Preprint of the paper published in Natural Hazards, 37: 315-329 (2006)

VALIDATION AND EVALUATION OF PREDICITIVE MODELS IN HAZARD ASSESSMENT AND RISK MANAGEMENT

Santiago Beguería[1]

UCEL - Utrecht Centre for Environment and Landscape Dynamics
Department of Physical Geography - Utrecht University
Heidelberglaan 2. PO Box 80115, 3508 TC Utrecht. The Netherlands

Abstract: The paper deals with the validation and evaluation of mathematical models in natural hazard analysis, with a special focus on establishing their predictive power. Although most of the tools and statistics available are common to general classification models, some peculiarites arise in the case of hazard assessment. This is due to the fact that the target for validation, the propensity to develop a dangerous characteristic, is not really known and must be estimated from a (usually) very small sample. This implies that the two types of errors (false positives and false negatives) should be given different meanings. Related to this, a very frequent situation is the presence of prevalence (different proportion of positive and negative cases) in the sample. It is shown that sample prevalence can have a dramatic effect in some very common validation statistics, like the confusion matrix and model efficiency. Here some statistics based on the confusion matrix are presented and discussed, and the use of threshold-independent approaches (especially the ROC plot) is shown. The ROC plot is also proposed as a convenient tool for decision-taking in a risk management context. A general scheme for hazard predictive modeling is finally proposed.

Keywords: geomorphological hazard modelling, probabilistic models, prediction errors, accuracy assessment, decision support, decision threshold, ROC plot

[1] fax: +31.302.531.145 email: S.Begueria@geog.uu.nl

1. Introduction

Model validation is a fundamental step in any natural hazards study. Validation refers to comparing the model predictions with a real-world data set, for assessing its accuracy or predictive power. Validation permits to establish the degree of confidence of the model, which is of great importance for transferring the results to the final users. Also, without a proper validation it is not possible to compare the model with other ones, or even with alternative sets of parameters or predictor variables. The evaluation of the model, on the other hand, refers to the assessment of its adequacy to the needs of the final users. In hazard analysis, this very often involves the delineation of zones with different hazard levels, that would lead to different management practices. In addition, a good validation and evaluation scheme can also provide feedback for improving the model.

The tipical case study in natural hazard analysis comprises a data set of study units (hillslopes, volcanos, grid pixels, etc) that can or can not develop a dangerous characteristic. This paper deals with predictive models that yield a continuous response variable expressing the degree of hazard or propensity to express a dangerous characteristic, what refers to different mathematical approaches:

- Bivariate analysis: a combined susceptibility index or a probability of occurrence is derived from the analysis of the influence of each explanatory variable. Several diferent methods have been published, from direct estimation (Clerici, 2002) to bayesian estimation or fuzzy-logic approaches (Lee et al., 2002).

- Multiple regression analysis: a linear relationship is used to predict a continuous characteristic of the dangerous phenomenon, like the

percentage of area affected, from a set of explanatory variables (eg. Carrara, 1983).

- Discriminant analysis: a function is determined that assigns discriminant scores to the study units. Usually, the units are classiffied according to the distances to the centroids of some a priori fixed response groups, but more refined rules can be used in a hazard analysis context. (eg. Lorente et al., 2002).

- General linear models: an extension of regression models allowing for non linear response functions. The mostly used example is the logistic regression, which yields directly a probability of occurrence of the dangerous phenomenon (eg. Bledsoe and Watson, 2001).

- The discussion is not only reduced to statistical approaches, as there are examples of physically based models with probabilistic components. Usually, probabilistic modules are included to account for uncertainty in parameter estimation (eg. Van Beek and Van Asch, 2004).

There are many examples of natural hazards analysed in a probabilistic way: volcanic eruptions (Perry et al., 2001), ice-jam induced flooding (Massie et al., 2002), channel instability (Bledsoe and Watson, 2001; Martínez-Casasnovas et al., 2003), gully erosion (Morgan and Mngomezulu, 2003), snow avalanches (Floyer and McClung, 2003). Among all natural hazards, the studies on slope instability have probably been the most commonly addressed by the methods mentioned above (i.e., Neuland, 1976; Rice and Pillsbury, 1982; Carrara, 1983; Furbish and Rice, 1983; Carrara et al., 1991; Chung et al., 1995; van Westen et al., 1997; Rowbotham and Dudycha, 1998; Chung and Fabbri, 1999; Dai and Lee, 2002; Santacana et al., 2003; etc).

Although a crucial step in predictive modelling, in many cases model validation is not given the necessary attention, and only very basic accuracy statistics are given. In most of the cases a classification threshold is set to allow the construction of confusion matrices and computation of classification statistics like the model efficiency (proportion of correctly classified observations). As it will be shown later on, this scheme is more adequate to pure classification studies than to predictive hazard models, where the meaning of false positives and false negatives (also known as error types I and II in many texts) can be significantly different. Also, as the subject of hazard analysis are by definition rare (unfrequent) processes, a very common situation is to deal with a great prevalence of negative cases (non observations of the dangerous phenomenon) in the sample. It will be shown that prevalence in the sample constitutes a great drawback for the use of some statistics widely used for model validation and model comparison. For this reason, an alternative set of statistics and the use of threshold-independent approaches like the ROC plot will be shown, and their use will be encouraged for the validation of hazard predictive models.

