Nat. Hazards Earth Syst. Sci., 13, 2209–2221, 2013
www.nat-hazards-earth-syst-sci.net/13/2209/2013/
doi:10.5194/nhess-13-2209-2013

Natural Hazards and Earth System Sciences
Open Access

Influence of management of variables, sampling zones and land units on LR analysis for landslide spatial prevision

R. Greco and M. Sorriso-Valvo

National Research Council – Research Institute for Geo-hydrological Protection, Cosenza Unit, Cavour st. 4–6, 87036, Rende, Cosenza, Italy

Correspondence to: R. Greco (roberto.greco@irpi.cnr.it)

Received: 1 August 2012 – Published in Nat. Hazards Earth Syst. Sci. Discuss.: –
Revised: 15 April 2013 – Accepted: 23 June 2013 – Published: 10 September 2013

Abstract. Several authors, according to different methodological approaches, have employed logistic Regression (LR), a multivariate statistical analysis adopted to assess the spatial probability of landslide, even though its fundamental principles have remained unaltered.

This study aims at assessing the influence of some of these methodological approaches on the performance of LR, through a series of sensitivity analyses developed over a test area of about 300 km^2 in Calabria (southern Italy).

In particular, four types of sampling (1 – the whole study area; 2 – transects running parallel to the general slope direction of the study area with a total surface of about 1/3 of the whole study area; 3 – buffers surrounding the phenomena with a 1/1 ratio between the stable and the unstable area; 4 – buffers surrounding the phenomena with a 1/2 ratio between the stable and the unstable area), two variable coding modes (1 – grouped variables; 2 – binary variables), and two types of elementary land (1 – cells units; 2 – slope units) units have been tested. The obtained results must be considered as statistically relevant in all cases (Aroc values $> 70\,\%$), thus confirming the soundness of the LR analysis which maintains high predictive capacities notwithstanding the features of input data.

As for the area under investigation, the best performing methodological choices are the following: (i) transects produced the best results ($0 < P(y) \leq 93.4\,\%$; Aroc $= 79.5\,\%$); (ii) as for sampling modalities, binary variables ($0 < P(y) \leq 98.3\,\%$; Aroc $= 80.7\,\%$) provide better performance than ordinated variables; (iii) as for the choice of elementary land units, slope units ($0 < P(y) \leq 100\,\%$; Aroc $= 84.2\,\%$) have obtained better results than cells matrix.

1 Introduction

According to the Centre for Research on the Epidemiology of Disasters (CRED), the number of geomorphological catastrophes registered on a yearly basis has increased fivefold, from 78 in 1975 to almost 450 in 2007. It has been estimated that the average loss per year accounts for 0.25 % of the global Gross Domestic Product (GDP). Over the last 20 yr, catastrophes registered in Europe have caused almost 90 000 casualties, 29 million injured people and economic loss equalling 211 billion euros. In Europe, the number of catastrophes caused by climate events has almost tripled: from 1280 between 1978 and 1987 to 3435 between 1998 and 2007 (Scheuren et al., 2008; Vos et al., 2010; Guha-Sapir et al., 2011).

This trend is essentially due to a greater exposure to hazard for properties and people (Scheuren et al., 2008; Vos et al., 2010; Guha-Sapir et al., 2011) and, very probably to the climate changes (AA.VV., 2009) accelerated by several factors such as greenhouse gases in the atmosphere and environmental degradation.

In Italy, one of the most geomorphologically unstable countries in Europe, slope instability phenomena are – immediately after earthquakes – one of the main sources of risk for individuals, built-up areas, infrastructures and architectural heritage. Moreover, over the last few years, an increase in extreme rainfalls (AA.VV., 2009), has determined a larger number of slope instability events. Within this perspective, assessment and mitigation of the risk implied by such phenomena play a key role in the scientific research for their impact. In particular, in terms of both civil defence against landslides and land management.

The assessment at a regional scale of the risk of mass movement poses relevant problems because of the lack or incompleteness of the historical series of events and the triggering phenomena and – as a consequence – of models on the trigger-event relationship, as well as the non-stationarity of landsliding phenomena in general. On the contrary, studies on the probability of occurrence of mass movements in a given area are at a good stage (Aleotti and Chowdhury, 1999; Chung and Fabbri; 1999; Guzzetti et al., 1999, 2005, 2006; Baeza and Corominas, 2001; Thiery et al., 2007; Van Westen et al., 2008; Sorriso-Valvo et al., 2009; Rossi et al., 2010; Yalcin et al., 2011; Choi et al., 2012); such probability, can also be defined as spatial hazard. Some authors, like us, prefer to employ the definition of susceptibility to mass movement when independent variables are exclusively related to the physical features of the land and do not include the triggering factors.

