On spatial uncertainty in hazard and risk assessment

A. G. Fabbri[1] & C.-J. Chung[2]
[1]*DISAT, University of Milano-Bicocca, Italy*
[2]*SpatialModels Inc., Canada*

Abstract

Exploiting the mathematical framework of favourability function modelling and of software designed for spatial target mapping, experiments are discussed to measure the effects of uncertainty of spatial support. A database that has been the focus of landslide hazard prediction and risk assessment of buildings, roads and land uses is reanalyzed gradually modifying the spatial characteristics of supporting patterns of the proposition of "*presence of landslide occurrences*." Boundary fuzziness of categorical mapping units and filtering of continuous fields, represent weakening of spatial relationships. Comparisons are made of ranges of ranks representing uncertainties of class membership in target patterns obtained by iterative cross-validations. The impacts of the different spatial uncertainties on risk assessment patterns are visualized using progressive combinations of uncertainties of target patterns to resolve risk equations and study the consequent changes in risk values.
Keywords: favourability modelling, target mapping, spatial uncertainty, spatial support, landslide hazard, landslide risk.

1 Introduction

It is now over fifty years that attention is being given to spatially distributed data for its integration into indices for development planning. Examples are resource exploration of mineral occurrences or locating hazardous sites such as zones likely to be affected by landslides or by floods. Most emphasis has been on methods of spatial statistics and on the introduction of expert's opinions to complement data scarcity or insufficiency.

The abundance of digital data today and the consequent feasibility of constructing well focused spatial databases for regional planning provide further


WIT Transactions on Information and Communication Technologies, Vol 47, © 2014 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/RISK140011

challenges: testing and interpreting the information content of the databases and navigating through the many options encountered to generate manageable representations. Instances of these are the spatiotemporal distribution of future discoveries or that of future hazardous occurrences. It means that not only a database must adequately and acceptably represent the distribution and dynamics of natural processes but also should allow measures of spatial uncertainty. In practice it must offer criteria for deciding on the relative significance and strength of the spatial support. The task becomes similar to the one in computer war games in military exercises where sequences of simulations are generated to predict consequences of modifying spatial relationships.

In this contribution the impacts of different spatial uncertainties on landslide risk assessment patterns are visualized using progressive combinations of uncertainties of target patterns to resolve risk equations and study the consequent changes in risk value distributions. A database that has been the focus of landslide hazard prediction and risk assessment of buildings, roads and land uses is reanalyzed gradually modifying the spatial characteristics of supporting patterns of the proposition of "*presence of landslide occurrences*." Following a description of the database, a brief introduction is provided to favourability modelling and its application software. Experiments are then discussed on landslide hazard prediction and risk assessment in the Deba Valley in northern Spain. The spatial data have been modified to weaken the spatial support and degrade the quality of prediction and assessment introducing levels of greater uncertainty. Conclusions are drawn on the importance of simulating spatial changes.

2 The Deba Valley spatial database

The Deba Valley is part of the Basque Province of Guipuzcoa up from the coast of northern Spain. The study area is located in the lower part of the valley and cover approximately 140 km^2, with maximum elevation just below 700 m a.s.l. Main annual rainfall reaches 1500 mm with episodes of over 100 mm/day every few years. Lithologies in the area are: limestone, marl, claystone, sandstone, flysch and volcanics, of the Cretaceous and Paleogene of the Basque-Cantabrian Pyrenees. Average slope gradient is about 22°, and regolith thickness ranges from 50 cm to 3 m. Shallow translational landslides and flows triggered by rainfalls are the most common types of mass movements in the area whose landscape is highly influenced by reforestation, cultivation, urbanizations, infrastructures and industrial activities. Population density reaches 500/km^2.

Remondo *et al.* [1–3], have constructed a digital database for landslide hazard studies later extended for risk assessment. Through photo-interpretation and field work, 1123 shallow translational landslides and associated flows were mapped and dated: 906 prior to 1997, and 217 for the period 1998–2001. The average size of their trigger zones is approximately 400 m^2, so that it was decided to represent each by a single picture element or pixel of 10 m resolution. The same digital resolution was used for the rest of the database that was to represent the typical setting of the landslides: 25 lithologic units, 9 land use classes (both

categorical), and elevation, aspect, curvature and slope (continuous fields). The study area falls within a rectangular raster of 1886 pixels by 1555 lines, and occupies 1,393,541 pixels.

