Environmental Geology

DOI 10.1007/s002540100440

Cases and Solutions

The use of predictive modeling techniques for optimal exploitation of spatial databases: a case study in landslide hazard mapping with expert system-like methods

Alberto Pistocchi( ) · Lucia Luzi · Paola Napolitano

A. Pistocchi
Studio di Ingegneria per l’Ambiente e il Territorio, viale G. Carducci, 15, 47023 Cesena, Italy

L. Luzi
IRRS-CNR, Milan, Italy

P. Napolitano
ACTA Studio Associato, Naples, Italy

E-mail: alberto.pistocci@libero.it
Phone: +39-054-7367336
Fax: +39-054-736-7336

Received: 21 June 2000 / **Accepted:** 5 September 2001 / **Published online:**

Abstract. A case study is presented in which different probabilistic prediction models (Bayesian probability, fuzzy logic "and", "or", "sum", "product", "gamma" operations, and certainty factors) are used to produce landslide hazard maps for a hilly and mountainous region in the northern Apennines, Italy. Seven data layers are exploited to detect most vulnerable areas: lithology, distance from the geological lineaments, annual rainfall amount, land cover type, topographic slope and aspect, and the distance from hydrographic network segments. The results of the different predictions are compared using the prediction rate index and critically discussed, to evaluate the possibility of using readily available databases for land planning.

Keywords. Favorability functions - integration modeling - Landslide hazard - Spatial database

Introduction and general issues

In the last few years, a great development has been seen in the setup of spatial databases by regional land planning authorities all over Europe. However, it seems that many databases are still ineffective for decision support, and the use of available data is very often quite naive. In particular, final data users and decision makers tend to have little consciousness about the modeling capabilities of geographic information systems (GIS; Campbell and Masser *1995*). Very few local government institutions make use of predictive models as effective support for everyday decisions.

Geographic information systems bring a great capability for detailed spatial features modeling, and most local administrations now have GIS technologies readily available. Can this great information potential be turned into an everyday usage of a more powerful method to have an insight in physical phenomena?

The necessity of participative planning and goal sharing, when deciding about common resources, has brought to the attention of geoscientists the importance of explaining how they make evaluations used in planning and decision support. Guzzetti and others (*2000*) stress the role of geoscientific maps in policy making and land-use planning. According to the authors, a major role of hazard maps is to provide policy makers with a correct perspective of the problem, oriented to the definition of regulations for correct land exploitation.

Predictive models, based on causal relationships between underlying physical phenomena, are in widespread use among hydrologists, geoscientists, environmental analysts, and engineers, for applications in the field of natural risk assessment, natural resource management, pollution prevention and reclamation, and environmental impact assessment. In the case of natural hazards such as landslides, however, it seems strikingly difficult to make reliable causal models applicable at a regional scale. Carrara and others (*1998*) discuss the main causes of this difficulty, invoking both model and data limitations. Different from other risk management contexts, little agreement has been reached about which quantitative models to use (Guzzetti and others *2000*).

The traditional approach to landslide hazard mapping relies on the expertise of geologists and geomorphologists to detect the features that account for the occurrence of landslides (by onsite inspection of field characteristics or by remote observations). This brings an appropriate recognition of past occurrences, but cannot support any prediction apart from the expert's subjective and qualitative judgement.

In more recent years, geotechnical model-based zonation has been proposed (Van Westen *1993*; Liener and Kienholz *1998*). Computations based on geotechnical models or physically based index overlaying methods, however, are often not reliable because of lacking or low quality data, despite their relatively strong physical basis.

On the other hand, a prediction made through "objective" replicable models, with limited arbitrary choices by the analyst, might be of interest, especially in the following cases:

- when social conflicts involved with planning assume a relevant importance;
- when phenomena are not easily observable;
- when detailed mapping of the phenomena over the whole area of interest is too expensive, and a "screening level" modeling is required for the identification of areas calling for further insights.

Generally speaking, the modeling process parallels decision making and works as a negotiation basis (Dutton and Kraemer *1985*). The rational justification of the hazard map, through methods in which the specialist's expertise is used in a reproducible modeling methodology, may contribute to the social construction of knowledge, i.e., the sharing of sound decision criteria between administrators, local communities, and scientists (Innes *1995*).

This reason leads to the investigation of the possibility of using probabilistic approaches to prediction. In this type of approach, prior knowledge of landslide occurrences is used to the full extent to make probabilistic predictions via well-established parametric, fuzzy, or stochastic map overlaying methods.

Over the last years, many approaches of this kind have been explored (Carrara *1983*; Carrara and others *1995*; Luzi and others *1998*; Massari and Atkinson *1998*).

All these methods have been extensively compared using sensitivity analysis or the performance of different methods on the same case study (van Westen *1993*). A major difficulty with these applications is the comparison of different maps (Guzzetti and others *2000*).

Chung and Fabbri (*1993*) proposed a framework approach to the problem, referred to as the favorability function mapping. In their work, the authors show how a wide range of probabilistic, fuzzy, and evidential functions can be used to detect the most favorable areas for the occurrence of phenomena, such as landslides or mineral deposits. These techniques allow inference calibration and validation, a feature that is common to other approaches such as neural and Bayesian networks (Duda and others *1976*; Stassopoulou and others *1998*), and to mathematical modeling in general. In this way, a unique criterion for the comparison of different prediction maps is easily found in what Chung and Fabbri (*1999*) call the prediction rate, i.e., a measure of the goodness of the model's validation, as explained below. Applications of favorability functions, as the one presented here, aim at producing maps that contain at least as much of correct predictions as an expert scientist might express using rule-based judgement derived from field experience. Of course it is necessary that experts closely follow the modeling because many choices that require deep knowledge and understanding of the phenomena are required during evaluation. In addition, spurious effects may arise because of mismatching the probabilistic hypotheses invoked for the variables, and lacking and unreliable data (Guzzetti and others *2000*). However, using quantitative techniques, calibration and validation of the models supports transparency and rationality of the prediction. The favorability function modeling approach has been recently applied in some case studies with ad hoc surveying (Fabbri and others *1998*; Remondo and others *1998*).

The aim of this paper is to discuss the applicability of the favorability function modeling in producing a hazard map for landslide occurrences using standard currently available databases, and to check how this method might improve the use of information in an existing database, compared with other techniques (for example, frequency mapping of landslides per lithological unit, or simple landslide inventory).

It is explained how favorability function modeling can be used as a conceptual scheme for the structuring of databases: data collection is strictly dependent on the theoretical framework in which information will be used.

Theoretical background

Many authors (Chung and Fabbri *1993*; Bonham Carter *1994*; Harris and Pan *1999*) have shown the use of numerical techniques to link the occurrence of a phenomenon of interest to the local value of some attributes deemed relevant for the phenomenon itself. The attributes are considered as evidence factors of the event, in the sense that the presence of each relevant attribute corresponds to a degree of "probability", "possibility" or "likelihood" to find the event (Chung and Fabbri *1993*).

Suppose that A is the domain on which the analysis is performed, and F is the phenomenon of which the occurrence was checked. If r data layers are available, for each one of the m_k classes of attributes, k=1,...r, are considered, a partition function can be defined for each data layer:

$$V_k\text{: } A \Rightarrow \{1,\ 2,\ ...,\ nk\} \tag{1}$$

That assigns each pixel in A to one of the classes in layer k. Furthermore, another function can be defined for each layer:

$$R_k\text{: } \{1,\ 2,\ ...,\ nk\} \Rightarrow [a,\ b] \tag{2}$$

which maps the occurrence of each layer in a value falling inside the interval [a,b], where a and b depend on further assumptions made by the analyst (as will be shown later). This value represents the degree of favorability, which is a measure of how reliable the assumption is that the phenomenon occurs once a particular class of the attribute is met.

The favorability function can then be defined as the functional composition of V and R for each data layer, i.e.,

$$F_k = R_k^\circ V_k \tag{3}$$

The interval extremes a, b must be assumed by the analyst according to his interpretation of the "reliability": if it is considered that reliability is the same as "probability", then a=0, b=1. If the measure of reliability is set equal to the certainty factor (Shortliffe and Buchanan *1975*; Heckermann *1986*), then a=-1, b=1. If a different technique is chosen, other values might be necessary.