Assessment of the potential existence of mass movements is generally carried out by following a conceptual model based on three fundamental steps: (a) landsliding inventory maps on the study area or in one of its subunits (test area); (b) thematic maps on territorial variables made up of instability factors and elements considered as directly or indirectly related to slope instabilities; (c) statistical multivariate models to estimate the contribution of each single variable to the slope instability and to organise the study area into domains with a different probability of being affected by mass movement phenomena.

The first two steps are particularly delicate, since the quality of assessments depends on the quality of input data.

Independent variables (either instability factors or features indirectly associated to the phenomena under investigation) that are used for these studies, are usually chosen based on the authors' direct experience, training and education.

Some authors, in fact, attach more importance to territorial factors related to geological and geo-morphological features (i.e. lithology, vegetation cover, type of soil, elevation, slope angle, aspect and curvature). Others tend to consider more quantitative factors and/or those factors related to geological-technical features (such as topographic humidity index, distance of the area from rivers, roads and faults, density of drainage, roads and faults, bedrock depth, soil porosity, etc.).

As for the number of variables to be employed, it is possible to recognise two main approaches: using a limited number of territorial variables considered as very relevant or in any case related to mass movement events; or employing as many variables as possible, so as not to exclude some of them, which may display high, despite spurious, correlation with the territorial distribution of mass movements.

Among the several statistical procedures adopted to assess the susceptibility to mass-movements, Logistic Regression (LR), has increasingly been adopted (Bernknopf et al., 1988; Gorsevski et al., 2000; Dai and Lee, 2002; Ohlmacher and Davis, 2003; Dai et al., 2004; Ayalew and Yamagishi, 2005; Can et al., 2005; Chau and Chan, 2005; Yesilnacar and Topal, 2005; Van Den Eeckhaut et al., 2006; Greco et al., 2007; Chen and Wang, 2007; Garcia-Rodriguez et al., 2008; Nefeslioglu et al., 2008; Mathew et al., 2009; Sorriso-Valvo et al., 2009; Falaschi et al., 2009; Nandi and Shakoor, 2009; Chauhan et al., 2010; Erener et al., 2010; Rossi et al., 2010; Yalcin et al., 2011; Choi et al., 2012). Different procedures of the LR application have been employed, and they mainly differ in terms of sampling modalities, variables transformation, and selection of reference land units.

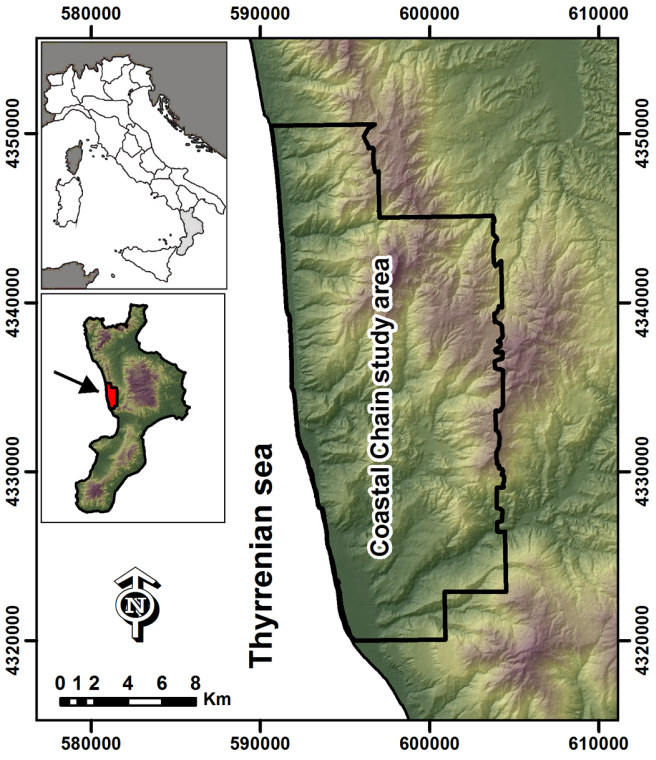

Fig. 1. Location of Coastal Chain study area (769 500 cells).