Remondo *et al.* [2] first modelled landslide hazard in the area to predict the distribution of future likely landslides using the spatiotemporal distribution of the 1123 landslide trigger zones and their spatial relationships with the categorical and continuous digital maps. Later, Remondo *et al.* [3] obtained a semi-quantitative risk-assessment augmenting the database with socioeconomic spatial elements: 5 types of roads, 3194 buildings and 9 types of land uses. They were made into co-registered digital maps for which values and vulnerabilities were estimated and compiled as 10 m pixels.

The Deba Valley spatial database that is used in this contribution has also been turned into a case study for training decision makers [4].

3 Favourability modelling: approach, software and strategies

Favourability modelling is based on constructing a proposition within the study area as a mathematical statement that has to be proven true or false given the spatial evidence. Such evidence is established for the study area as a function of the spatial relationships established between a direct supporting pattern of the proposition, DSP, and indirect supporting patterns, ISPs. In the Deba Valley study area, for instance, the DSP is the spatial distribution of a well defined set of occurrences, such as the 1123 shallow translational landslides and associated flows. The ISPs are then the distributions of the various categorical mapping units and the values of the continuous fields, used to represent the typical settings of the DSP: lithologies, land use classes and values of continuous fields of topographic aspect, curvature, digital elevation and slopes.

The spatial relationships are conveniently calculated from co-registered digital images with a given common spatial resolution: in our case of 10 m for all DSP and ISPs. Example of a proposition is: "P_i*: a point i in the study area is affected by a part of a future landslide of type shallow translational landslide and associated flows | given the presence of the classes and values of spatial evidence.*" Different interpretations are possible for such a favourability function: possibility, likelihood, certainty, belief and plausibility, conditional probability and more, using the corresponding modelling assumptions. Here we will limit our analyses to the empirical likelihood ratio function or ELR, amply discussed elsewhere [5]. The categorical ISPs are transformed into normalized frequencies and the continuous fields into density functions, so that the ELR ratios are computed between the functional values in the presence of the landslide trigger zones with those in their absence. Ratios for overlapping ISPs are combined by means of the rules and assumptions of the model. ELR values range from 0 to ∞.

The application of a favourability function must be properly structured in time and space and can be used as predictor of future landslide occurrences via assumptions or scenarios of spatiotemporal nature, e.g., similarity of settings through the study area, of database sufficiency, or similarity of frequency of occurrence through time etc.

Criticism on applications of favourability modelling [6] apparently justified the programming of STM, a software for spatial target mapping developed for research and training [4, 7] as a tool accessory to conventional geographic information systems, GIS. Besides data transfer to and from GIS, STM provides several prediction models: fuzzy sets, likelihood ratio, linear and logic regression, and Bayesian probability functions. In addition it allows: spatial data input modifications, modes of iterative cross-validations, generation of prediction, target, uncertainty and uncertainty/target combination patterns, with associated statistics and prediction-rate tables. We will describe these in the application section that follows.

In particular, with the selection of ISPs as input to modelling, transformation parameters are available. Linked with the digital maps of categorical ISPs, such as lithology and land use, inputs are the respective boundary images of the units as thin lines of one-pixel width. A *number of pixels in neighbourhood* parameter allows to generate fuzzy boundaries of desired width, say 5, 10, 15 etc. For the continuous fields ISPs, a *spread parameter* of continuous values is used to obtain an appropriately smooth distribution before proceeding with further analyses. These two parameters are indicated as N and S, respectively, used for artificially fuzzyfying categorical boundaries and smoothing continuous fields.

Of a study area, STM generates equal-area rankings as classifications termed prediction patterns. The statistics for the DSP/ISPs spatial relationships can also be extended from a study area to another, assumed to have similar settings and to generate a prediction pattern there. Within a study area, the DSP can be partitioned to obtain sequences of prediction patterns and of associated statistics by iterative cross-validations according to one of the following strategies:

(a) Sequential selection of a given number of occurrences as DSP;
(b) Sequential exclusion of a given number of occurrences as DSP;
(c) Random selection of a given number of occurrences as DSP.

