Nat. Hazards Earth Syst. Sci., 12, 1937–1947, 2012
www.nat-hazards-earth-syst-sci.net/12/1937/2012/
doi:10.5194/nhess-12-1937-2012

Natural Hazards and Earth System Sciences

Logistic regression applied to natural hazards: rare event logistic regression with replications

M. Guns[1,2] **and V. Vanacker**[1,*]

[1]Université catholique de Louvain, Earth and Life Institute, Georges Lemaître Centre for Earth and Climate Research, Place Louis Pasteur 3 boîte L4.03.08, 1348 Louvain-la-Neuve, Belgium
[2]Fund for Scientific Research – FNRS, Rue d'Egmont 5, 1000 Brussels, Belgium

[*]*Invited contribution by V. Vanacker, recipient of the EGU Division Outstanding Young Scientists Award 2012.*

Correspondence to: M. Guns (marie.guns@uclouvain.be)

Received: 11 July 2011 – Revised: 4 April 2012 – Accepted: 20 May 2012 – Published: 18 June 2012

Abstract. Statistical analysis of natural hazards needs particular attention, as most of these phenomena are rare events. This study shows that the ordinary rare event logistic regression, as it is now commonly used in geomorphologic studies, does not always lead to a robust detection of controlling factors, as the results can be strongly sample-dependent. In this paper, we introduce some concepts of Monte Carlo simulations in rare event logistic regression. This technique, so-called rare event logistic regression with replications, combines the strength of probabilistic and statistical methods, and allows overcoming some of the limitations of previous developments through robust variable selection. This technique was here developed for the analyses of landslide controlling factors, but the concept is widely applicable for statistical analyses of natural hazards.

1 Introduction

Natural hazards and risks are increasing, especially in developing countries (Alcántara-Ayala et al., 2006). Landslides are particularly affecting human occupation and socio-economic development in mountainous areas of developing countries. Given the current population growth with increasing occupation of steep uplands, landslide risks are expected to increase in the future. Understanding the causal and controlling factors of landsliding is therefore important. It is known that extreme rainfalls, rapid snowmelt or seismic activities are the primary triggers of landslides (Brunetti et al., 2010; Tatard et al., 2010). Prediction of the timing of future landslide occurrence is rare, as landslide records often do not contain detailed information on the date of occurrence (Baum and Godt, 2010; Larsen and Torres-Sánchez, 1998).

Prediction of the areas that are particularly sensitive to landsliding through the development of stochastic or process-based susceptibility models has been the goal of extensive research (Brenning, 2005; Dai et al., 2002; Guzzetti et al., 2006; Komac, 2006). Various techniques have been used in the past to analyse landslide controlling factors (see overviews of Dai et al., 2002; Guzzetti et al., 1999; Huabin et al., 2005). Process-based susceptibility models often focus on the rheological parameters of the sliding mass, while stochastic models are mainly based on the biophysical site conditions. Frequently used stochastic techniques are discriminant analysis and regression techniques (Atkinson et al., 1998; Guzzetti et al., 2006). Logistic regression is particularly interesting for landsliding susceptibility analysis as it models the relationship between a dichotomous variable (presence/absence of landslide) and a set of independent biophysical site variables (controlling factors). This technique allows evaluating the probability of landslide occurrence and its significance, and has been widely used for landslide susceptibility mapping (e.g. Atkinson and Massari, 1998; Ayalew and Yamagishi, 2005; Dai and Lee, 2002, 2003; Dai et al., 2001; Vanacker et al., 2003).

Logistic regression techniques have to be adapted to the specificities of landslide analysis, as landslides (like many other natural hazards) can be considered to be rare events (Demoulin and Chung, 2007). Rare events have occurrence frequencies that are low (Maalouf and Trafalis, 2011), with the number of events in the dataset dozens to thousands times smaller than the number of non-events. King and Zeng (2001a, b) have shown that rare events are difficult

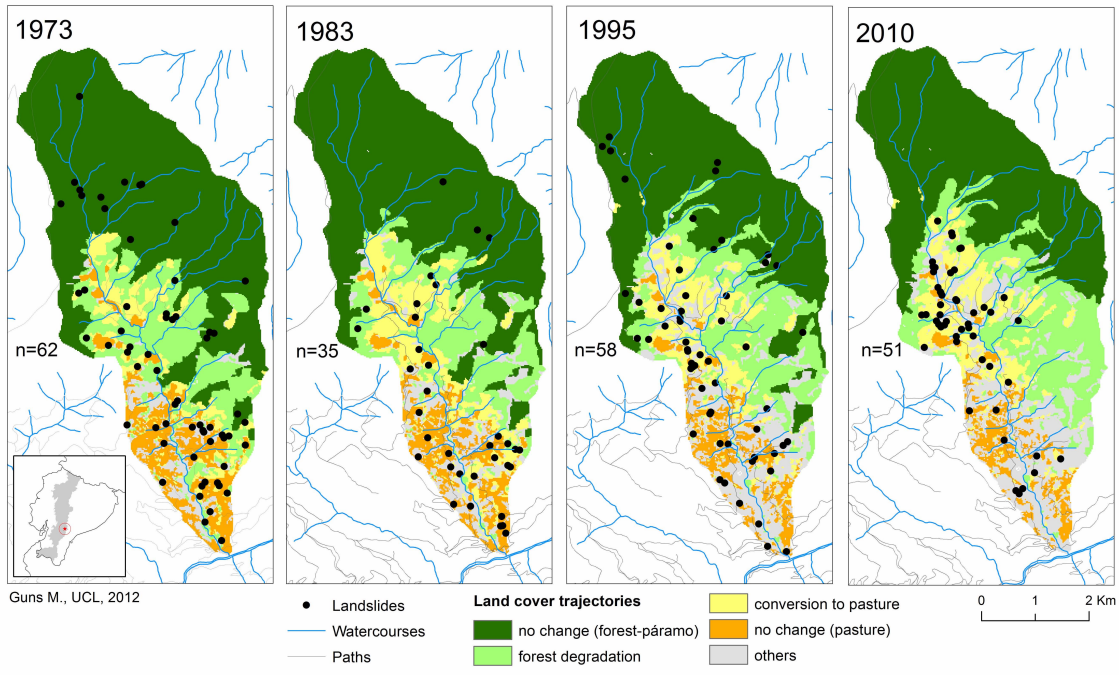

Fig. 1. Time series (1973–2010) of landslide occurrence and land cover in the Llavircay catchment (southern Ecuadorian Andes).

to predict as the standard application of logistic regression techniques can sharply underestimate the probability for rare events. Rare-event logistic regression was proposed by King and Zeng (2001a, b) to correct this bias by (i) an endogenous stratified sampling of the dataset, (ii) a prior correction of the intercept and (iii) a correction of the probabilities to include the estimation uncertainty.