Some authors have carried out studies to compare the different modalities of assessment of spatial probabilities (Discriminant Analysis, Logistic Regression, Artificial Neural Networks, etc.), to evaluate their predictive capacities (Yesilnacar and Topal, 2005; Carrara et al., 2008; Van Den Eeckhaut et al., 2010; Das et al., 2010; Rossi et al., 2010; Yalcin et al., 2011; Choi et al., 2012), while, less numerous are the studies on the evaluation of the relevance of the several methodological choices adopted in the implementation of the procedure on LR performance (Guzzetti et al., 1999, 2006; Carrara et al., 2008; Sorriso-Valvo et al., 2009; Van Den Eeckhaut et al., 2009). This study aims a providing a contribution to fill this gap by means of a comparison of the results from a series of analyses developed over a study area of about 300 km^2 in Calabria (southern Italy; Fig. 1).

The predictive capacities of carried out regressions have been compared through the ROC analysis (Hosmer and Lemeshow, 1989).

2 Logistic regression in the assessment of susceptibility to mass movements

Given for granted the theory of the LR analysis (Pampel, 2000), the numerous methodologies in the scientific literature on the application of Logistic Regression to assess susceptibility to mass movements differ in three main aspects: (1) sampling modalities to select the reference sample in order to calculate regression coefficients; (2) variables transformations; and (3) the type of land unit to carry out the analysis.

2.1 Sampling

Sample collection is one of the most delicate steps of LR analysis, as far as the sampled population is employed to calculate regression coefficients (weights) of independent variables. It follows that the sample must be sufficiently representative of the whole study area, and – in particular – it must take into account all the territorial fields and contexts that best summarise the relationships between the landsliding phenomena and the predisposing factors.

In this perspective, the approaches adopted by the several authors are very different: some employ samples made up of the same number of stable and landsliding prone cells ($0/1 = 1$) (Dai and Lee, 2002; Dai et al., 2004; Chau and Chan, 2005; Yesilnacar and Topal, 2005; Garcia-Rodriguez et al., 2008; Nefeslioglu et al., 2008; Mathew et al., 2009; Nandi and Shakoor, 2009). Among this group, the majority of authors employs a sample population randomly extracted from the whole study area (Dai and Lee, 2002; Dai et al., 2004; Chau and Chan, 2005; Yesilnacar and Topal, 2005; Garcia-Rodriguez et al., 2008; Mathew et al., 2009; Choi et al., 2012), other researchers select only a portion of the surface to be used (Nefeslioglu et al., 2008; Nandi and Shakoor, 2009). Further approaches consist in using a different number of stable and landslide prone cells ($0/1 \neq 1$), considering as a sample population either the whole study area (Bernknopf et al., 1988; Ohlmacher and Davis, 2003; Ayalew and Yamagishi, 2005; Chen and Wang, 2007; Falaschi et al., 2009; Chauhan et al., 2010; Erener et al., 2010; Rossi et al., 2010; Yalcin et al., 2011) or a part of the same (Gorsevski et al., 2000; Can et al., 2005; Van Den Eeckhaut et al., 2006; Greco et al., 2007; Sorriso-Valvo et al., 2009).

2.2 Coding of variables

LR allows researchers to employ also categorical variables which, indeed, have to be codified to allow the interpolation algorithm to calculate the regression coefficients. Consequently, variables transformation is a common aspect of the application of Logistic Regression analyses.

The dependent variable is always coded according to the logit model proposing two potential values for a dichotomic variable, i.e. 1 = existing (or true) and 0 = non existing (or false).

Coding of independent variables has been dealt with by several authors according to two different approaches; the first approach, which is used by the vast majority of authors (Bernknopf et al., 1988; Dai and Lee, 2002; Ohlmacher and Davis, 2003; Dai et al., 2004; Ayalew and Yamagishi, 2005; Can et al., 2005; Chau and Chan, 2005; Van Den Eeckhaut et al., 2006; Chen and Wang, 2007; Garcia-Rodriguez et al., 2008; Nefeslioglu et al., 2008, Mathew et al., 2009; Falaschi et al., 2009; Chauhan et al., 2010; Rossi et al., 2010), creates layers having binary variables (dummy variables) for all the categories or classes of each dependent variable (a variable having ten categories or classes, generates the same number of binary variables). Such approach is sensible when a limited number of independent variables is available and when such variables are – in turn – articulated in a few categories or classes. When analyses imply a high number of independent variables and consequently of categories, the risk is to produce a too long regression equation which may generate calculation problems. For this reason, some authors (Yesilnacar and Topal, 2005; Greco et al., 2007; Sorriso-Valvo et al., 2009; Nandi and Shakoor, 2009; Erener et al., 2010; Yalcin et al., 2011; Choi et al., 2012), have adopted an approach according to an ascending ordination of the different classes of variables, based on the observed frequency of mass movement in sampling zones (grouped variables). Values of such relative-scale variables are grouped into classes. Such an approach avoids calculation problems generated by too long equations and operates a linearization of independent variables.