The different interpretations of the favorability functions as used in the present work are reported.

If the favorability is assumed to coincide with the probability that a certain phenomenon F occurs given the occurrence of a set of attribute classes $E1$,...En, then according to Bayes' theorem, under the hypothesis that $E1$,...En are conditionally independent, it can be written (Chung and Fabbri *1999*):

$$\begin{gathered}\text{Probability}\{\text{F occurs given E1,..., En}\} = \\ = \text{P}\{\text{F/E1...En}\} = (\mathbf{pps}_1 *...* \mathbf{pps}_\text{n}) * (\mathbf{ppa}_1 *...* \mathbf{ppa}_\text{n}) / (\mathbf{psF})^{\text{n}-1} * \mathbf{pps}_{\text{1ton}}\end{gathered} \tag{4}$$

where pps_I, I=1,...n is the prior probability that a certain attribute class occurs, and can be estimated by the percentage of the total area where an attribute class occurs. $pps_{1\text{ to }n}$ is the prior joint probability of the attribute classes considered; this can be assumed as the percentage of the total area where all the classes occur together. ppa_I, I=1,...n is the probability of finding F given the occurrence of the attribute class Ei; this can be computed according to the formula. Ppa_I=1-(1-($area_I$)$^{-1}$)$^{nb(I)}$ (Chung and Fabbri *1993*) where $area_I$ is the area where class i is met, and $nb(i)$ is the area of class i where F is also met. psF is the prior probability of F all over the area, and can be evaluated as the percentage of the total area where F is met.

According to this rule, a map can be computed for each combination of the attribute classes that occur. This can be done through a routine cross operation within the raster GIS used.

If certainty factors are used, then rules change according to the following:

1. the certainty factor for an attribute class can be defined as:

$$\mathrm{CF_I} = [\mathrm{P}\{\mathrm{F}/\mathrm{E_I}\} - \mathrm{P}\{\mathrm{F}\}]/[\mathrm{P}\{\mathrm{F}/\mathrm{E_I}\}(1 - \mathrm{P}\{\mathrm{F}\})] \ \text{ if } \ \mathrm{P}\{\mathrm{F}/\mathrm{E_I}\} > \mathrm{P}\{\mathrm{F}\}$$
$$\mathrm{CF_I} = [\mathrm{P}\{\mathrm{F}/\mathrm{E_I}\} - \mathrm{P}\{\mathrm{F}\}]/[(1 - \mathrm{P}\{\mathrm{F}/\mathrm{E_I}\})\mathrm{P}\{\mathrm{F}\}] \ \text{ if } \ \mathrm{P}\{\mathrm{F}/\mathrm{E_I}\} < \mathrm{P}\{\mathrm{F}\}$$

with I=1,...n, n being the number of thematic data classes of the causal factors;

2. for two classes, the certainty factor is computed according to the following rules:
   $CF_{1+2}=CF_1+CF_2-(CF_1 \times CF_2)$, if both CF_1 and CF_2 are non-negative

   $CF_{1+2}=CF_1+CF_2/\{1-\min(|CF_1|,|CF_2|)\}$, if CF_1 and CF_2 have opposite sign
   $CF_{1+2}=CF_1+CF_2+(CF_1 \times CF_2)$, if both CF_1 and CF_2 are negative;

3. the procedure applies iteratively for more maps by computing first the $CF_{1+2}=CF_{12}$, then $CF_{13}=CF_{12+3}$ and so on.

As the last method, fuzzy sets theory (Zadeh *1965*) was applied by calculating the "fuzzy sum", "fuzzy product", "fuzzy and", "fuzzy or", and "fuzzy gamma function". The membership functions were assumed to be equal to the estimates of the probability of finding F given the class E_I (Zadeh *1968*), i.e., ppa_I, being:

- "fuzzy and" = min. (ppa_I), I=1,...n
- "fuzzy or" = max. (ppa_I), I=1,...n
- "fuzzy product" = $\prod$ (ppa_I), I=1,...n
- "fuzzy sum" = 1-$\prod$ (1-ppa_I), I=1,...n
- "fuzzy gamma operation" = (fuzzy sum)$^{\gamma}$ (fuzzy product)$^{1-\gamma}$, γ being a parameter in the range 0:1.

In this way, rules for map overlaying are defined so that the analyst can evaluate the influence of the different occurrences in attribute data all over the study area to detect the favorability of sites for the occurrence of further phenomena. These calculations result in numbers that represent indices of favorability with respect to the considered phenomenon. It must be noticed that, in addition to the described ones, different techniques can be used as the weights of evidence (Cheng and Agterberg *1999*), the belief functions (Shafer *1976*), linear regression over probabilities (Chung and others *1995*), and many others.

It must be noticed that prior probability *psF* needs to be estimated for the computation of certainty factors, but its use in absolute terms is meaningless because it is practically impossible to know the probability of future landslide occurrences. The prediction score must then be taken as an indicator of favorability to the phenomenon in general terms, and not as a numerical estimate of the hazard.

Application

The area used for the case study is the Savio River catchment, a province of Forlì-Cesena, Emilia Romagna in northern Italy (Fig. 1). A geological outline of the area indicates a basically sedimentary basin with a dominance of marls and sandstones (the "Marnoso Arenacea Romagnola" formation). In more detail, the three main following geological formations can be distinguished:

1. "Formazione marnoso-arenacea romagnola" (Serravallian-Tortonian age) is made of arenaceous and pelitic gray turbidites, is the main geological formation, and lies on both sides of the main stream.

2. "Formazione gessoso-solfifera" (Medium Messinian) is made of microcrystalline gypsum interbedded with marly-clayey and sandy layers, and of basal sulfur-bearing limestone.

3. "Formazione a colombacci" (Medium-Upper Messinian) is made of three facies: a pelitic, an arenaceous, and a conglomerate, all containing limestone horizons.

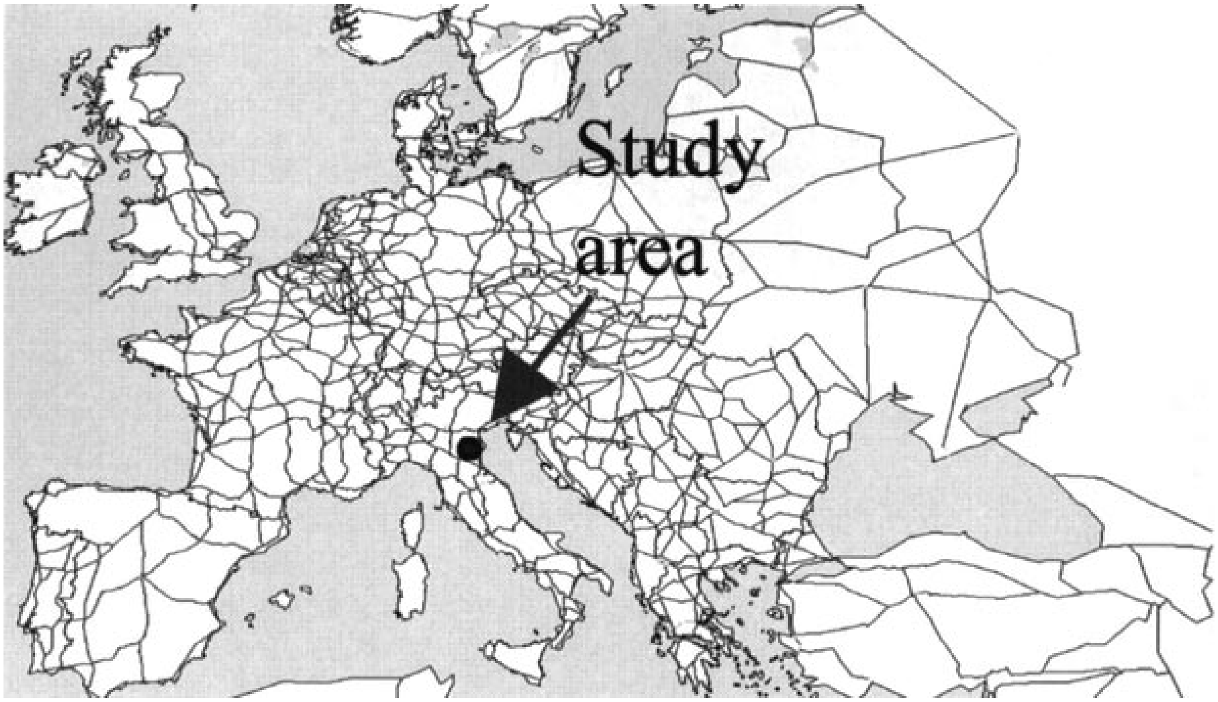

Fig. 1. Location of the area

Also, the " Marne di Verghereto" Formation is present, made of gray silty marls; the "Arenarie del M. Comero" Formation (upper Eocene), and outcrops of sandstones and Pliocene clays, and chaotic "Argille Varicolori" clays.