In this paper, we first evaluate the use of probabilistic approaches to detect landslide controlling factors. Then, we build some concepts from probabilistic theory into rare event logistic regression analysis. This technique, called rare event logistic regression with replications allows overcoming some of the limitations of previous developments, and offers a robust variable selection. We apply this technique here for the analyses of landslide controlling factors, but the concept is widely applicable for statistical analyses of natural hazards.

2 Landslide occurrence in Llavircay, as a case study

The Llavircay catchment was selected for the development of the rare event logistic regression technique with replications. The study area of 24 km^2 is located in the tropical Andes (Fig. 1), and is subjected to a warm and humid tropical climate (Winckell et al., 1997). The mean annual precipitation is about 1330 mm, the average temperature is 10 °C and the atmospheric humidity is high, 87 % on average (Acotecnic, 2006; INAMHI, 2008). The elevation varies from 2017 m to 3736 m and slopes reach up to 55°. With a mean slope angle of 26°, the topography can be considered as very rough. About one third of the area has slope angles that are above the mean angle of internal friction (estimated at 30° according to Basabe, 1998). Landslides and creep are abundantly present in the area. Inventories of mass movements created from aerial photo interpretation and field campaigns revealed 206 landslides (reactivation excluded) between 1973 and 2010. They are mainly earth slides (translational slides) and earth slumps (rotational slides) according to the classification of Varnes (Summerfield, 1991).

Land cover change (1963–2010) was documented using four sets of archived aerial photographs for the time period 1963–1995, complemented with a field survey in 2010. Because of significant differences in quality and scale, between and within the aerial photographs, the land cover classification was performed manually using a WILD stereoscope. Six land cover classes were identified: (i) dense forest; (ii) degraded forest, as a result of selective logging (Sierra and Stallings, 1998); (iii) bushes, as a result of natural regeneration or so called matoral in Ecuador; (iv) pasture with sporadic trees; (v) pasture; and (vi) subpáramo and páramo corresponding to the natural shrub and grassland found at high altitudes in the Andes (Luteyn, 1999). More details on the land cover classification are given in Vanacker et al. (2000). Based on the time series of land cover data from 1963–2010, land cover change trajectories were created (Fig. 1). Land cover changed rapidly in this area, with half of the primary forest disappearing since 1963. In 2010, about half of the catchment was covered by trees, a quarter by páramo, subpáramo and bushes, and a quarter of the area is covered by pastures.

3 Materials and methods

3.1 Database creation

Potential anthropogenic and biophysical explanatory variables of landslide occurrence have been selected based on literature and data availability. The following explanatory variables were included in the analyses: slope, distance to watercourse, distance to path, curvature and different trajectories of land cover change. The first three variables are quantitative, while the two last ones are qualitative variables composed of respectively three and five classes (Table 1). All our data are spaced in maps in a grid-cell mapping unit, as it is very common nowadays with GIS utilisation (Guzzetti et al., 1999). The GIS grid-data has been transformed into a matrix format: an attribute table in which the lines correspond with the 20 m resolution pixels of the catchment and the columns with the 11 potential explanatory variables. A similar attribute table was made for the landslide inventories. In order to avoid auto-correlation, we represented every landslide by one grid-cell (pixel) located in the centre of the shear plane. For the logistic regressions, one matrix was established including the matrix of GIS grid-data and spatial information on the observed landslide occurrence in the catchment. For all grid-cells, the value of a dichotomous dependent variable landslide indicates the presence (landslide = 1) or absence (landslide = 0) of a landslide. The matrices were imported in R software for the probabilistic and statistical analyses.

3.2 Probabilistic approach based on Monte Carlo methods

Probabilistic approaches are useful in landslide analyses, as they allow determining the probabilities of sliding for different biophysical and anthropogenic site conditions. The basic principle that is behind these analyses is the hypothesis that landslides have specific site characteristics that differ from the overall environmental setting of the area. A statistical nonparametric test that is commonly used to compare differences between groups is the Wilcoxon rank-sum test also called Mann-Whitney U test (Crawley, 2005). Such test works with ordinal data and with groups of more or less similar size. In our case, some explanatory variables, such as land cover trajectories, are nominal. Moreover, the group "event" (presence of landslide) is much smaller than the group "non-event" (absence of landslide). A comparison of the two groups based on their distribution is thus not appropriate. So we apply a probabilistic approach based on Monte Carlo methods (Sawilowsky, 2003; Vanacker et al., 2001).

The main idea is to test if significant associations exist between explanatory variables and the location of landslides by comparing our landslide sample to a bundle of randomly selected samples in the study area.

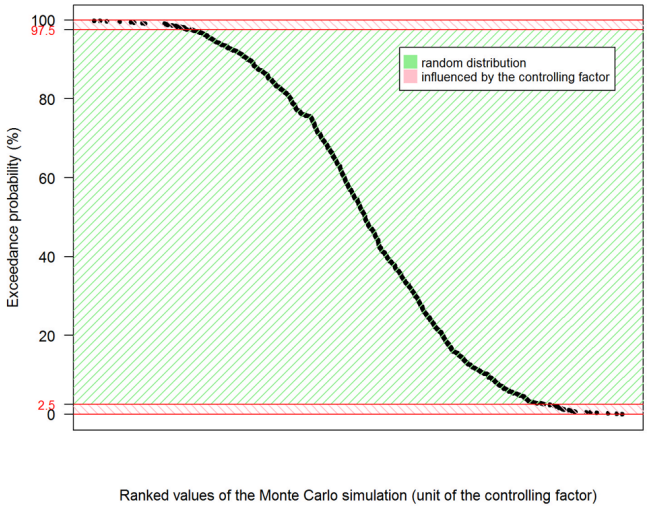

Fig. 2. Plot of the exceedance probability against theoretical ranked randomly sampled values with a confidence interval of 95.

The probability of having a sample with a given distribution of observations over each class of a qualitative variable (or with a median value for a quantitative value) is given by its exceedance probability. So, if we randomly selected enough samples to approximate the reality in the catchment (according to Monte Carlo methods, Sawilowsky, 2003), a plot of their exceedance probability against any potential explanatory variable will give a curve that represents the distribution of randomly selected samples in the study area (Fig. 2). Thus, for every variable, the exceedance probability of the landslide sample can be derived from the plot and be compared with a given significance level. If the exceedence probability lies outside the probabilities for the confidence interval, we can conclude that the distribution of landslides over this explanatory variable is not random (Vanacker et al., 2001).