2.3 Elementary land units

Land units (described in the scientific literature) employed in studies on diffused mass movements can be ascribed to four main types:

- *geo-morphologic units*: morphologic units representing a territory, such as slopes, talweg, flat valley floors, basin heads, ridges, noses, etc.;
- *cells matrix*: land units generated by sorting the area into similar or different cells in terms of shape and dimension;
- *homogeneous land units*: units that are generated starting from a series of thematic maps on the relevant instability factors, where each factor is described trough few classes that are sufficient to express internal variability; the intersection among thematic maps points out homogeneous land sections in terms of instability factors.
- *slope units*: territorial units automatically derived from high definition digital land models.

Among these units, except for a less number of authors who make use of slope units (Carrara et al., 1991, 1995; Komac, 2006; Rossi et al., 2010) and homogeneous land units (Can et al., 2005; Falaschi et al., 2009), cells matrix is the most widely used land unit by the authors considered (Bernknopf et al., 1988; Gorsevski et al., 2000; Dai and Lee, 2002; Ohlmacher and Davis, 2003; Dai et al., 2004; Ayalew and Yamagishi, 2005; Chau and Chan, 2005; Yesilnacar and Topal, 2005; Van Den Eeckhaut et al., 2006; Greco et al., 2007; Chen and Wang, 2007; Garcia-Rodriguez et al., 2008; Nefeslioglu et al., 2008; Mathew et al., 2009; Sorriso-Valvo et al., 2009; Nandi and Shakoor, 2009; Chauhan et al., 2010; Erener et al., 2010; Yalcin et al., 2011; Choi et al., 2012).

Authors almost homogeneously perform the fitting and weight assessment steps for independent variables and for the variables of analysis application (assessment of $P(y)$).

To facilitate readability of the results, the probability values obtained are then grouped in classes with ranges varying according to the aim of the analysis. In this study $P(y)$ values are re-classified in five different susceptibility classes: Null ($P(y) \leq 5\,\%$); Low ($5\,\% < P(y) \leq 25\,\%$); Medium ($25\,\% < P(y) \leq 50\,\%$); High ($50\,\% < P(y) \leq 75\,\%$); Very high ($P(y) > 75\,\%$).

3 Study area

The data necessary to carry out this study have been drawn from a study area of about 308 km^2 in the southwest area of the Coastal Chain, Calabria, Italy (Fig. 1).

In this area, we have collected data relevant to independent and dependent variables. These are mainly geological and geo-morphological variables. Such variables have been recognised as associated through a cause-effect relation with landsliding phenomena. Given the extension of potentially target-study areas, selected variables fulfil the fundamental requisite of being cheap in terms of variables collection.

The data relevant to independent variables and the dependent variable have been obtained from pre-existing thematic maps, ad hoc survey and the processing of other variables (DEM).

A 1 : 10 000 (Fig. 2) mass movement map has been produced with the aim of both characterising sample areas and assessing performance. This map, which represents the dependent variable, has been generated though photo interpretation carried out on both 1 : 33 000 aerial photos (IGM flights) and on field surveys. As a whole, 1206 phenomena covering a surface of about 122 km^2, i.e. 39 % of the study area, have been mapped.

Nine independent variables have been identified. Two of them are of categorical type: lithology and land use. Lithology (LTU) has been obtained digitalizing a pre-existing thematic map (Carta Geologica della Calabria, Burton, 1971); land use (LUS) has been obtained by means of the interpretation of aerial photos shot in 2006 on a 1 : 5000 scale. Such