By cross-validation the statistics of the relative distribution of the occurrences in the equal area ranked classes of the prediction pattern is obtained as a prediction-rate table. The selected subset of the available occurrences is used to apply the modelling and the remainder is used to study their distribution across the equal area ranked classes obtained by modelling. From the iterative processing strategies in (a), (b) or (c), or other combinations of those, a sequence of prediction patterns is generated. A target pattern combines them by means of a number of optional statistics: basic statistics such as sample average and variance or sample median and range, jackknife statistics as average and range, or rank-based statistics as median range and range of ranks.

For instance, a target pattern can be a digital image with pixel values as means of a sequence of prediction patterns from iterations in (a), (b) or (c). An associated uncertainty pattern of class membership in the target pattern has ranked values that represent the standard deviation at each pixel. Furthermore, a given % of the lower rank values, corresponding to a narrower range in an uncertainty pattern, can be used to select only the values in a target pattern that

correspond to those lower uncertainty values, thus obtaining a combination pattern of uncertainty and target.

Iterative cross-validation strategies allow establishing the relative reliability, stability or acceptability of the prediction and target patterns. These are represented as prediction-rate tables, histograms and cumulative curves.

The associated SRA software for spatial risk analysis [6, 7] is complementary to STM in that it takes as part of the inputs a prediction pattern and a prediction-rate table. The table contains the cumulative frequency of occurrences for all equal area ranked classes of predicted hazard and is then converted, via a scenario, into a probability of occurrence histogram either directly or via a monotonically decreasing transformation or as a fitted function.

In addition SRA requires as inputs co-registered categorical images of socioeconomic indicators representing elements exposed to risk, such as roads, buildings, land uses (and persons in case of casualties), with accompanying tables of vulnerabilities and estimated values. From all those inputs SRA calculates Risk Assessment patterns and associated statistics, either for single element type or as aggregation of them.

The following section deals with experiments that exploit many of the features of STM and SRA described here. The purpose is to expose aspects that contain more general significance out of the Deba Valley database.

4 Weakening of spatial support

Experiments on landslide hazard prediction and risk assessment in the Deba Valley study area have been discussed in a number of studies [2–4]. Here we want to explore situations in which the support for spatial relationships becomes gradually weaker to simulate poor quality of prediction/risk patterns. This is to provide insight on the question: "when to decide that a database is too poorly representing future hazardous occurrence distribution?" Or, in other words: "how to compare a poor prediction with a good one?"

The application performed here uses the digital image with the one-pixel distribution of the *906* pre-'97 landslides and that of the *217* post-'97 as DSP that are shown in Figure 1a and 1b, respectively, with the study area as background. As ISPs to characterize the settings of the landslides it uses two categorical digital images, lithology and land use, termed *l* and *u*, and four continuous ones, aspect, curvature, digital elevation and slope, not shown here, termed *A, C, D* and *S*. The ELR model is first applied using the entirety of *1123* landslides as DSP and the *luACDS* as ISPs. This is because the prediction pattern obtained from them is the most detailed as it uses all the DSP data available. However, we do not know as yet the relative quality of its classes without performing their cross-validation. This is done by repeating the modelling using only the *906* pre-'97 landslides and verifying the ranking of its classes with the distribution of the *217* post-'97 landslides. Such a cross-validation is based on the time partitioning of the DSP and is the most natural when time partitioning is available.