As it has been said, after building the model a decision threshold (cutoff value) is frequently set to divide the continuous response variable in two or more hazard classes. Although this is not strictly necessary (a continuous variable is certainly more informative than a sorted categorical scale), most of the final users will better handle a map with a legend with labels like ‘safe’, ‘probably safe’ and ‘unsafe’ than a cryptic numeric value. Although a continuous variable can be more meaningful to the researcher, in many occasions he will be requested to provide a threshold to discriminate between safe and potentially unsafe locations, for the model to be usefull in a decision-taking context (this is why I suggest the use of the term ‘decision threshold’, opposing to the word ‘classification threshold’ used above). To avoid

subjective thinking, setting a decission threshold should include an analysis of the costs of commiting positive and negative errors. The topic is also dealt with in this paper, and a modification of threshold-independent plots for error cost and decision analysis is proposed.

Finally, an alternative methodological scheme is proposed, that clearly separates validation and evaluation steps.

2. The confusion matrix and derived statistics

A common methodological scheme in hazard modelling is depicted in figure 1. As it can be seen, once a continuous reponse variable expressing the degree of hazard has been obtained, a classification threshold is set to divide the continuous variable into two or more classes. This categorical solution is normally considered the final product of the model, and validation is performed by comparing this prediction with the observations in a validation data set, different from the one used to build the model (for a complete discussion of sample partition for model validation, see Chung and Fabbri, 2003).

Note that in the this scheme the setting of a classification threshold is considered an integral part in the construction of the model. As the response variable yielded by the mathematical model has a continuous nature, this cutoff value is necessary to obtain a dichotomous variable (for the ongoing discussion more than two classes can be considered a set of dichotomous variables) that can be compared with the validation sample, that by definition has a binary nature. This is described in figure 2, where are plotted the frequency distributions corresponding to the two cohorts in the sample (cases with and without the dangerous characteristic, X_1 and X_0). For each one of the two cohorts in the validation sample one obtains a frequency distribution, according to

the scores given by the model. The classification threshold (dotted vertical line), that separates the cases predicted as safe (X'_0) and unsafe (X'_1), is usually set equal to the proportion of positive cases in the model sample. In an ideal situation, with perfect discrimination between the two groups, the two frequency distributions would appear separated in the plot. In most of the cases, however, a different degree of overlapping will occur, leading to prediction errors. In the figure, prediction errors have been marked with letters *b* and *c*. The set *b* are the false positives, or error type I in common statistical literature; the set *c*, on the other hand, represents the false negatives, or error type II. Sets *a* and *b*, respectively, group the true positives and true negatives.

As in many classification studies the cohorts tend to be more or less balanced, the threshold frequently has a value around 0.5 (figure 2 A). This is not the case, however, in most hazard studies, where the size of the two cohorts in the sample can differ in several orders of magnitude. This is the case described in figure 2 B. Note also the different meaning of both types of errors in hazard analysis. In most common classification studies (i.e. land use type from satellite imagery) false negatives (c) and false positives (b) are more or less equivalent (just something was classified in the wrong group). In hazard studies, however, one deals with a rare phenomenon, that can or can not have happened within the study period, but can happen in the future. False positives, in this context, can be either genuine assignment errors, or else real hazard-prone areas that have not yet developed the dangerous phenomenon. This is a very important fact that has to be kept in mind in predictive hazard analysis, and the discussion will reappear later on in this paper.Once a prediction threshold has been adopted, the binary predictions can be compared with the validation sample, allowing the construction of a confusion matrix (table 1). The confusion matrix shows the number of correctly and incorrectly predicted observations, for both positive and

negative cases. The letters in the cells correspond to that of figure 2 (see explanation above).

In table 2 are defined some statistics commonly used in classification and prediction models. Between them, the model efficiency (also referred as success rate) is the most frequent in the literature; it can be defined as the proportion of correctly classified observations, and for this reason it is sometimes considered equivalent to the R^2 statistic. Its opposite (rate of incorrect classified observations) is the missclassification rate. The positive predictive power is the proportion of true positives in the total of positive predictions, the negative predictive power being the contrary. The odds ratio (ratio between correctly and incorrectly classified observations) is the only statistic that makes use of all the values in the confusion matrix.

A very important drawback of the statistics presented in table 2 is that they are highly dependent on the proportion of positive and negative cohorts in the validation sample. If the sample presents high prevalence of one of the cohorts, as is normally the case in hazard studies, then columns X_1 and X_0 of the confusion matrix are not directly comparable, as their sums are not equal. For example, consider the case where the sample contains a very low proportion of positive cases (X_1). This will make values *a* and *c* of the confusion matrix much lower than their counterparts, *b* and *d*, thus affecting all the statistics presented in table 2 in the sense of making them more optimistic or 'liberal'. Paradoxically, in such a case the most efficient model would be to predict all places as safe (X'_0), as the true positives will be irrelevant compared to the true negatives! Despite this, the model efficiency is the only accuracy statistic reported in many studies, what constitutes an important drawback to evaluate and compare the different approaches.

For this reason, an alternative set of statistics, not relying in prevalence, is recommended (table 3). It can be seen that in their calculation the two groups in the validation sample are kept separated (columns X_0 and X_1 of the confusion matrix). The model's sensitivity expresses the proportion of positive cases correctly predicted, and can be considered the main statistic for expressing the predictive power of the model. It is analogous to the 'success rate' and 'prediciton rate' statistics defined by Chung and Fabbri (1999), and its use should be recommended instead of the more spreaded model efficiency. Specificity, on the other hand, is the proportion of negative cases correctly predicted. The false positive rate is defined as the proportion of false positives in the total of negative observations, and the false negative rate as the proportion of false negatives in the total of positive observations. The likelihood ratio makes use of all the values present at the confusion matrix.