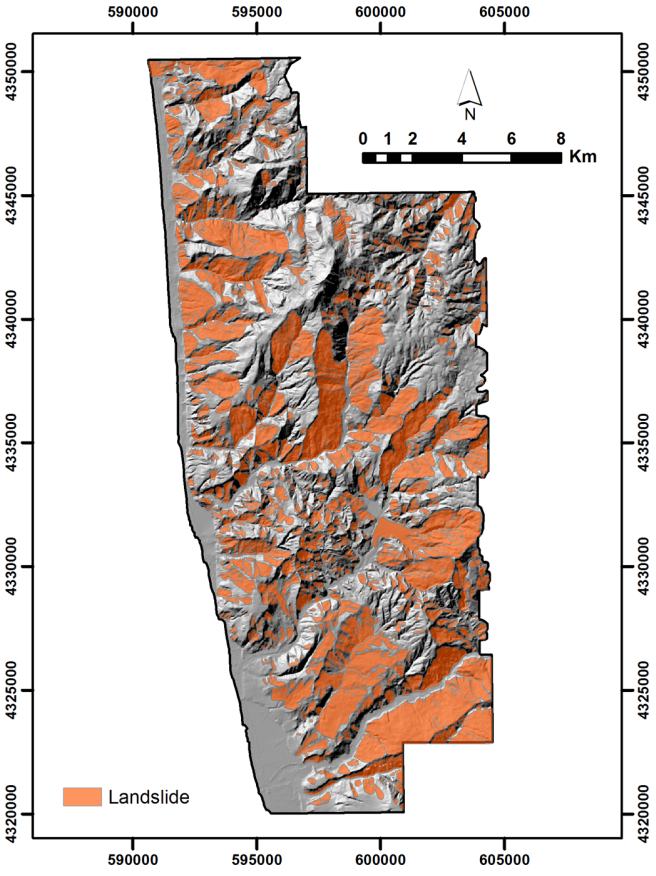

Fig. 2. Mass movement inventory map of study area.

photos have been corrected in plane projection and are available on the Portale Cartografico Nazionale (PCN), i.e. the national cartographic portal, in the section WMS-Server of the Ministero dell'Ambiente e Territorio (Italian Ministry for the Environment and Territory).

Seven variables are parametric; six of them have been derived from a 20 m-square cells DTM, i.e. elevation (ELEV), slope (SLO), aspect (ASP), curvature of land surfaces calculated both perpendicularly (ACUR) and in parallel (DCUR) to the maximum slope, and topographic wetness index (TWI) (Moore et al., 1991); one variable, the distance to the closest fault (FDIST), has been obtained by applying Euclidean Distance Operator (an Arc-Info tool) to a tectonic lineation map detected by interpreting aerial photographs and field surveys (Gullà et al., 2010). Parametric territorial factors have been classified according to personal experience gained in previous studies (Greco et al., 2007; Sorriso-Valvo et al., 2009). Table 1 shows the distribution of the territorial variables considered in the study area. All acquired variables, that have been georeferenced according to the Gauss-Boaga reference system (Monte Mario Italy 2), have been stored in grid format (20 m square cells) in an Arc-Info database (ver. 9.3), through which all operations of data management, processing and graphic outline have been carried out.

Table 1. Frequency of territorial factors categories or classes and relative dummy variables.

Factors	Category or class	(%)	Dummy variable
Lithological Unit	CDS	8.22	LTU1
	GS	17.51	LTU2
	CA	12.07	LTU3
	CL	6.82	LTU4
	LD	7.45	LTU5
	LGMR	39.00	LTU6
	HGMR	7.65	LTU7
	IR	1.29	LTU8
Land Use	UA	4.47	LUS1
	BB	4.67	LUS2
	PF	21.52	LUS3
	PA	10.95	LUS4
	OFV	4.67	LUS5
	FOR	53.73	LUS6
Elevation	<100 m	11.85	ELEV1
	100–400 m	34.15	ELEV2
	400–800 m	30.98	ELEV3
	800–1200 m	22.27	ELEV4
	>1200 m	0.74	ELEV5
Slope angle	<8°	16.10	SLO1
	8°–15°	18.06	SLO2
	15°–30°	46.31	SLO3
	30°–45°	18.45	SLO4
	45°–60°	1.04	SLO5
	>60°	0.04	SLO6
Aspect	Flat	0.58	ASP1
	North	9.41	ASP2
	East	6.72	ASP3
	South	11.60	ASP4
	West	17.23	ASP5
Across slope curvature	Concave	37.88	ACUR1
	Plane	20.00	ACUR2
	Convex	42.12	ACUR3
Down slope curvature	Concave	41.17	DCUR1
	Plane	17.89	DCUR2
	Convex	40.95	DCUR3
Topographic wetness index	<2	5.79	TWI1
	2–3	15.86	TWI2
	3–5	27.55	TWI3
	>5	50.79	TWI4
Distance to fault	<20 m	16.07	FDIST1
	20–80 m	30.96	FDIST2
	80–200 m	36.28	FDIST3
	>200 m	16.70	FDIST4

Key to the acronyms. Lithological unit: CDS = colluviums, debris and soil; GS = gravel and sand; CA = conglomerate and arenite; CL = clay; LD = limestone and dolomite; LGMR = low grade metamorphic rock; HGMR = high grade metamorphic rock; IR = igneous rock. Land Use: UA = urban areas; BB = bare rocks and beaches; PF = plowing field; PA = pastures; OFV = olive groves, fruit plantations and vineyards; FOR = forests.