The first stage is to create the exceedance probability curve of the explanatory variable analysed. For a given year Y and an explanatory variable X, a simulation is composed of k samples of N randomly selected points. k is the number of samples needed to obtain a stable empirical probability distribution of the population, and equals 1000 according to our sensibility analysis for the Llavircay case-study (Fig. 3). N is the number of landslides observed in year Y. Samples are considered to be independent, as each sample contains less than 0.002 % of the entire population. Note that the explanatory variable X can be quantitative (e.g. slope) or qualitative (e.g. land cover trajectory). Code was written for R software to automate the procedure for simulations, and consists of the following steps: (1) import the matrix with the anthropogenic and biophysical explanatory variables for all grid cells in the catchment, (2) randomly select from this matrix N points, (3) calculate the median value of the explanatory variable X for the N points (if the variable is quantitative)

Table 1. Set of anthropogenic and biophysical variables included in the probabilistic and statistical analyses.

Explanatory variables		Range of values	% of total area (variation between 1973 and 1995 for trajectory)
Slope		0–55°	
Distance to path		0–2721 m	
Distance to watercourse		0–1219 m	
Plan curvature	concave	0 or 1	42 %
	rectilinear	0 or 1	22 %
	convexe	0 or 1	36 %
Land cover trajectory	no change (forest-páramo)	0 or 1	−16 %
	forest degradation	0 or 1	+3 %
	conversion to pasture	0 or 1	+3 %
	no change (pasture)	0 or 1	−4 %
	others	0 or 1	+14 %

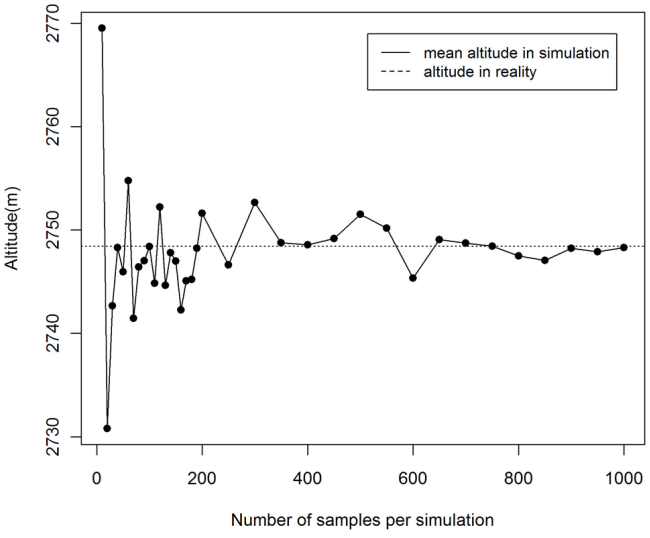

Fig. 3. According to the Monte Carlo principle, a sufficient number of randomly selected samples are needed to approximate correctly the true population (Llavircay catchment in this case). A simulation with 1000 samples provides a stable empirical approximation.

or the frequency of each class (if the variable is qualitative), (4) repeat steps 2 and 3, k times, (5) summarise in a table the k median values (or class frequencies) obtained in step 3 and rank them in an ascending order, (6) calculate the exceedance probability of the k median values or class frequencies as follows:

$$P(X \geq X_j) = 1 - F_X(X_j) = \frac{(k-j)}{(k+1)} \quad (1)$$

where $F_X(\mathrm{X}_j)$ is the cumulative density function of X, k is the number of samples created for the simulation, X_j is a given sample of the population, and j is the rank number after ordering the randomly selected samples, (7) plot the exceedance probabilities calculated in step 6 against the ranked sampled values calculated in step 5 (an example of such a plot can be seen in Fig. 2).

The second stage is to see if landslides are randomly distributed over the explanatory variable X. We derive the exceedance probability of the N landslides observed in year Y by transferring on the plot created in step 7 the median value (or class frequency) of the landslide inventory for variable X. We compare this exceedance probability with a given significance level. As frequently used in literature, the critical value (also called p-value) is here fixed at 5 %. If the exceedence probability lies outside the probabilities for the confidence interval, we can conclude that the distribution of landslides over this explanatory variable is not random (Fig. 2). We can thus assume that the explanatory variable X could be a controlling factor of landslide occurrence. This procedure is repeated for every explanatory variable X and every year Y.

3.3 Ordinary rare event logistic regression

Logistic regression is commonly used to analyse the dependency of a dichotomous variable, here landslide presence/absence, on a set of explanatory variables (Atkinson and Massari, 1998; Vanacker et al., 2003). The ordinary logistic model can be written as Eq. (2) (Kleinbaum and Klein, 2010):

$$p_i = \frac{1}{1 + e^{-(\hat{\alpha} + \sum \hat{\beta}_i X_i)}} \quad (2)$$

where p_i denotes, in our case, the probability of an event as a function of m independent variables X and i ranges from 1 to m. The terms $\hat{\alpha}$ and $\hat{\beta}$ are unknown parameters that are estimated from the data by the maximum likelihood method. This equation is often linearized by a logit transformation, the natural logarithm of the odd which is the ratio of the probability of events divided by the probability of

non-events. The logit form of the model can be expressed as Eq. (3) (Kleinbaum and Klein, 2010):

$$\text{logit}\, p_i = \ln \frac{p_i}{1-p_i} = \hat{\alpha} + \sum \hat{\beta}_i X_i \quad (3)$$

In the case of natural hazards, the total number of grid-cells that are affected by an event (such as landslide occurrence) is often much smaller than the total number of grid-cells in the study area. It is common that less than 1 % of the study area is affected by a natural hazard event. King and Zeng (2001a, b) have shown that ordinary logistic regression strongly underestimates the probability of occurrence of rare events. They developed for political sciences an adapted version of the logistic regression technique, so-called "rare event logistic regression", that includes three corrections measures for rare event data. They first recommended the utilisation of a choice-based (or case-control) sampling design based on endogenous stratified sampling (Ramalho, 2002). It consists of taking all the events (1 s) and a random selection of the non events (0 s). The proportion of events to non events is often set at one to ten (Beguería, 2006a). The use of choice-based sampling designs might significantly bias the estimation of the intercept term $\hat{\alpha}$. Therefore, a prior correction is needed to avoid sampling bias (King and Zeng, 2001a). The corrected intercept term, α_0, is calculated based on the intercept estimate, $\hat{\alpha}$, and the fraction of 1 s in the population, τ, and the fraction of 1 s in the sample, $\bar{\gamma}$ as in Eq. (4):

$$\alpha_0 = \hat{\alpha} - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{\gamma}}{1-\bar{\gamma}}\right)\right] \quad (4)$$

The second adaptation aims to correct for the underestimation of the probabilities when using the corrected intercept α_0 in Eq. (3). A correction factor C_i is thus added to the estimated probability $\tilde{p}_i$ (Eq. 5):

$$p_i = \tilde{p}_i + C_i \quad (5)$$

For each observation, C_i can be calculated from Eq. (6) (King and Zeng, 2001a; Van Den Eeckhaut et al., 2006):

$$C_i = (0.5 - \tilde{p}_i)\tilde{p}_i(1-\tilde{p}_i)X_0 V(\beta) X_0' \quad (6)$$

where $\tilde{p}_i$ is the event probability estimated using the bias-corrected coefficient α_0, X_0 is a $1 \times (m+1)$ vector of values for each explanatory variable, X_0' is the transpose of X_0 and $V(\beta)$ is the variance-covariance matrix.