The use of these accuracy measures is well established in other disciplines like Medicine (see, i.e., Forbes, 1995) or Ecology (i.e. Fielding and Bell, 1997), also dealing with predictive models of rare events. The particular meaning of false positives in this kind of models has to be emphasized again. As it has been explained above, false positives have to be thought as cases highly propense to develop the dangerous characteristic in the future, and not merely as classification errors. Reporting the model specificity is therefore very important, as it permits to describe a model as being pesimistic or 'conservative', if it is low specific (a big part of the study units are given high hazard susceptibility rates), or else optimistic or 'liberal', when it is high specific (only a small part of the units are predicted unsafe). Sensitivity and specificity are thus complementary statistics, as can be seen in the following example: Consider two different samples containing 100 study units each one, 50 of what present a dangerous characteristic (volcanic activity, slope instability, etc) in sample A, and only 5 in sample

B. Suppose that we build a predictive model for each one of the samples, yielding the same prediction rate of 0.8 (80% of positive cases correctly predicted, or sensitivity) when 50% of the cases are predicted as unsafe. This can be represented by the confusion matrices shown in table 4. In case A, also 80% of the negative cases would be correctly predicted, whereas in case B 98% of the negative cases would be predicted as potentially dangerous. It is clear that these two models could not be considered equal, but this difference is very difficult to express if a measure of specificity is not provided.

3. Threshold-independent methods: the ROC plot

The above defined statistics have in common that they need the establishment of a threshold value for their calculation. It should be stressed at this point that the selection of a threshold and the categorization of the response variable should not be a characteristic of the model itself, but a result of the use of the model in a specific context. For this reason, the validation of the model should not be based on one pre-determined threshold.

One way of achieving this would be plotting the different accuracy values obtained against the whole range of possible threshold values. That is exactly what a ROC (receiver-operating characteristic) plot does. The ROC plot was first introduced by Deleo (1993) in the field of signal processing to designate the performance of a system for classifying a variable into dichotomous classes.

An example of ROC plot is given in figure 3. The dots represent all the possible cutoff thresholds, corresponding to the cases in the sample. Although the threshold values are not represented directly in the ROC plot, it is easy to obtain them from the

data base. In figure 3 some threshold values (probabilities) have been marked for guidance.

For every point a different sensitivity / specificity pair of values is obtained (in some texts the value plotted in abscissas is 1 - specificity). These values indicate the ability of the model to correctly discriminate between positive and negative observations in the validation sample. See that they are directly related to the two errors, type II error being opposite of model's sensitivity and type I error opposite of model's specificity. In this sense, it is equivalent to speak in terms of error II / error I pairs, and for this reason secondary labels have been added to the plot.

It can be seen in the figure that for a low threshold the model will yield a high number of true positives (will be highly sensitive), but at the expense of having a high type I error. In the example presented in figure 3 we will obtain around 90% of true positives at a 0.05 threshold, but the type I error (false positives) will be also very high, around 65%. The opposite will occur if we take a high threshold. As stated above, the first case represents a conservative model, with the emphasis put on covering all the potentially dangerous study units, at the expense of including also some units that could not be really dangerous.

The area-under-ROC can serve as a global accuracy statistic for the model, independent of a single prediction threshold. This statistic varies between 0.5 (no improvement over random assignment, represented by the diagonal straight line) and 1 (perfect discrimination). It is clearly seen that the most separated the ROC curve appears in relation to the diagonal straight line, the better the model discriminates between safe and unsafe locations. This value can be approximated by finite differences:

$$S = \sum_{i=1}^{n+1} \frac{1}{2} \cdot \sqrt{(x_i - x_{i+1})^2} \cdot (y_i + y_{i+1}) \qquad \text{(eq. 1)}$$

being x_i the specificity and y_i the sensitivity at threshold i, and $x_{n+1} = 0$, $y_{n+1} = 1$. If the number of points in the sample is not enough to use this procedure, the area under ROC-curve can also be estimated by adjusting a polynomial curve and integrating.

Developing the idea of the success rate, Chung and Fabbri (1999) have proposed the so-called prediction rate curve (PRC), that has been used also by other authors (Lee et al., 2002, Remondo et al., 2003). Similarly to the ROC plot, the PRC shows the success rate (equivalent to the sensitivity) in ordinates, against the proportion of the total cases (map area, in the original work) predicted as positive in abscissa, for the whole range of possible thresholds. Like the ROC plot, the area under the curve can be used as a threshold-independent statistic, ranging in this case from 0 to 1. The PRC approach lacks an explicit representation of the model specificity, although it is implicit in the proportion of total cases cases predicted as positive, if one knows the proportion of positive and negative cases in the validation sample.

4. Error cost analysis and the use of the ROC plot for optimum decision threshold selecting

When evaluating the hazard of a dangerous natural phenomenon, the continuous response provided by the model should fulfill the requirements of the researcher. In this sense, the ROC plot and the area-under-ROC statistic permit to evaluate the model's performance independently of a determined cutoff value. In a risk management context, however, researchers are often asked for a decision threshold to determine if a given place is safe or unsafe, what will determine the prevention measures undertaken.

Contrary to what has been explained before, here the selection of a cutoff value belongs to the practical use of the hazard model.