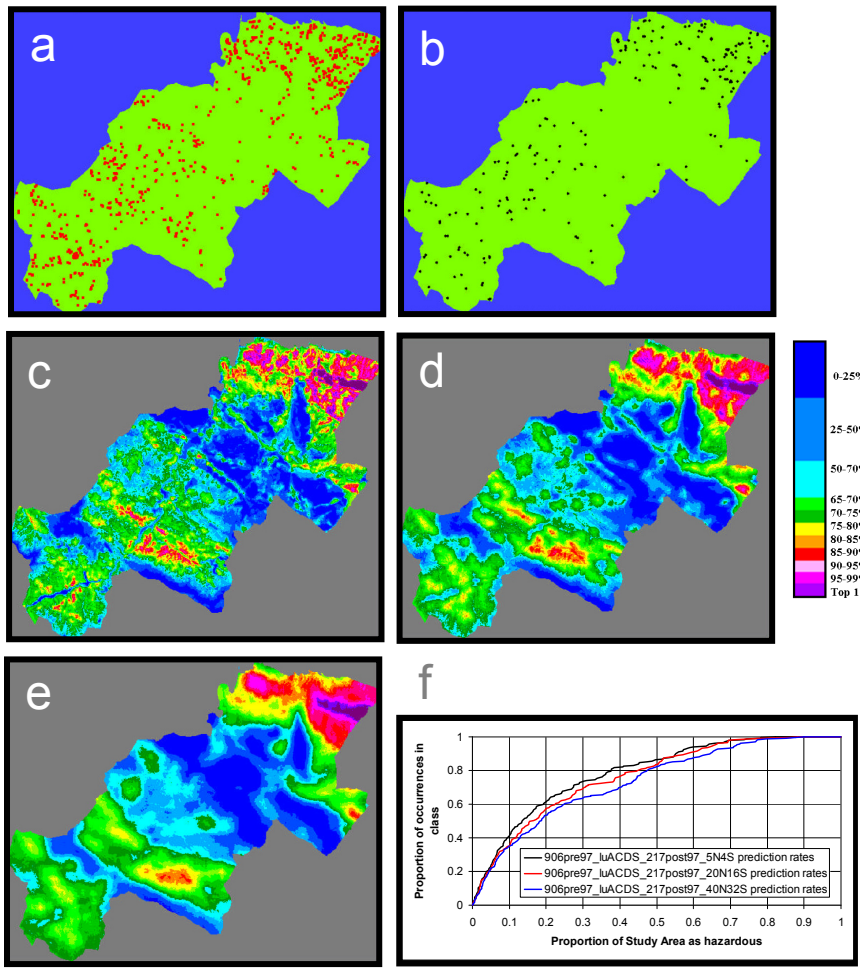

Figure 1: Deba Valley study area. Distribution is shown of the *906*-pre'97 shallow translational landslides and associated flows in red (a) and that of the *217*-post'97 in black (b). Sizes of trigger zones are exaggerated for visibility. Prediction pattern obtained using the entire distribution of the *1123* landslides as DSP and the *luACDS* as ISPs, with *5N4S* (c), *20N16S* (d), and *40N32S* (e). The legend shows pseudo-colours for different groupings of ranked classes. In (f) the three cumulative prediction-rate curves obtained using only the *906*-pre'97 landslides as DSP and the *luACDS* as ISPs, to predict the *217*-post'97.

Three analyses were made in which the sets of ISPs were modified as follows:

(i) Using 5 as *number of pixels in the neighbourhood* of categorical boundaries and 4 as *spread parameter* for continuous fields, *5N4S*;
(ii) Using 20 and 16, respectively, *20N16S*; and
(iii) Using 40 and 32, respectively, *40N32S*.

The corresponding ELR prediction patterns are shown in Figures 1c, 1d and 1e, as pseudo-colour images with selected groupings, fixed for visibility, of 200 ranked classes of equal area, each of 0.5% of the study area as default. We can see the loss of detail due to the parameters *5N4S*, *20N16S* and *40N32S*.

To study the relative quality of those prediction patterns, three more prediction patterns were obtained using only the *906* pre-'97 as DSP and representing the distribution of the remaining *217* post-'97 in them as cumulative prediction-rate curves, as shown in Figure 1f. On the horizontal axis we have the proportion of study area classified as hazardous in decreasing order of 200 ranked values and on the vertical axis the corresponding proportion of the *217* post-'97 landslides. The steeper is the curve at the origin the better is the prediction pattern: the cumulative curve can be interpreted in terms of cost-benefit. We can easily see that prediction-rate curve with *5N4S* is better than that with *20N16S*, which in turn is better that that with *40N32S*. However, we can see that even prediction-rate curve with *5N4S* indicates a not particularly good prediction. It is due to the variety of geological/land use settings of the study area elongated from the sea coast line to the northeast to inland towards southwest. ELR values are higher for muddy flysh, marly limestone and calcareous flysh lithologies, and lower for other marly units, poorly graded gravel and silty sands, and again higher for grasslands and cultivated areas. All other categorical units and all continuous fields contribute marginally to the ELR values.

In our analyses, for instance, 0.1 of the study area with the highest predicted values contains 0.42 of the *217*, 0.2 contains 0.61 and 0.3 contains 0.74 with *5N4S*. The corresponding proportions with *20N16S* and *40N32S* are 0.37, 0.58, 0.70, and 0.36, 0.54, 0.60, respectively. The cumulative prediction-rate curves in Figure 1f are critical to estimate the probability of occurrence for each class and each pixel for risk assessment later on using an appropriate scenario. We can consider this part of the analysis as preparation of hazard prediction for risk assessment. Let us now consider a complete procedure for it in eight steps. Table 1 describes the steps to be considered of a general framework for favourability modelling of hazard and risk.