The area is covered by a large number of landslides, occurring in most cases as slide type movements or debris flows, in different geological units. Furthermore, areas with rock falls and block translational movements are present but they were not analyzed . The data used in the research was made available by the Regione Emilia Romagna Geological Survey (Regione Emilia Romagna *1991*).

The database used for the case study consists of thematic layers concerning:

- structural lineaments (faults, synclines, and anticlines), scale 1:50,000;
- lithological units, scale 1:50,000;
- land cover obtained from TM Landsat imagery according to the CORINE European Project Guidelines (Briggs and Martin *1988*), scale 1: 50,000;
- digital terrain model (DTM) obtained from 50-m equidistance contour lines available from the cartographic database of the Regione Emilia Romagna through standard linear interpolation;
- rainfall measurements at seven gauge stations over the area (data published by the Regione Emilia Romagna *1996*);
- digitized hydrographic network, scale 1: 10,000.

It must be underlined that the database has a very poor resolution and, in addition, data are very inhomogeneous in scale. One should think that, in particular, topographic information is clearly too coarse when compared with average landslide dimension, thus becoming often not representative of actual sliding kinematics. It must be recalled that the aim of this study was to evaluate the predictive capability (in the sense defined above) of a real world database, rather than producing reliable hazard maps, and thus the best available information has been used without any further field investigation and data capturing. As will be highlighted in the following, the results of this evaluation give more input for database improvement than predictions for land planning.

From the DTM, a slope and an aspect map have been calculated and classified using a constant value range slicing interval.

A distance from the geological lineaments was also calculated in order to appreciate the possible effect of structural disturbance on slope stability. The results were gridded on a raster map.

Rainfall data was analyzed to detect a relationship between elevation and annual rainfall amount. A regression equation for the two variables was found to be y=0.7086x+708.19 (R^2=0.66), x being elevation above sea level (m) and y the total mean annual rainfall for more than 30 years long time series (mm/year; Regione Emilia Romagna *1996*). The equation was later used to obtain a continuous rainfall map, thus precisely reflecting the DTM: this is both a rainfall indicator and an indicator of the relief potential energy.

It must be noticed that the correlation coefficient for rainfall over elevation is quite poor, and further analysis is required to better delineate the actual rainfall distribution over the region. With the available data, however, it can be said that a general trend in rainfall distribution is correctly detected.

However, a conceptual distinction can be made between factors needed for landslide occurrence, and factors just triggering landslides in the presence of other necessary features. It was assumed that all these data layers may be meaningful a priori.

As for the available landslide occurrence data, it has been possible to rely on the land instability inventory of the Regione Emilia Romagna (*1991*). It is important to point out that the database made available was structured for a long time oriented to GIS analysis (Artioli and others *1997*). Moreover, data density and distribution was statistically representative of the actual landslide distribution. It can be argued that, in fact, when the training data set for Bayesian procedures is not large enough (and randomized enough) to be regarded as a realization of a regionalized random variable, probabilistic integration modeling is meaningless under the point of view of quantitative estimates. In this case, a priori total and conditional (class-specific) probabilities of landslide occurrence, ppa_I and psF, need to be estimated by an expert’s judgement. Furthermore, It is important to choose the type and the age of landslides to be used as a training set, so that the set can be considered homogeneous. The analysis was performed on the "earth flow"- and "slumped earth flow"-type landslides (Varnes *1958*), which most frequently occurred in the considered area. In the present study, only the ones mapped as active were used.

The regional land instability inventory also considers rock falls, block sliding, and potentially unstable areas, but these were not included in the analysis. Figure 2 shows the data layers used for the analysis.

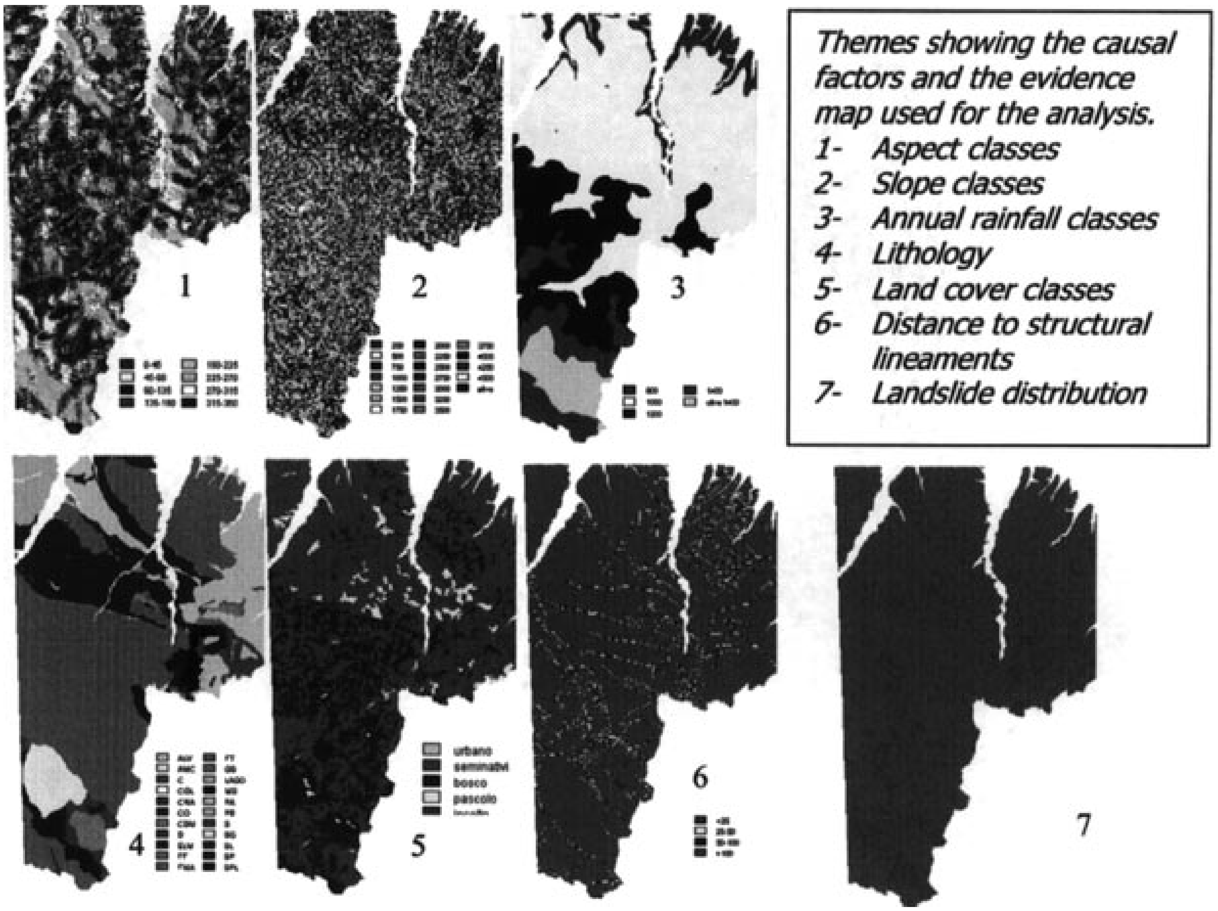

Fig. 2. Thematic maps showing the causal factors used for prediction

All thematic data considered have, in principle, the possibility of mutual association. This would lead to a computational effort that might give useless results because of redundant information. An association test has been performed over the seven themes used for the analysis (i.e., rainfall map, lithology, land cover, slope, aspect, distance to hydrographic network, and distance from the lineaments). The seven themes have been classified into discrete legends and tested with respect to the map of active landslides.