The rare-event logistic regression was first applied in landslide susceptibility analysis by Van Den Eeckhaut et al. (2006). To our knowledge, this method has been applied in natural hazard analyses since then only by Bai et al. (2011) and Vanwalleghem et al. (2008). We slightly modified the methodological description of Van Den Eeckhaut et al. (2006) and automated the statistical procedure entirely in R software. For the endogenous stratified sampling, a proportion of 1:10 for the ratio of events to non events was used following Beguería (2006a). To avoid multi-collinearity among the independant variables, we calculated the Variation Inflation (VIF) and Tolerance (TOL) factors. All explanatory variables with a VIF > 2 and TOL < 0.4 were excluded from the stepwise logistic regression (Allison, 2001). From this selection, only the explanatory variables that significantly explain the landslide distribution pattern (at a significance level of 0.05) were included in the rare event logistic regression. The "relogit" function from the R package Zelig (Imai et al., 2009) was used to implement the rare event logistic regression.

3.4 Rare event logistic regression with replications

Rare event logistic regression with replications combines the strength of probabilistic and statistical methods. It is based on the statistical method of rare-event logistic regression (King and Zeng, 2001a; Van Den Eeckhaut et al., 2006), but it includes probabilistic techniques to estimate the robustness of the regression estimates (Beguería, 2006a). The main idea is to average the results of 50 replications of an ordinary rare event logistic regression made with 50 different endogenous stratified samples. A similar methodological step has been used in Van Den Eeckhaut et al. (2009) for improving the model reliability of a discriminant analysis. We could also see a resemblance with the bootstrapping aggregation (bagging) method (Breiman, 1996) even thought, in our case, we do not resample with replacement using the obtained sample of the population as a basis.

In our approach, we create new sub-samples of non-events (0 s) using the entire population as a basis. In this study, we select 50 sub-samples as a trade-off between model reliability and computational time (Andresen, 2009). This conforms to previous geomorphic studies (see for example Beguería, 2006a; Davis and Keller, 1997; Van Den Eeckhaut et al., 2009). In this case-study, for each of the 50 endogenous stratified samples, 10 N points of non-events (0 s) were randomly selected from the population (matrix of grid-cells) and joined to the N events. The procedure was automated in R software.

The first steps of this method are similar to the ordinary rare event logistic regression technique described above; and include the selection of explanatory variables based on collinearity criteria and significance level. The ordinary rare event logistic regression was repeated 50 times with the different samples. For the final results, only variables with a p-value of 0.05 and present in at least 5 replications (10 %) are kept (following Beguería, 2006a). The final regression equation for landslide susceptibility is based on the explanatory variables that are robustly detected, and the regression parameters estimates are calculated as the average from the q parameter estimates from the repeated rare event logistic regressions (q being the number of replications for which the variables were significant).

Table 2. Exceedance probabilities of the landslide inventories of 1973, 1983, 1995 and 2010 for the 11 potential explanatory variables. The values that are significant at 5 % are highlighted in bold.

Variable	Classes	1973 ($n = 62$)	1983 ($n = 35$)	1995 ($n = 58$)	2010 ($n = 51$)
Quantitative variables:					
Slope		**<0.001**	**<0.001**	**<0.001**	**0.003**
Distance to path		**>0.998**	**>0.998**	**>0.998**	**>0.998**
Distance to watercourse		**0.990**	**0.987**	**>0.998**	**>0.998**
Qualitative variables :					
Plan curvature	concave	0.346	0.155	0.458	0.058
	rectilinear	0.952	0.712	0.670	0.953
	convexe	0.050	0.506	0.221	0.348
Land cover trajectory	no change (forest-páramo)	**>0.998**	**>0.998**	**>0.998**	**>0.998**
	forest degradation	0.436	0.949	**0.997**	**0.998**
	conversion to pasture	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	no change (pasture)	**0.002**	**<0.001**	0.031	**<0.001**
	others	**0.012**	**0.010**	**0.015**	**0.005**

3.5 Validation of the landslide susceptibility analyses

By definition, model validation allows assessing the accuracy and prediction power of a predictive model. It also allows comparing the performance of various models (Beguería, 2006b). Multivariate statistical models are frequently used for landslide susceptibility analyses, and a classification threshold or so-called cut-off value is often selected to classify the landslide susceptibility and assess hazards. The selection of the cut-off value is not straightforward, and different methodologies actually exist (Beguería, 2006b; Greiner et al., 2000). Receiver-operating characteristic plots (ROC plots) are an alternative solution to evaluate the model performance, as ROC plots contain information on the different model accuracies for a range of possible threshold values (Beguería, 2006b). They are constructed based on two statistical evaluation criteria that are not relying on the prevalence of events (1 s) in the sample: (i) the sensitivity (true-positive fraction) and (ii) the specificity (false-positive fraction) (Beguería, 2006b). The area-under-ROC (AUC) statistic allows evaluating the model's performance independently of a determined threshold value (Beguería, 2006b) so it gives rapidly an overall idea of the model goodness of fit.

4 Results and discussion

4.1 Probabilistic approach using Monte Carlo simulations

The results for the Monte Carlo simulations are shown in Table 2 that gives the exceedance probabilities for 4 landslide inventories for the 11 explanatory variables. From Table 2, it is clear that the spatial distribution of the landslides is not random, and that systematic association with morphological and anthropogenic factors occurs. Landslides are significantly associated with steep slopes (exceedance probability ≤ 0.003), and tend to cluster close to paths and watercourses (Table 2). In Llavircay, plan curvature is not significantly associated with the landslide pattern. Land cover change trajectories significantly control the landslide pattern (Table 2): landslides are significantly rare where tree cover is present (such as the trajectories no change (forest – páramo) and forest degradation), but are significantly overrepresented in pastures (conversion to pasture or no change (pasture)) and area with strong inter-annual changes in vegetation cover (others).