In the procedure described in Table 1, the eight steps require decisions, selection of parameters, testing of different alternatives, assumptions and scenarios. For instance, besides the selection of the study area, what becomes critical are those of choosing the DSP, the ISPs, of partitioning of the study area and/or the DSP, of selecting a threshold for the uncertainty of class membership and of interpreting the classes of costs in the Risk pattern.

Table 1: Eight steps for modelling hazard prediction and risk assessment with favourability modelling, applied to *luACDS* ISPs with parameters *5N4S*, *20N16S* and *40N32S*, using *1123*, *906* and *217* landslide distributions as DSP.

Step	Description
1	**Prediction pattern 1**: use distribution of set of **1123** landslides as DSP and the six ISPs, **luACDS,** to be used in all predictions. Figs. 1a, 1b.
2	**Prediction pattern 2**: use the distribution of only the **906** pre-'97 landslides and cross-validate with that of the **217** post-'97 to get the prediction-rate curve. It provides a measure of relative quality. Fig. 1f.
3	**Target and Uncertainty patterns**: iterative predictions and cross-validations to establish prediction robustness. Example: exclude successively 50 landslides from the **906** pre-'97 and generate 18 prediction patterns. Using rank-based statistics, the median values of the 18 values per pixel provides the Target and the range of ranks the Uncertainty pattern of class membership. Figs. 2ab, 3ab and 4ab.
4	**Combination pattern of Uncertainty and Target patterns**: selection of reasonable or convenient proportion of lower values of Uncertainty to isolate the corresponding Target values, e.g., 50%. Figs. 2c, 3c and 4c.
5	**Introduction of socioeconomic elements at risk**: categorical digital images of identical resolution with a associated tables of values and vulnerabilities, for roads, buildings and land uses. Not shown here.
6	**Modelling the prediction-rate table**: table from **Step 2** is transformed into a table of probability of occurrence for each class and for each pixel, if necessary converted to a monotonically non decreasing function and/or by a fitted exponential function. Figs. 2d, 3d and 4d. The example of scenario used assumes that an occurrence rate of **217** one-pixel landslides in 4 years remains constant for the next 50 years since 2001 (217/4 = 54.25 x 50 = 2712.5 ≈ 2713 one-pixels landslides expected).
7	**Risk pattern**: it is generated using the prediction pattern from **Step 1**, the elements exposed, values and vulnerabilities from **Step 5**, and the modelled probability of occurrence function from **Step 6**. The risk equation is computed and aggregated for roads, building and land uses. In it classes of expected damage per 10 m pixels in € are tentatively reclassified in 10 groups from 100 classes. Figs 2e, 3e and 4e. Only the highest 4 1% classes have values > 1 € and are indicated by black pixels.
8	**Combination pattern of Uncertainty and Risk pattern**s: use the Combination pattern of 50% lower values of uncertainty in **Step 4** to obtain the corresponding Combination Uncertainty/Risk pattern with values corresponding to relatively low uncertainty. Figs. 2f, 3f and 4f.

The Risk patterns in the three experiments are visibly different, however, in all of them it is the highest 4 classes that represent risk values greater than 1.00 €/pixel. Table 2 shows the respective expected cost of a pixel for those classes. The expected total costs for the Risk patterns in the study area are as follows: 9,104,146 € (*5N4S*), 9,649,109 (*20N16S*) and 12,816,878 (*40N32S*).

Table 2: Expected cost of a pixel > 1 € for the four highest risk classes in experiments *5N4S*, *20N16S* and *40N32S*.