Four association indices were calculated for each map pair (Press and others *1986*; Bonham Carter *1994*):

- the chi-square index;
- the Cramers index;
- the contingency index;
- the joint information uncertainty score.

The first one is defined as :

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(T_{ij} - T^*_{.ij})^2}{T^*_{.ij}} \tag{5}$$

where

$$T^*_{.ij} = \frac{(T_{.j} * T_{.i.})^2}{T..} \tag{6}$$

and *T*= total number of pixels, *Ti*= number of pixels in class *i* of map 1, *Tj*= number of pixels in class *j* of map 2. The indices *n* and *m* are the number of classes in maps 1 and 2, respectively.

The Cramers index (*V*) and the contingency index (*C*) are defined as:

$$\begin{aligned} V &= \sqrt{\frac{\chi^2}{T..M}} \\ C &= \sqrt{\frac{\chi^2}{T..+\chi^2}} \end{aligned} \tag{7}$$

respectively, where the meaning of the symbols holds as before and *M* is the minimum of (*n*-1, *m*-1), with *n* and *m* being the number of data classes in each of the two maps.

The joint information uncertainty score for the pair of maps A and B is defined as:

$$U(A,B) = 2\left[\frac{H(A) + H(B) - H(A,B)}{H(A) + H(B)}\right] \tag{8}$$

where

$$\begin{aligned} H(A) &= -\left[\sum_{j=1}^{n} p_{.j} \ln p_{.j}\right] \\ H(B) &= -\left[\sum_{i=1}^{m} p_{i.} \ln p_{i.}\right] \\ H(A,B) &= -\left[\sum_{j=1}^{n} \sum_{i=1}^{m} p_{.j} \ln p_{.j}\right] \end{aligned} \tag{9}$$

n being the number of classes in map A, *m* the one in map B, and *Pij* the ratio of the number of pixels at the intersection of classes *i* and *j* in maps A and B, respectively, to the total number of pixels. *Pj* denotes the total pixel number in class *j* of map A, and *Pi* the total pixel of class *i* in map B.

The above indices can be interpreted as a measure of association between map pairs. The chi-square index gives an absolute measure (upper unbounded) of the association and is useless in itself; the *V* and *C* indices provide a standardized measure in the range [0,1] so that the closer they are to 1, the stronger the association is between the two maps. When used together, these three indices provide an overall measure of association and allow us to compare the different degrees of association in pairs over a set of maps. In general, it can be noticed that the three indices give very similar responses as expected. The joint information uncertainty score should also be used to confirm the pattern of association as detected by the previous indices, and is supposed to vary between 0 (completely independent maps) and 1 (completely associated maps). Table 1 shows the indices as computed for the maps described above.

Table 1. Association measures between data layers

	Lithology	Land cover	Aspect	Slope	Rainfall	Distance from rivers	Distance to lineaments	Index
Landslides	14,323.84	3,896.73	1,316.985	1,122.374	5,638.054	2,015.259	105.83	Chi-square
	0.2169	0.1131	0.0658	0.0609	0.1361	0.0814	0.0186	Cramers V
	0.212	0.1124	0.0657	0.0607	0.1348	0.0811	0.0186	Contingency
	0.02	-	-	-	-	-	-	Uncertainty
Lithology		73,615.45	23,898.36	37,642.06	225,075.7	4,254.247	106,13.059	Chi-square
		0.2459	0.106	0.0855	0.4308	0.1182	0.1078	Cramers V
		0.4413	0.27	0.3324	0.6528	0.1174	0.1835	Contingency
		0.1	0.02	0.04	0.23	0.01	0.02	Uncertainty
Land cover			6,060.324	25,611.87	61,260.3	7,634.952	821.463	Chi-square
			0.1398	0.1454	0.2243	0.1584	0.03	Cramers V
			0.0706	0.2792	0.4093	0.1564	0.0519	Contingency
			-	0.03	0.07	-	0.04	Uncertainty
Aspect				6,289.839	10,158.14	1,493.088	300.815	Chi-square
				0.0431	0.0914	0.0554	0.0182	Cramers V
				0.1132	0.1798	0.0553	0.0314	Contingency
				-	-	-	-	Uncertainty
Slope					41,613.68	1,230.77	1,245.942	Chi-square
					0.1853	0.0504	0.037	Cramers V
					0.3475	0.0503	0.064	Contingency
					0.04	-	-	Uncertainty
Rainfall						501.429	2,593.115	Chi-square
						0.0406	0.0533	Cramers V
						0.0405	0.0919	Contingency
						-	-	Uncertainty
D.f. rivers							60.929	Chi-square
							0.0141	Cramers V
							0.0141	Contingency
							-	Uncertainty

Although the computed indices cannot be used, in rigorous terms, to assess Bayesian conditional independence (which is a property much stronger than non-association), the measures of association they provide suggest that all data layers can be supposed to be independent.

As comes from the analysis, it must be noticed that landslides show some association with lithology (the only theme that also has non-irrelevant joint information uncertainty, *U*, with the landslide one), and a spatial trend associated with elevation/rainfall and land cover.

If one looks for association among causal factors, it must be noticed that lithology is associated with elevation/rainfall and land cover, whereas the association is weaker with slope, and even less with the other themes. The inadequate DTM available for the study seems to be a prime cause of this. All other associations can be considered irrelevant, apart from a weak association between slope and rainfall/elevation.

The analysis seems to come to the same conclusion drawn by the Regione Emilia Romagna Geological Survey, which used the lithological factor only to map the landslide hazard, and used the frequency of landslides per lithological unit as an index of landslide hazard (Preti and Tomassetti, personal communication).

During each run, only one half of the known landslides (chosen by random sampling) was used to make the prediction map, whereas the remaining ones were considered as a validation data set. To predict the landslide hazard pattern, potential causal factors were initially used, and in a second attempt only the three most relevant factors were used, as explained in the following section.

Discussion of the results

The computation of favorability functions was performed for different modeling hypotheses as described below. The test on the predictive power of each computed favorability map was done using a prediction rate curve (Chung and Fabbri *1999*). This curve was obtained by plotting the cumulative percentage of the study area sorted by decreasing value of the favorability value (according to the various rules as presented above) as the abscissa against the cumulative percentage of the landslide area as the ordinate. The percentage of predicted landslides corresponding to, say, 20% of the highest value area is a good estimator of the predictive capability of the model. In a wider sense, the closer the curve approaches the ordinate axis, the more the prediction fits. In contrast, the more the curve approaches the 45° straight line, the less it is useful to combine factors because the prediction is close to a random distribution of the favorability values. Among the causal factors, it has been recognized that the hydrographic network plays a minor role because the detail to which it has been mapped is much higher than the precision of the other factors, and river segments appear to be "pervasive", thus not allowing the linkage of the landslide distribution with a distance from the hydrographic network. For this reason, it has been decided not to include rivers among the causal factors.

In Fig. 3 the prediction rate of the six causal factors considered is shown one by one. In this case, the predictor is the conditional frequency estimating ppa_I, I=1,...n (conditional probability of finding the landslide occurrence, given class *i*) for each class in each theme.