Even though the sample size is sometimes small ($n = 35$ for the 1983 landslide inventory), this probabilistic approach is able to identify the explanatory variables that are significantly associated with the landslide pattern. Besides, this approach is widely applicable as it does not require any a priori distribution of the independent or dependent variables. The major drawback of this univariate probabilistic approach is the lack of information on the relative influence of the different explanatory variables on the landslide pattern. Besides, it is not possible to account for multi-collinearity, which makes it difficult to use the results of the Monte Carlo simulations directly as an input for landslide susceptibility maps.

4.2 Ordinary rare event logistic regression

All explanatory variables were included in the logistic regression, as no multi-collinearity was detected based on the VIF and TOL values. Variables that are significant at 5 % were included in the rare event logistic regression. Results can be written in the form of Eq. (7) where p_i is the probability of landslide occurrence, here based on the landslide inventory of 1995 as an example (Table 3, Trial 1):

Table 3. Rare event logistic regression with landslide sample of 1995 for three trials, each of them having a different sample of non event; logistic coefficient (β); standard error on β (S.E); Wald statistic; variable significance (Pr($>|z|$)); Odd ratio; maximum value of explanatory variable in dataset (MPV); measure of parameter importance (MPI).

	β	S.E	Wald	Pr($>|z|$)	Odd ratio	MPV	MPI
Trial 1							
Intercept	−7.708	0.557	−13.852	<0.001	0.000		
Conversion to pasture	0.980	0.337	2.909	0.004	2.664	1.000	0.980
Distance to path	−0.001	0.000	−2.775	0.006	0.999	2721	−2.721
Slope	0.045	0.016	2.717	0.007	1.046	55	2.475
Forest degradation	−1.219	0.616	−1.979	0.048	0.296	1.000	−1.219
Trial 2							
Intercept	−7.914	0.603	−13.127	<0.001	0.000		
Conversion to pasture	1.092	0.332	3.291	0.001	2.980	1.000	1.092
Distance to path	−0.001	0.000	−2.723	0.006	0.999	2721	−2.721
Slope	0.046	0.018	2.590	0.010	1.047	55	2.530
Trial 3							
Intercept	−8.893	0.516	−17.248	<0.001	0.000		
Conversion to pasture	1.984	0.371	5.347	<0.001	7.272	1.000	1.984
Others	1.396	0.388	3.600	<0.001	4.039	1.000	1.396
No change (pasture)	1.571	0.497	3.163	0.002	4.811	1.000	1.571
Slope	0.039	0.016	2.378	0.017	1.040	55	2.145

$$\begin{aligned}\text{logit}(p_i) = &-7.708 \\ &+ (0.980 \times \text{conversion to pasture}) \\ &+ (0.045 \times \text{slope}) \\ &- (0.001 \times \text{distance to path}) \\ &- (1.219 \times \text{forest degradation}) \end{aligned} \quad (7)$$

Most of the shortcomings of the probabilistic methods can be solved with the rare event logistic regression. This multivariate analysis can combine a wide range of explanatory variables into one statistical analysis. The coefficients of the logistic regression allow to predict a logit transformation of event's presence probability and to create susceptibility maps (Van Den Eeckhaut et al., 2006; Vanwalleghem et al., 2008). Moreover, it is possible to determine the most important controlling factors by multiplying the maximum value of the variable in the dataset (MPV) with its regression coefficient (Vanwalleghem et al., 2008). This measure of parameter importance (MPI) indicates that distance to path and slope are the most important variables for predicting the landslide occurrence in 1995 (Table 3, Trial 1).

The major drawback of the rare event logistic regression is the dependency of the results on the endogenous stratified sampling. In Table 3, we give the outcome of the rare event logistic regression for landslide prediction based on the landslide inventory of 1995 for three different random samples of non-events (Table 3, Trials 1 to 3). When comparing the regression coefficient estimates, their standard errors and significance levels for the explanatory variables, we observe clear differences between the three predictive models. This example shows that ordinary rare event logistic regression can be strongly sample-dependent, and does not always lead to a stable detection of the controlling variables (Table 3). This sample dependence was also highlighted by Demoulin and Chung (2007).

4.3 Rare event logistic regression with replications

The predictive models for landslide occurrence based on the landslide inventories of four different years are resumed in Table 4. These results are obtained using the rare event logistic regression technique with replications (here 50 replications). The column "count" indicates the percentage of replications in which the explanatory variable was significant; and, thus, the number of values used for averaging the regression parameters (100 % = 50 replications). Table 4 clearly shows that many explanatory factors do not appear in every replication, although the fact that they are highly significant. For example, the trajectory no change (pasture) is present in only 18 % of the replications in 1995, although the variable is significant at 0.01. In the case of an ordinary rare event logistic regression, it would have been very likely that this variable was not included in the final regression model.

All studied years confounded, six out of the eleven potential explanatory variables were identified as being significant: slope, distance to watercourse, distance to path, conversion to pasture, no change (pasture) and the trajectory others

Table 4. Summary of the rare event logistic regression with replications showing, for every year, percentage of replications in which the variable entered (count), logistic coefficient (β) and its standard error (S.E), Wald statistic, variable significance (Pr($>|z|$)), Odd ratio, maximum value of explanatory variable in dataset (MPV) and measure of parameter importance (MPI).

1973 ($n = 62$)	Count (%)	β	S.E	Wald	Pr($>\|z\|$)	Odd ratio	MPV	MPI
intercept	100	−8.9566	0.6106	−14.6845	<0.001	0.0001		
slope	100	0.0799	0.0177	4.5201	<0.001	1.0832	55	4.395
conversion to pasture	94	1.1671	0.3676	3.1619	0.008	3.2126	1	1.167
distance to path	78	−0.001	0.0004	−2.7355	0.010	0.999	2721	−2.721
no change (pasture)	32	1.049	0.384	2.7416	0.014	2.8549	1	1.049
Variables not included: distance to watercourse, concave, rectilinear, convexe, no change (forest-páramo), forest conversion, others								
1983 ($n = 35$)	Count (%)	β	S.E	Wald	Pr($>\|z\|$)	Odd ratio	MPV	MPI
intercept	100	−9.8846	0.691	−15.3609	<0.001	0.0001		
conversion to pasture	100	2.6599	0.5572	4.7722	<0.001	14.2943	1	2.660
no change (pasture)	100	2.681	0.5857	4.5751	<0.001	14.5992	1	2.681
others	100	2.198	0.6238	3.5246	0.001	9.0066	1	2.198
slope	54	0.0584	0.0249	2.3419	0.022	1.0601	55	3.212
Variables not included: distance to watercourse, distance to path, concave, rectilinear, convexe, no change (forest-páramo), forest conversion								
1995 ($n = 58$)	Count (%)	β	S.E	Wald	Pr($>\|z\|$)	Odd ratio	MPV	MPI
intercept	100	−8.0335	0.5496	−14.9889	<0.001	0.0003		
slope	98	0.0448	0.0167	2.6899	0.012	1.0458	55	2.464
conversion to pasture	98	1.2547	0.3383	3.6678	0.004	3.5069	1	1.255
distance to path	76	−0.0011	0.0004	−2.8668	0.005	0.9989	2721	−2.994
others	20	1.2986	0.3841	3.3751	0.001	3.6643	1	1.299
no change (pasture)	18	1.5109	0.4893	3.0842	0.003	4.5306	1	1.511
Variables not included: distance to watercourse, concave, rectilinear, convexe, no change (forest-páramo), forest conversion								
2010 ($n = 51$)	Count (%)	β	S.E	Wald	Pr($>\|z\|$)	Odd ratio	MPV	MPI
intercept	100	−8.0581	0.5931	−16.2145	<0.001	0.0003		
conversion to pasture	100	2.929	0.5659	5.3274	<0.001	18.7087	1	2.929
no change (pasture)	92	2.7379	0.6675	4.0646	0.003	15.4549	1	2.738
distance to watercourse	86	−0.0046	0.0016	−2.8451	0.008	0.9954	1219	−5.608
others	52	2.7761	0.7691	3.6215	0.002	16.0571	1	2.776
distance to path	24	−0.0026	0.0008	−3.0668	0.015	0.9974	2721	−7.076
slope	22	0.0465	0.0196	2.366	0.023	1.0476	55	2.558
Variables not included: concave, rectilinear, convexe, no change (forest-páramo), forest conversion								