Class	€	5N4S €	20N16S €	40N32S €
97	> 1	1.64	1.08	1.50
98	> 2	2.94	2.33	2.68
99	> 5	5.29	5.23	5.04
100	> 10	11.98	16.39	12.33

5 Concluding remarks

Many aspects of spatial prediction modelling provide challenges that demand experimentations on representative databases. Generalizing the procedure applied we can question some or all of the following points:

(i) *Database*: Study area, direct and indirect supporting patterns, homogeneity of occurrences and settings, time/space partitions for cross-validation.
(ii) *Prediction patterns*: How good are they? Did we overfit? Is it enough to use prediction-rate tables via relative ranking?
(iii) *Modelling*: What happens if conditional independence assumptions, required by most models, are not respected?
(iv) *Scenarios*: How to formulate and construct them to estimate the probability of occurrence essential for risk assessment? How far back in time should we look at the statistics of occurrences and to predict how far in the future?
(v) *Expected costs*: How representative are such costs? How to classify and interpret them, and display their spatial configuration?

Because of these many challenging aspects due to variability of all points discussed, we consider empirical analyses of databases more in demand than the development of new mathematical models. The distribution and sharing of spatial databases among expert organizations or researchers is perhaps a more promising way to deepen insight on how to manage spatial data for modelling the future as to the discovery of new natural resources or to the identification of future hazardous zones. Such a sharing is on its own a considerable challenge.

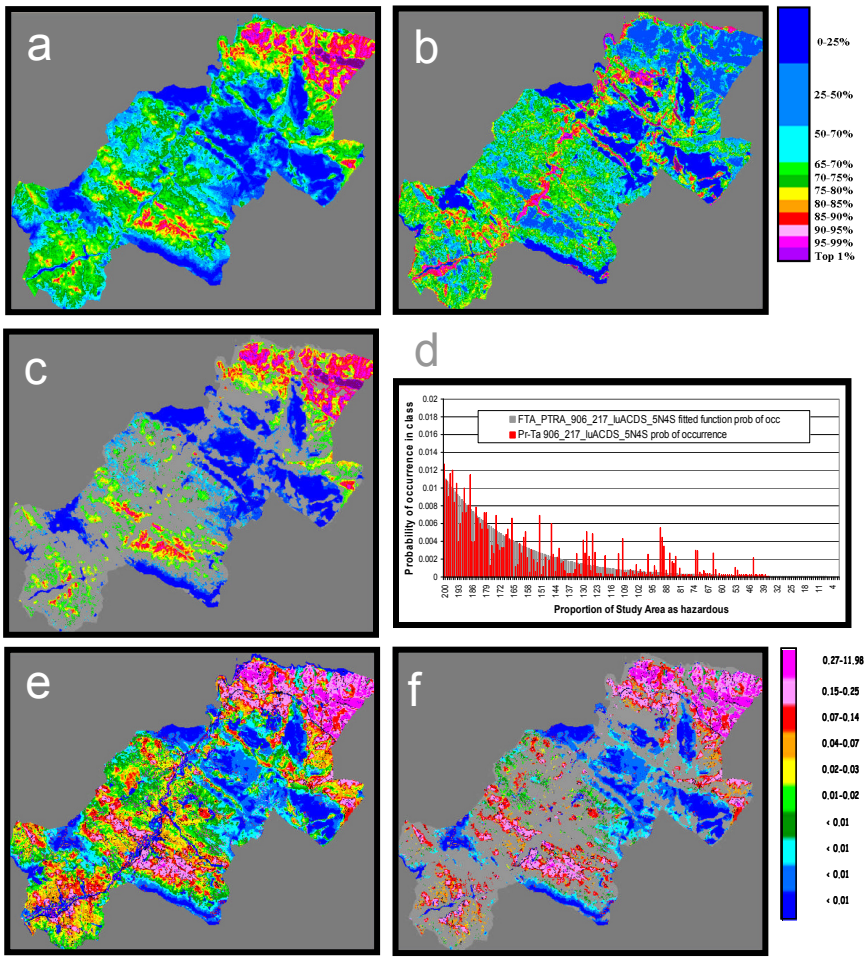

Figure 2: Target, Uncertainty and Risk patterns with probability of occurrence histograms for the Deba Valley study area for analyses with *5N4S* parameters for the *luACDS* ISPs. Target pattern obtained using the distribution of the *906*-pre'97 landslides and the 906-50 x 18 iterative cross-validation (a), the Uncertainty pattern (b) and the 50% Uncertainty/Target Combination pattern (c). In (d) is the transformation of the corresponding prediction-rate in Figure 1d, into a probability of occurrence histogram in red, and in gray the fitted function to its monotonically decreasing transformation. The Risk pattern is in (e), with the four classes with estimated costs > 1 € in black, and the corresponding 50% Uncertainty/Risk Combination pattern of (c) in (f). In the risk legend are ten classes with expected costs in €.