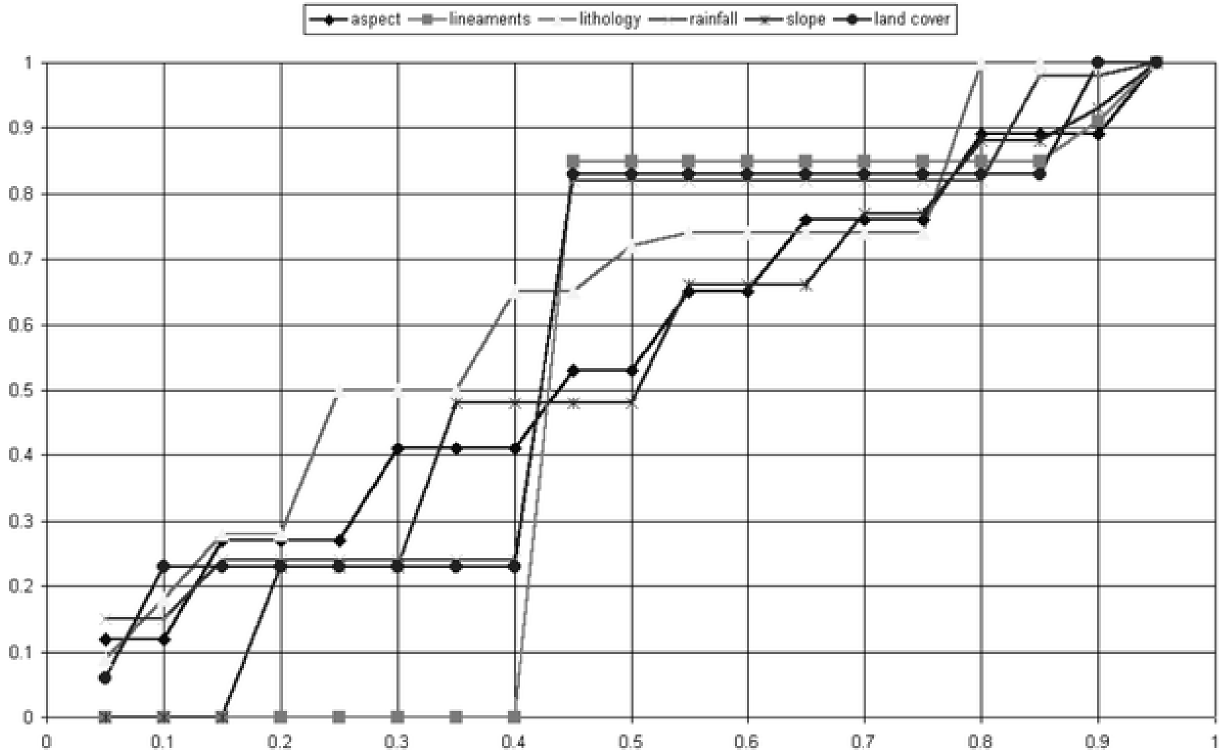

Fig. 3. Prediction rate of the considered causal factors, using the whole landslide polygons and conditional frequency

The first computation was performed using, as evidence, data from all polygons of the mapped active landslides. The landslides were divided into two randomly sampled groups, one of which was used for calibration and the other for validation. The computation was performed using the three most relevant themes (lithology, land cover, and elevation/rainfall) according to the previously described indices. The prediction rate curves are shown in Fig. 4.

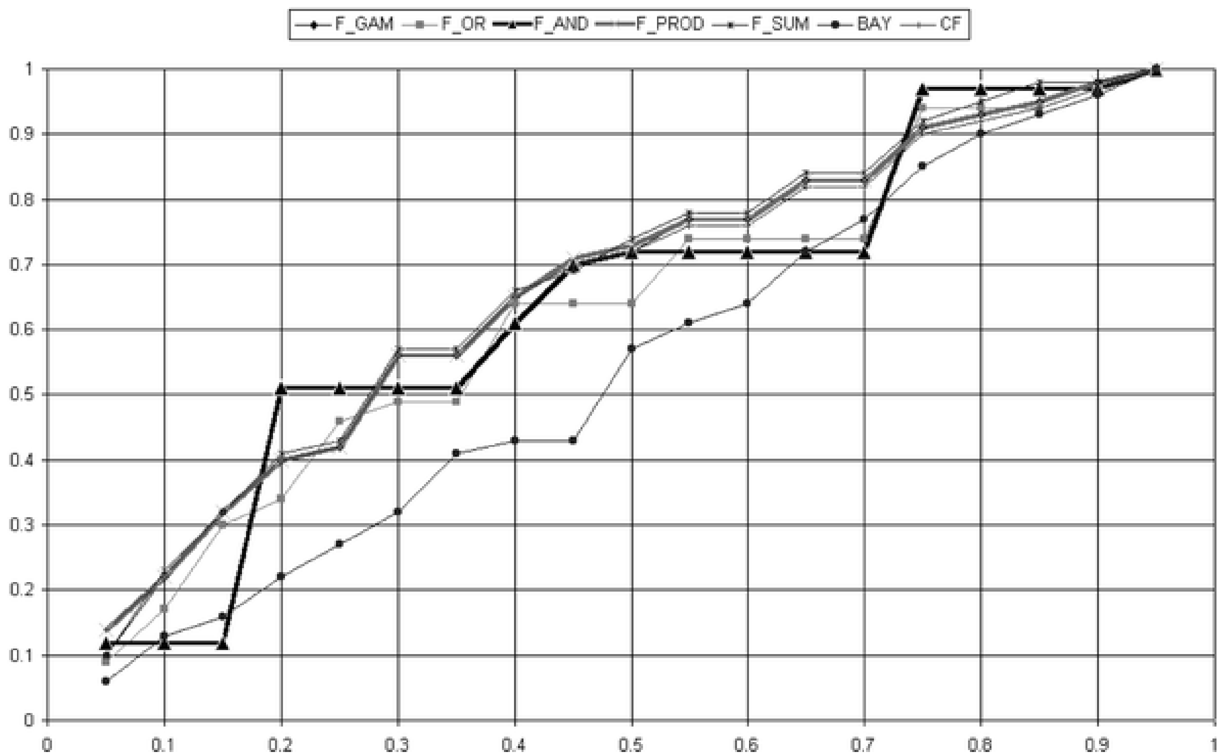

Fig. 4. Prediction rate of the seven predictors, using three causal factors (lithology, rainfall, and land cover) and the whole landslide polygons

---

A further computation was performed using all six factors, and the prediction rates are shown in Fig. 5.