(Table 4). Only three variables are systematically and significantly associated with the landslide pattern for all years: the two land cover trajectories that are directly linked with pastures and the slope gradient. The variables distance to path and the trajectory others are present 3 out of 4 times. The variable distance to watercourse is only present in 2010. The number of significant explanatory variables is increasing with time, which might be due to the fact that the land cover heterogeneity increases with time (Fig. 1). As both the number and the spatial repartition of landslides change with time (Fig. 1), it is not abnormal to find changes in the explanatory variables with time (Table 4).

An analysis of the influence of each explanatory variable on the landslide probability, MPI (Vanwalleghem et al., 2008), indicates that the relative importance of the different controlling factors changes with time (Fig. 4). The most obvious change is observed for the topographical factor slope. For the early years 1973 and 1983, the slope gradient was detected as the most important controlling factor for landslide occurrence. In 2010, the slope gradient is only marginally important for explaining the spatial distribution of landslides within the catchment, while the anthropogenic variables such as the land cover trajectories that are linked with human disturbance are ranked higher in terms of variable influence (Fig. 4). The fact that two explanatory variables (distance to watercourse and path) have an exceptionally high variable influence in 2010 might be linked to data collection bias, as the 2010 landslide repertory is based on fieldwork (in contrast

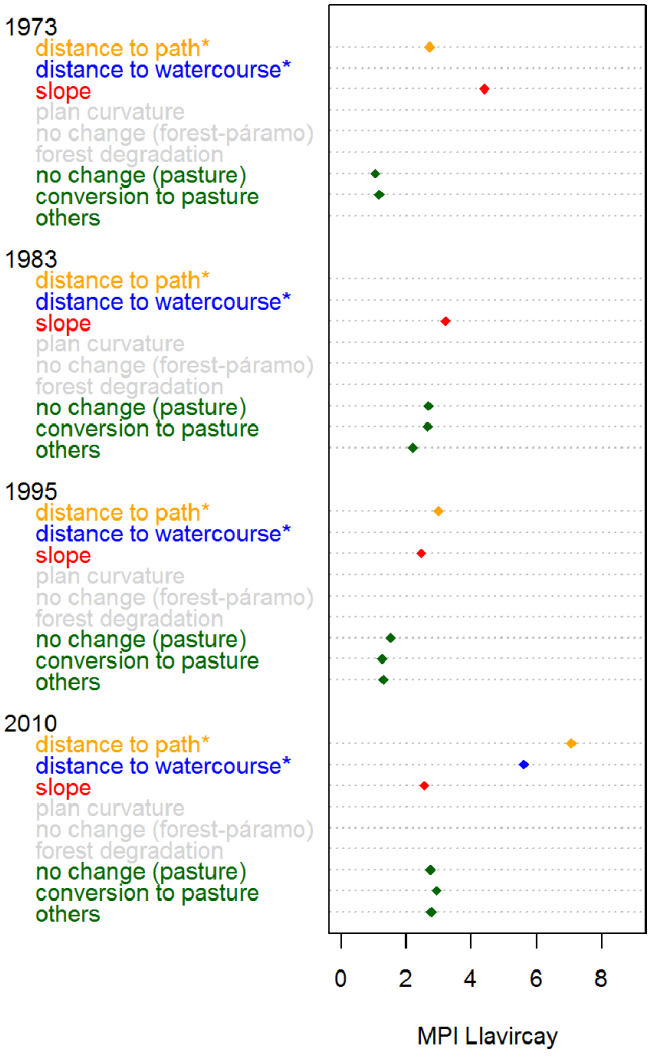

Fig. 4. Evolution of the most important controlling variables through time. For the variables with * we took the absolute value of measure parameter importance (MPI).

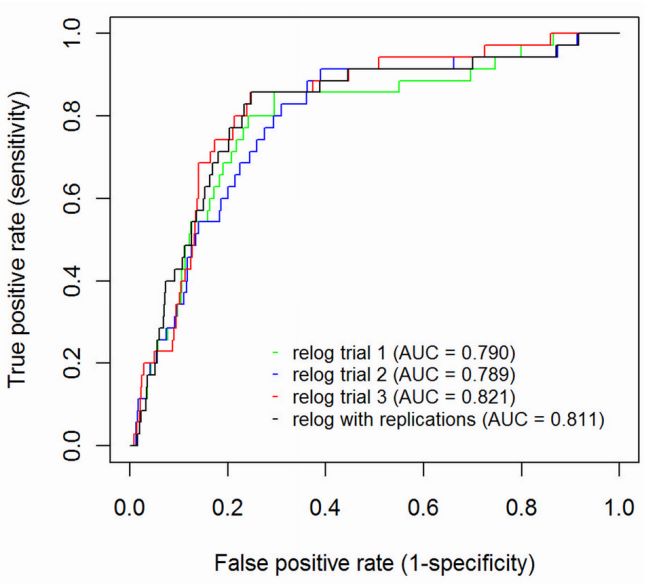

Fig. 5. ROC plot and AUC of validation datasets ($n = 19\,250$) for the different rare event logistic regressions (details of them in Tables 3 and 4 – 1995).

to the landslide inventories of 1973, 1983 and 1995 that are based on aerial photographs). This might explain why more landslides were observed in 2010 close to paths and watercourses, which are the two most important accessibility corridors in this remote area.