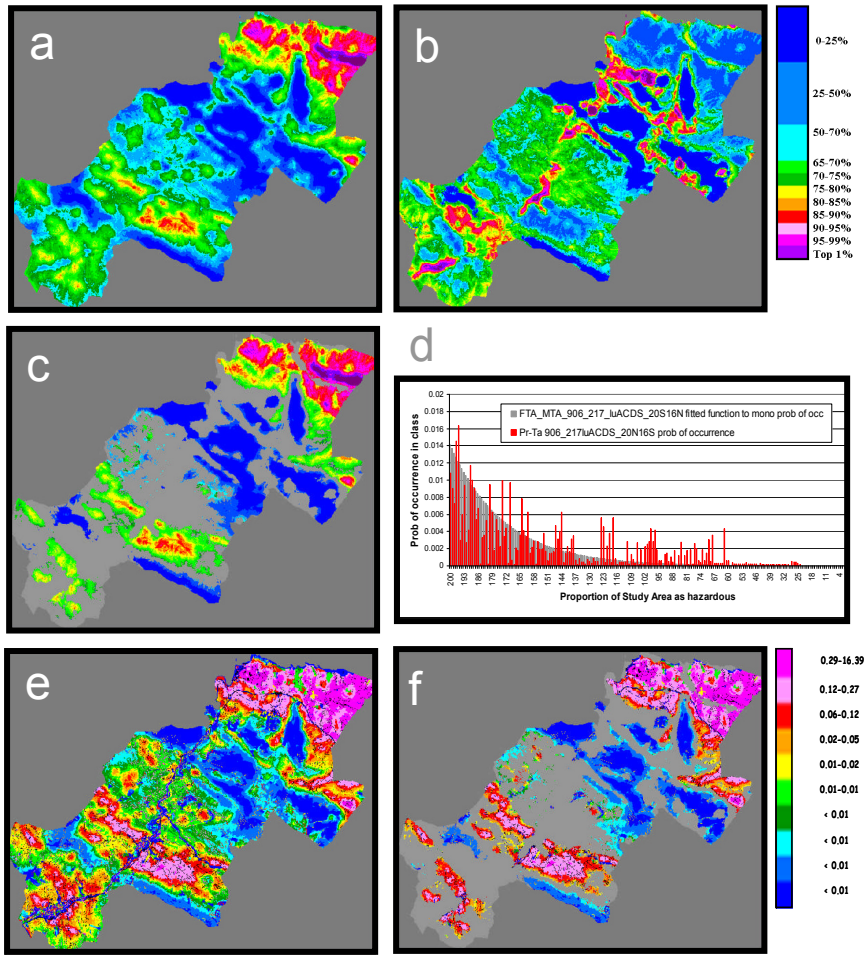

Figure 3: Target, Uncertainty and Risk patterns with probability of occurrence histograms for the Deba Valley study area for analyses with *20N16S* parameters for the *luACDS* ISPs. Target pattern obtained using the distribution of the *906*-pre'97 landslides and the 906-50 x 18 iterative cross-validation (a), the Uncertainty pattern (b) and the 50% Uncertainty/Target Combination pattern (c). In (d) the transformation of the corresponding prediction-rate in Figure 1d, into a probability of occurrence histogram in red and in gray the fitted function to its monotonically decreasing transformation. The Risk pattern is in (e), with the four classes with estimated costs > 1 € in black, and the corresponding 50% Uncertainty/Risk Combination pattern of (c) in (f). In the risk legend are ten classes with expected costs in €.