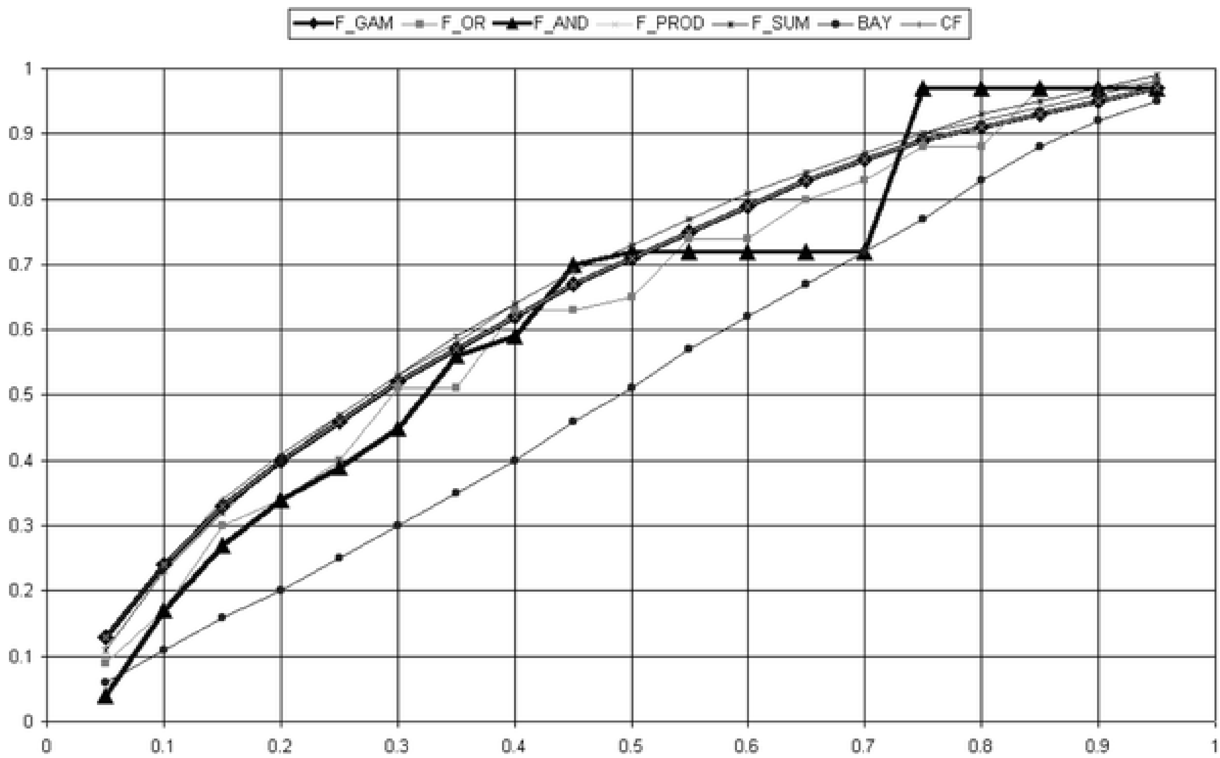

Fig. 5. Prediction rate of the seven predictors, using all six relevant causal factors and the whole landslide polygons

---

It was noticed that all landslide bodies were mapped, and this might have induced some loss of precision, in that the combination of the causal factors that is met at the landslide trigger point is not the same as in the landslide toe. For this reason, a prediction was drawn using only the highest points within each of the landslide polygons, assuming that, given the kinematics of the mass movements considered, the trigger point must have been at the top. Under this assumption the prediction rate of the six causal factors has been computed as shown in Fig. 6. The prediction rates of the seven predictors using three and six causal factors, respectively, are shown in Figs. 7 and 8.

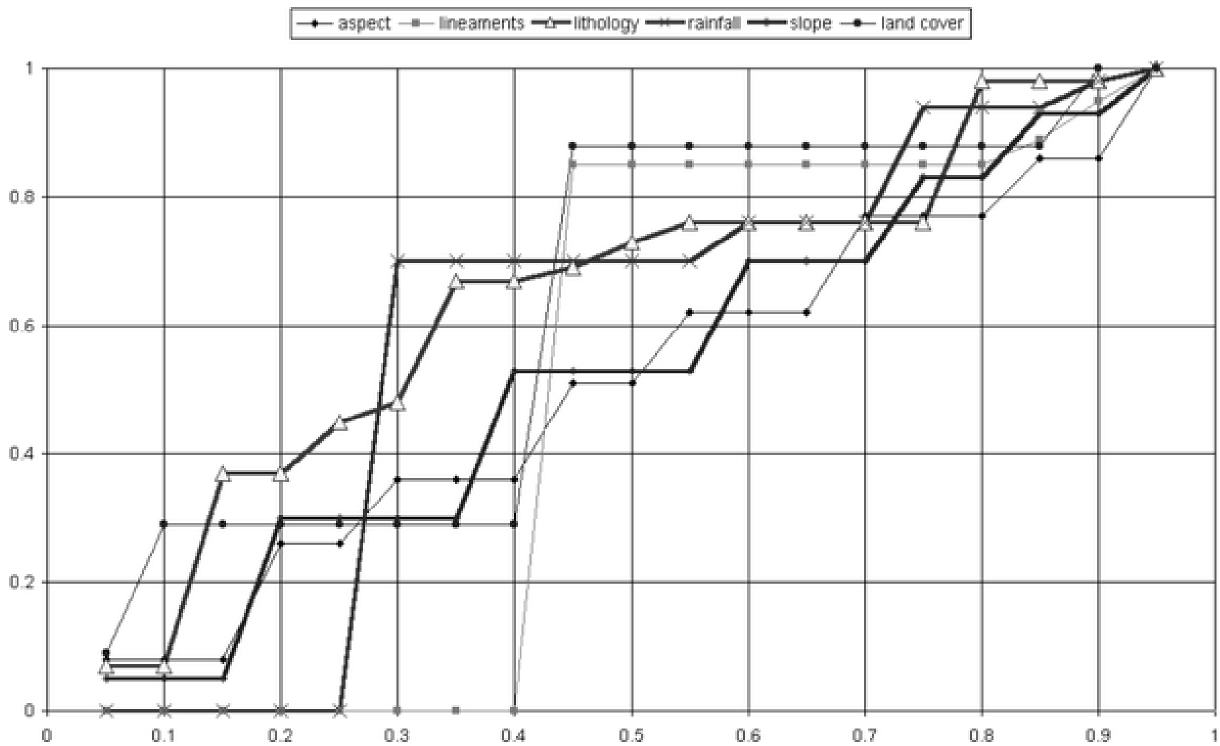

Fig. 6. Prediction rate of the causal factors, using the trigger points only

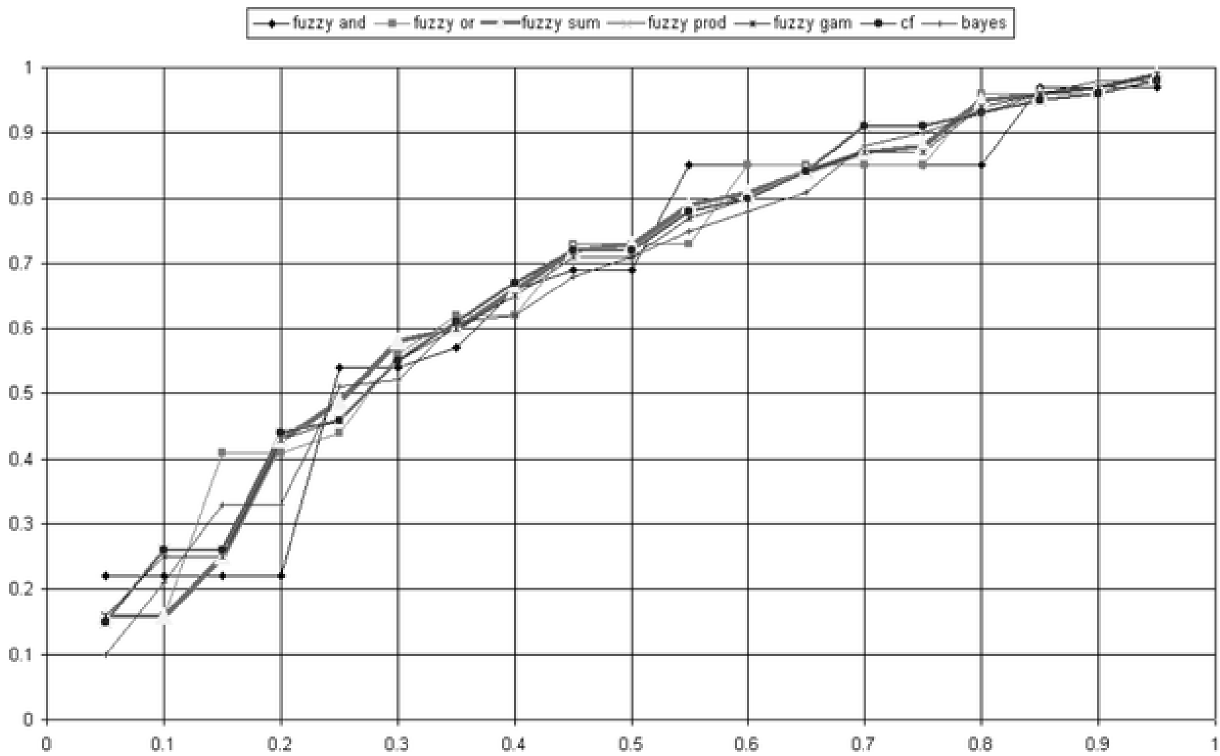

Fig. 7. Prediction rate of the seven predictors using three causal factors (lithology, land cover, and rainfall) and the trigger points only