In a purely theoretical hazard study the two types of errors are perfectly equivalent in importance, although they mean different things. For the researcher on natural hazards, a false positive may mean that a given place is potentially dangerous, even that no dangerous activity has been observed there. A false negative, on the other hand, means that the model has not been capable of predicting the potential hazard. The analysis of the two types of errors should provide useful information for improving the model. For the risk manager, however, the two types of errors have a very different meaning. A very pesimistic model, containing a great number of false positives, can imply the loss of a potentially safe space, or even the uselessness of the investments made for prevention. But a false positive error may signify the loss of lives or the destruction of infrastructure.

The ROC plot can be used to support decision taking for a given place. Suppose, for example, that a certain slope is given a probability of failure of 0.1 within a given time period (see figure 3). From the ROC plot we see that we have two choices: we can state that it is unsafe with 80% probability of being right (true positive), or we can say that the slope is safe, with 63% probability (true negative). Suppose now that we decide that the slope is unsafe, so we are to recommend some prevention measures, with a total cost of 2000€. The probability of making a type I error (false positive) is 37%, so the net cost at risk would be: 2000 * 0.37 = 740. Otherwise, we can declare the slope as safe, and do not recommend any correction measures. Despite this, a landslide can still occur, with probability 20% (error type II). Suppose that this landslide would bury some infrastructure, with a total cost of repairing it of 1000€. This makes: 1000 * 0.2 = 200. Comparing the net costs of making type I and type II errors, the less expensive option is

to consider the slope as safe and do not take any correction measures, even if the initial probabilities were favourable to the unsafe option.

When considering a greater area instead of a single point, the ROC plot can also be modified into an error-cost plot to select the most convenient decision threshold for the whole zone, as shown in figure 4. The secondary axes (errors) have been modified to express the net costs of both error types. The optimum threshold should be the value that minimises the total error costs, integrated over the whole area. In the ROC plot, this is the value where the two cost-weighted errors are approximately equal (a value of 0.1475 in the example shown in figure 4).

5. Alternative methodology

From the discussion above an alternative methodology based on the use of threshold-independent methods can be proposed (see flowchart in figure 5). As it has been stressed along this paper, the construction and the use of the model should be separated. The validation of the model can be done without the need of a predefined threshold, by a threshold-independent method like the ROC plot. The evaluation of the model is done afterwards, including the selection of one or more decision thresholds, also with the aid of the ROC plot. The evaluation step should answer the question of how good is the model in stating the security or safety of a given place or study area. An idea of this can be obtained by the confusion matrix and derived accuracy statistics, once one or more decision thresholds have been set.

6. Conclusions

In this paper the importance of the validation and evaluation steps in model design is encouraged. It has been shown that validation can provide the researcher with very useful information for improving the model, but it is also important to give the final users of the model an idea about the confidence of the model results. The evaluation of the model permits to adapt the model results to the needs of the final users.

After fitting the mathematical model, the usual methodology consists in establishing a threshold or cutoff value to divide the response variable into dichotomous classes. Then, one or more statistics based in the confusion matrix are calculated for validating the model. The threshold value is normally fixed equalling the prior probabilities for the dangerous phenomenon, estimated by its sampling rate. However, the setting of a threshold is more a question of use of the model than a characteristic of the probabilistic model itself. For this reason, the use of threshold-independent validation methods is proposed.

The construction and the use of ROC (receiver-operating characteristic) plots has been shown. The ROC plot, and the area-under-ROC statistic, provide a complete validation scheme without depending on a pre-defined threshold. The ROC plot can also be used afterwards as an error cost analysis tool to assist in selecting a decision threshold for risk management.

An alternative methodology for probabilistic hazard analysis has been proposed. The use of threshold independent methods is recommended in the validation step. They can also be used during the evaluation step to provide the final users with one or more alternative decision thresholds. After that, several accuracy statistics based in the confusion matrix can be calculated to express the confidence of the model at this specific thresholds.

The influence of sample prevalence (different proportion of positive and negative cases) in several very common accuracy statistics has been shown, and alternative measures have been proposed.

Acknowledgements

This study has been supported by the research projects "Procesos hidrológicos en cuencas pirenaicas en relación con los cambios de uso del suelo y las fluctuaciones climáticas" (PIRIHEROS, REN2003-08678/HID), and "Procesos hidrológicos en áreas seminaturales mediterráneas" (PROHISEM, REN2001-2268-C02-01/HID), funded by the Spanish Government (CICYT). I also wish to acknowledge personal support by a post-doctoral grant funded by the Spanish Government Secretary for Education and Universities and the European Social Fund.

References

Bledsoe, B.P. and Watson, C.C., 2001. Logistic analysis of channel pattern thresholds: meandering, braiding, and incising. Geomorphology, 38(3-4): 281-300.

Carrara, A., 1983. Multivariate models for landslide hazard evaluation. Mathematical Geology, 15(3): 403-426.

Carrara, A., Cardinali, M., Detti, R., Guzzetti, F., Pasqui, V., Reichenbach, P., 1991. GIS techniques and statistical models in evaluating landslide hazard. Earth surface processes and landforms, 16: 427-445.

Chung, C.F. and Fabbri, A.G., 1999. Probabilistic Prediction Models for Landslide Hazard Mapping. Photogrammetric Engineering & Remote Sensing, 65(12): 1389-1399.

