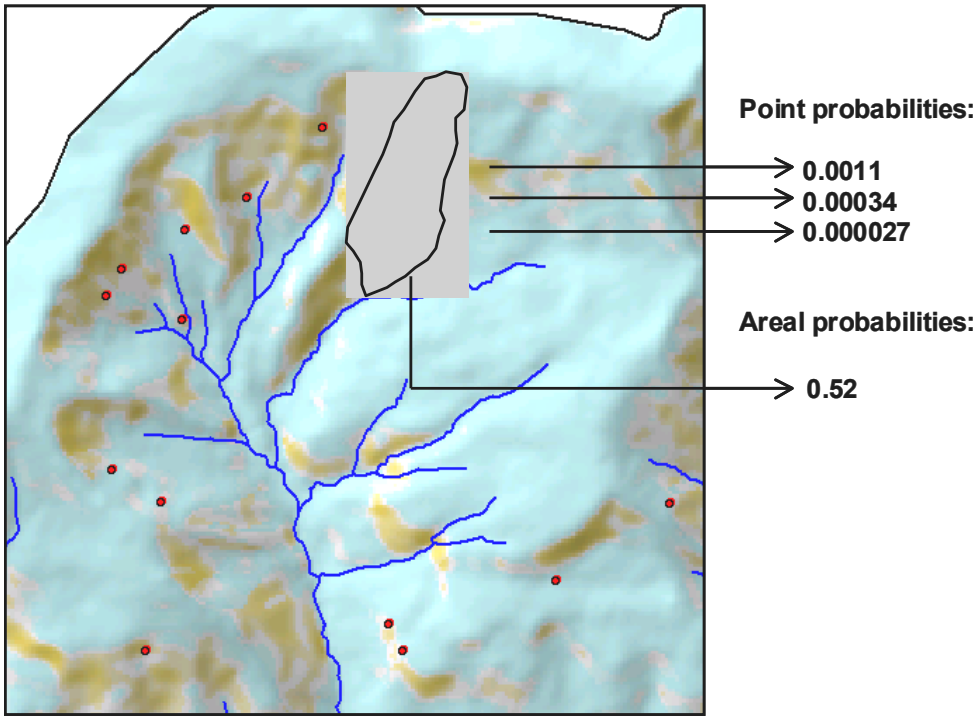


Figure 7. Point and areal probabilities estimated by the model (values referred to a 33 years period).

In the maps shown in figures 6 and 7 a distinction is made between stable (in blue) and unstable (values of brown) locations. This distinction has been done for visualization purposes, and does not imply a discrete zonation of the study area in safe and unsafe places. In fact, completely safe places do not exist, and every cell has a certain probability of experiencing landsliding (even though this probability can be very low). A confusion arises at this point, as logistic models are frequently used in a classification approach. This implies selecting a given value of the response variable (the probability of debris flows, in our case) and classifying all the cells in one of two groups according to it. The threshold value is normally the 0.5 probability, as usually the two sample groups are similar in size. Sometimes a third group, 'unclassified', is added, for the values around the threshold probability. For the case where the two groups are very dissimilar, the proportion of ones in the sample (y) should be used instead of the 0.5 value.

In table 4 a confusion matrix for the model is presented. The confusion matrix is used in classificatory approaches as a way to test the model performance, being very

similar to the *r* squared statistic in a linear regression model. The threshold value used has been the proportion of ones in the model sample. The average ratio is 0.68, what is a quite good value (values higher than 0.7 are considered good in most classification applications). The proportion of observed debris flows that have been correctly classified (0.76) is much higher than that of well classified zeroes (0.67). This would be normally considered a conservative model, as it has a higher number of type I errors (false positives). Specially in rare events modelling, type I errors can be considered as cases where a high (relative to the mean) probability of experiencing debris flow exists, but no events have been observed within the sample period, due to the rarity of the process. For this reason, the 0.67 proportion of correctly classified zeroes should not be considered a model flaw.

		Predicted 0	Predicted 1		
Actual	0	822	402	1224	**67%**
	1	33	103	136	**76%**
		855	505	1360	**68%**

Table 4. Confusion matrix.

An integration (addition) of the probabilities of debris flows triggering in the whole catchment can easily be made in a GIS. This value equals the expected number of debris flows in the study area during a period of time equivalent to the sample period. The integration of the probabilities of debris flow in all the Garcipollera valley yields an expected number of 150.52 debris flows, a value that is very close to the observed number, that is 136.

References

Aleotti P. & Chowdbury R. (1999), Landslide hazard assessment: sumary review and new perspectives, *Bulletin of Engineering Geology and Environment*, 58:21-44.

Carrara A., Cardinali M. & Guzetti F. (1992), Uncertainty in assessing landslide hazard risk, *ITC Journal*, 1992(2):172-183.

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

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

King G. & Zeng L. (2001), Logistic regression in rare events data, *Political analysis*, 9(2):137-163.

Lorente A., García-Ruiz J.M. & Beguería S. (2002), Factors explaining the spatial distribution of hillslope debris flows. A case study in the flysch sector of the Central Spanish Pyrenees, *Mountain Research and Development*, 22(1):32-39.

Manski Ch.F. & Lerman S.R. (1977), The estimation of choice probabilities from choice based samples, *Econometrica*, 45(8):1977-1988.

O'Loughlin E.M. (1986), Prediction of surface saturation zones in natural catchments by topographic analysis, *Water Resources Research*, 22(5):794-804.

Prentice R.L. & Pyke R. (1979), Logistic disease incidence models and case-control studies, *Biometrika*, 66:403-411.

Van Westen C.J., Seijmonsbergen A.C. & Mantovani F. (1999), Comparing landslide hazard maps, *Natural Hazards*, 20:137-158.

Varnes, D.J. (1984). *Landslide hazard zonation: a review of principles and practice,* Commission of Landslides of the IAEG, UNESCO, Natural Hazards No. 3, 61 pp.