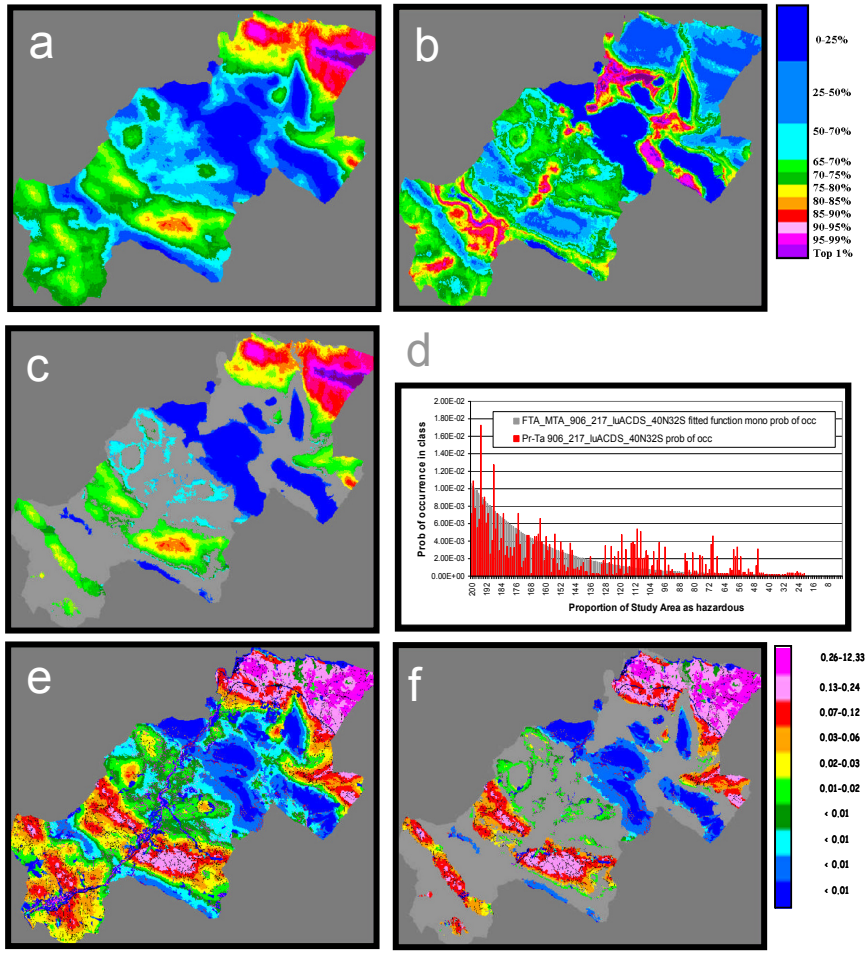

Figure 4: Target, Uncertainty and Risk patterns with probability of occurrence histograms for the Deba Valley study area for analyses with *40N32S* parameters for the *luACDS* ISPs. Target pattern obtained using the distribution of the *906*-pre'97 landslides and the 906-50 x 18 iterative cross-validation (a), the Uncertainty pattern (b) and the 50% Uncertainty/Target Combination pattern (c). In (d) the transformation of the corresponding prediction-rate in Figure 1d, into a probability of occurrence histogram in red and in gray the fitted function to its monotonically decreasing transformation. The Risk pattern is in (e), with the four classes with estimated costs > 1 € in black, and the corresponding 50% Uncertainty/Risk Combination pattern of (c) in (f). In the risk legend are ten classes with expected costs in €.

References

[1] Remondo, J., González-Díez, A., Díaz de Terán J. R. and Cendrero, A., 2003a, Landslide susceptibility models utilising spatial data análisis techniques. A case study from the lower Deba valley, Guipúzcoa (Spain). Natural Hazards, v. 30, p. 267-279.
[2] Remondo, J., González-Díez, A., Díaz de Terán J. R., Cendrero, A., Fabbri, A. and Chung, C. F., 2003b, Validation of landslide susceptibility maps; examples and applications from a case study in northern Spain. Natural Hazards, v. 30, p. 437-449.
[3] Remondo, J., Bonachea, J. and Cendrero, A., 2005, A statistical approach to landslide risk modelling at basin scale: from landslide susceptibility to quantitative risk assessment. Landslides, v. 2, n. 4, p. 321-328.
[4] Fabbri A.G., Chung C.-J., 2009, Training decision-makers in hazard spatial prediction and risk assessment: ideas, tools, strategies and challenges. In K. Duncan and C. A. Brebbia, eds., Disaster Management and Human Health Risk. Southampton, WIT Press, p. 285-296.
[5] Chung, C. F. (2006): Using likelihood ratio functions for modelling the conditional probability of occurrence of future landslides for risk assessment**.** Computers and Geosciences, Vol. 32, 1052-1065.
[6] Fabbri, A. G. and Chung C.-J., 2008, On blind tests and spatial prediction models. Natural Resources Research, 17(2), 107-118.
[7] http://www.spatialmodels.com