Chung, C.F. and Fabbri, A.G., 2003. Validation of spatial prediction models for landslide hazard mapping. Natural Hazards, 30: 451-472.

Chung, C.F., Fabbri, A. and Van Westen, C.J., 1995. Multivariate regression analysis for landslide hazard zonation. In: A. Carrara and F. Guzzetti (Editors), Geographical Information Systems in assessing natural hazards. Kluwer Academic Publishers, The Nederlands, pp. 107-133.

Clerici, A., Parego, S., Tellini, C. and Vescovi, P., 2002. A procedure for landslide susceptibility zonation by the conditional analysis method. Geomorphology, 48, 349-364.

Dai, F.C. and Lee, C.F., 2002. Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. Geomorphology, 42(3-4): 213-228.

Deleo, J.M. (1993). Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. In: Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis, pp. 318-325. College Park, Computer Society Press.

Fielding, A.H. and Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence / absence models. Environmental Conservation, 24(1): 38-49.

Floyer, J.A. and McClung, D.M., 2003. Numerical avalanche prediction: Bear Pass, British Columbia, Canada. Cold Regions Science and Technology, 37(3): 333-342.

Forbes, A.D., 1995. Classification-algorithm evaluation: five performance measures based on confusion matrices. Journal of Clinical Monitoring, 11(3): 189-206.

Furbish, D.J. and Rice, R.M., 1983. Predicting landslides related to clearcut logging, Northwestern California, USA. Mountain Research and Development, 3(3): 253-259.

Lee, S., Choi, J. and Min, K., 2002. Landslide susceptibility analysis and verification using the Bayesian probability model. Environmental Geology, 43: 120-131.

Lorente, A., García-Ruiz, J.M., Beguería, S. and Arnáez, J., 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.

Martínez-Casasnovas, J.A., Ramos, M.C. and Poesen, J., 2003. Assessment of sidewall erosion in large gullies using multi-temporal DEMs and logistic regression analysis. Geomorphology, In press.

Massie, D.D., White, K.D. and Daly, S.F., 2002. Application of neural networks to predict ice jam occurrence. Cold Regions Science and Technology, 35(2): 115-122.

Morgan, R.P.C. and Mngomezulu, D., 2003. Threshold conditions for initiation of valley-side gullies in the Middle Veld of Swaziland. Catena, 50(2-4): 401-414.

Neuland, H., 1976. A prediction model of landslips. Catena, 3: 215-230.

Perry, F.V., Valentine, G.A., Desmarais, E.K. and WoldeGabriel, G., 2001. Probabilistic assessment of volcanic hazard to radioactive waste repositories in Japan: Intersection by a dike from a nearby composite volcano. Geology, 29(3): 255-258.

Remondo, J., González-Díez, A., Terán, J.R.D.D. and Cendrero, A., 2003. Landslide Susceptibility Models Utilising Spatial Data Analysis Techniques. A Case Study from the Lower Deba Valley, Guipuzcoa (Spain). Natural Hazards, 30(4): 233-249.

Rice, R.M. and Pillsbury, N.H., 1982. Predicting landslides in clearcut patches, Symposium on recent development in the explanation and prediction of erosion and sediment yield. Internation Association of Hydrological Sciences, pp. 303-311.

Rowbotham, D.N. and Dudycha, D., 1998. GIS modelling of slope stability in Phewa Tal watershed, Nepal. Geomorphology, 26(1-3): 151-170.

Van Beek, R. and Van Asch, T., 2004. Regional assessment of the effects of land-use change on landslide hazard by means of physically based modelling. Natural Hazards, 31: 289–304.

Van Westen, C.J., Rengers, N., Terlien, M.T.J. and Soeters, R., 1997. Prediction of the occurrence of slope instability phenomenal through GIS-based hazard zonation. Geologische Rundschau (International Journal of Earth Sciences), 86(2): 404 - 414.

		Observed	
		X_1	X_0
Predicted	X'_1	*a*	*b*
	X'_0	*c*	*d*

Table 1. Confusion matrix. *a*, true positives; *b*, false positives (error type I); *c*, false negatives (error type II); *d*, true negatives.

Efficiency	$(a + d) / N$	Proportion of correctly classified observations
Misclassification rate	$(b + c) / N$	Proportion of incorrectly classified observations
Odds ratio	$(a + d) / (b + c)$	Ratio between correctly and incorrectly classified cases
Positive predictive power	$a / (a + b)$	$p(X_1\|X'_1)$, or the proportion of true positives in the total of positive predictions
Negative predictive power	$d / (c + d)$	$p(X_0\|X'_0)$, or the proportion of true negatives in the total of negative predictions

Table 2. Accuracy statistics derived from the confusion matrix

Sensitivity	*a / (a + c)*	$p(X'_1\|X_1)$, or the proportion of positive cases correctly predicted
Specificity	*d / (b + d)*	$p(X'_0\|X_0)$, or the proportion of negative cases correctly predicted
False positive rate	*b / (b + d)*	$p(X'_1\|X_0)$, or the proportion of false positives in the total of negative observations
False negative rate	*c / (a + c)*	$p(X'_0\|X_1)$, or the proportion of false negatives in the total of positive observations
Likelihood ratio	*sensitivity / (1- specificity)*	Ratio between true positive and false negative fractions

Table 3. Some accuracy statistics not depending on prevalence

A)

		Observed X_1	Observed X_0	
Predicted	X'_1	*40*	*10*	*50*
	X'_0	*10*	*40*	*50*
		50	*50*	

B)