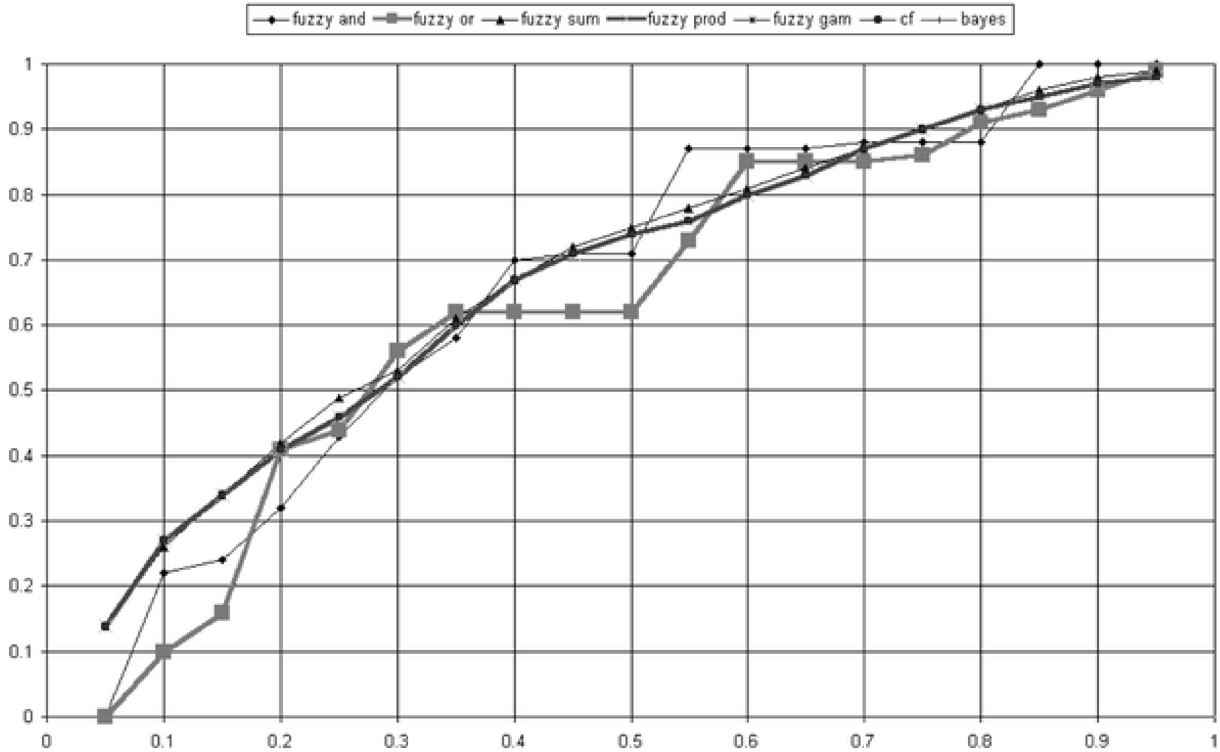

Fig. 8. Prediction rate of the seven predictors using the six relevant causal factors and the trigger points of landslides

---

As far as the input data are concerned, one can argue that using a better representation of slope, aspect, and rainfall distribution (i.e., a more precise DTM, and a better regionalization of the rain gauge data) will allow improvement in the results. As soon as new data are available, the analysis can be repeated to appreciate the potential effects on the prediction.

From the comparison of the prediction rates one can assess that:

- There seems to be no significant improvement when using six causal factors instead of the three more strictly related to landslides (lithology, land cover, and rainfall); the prediction behaves very similarly in the two cases, being just smoother in classification, the more factors are used.

- A further smoothing effect can be observed when all landslide bodies are used as evidence, instead of the trigger points only. This does not bring a worsening of the overall predictive capability of the maps, but a less reliable mapping can be expected because of over-smoothing.

- From the prediction rate diagrams it can be seen that the causal factors that have higher predictive capability are in any case lithology (thus confirming the choice of the Regione Emilia Romagna Geological Survey to use this layer alone for hazard mapping), and then land cover and rainfall. All other themes are not relevant for prediction.

- The seven predictors used in this case study behave very similarly in prediction, except for the case of Bayesian probability - which is very sensitive to the actual dispersion of the data, and approaches a random prediction when all landslide bodies are used as evidences - and fuzzy "or", "and" in some cases. In general, it seems that the certainty factor predictor is the most useful method in this specific case study, although in each case some of the predictors are virtually identical as denoted both by the prediction rate graphs and the prediction maps.

Figure 9 shows the seven predictions in the case when the three more significant factors are used together with only the trigger points as evidence themes. This is the case among those explored in this case study, which shows better prediction rates and should be considered as the best basis for landslide hazard zonation, given the present state of knowledge.

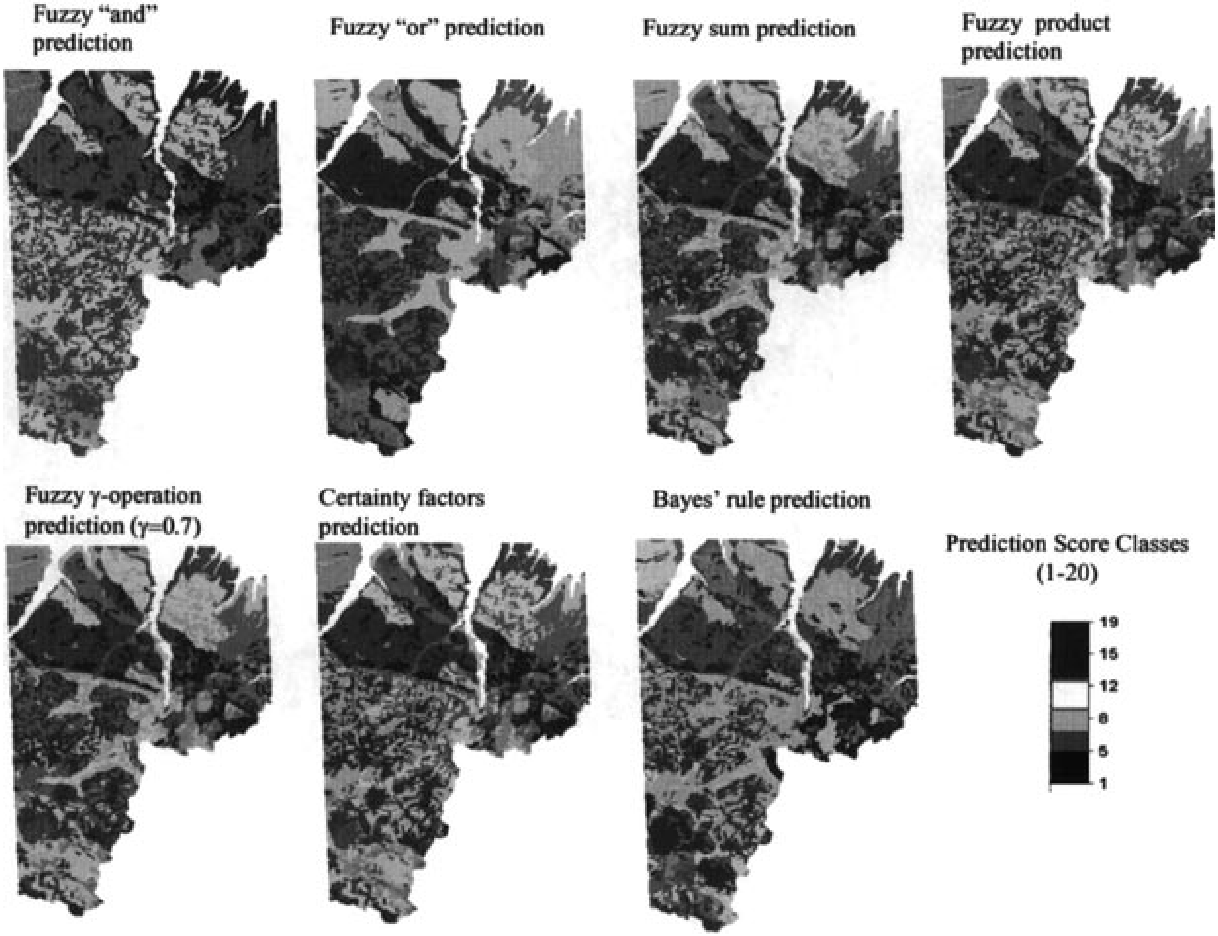

Fig. 9. Prediction maps according to the seven predictors, using rainfall, lithology and land cover themes, and the trigger points as evidences. The region is subdivided into equal area classes in increasing order of favorability (class 20=max. favorability) according to the legend

Conclusions

The approach hereby discussed allows land classification according to landslide hazard using numerical modeling (requiring less subjective expert's judgements). It seems that, when an objective prediction can be extracted from a spatial database, then it can be said that the themes have some "system"-added value, i.e., using all data together is better than using just some of them.