4.4 Validation of the landslide susceptibility analyses

For model validation, various methodologies exist to select the validation data (Chung and Fabbri, 2003). As the number of landslides in our database is small, we decided not to split our datasets in a calibration and validation set. Instead, we used the landslide inventories from the closest time period to evaluate the performance of the predictive models. As the landslide controlling factors are slightly changing through time, we hypothesise that the use of the landslide inventory of the closest time period has only a minor influence on the model evaluation.

In Fig. 5, we present the results of the evaluation of the four predictive landslide susceptibility models based on the landslide inventory of 1995. It includes the ordinary rare event logistic regressions based on three different random samples of non-events (see Table 3, Trials 1 to 3), and the rare event logistic regression with replications (see Table 4, year 1995). The AUC for all four models varies between 0.79 and 0.82 (Fig. 5), and we can consider that the predictive models are moderately accurate according to the arbitrary guideline of Swets (1988). As observed from the ordinary rare event logistic regression, the model performance of the three predictive models is sample-dependent (Fig. 5). The ROC and AUC vary between the three replications of the ordinary rare event logistic regression. The performance of the rare event logistic regression with replication is not significantly better than the ordinary rare event logistic regression models, but a conceptual improvement is made on the identification of the landsliding controlling factors.

5 Conclusions

Statistical analysis of natural hazards needs particular attention, as most of these phenomena are rare events. This specificity of natural hazards was only taken into account recently by adapting the ordinary logistic regression techniques for the analysis of rare events. This study shows that the ordinary rare event logistic regression, as it is now commonly used in geomorphologic studies, does not always lead to a

robust detection of controlling variables as the results can be strongly sample-dependent.

In this study, we developed a modified version of the rare event logistic regression technique. Our so-called rare event logistic regression with replications builds some concepts from probabilistic theory into rare event logistic regression analysis. It is based on the statistical method of rare-event logistic regression, but it includes Monte Carlo simulations to estimate the robustness of the regression estimates. The use of replications in the rare event logistic regressions allows avoiding instability of the results due to sampling bias. Our results demonstrate that rare event logistic regression with replications has a similar modelling quality as the ordinary rare event logistic regression techniques. It allows having a more robust selection of factors that are significant for explaining the spatial variation in the occurrence of natural hazards. This new technique was here developed for landslide spatial pattern analyses, but the concept is widely applicable for statistical analyses of natural hazards.

Acknowledgements. Funding for this research was provided by the Fonds National de la Recherche Scientifique (FNRS, Brussels). Landslide inventories were realised in 2009 and 2010, and were funded through the CUD-PIC project "Strengthening the scientific and technological capacities to implement spatially integrated land and water management schemes adapted to local socio-economic and physical settings" between the Faculty of Agricultural Engineering at the Universidad de Cuenca (Ecuador), the Université catholique de Louvain, the KULeuven and the FUNDP (Belgium). The authors would like to thank Dario Alvarado Moncayo and Pablo Borja Ramon of the Universidad de Cuenca (Ecuador), colleagues at CGPaute (Ecuador), Luis Jerves at Celec – Hidropaute (Ecuador) and Armando Molina for their precious help during the field work.

Edited by: T. Glade
Reviewed by: three anonymous referees

References

Acotecnic: Proyecto Hidroelectric Mazar: Estudios de impacto ambiental definitivos – Informe Final, Ecuador, 2293 pp., 2006.

Alcántara-Ayala, I., Esteban-Chávez, O., and Parrot, J. F.: Landsliding related to land-cover change: A diachronic analysis of hillslope instability distribution in the Sierra Norte, Puebla, Mexico, CATENA, 65, 152–165, doi:10.1016/j.catena.2005.11.006, 2006.

Allison, P. D.: Logistic regression using SAS System: Theory and application, Wiley Interscience, New York, 2001.

Andresen, M. A.: Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach, Appl. Geogr., 29, 333–345, doi:10.1016/j.apgeog.2008.12.004, 2009.

Atkinson, P. M. and Massari, R.: Generalised linear modelling of susceptibility to landsliding in the Central Apennines, Italy, Comput. Geosci., 24, 373–385, doi:10.1016/s0098-3004(97)00117-9, 1998.

Atkinson, P. M., Jiskoot, H., Massari, R., and Murray, T.: Generalized linear modelling in geomorphology, Earth Surf. Proc. Land., 23, 1185–1195, doi:10.1002/(sici)1096-9837(199812)23:13<1185::aid-esp928>3.0.co;2-w, 1998.

Ayalew, L. and Yamagishi, H.: The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan, Geomorphology, 65, 15–31, doi:10.1016/j.geomorph.2004.06.010, 2005.

Bai, S., Lü, G., Wang, J., Zhou, P., and Ding, L.: GIS-based rare events logistic regression for landslide-susceptibility mapping of Lianyungang, China, Environ. Earth Sci., 62, 139–149, doi:10.1007/s12665-010-0509-3, 2011.

Basabe, P.: Prevención de desastres naturales en la Cuenca del Paute – Informe final : Proyecto Precupa, Swiss Disaster Relief Unit (SDR/CSS), Cuenca, Ecuador, 1998.

Baum, R. and Godt, J.: Early warning of rainfall-induced shallow landslides and debris flows in the USA, Landslides, 7, 259–272, doi:10.1007/s10346-009-0177-0, 2010.

Beguería, S.: Changes in land cover and shallow landslide activity: A case study in the Spanish Pyrenees, Geomorphology, 74, 196–206, doi:10.1016/j.geomorph.2005.07.018, 2006a.

Beguería, S.: Validation and Evaluation of Predictive Models in Hazard Assessment and Risk Management, Nat. Hazards, 37, 315–329, doi:10.1007/s11069-005-5182-6, 2006b.

Breiman, L.: Bagging predictors, Mach. Learn., 24, 123–140, 1996.

Brenning, A.: Spatial prediction models for landslide hazards: review, comparison and evaluation, Nat. Hazards Earth Syst. Sci., 5, 853–862, doi:10.5194/nhess-5-853-2005, 2005.

Brunetti, M. T., Peruccacci, S., Rossi, M., Luciani, S., Valigi, D., and Guzzetti, F.: Rainfall thresholds for the possible occurrence of landslides in Italy, Nat. Hazards Earth Syst. Sci., 10, 447–458, doi:10.5194/nhess-10-447-2010, 2010.

Chung, C.-J. F. and Fabbri, A. G.: Validation of Spatial Prediction Models for Landslide Hazard Mapping, Nat. Hazards, 30, 451–472, doi:10.1023/B:NHAZ.0000007172.62651.2b, 2003.

Crawley, M. J.: Statistics: an introduction using R, John Wiley and Sons, England, 2005.