		Observed X_1	Observed X_0	
Predicted	X'_1	*4*	*46*	*50*
	X'_0	*1*	*49*	*50*
		5	*95*	

Table 4. Confusion matrices of two models exhibiting same sensitivity but greatly differing in specificity

Figure captions

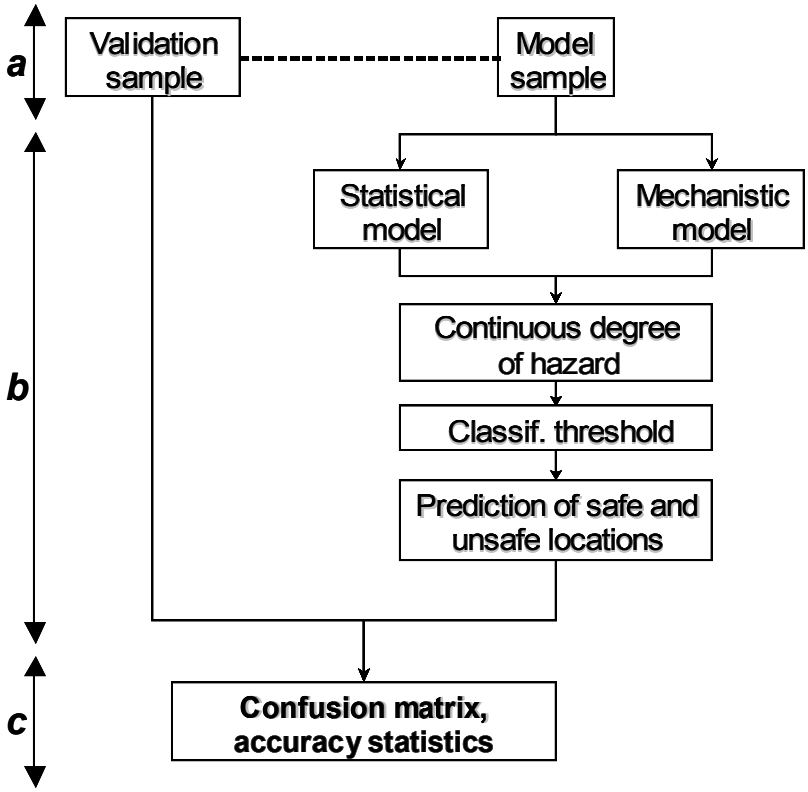
Fig. 1. Flowchart of a common probabilistic model design. *a*: sampling; *b*: model construction; *c*: model validation

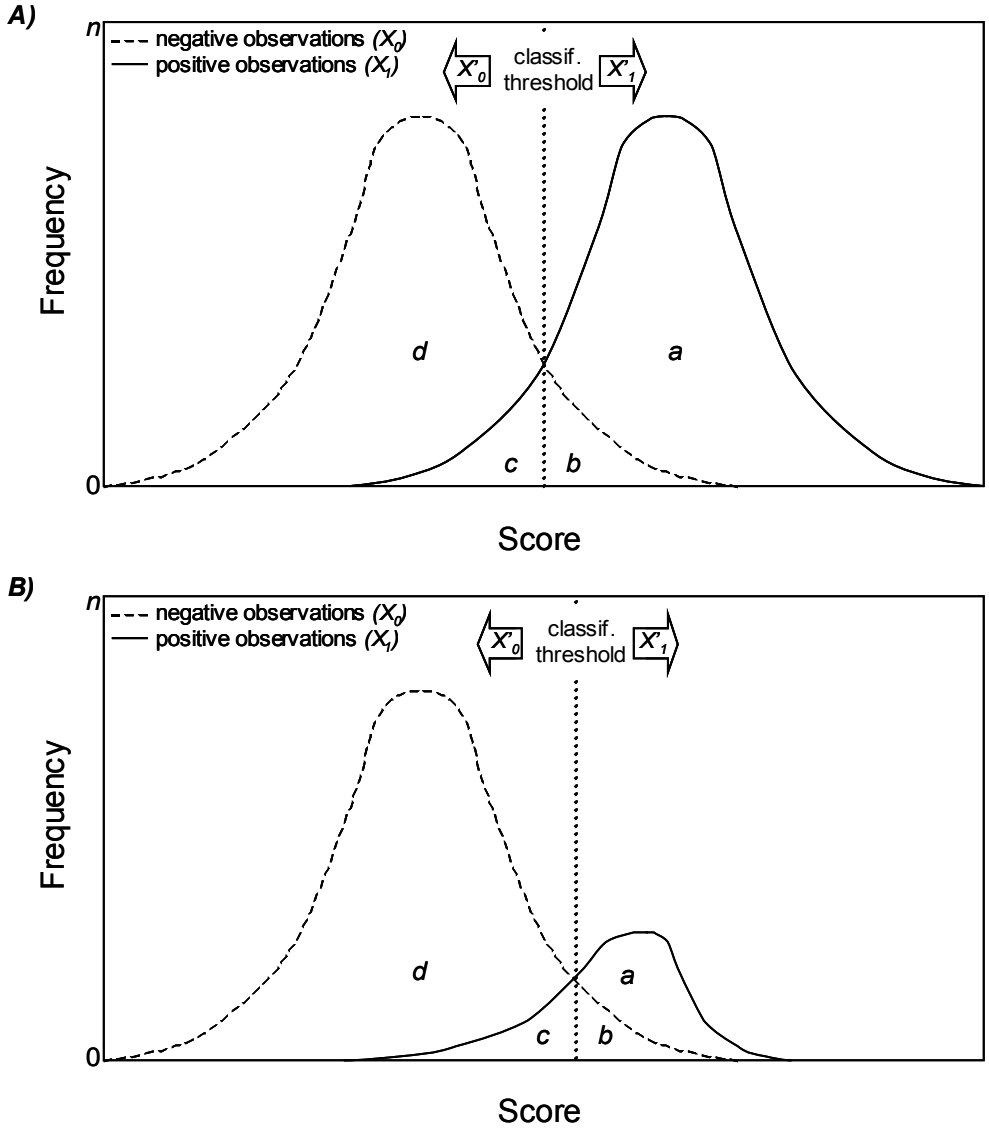
Fig. 2. Frequency distributions for the negative and positive groups, and the role of the prediction threshold. A) Equal groups. B) Unequal groups. *a*, true positives; *b*, false positives (error type I); *c*, false negatives (error type II); *d*, true negatives.