It must be highlighted that this approach starts from existing databases but remains open to improvements with knowledge of each theme. The best predictor among the various ones tested (certainty factors, Bayesian probability, fuzzy operations, and other possible techniques) is simply chosen on the basis of its predictive capability, measured a posteriori using the prediction rate curves.

This analysis has led to the recognition that the existing database is not fit, however, for predictive modeling because of inadequate topographic data. This constitutes an input for addressing future survey and data capturing to define a better digital terrain model. As soon as the improved causal factor map is produced, or a new causal factor is supposed to be relevant for the phenomenon, the calculations can be repeated and a new prediction map can be produced. The validation using prediction rates allows a check for actual and effective improvements, and can be used to orient further efforts in data acquisition and geotechnical monitoring. For instance, in the present case study it was shown that lithology, land cover, and rainfall (represented by elevation, as described above) are the most relevant factors for landslides, and thus further analyses should be devoted to the investigation and mapping of these factors. Moreover, it appears necessary to prepare and use a DTM with appropriate resolution, in order to check the influence of more detailed topographic data. The analysis disregarded other themes, such as the water table elevation, which, in turn, might become very important for improved hazard mapping.

The hereby discussed approach is appealing because of its capability to be integrated in real world policy making because modeling closely follows the stepwise decision process. The decision support capitalizes on the gradual accumulation of knowledge, which is commonly observed in physical planning.

References

Artioli GP, Baldini U, De Nardo MT, Farina M, Palumbo A, Pignone R, Tomassetti C, Viel G (1997) Area metropolitana di Bologna stato di avanzamento delle ricerche. In: Atti del Convegno Geologia delle Grandi aree urbane, Progetto Strategico del CNR, Bologna, pp 3-48

Bonham-Carter G (1994) GIS for geoscientists - modeling with GIS. Elsevier, New York

Briggs DJ, Martin DM (1988) CORINE: an environmental information system for the European Community. Eur Environ Rev 2(1):29-34

Campbell H, Masser I (1995) GIS and organizations: how effective are GIS in practice? Taylor and Francis, Bristol

Carrara A (1983) Multivariate models for landslide hazard evaluation. Math Geol 15(3):403-427

Carrara A, Cardinali M, Guzzetti F, Reichenbach P (1995) GIS based techniques for mapping landslide hazard. In: Carrara A, Guzzetti F (eds) Geographical information systems in assessing natural hazards. Kluwer, Dordrecht, pp 135-175

Carrara A, Guzzetti F, Cardinali M, Reichenbach P (1998) Current limitations in modeling landslide hazard. In: Buccianti A, Nardi G, Potenza R (eds) Proceedings of the fourth annual conference of the International Association for Mathematical Geology. De Frede, Naples, pp 195-203

Cheng Q, Agterberg FP (1999) Fuzzy weights of evidence method and its application in mineral potential mapping. Nat Resour Res 8(1):27-35

Chung CF, Fabbri AG (1993) The representation of geoscience information for data integration. Non Renew Resour 2(2):122-139

Chung CF, Fabbri AG (1999) Probabilistic prediction models for landslide hazard mapping. Photogrammet Eng Remote Sensing 65(12):1389-1399

Chung CF, Fabbri AG, van Westen CJ (1995) Multivariate regression analysis for landslide hazard zonation. In: Carrara A, Guzzetti F (eds) Geographical information systems in assessing natural hazards. Kluwer, Dordrecht, pp 107-133

Duda R, Hart P, Nilsson N (1976) Subjective Bayesian methods for rule-based inference systems. Proc AFIPS Natl Comp Conf 45:1075-1082

Dutton WH, Kraemer KL (1985) Modeling as negotiation: the political dynamics of computer models in the policy process. Ablex Publishing Corporation, Norwood

Fabbri AG, Napolitano P, Chung CF, Zezere L, Fereira AB, Reis E (1998) Modelling landslide hazard north of Lisbon, Portugal. Using favourability functions. In: Buccianti A, Nardi G, Potenza R (eds) Proceedings of the fourth annual conference of the International Association for Mathematical Geology. De Frede, Naples, pp 254-255

Guzzetti F, Cardinali M, Reichenbach P, Carrara A (2000) Comparing landslide maps: a case study in the upper Tiber river basin, central Italy. Environ Manage 25(3):247-263

Harris DV, Pan G (1999) Mineral favorability mapping: a comparison of artificial neural networks, logistic regression, and discriminant analysis. Nat Resour Res 8(2):93-109

Heckermann D (1986) Probabilistic interpretation of the MYCIN's certainty factors. In: Kanal LN, Lemmer JF (eds) Uncertainty in artificial intelligence. Elsevier, New York, pp 167-196

Innes J (1995) Information in communicative planning. Working Paper 679, University of California at Berkeley

Liener S, Kienholz H (1998) Predicting the spatial distribution of shallow landslides - the improved model SLIDISP. In: Buccianti A, Nardi G, Potenza R (eds) Proceedings of the fourth annual conference of the International Association for Mathematical Geology. De Frede, Naples, pp 229-234

Luzi L, Pergalani F, Terlien MTJ (1998) A probabilistic approach to assess uncertainty of landslide vulnerability to earthquakes. In: Buccianti A, Nardi G, Potenza R (eds) Proceedings of the fourth annual conference of the International Association for Mathematical Geology. De Frede, Naples, pp 235-240

Massari R, Atkinson PM (1998) Using the Gibbs sampler in mapping susceptibility to landsliding. In: Buccianti A, Nardi G, Potenza R (eds) Proceedings of the fourth annual conference of the International Association for Mathematical Geology. De Frede, Naples, pp 247-252

Press WH, Flannery BP, Teukolsky SA, Vetterling VT (1986) Numerical recipes: the art of scientific computing. Cambridge University Press, Cambridge

Regione Emilia Romagna (1991) Piano Territoriale Paesistico Regionale, Bologna

Regione Emilia Romagna-Servizio Meteorologico Regionale (1996) Tavole climatologiche dell'Emilia Romagna 1951-94, Bologna

Remondo J, Diaz de Teran JR, Marquinez J, Menendez R, Napolitano P, Fabbri AG, Chung CF (1998) Geologic hazard zonation by a fuzzy set favourability function applied in the Basque Country, Spain. In: Buccianti A, Nardi G, Potenza R (eds) Proceedings of the fourth annual conference of the International Association for Mathematical Geology. De Frede, Naples, pp 253-254

Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton

Shortliffe EH, Buchanan BG (1975) A modeling of inexact reasoning in medicine. Math Biosci 3:351-379

Stassopoulou A, Petrou M, Kittler J (1998) Application of a Bayesian network in a GIS based decision making system. Intl J Geogr Inform Sci 12(1):23-45

Van Westen CJ (1993) Application of geographical information systems to landslide hazard zonation. PhD Thesis, Technical University of Delft, ITC publication no 15

Varnes DJ (1958) Landslide types and processes. In: Eckel EB(ed) Landslides and engineering practice. Special report 29, Highway Research Board, Washington

Zadeh LA (1965) Fuzzy sets. IEEE Inform Control 8:338-353

Zadeh LA (1968) Probability measures of fuzzy events. J Math Anal Appl 10:421-427