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

Dai, F. C. and Lee, C. F.: A spatiotemporal probabilistic modelling of storm-induced shallow landsliding using aerial photographs and logistic regression, Earth Surf. Proc. Land., 28, 527–545, doi:10.1002/esp.456, 2003.

Dai, F. C., Lee, C. F., Li, J., and Xu, Z. W.: Assessment of landslide susceptibility on the natural terrain of Lantau Island, Hong Kong, Environ. Geol., 40, 381–391, 2001.

Dai, F. C., Lee, C. F., and Ngai, Y. Y.: Landslide risk assessment and management: an overview, Eng. Geol., 64, 65–87, doi:10.1016/s0013-7952(01)00093-x, 2002.

Davis, T. J. and Keller, C. P.: Modelling uncertainty in natural resource analysis using fuzzy sets and Monte Carlo simulation: slope stability prediction, Int. J. Geogr. Inf. Sci., 11, 409–434, 1997.

Demoulin, A. and Chung, C.: Mapping landslide susceptibility from small datasets: A case study in the Pays de Herve (E Belgium), Geomorphology, 89, 391–404, doi:10.1016/j.geomorph.2007.01.008, 2007.

Greiner, M., Pfeiffer, D., and Smith, R. D.: Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests, Prev. Vet. Med., 45, 23–41, doi:10.1016/s0167-5877(00)00115-x, 2000.

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

Guzzetti, F., Galli, M., Reichenbach, P., Ardizzone, F., and Cardinali, M.: Landslide hazard assessment in the Collazzone area, Umbria, Central Italy, Nat. Hazards Earth Syst. Sci., 6, 115–131, doi:10.5194/nhess-6-115-2006, 2006.

Huabin, W., Gangjun, L., Weiya, X., and Gonghui, W.: GIS-based landslide hazard assessment: an overview, Prog. Phys. Geogr., 29, 548–567, 2005.

Imai, K., King, G., and Lau, O.: Zelig: Everyone's Statistical Software, available at: http://gking.harvard.edu/zelig (last access: 14 April 2011), 2009.

INAMHI: Anuario hidrologico y meteorológico 1960–2008, Technical reports, 2008.

King, G. and Zeng, L.: Logistic Regression in Rare Events Data, Polit. Anal., 9, 137–163, 2001a.

King, G. and Zeng, L.: Explaining Rare Events in International Relations, Int. Organ., 55, 693–715, doi:10.1162/00208180152507597, 2001b.

Kleinbaum, D. G. and Klein, M.: Introduction to Logistic Regression, in: Logistic Regression, Statistics for Biology and Health, Springer New York, 1–39, 2010.

Komac, M.: A landslide susceptibility model using the Analytical Hierarchy Process method and multivariate statistics in perialpine Slovenia, Geomorphology, 74, 17–28, doi:10.1016/j.geomorph.2005.07.005, 2006.

Larsen, M. C. and Torres-Sánchez, A. J.: The frequency and distribution of recent landslides in three montane tropical regions of Puerto Rico, Geomorphology, 24, 309–331, 1998.

Luteyn, J. L.: Páramos : A checklist of plant diversity, geographical distribution, and botanical literature, Memoirs of the New York Botanical Garden, The New York Botanical Garden, Bronx, New York, 1999.

Maalouf, M. and Trafalis, T. B.: Robust weighted kernel logistic regression in imbalanced and rare events data, Comput. Stat. Data An., 55, 168–183, doi:10.1016/j.csda.2010.06.014, 2011.

Ramalho, E. A.: Regression models for choice-based samples with misclassification in the response variable, J. Econometrics, 106, 171–201, doi:10.1016/s0304-4076(01)00094-x, 2002.

Sawilowsky, S. S.: You think you've got trivials?, Journal of Modern Applied Statistical Methods, 2, 218–225, 2003.

Sierra, R. and Stallings, J.: The Dynamics and Social Organization of Tropical Deforestation in Northwest Ecuador, 1983–1995, Hum. Ecol., 26, 135–161, doi:10.1023/a:1018753018631, 1998.

Summerfield, M. A.: Global geomorphology – An introduction to the study of landforms, Pearson, Prentice Hall, England, 560 pp., 1991.

Swets, J. A.: Measuring the Accuracy of Diagnostic Systems, Science, 240, 1285–1293, 1988.

Tatard, L., Grasso, J. R., Helmstetter, A., and Garambois, S.: Characterization and comparison of landslide triggering in different tectonic and climatic settings, J. Geophys. Res.-Earth, 115, 18 pp., doi:10.1029/2009JF001624, 2010.

Van Den Eeckhaut, M., Vanwalleghem, T., Poesen, J., Govers, G., Verstraeten, G., and Vandekerckhove, L.: Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium), Geomorphology, 76, 392–410, doi:10.1016/j.geomorph.2005.12.003, 2006.

Van Den Eeckhaut, M., Reichenbach, P., Guzzetti, F., Rossi, M., and Poesen, J.: Combined landslide inventory and susceptibility assessment based on different mapping units: an example from the Flemish Ardennes, Belgium, Nat. Hazards Earth Syst. Sci., 9, 507–521, doi:10.5194/nhess-9-507-2009, 2009.

Vanacker, V., Govers, G., Tacuri, E., Poesen, J., Dercon, G., and Cisneros, F.: Using sequential aerial photographs to detect land-use changes in the Austro Ecuadoriano, Revue de Géographie Alpine, 88, 65–75, 2000.

Vanacker, V., Govers, G., Van Peer, P., Verbeek, C., Desmet, J., and Reyniers, J.: Using Monte Carlo Simulation for the Environmental Analysis of Small Archaeologic Datasets, with the Mesolithic in Northeast Belgium as a Case Study, J. Archaeol. Sci., 28, 661–669, doi:10.1006/jasc.2001.0654, 2001.

Vanacker, V., Vanderschaeghe, M., Govers, G., Willems, E., Poesen, J., Deckers, J., and De Bievre, B.: Linking hydrological, infinite slope stability and land-use change models through GIS for assessing the impact of deforestation on slope stability in high Andean watersheds, Geomorphology, 52, 299–315, 2003.

Vanwalleghem, T., Van Den Eeckhaut, M., Poesen, J., Govers, G., and Deckers, J.: Spatial analysis of factors controlling the presence of closed depressions and gullies under forest: Application of rare event logistic regression, Geomorphology, 95, 504–517, doi:10.1016/j.geomorph.2007.07.003, 2008.

Winckell, A., Zebrowski, C., and Sourdat, M.: Las regiones y paisages del Ecuador, Geografía básica del Ecuador, edited by: Geográfica, C. C. E. d. I., Quito, Ecuador, 417 pp., 1997.