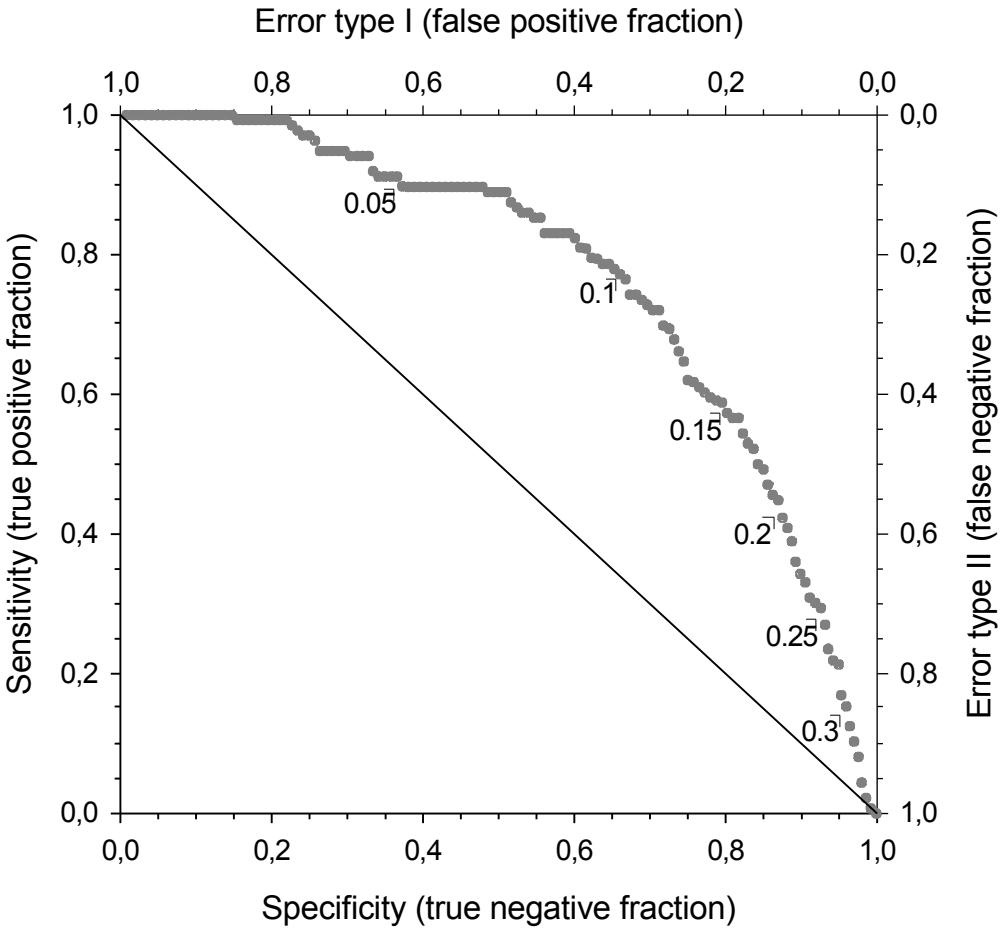
Fig. 3. Example of a ROC plot

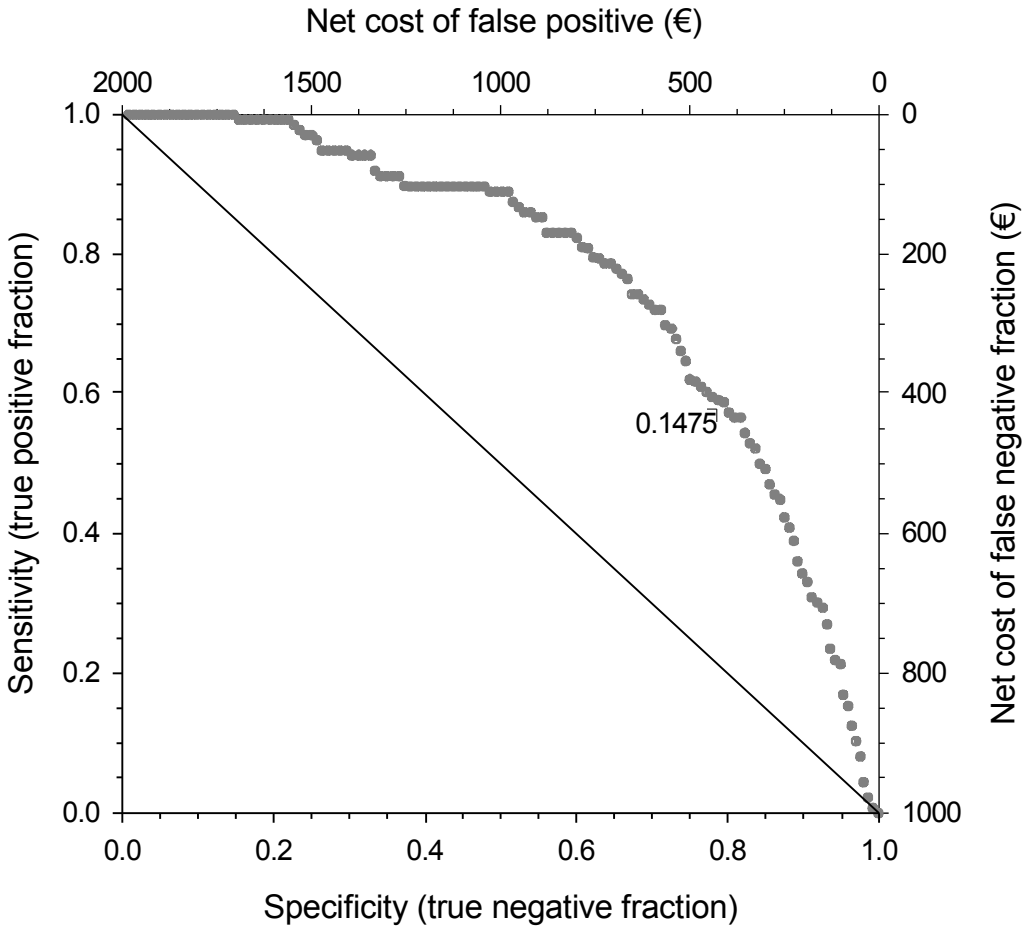
Fig. 4. Cost / benefit ROC plot

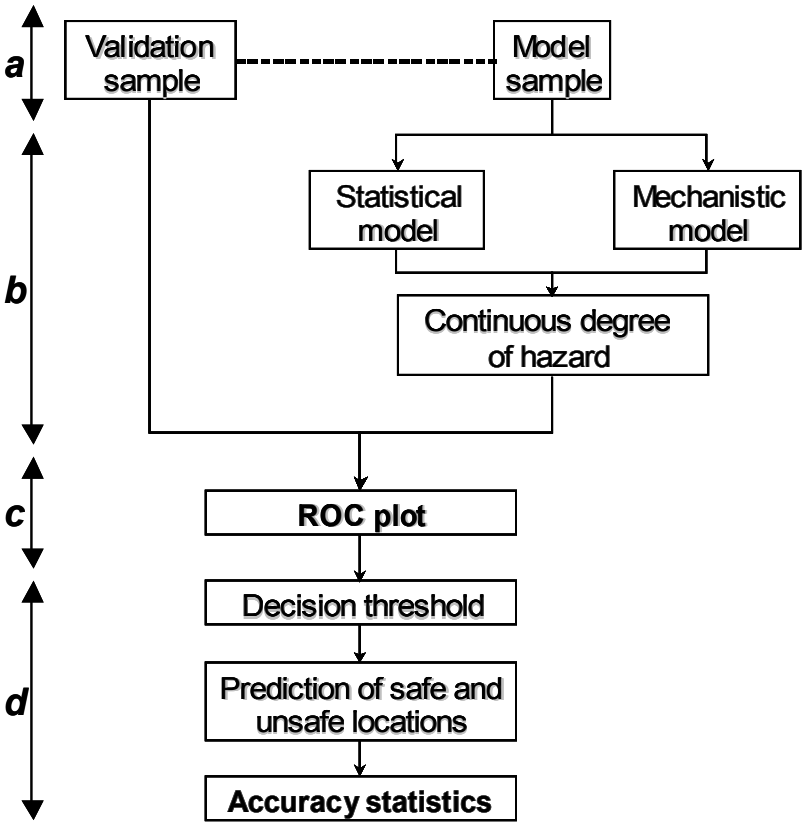
Fig. 5. Flowchart of alternative methodology based in threshold-independent methods. *a*: sampling; *b*: model construction; *c*: model validation; *d*: model evaluation


a
b
c
Validation sample
Model sample
Statistical model
Mechanistic model
Continuous degree of hazard
Classif. threshold
Prediction of safe and unsafe locations
Confusion matrix, accuracy statistics


A)
n
negative observations (X0)
positive observations (X1)
X'0
classif.
threshold
X'1
Frequency
d
a
c
b
0
Score
B)
n
negative observations (X0)
positive observations (X1)
X'0
classif.
threshold
X'1
Frequency
d
a
c
b
0
Score


Error type I (false positive fraction)
1,0
0,8
0,6
0,4
0,2
0,0
Sensitivity (true positive fraction)
Error type II (false negative fraction)
0.05
0.1
0.15
0.2
0.25
0.3
Specificity (true negative fraction)


Net cost of false positive (€)
2000
1500
1000
500
0
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Sensitivity (true positive fraction)
Net cost of false negative fraction (€)
0.1475
0.0
0.2
0.4
0.6
0.8
1.0
Specificity (true negative fraction)


a
b
c
d
Validation sample
Model sample
Statistical model
Mechanistic model
Continuous degree of hazard
ROC plot
Decision threshold
Prediction of safe and unsafe locations
Accuracy statistics