4 Evaluation of the effectiveness of methodological approaches

4.1 Premise

To test which methodological choices concerning sampling, variable transformation and adopted land units, provide the best results in terms of assessment of susceptibility, a certain number of LR analyses have been carried out by processing data according to the several approaches proposed.

The results obtained through the different applications have been assessed by means of ROC (Receiver Operating Characteristics) analysis (Hosmer and Lemeshow, 1989). ROC analysis consists in plotting, on a binary diagram, in the y-axis the Sensitivity value [sensitivity = number of land units correctly assessed as unstable (true positive)/total number of land units really unstable (true positive + false negative)], and in the x-axis the values of 100-Specificity [specificity = number of land units correctly assessed as stable (true negative)/total number of really stable territorial units (true negative + false positive)].

LR and ROC analyses have been performed in the IBM-SPSS ver. 19 statistical package.

4.2 Sampling modes

In order to verify which mode provides the best results, four different types of sampling modes have been tested (Fig. 3a–d). These are:

1. The whole study area (Fig. 3a).
2. A series of transects covering a surface of about 1/3 of the whole study area (Fig. 3b). The study area has a westward prevailing aspect. The same trend characterises territorial distribution of lithology (LTU), land use (LUS) and distance to fault (FDIST). Thus, transect are set orthogonal to the N–S trend of the topographic relief of the study area, i.e. in a way that almost the whole range of elevation and its eventual trend be followed. Such arrangement criterion for transects is commonly adopted in geological and geomorphological studies, when it is not possible to survey completely the study area. In our field of interest, transect with different arrangements were tested by Guarascio et al. (2005) confirming tradition geological procedure. Transects are representative for categorical variables (Lithology – LTU and Land Use – LUS) perspective.
3. Buffers surrounding the detected phenomena with a 1/1 ratio between the stable and the unstable area (Fig. 3c).
4. Buffers with a 1/2 ratio between the stable and the unstable area (Fig. 3d).

These last two sampling modes consists in generating a buffer around each landslide of the training set. The width

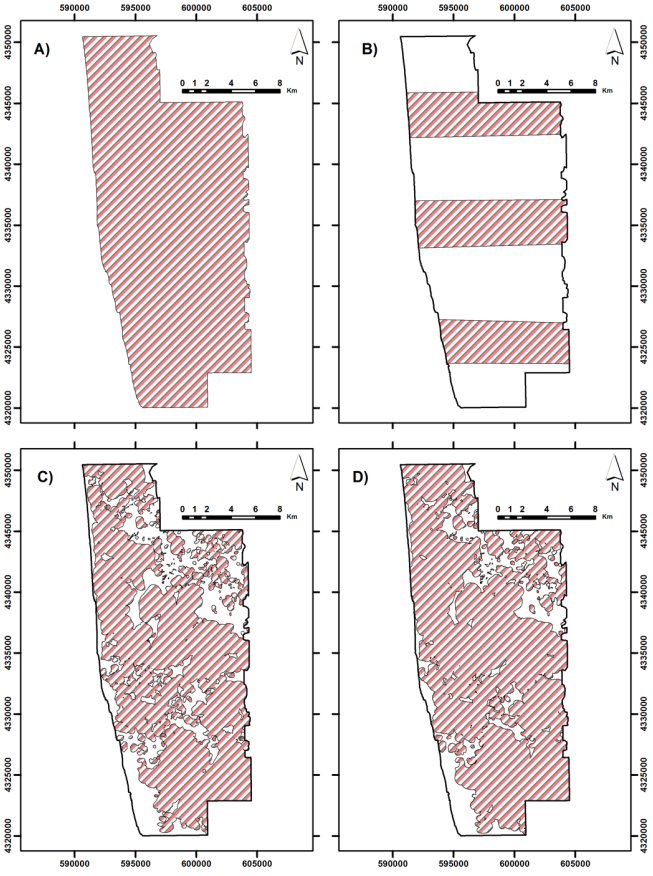

Fig. 3. Tested sampling area: **(A)** All area; **(B)** Transects; **(C)** Buffer 1/1; **(D)** Buffer 1/2.

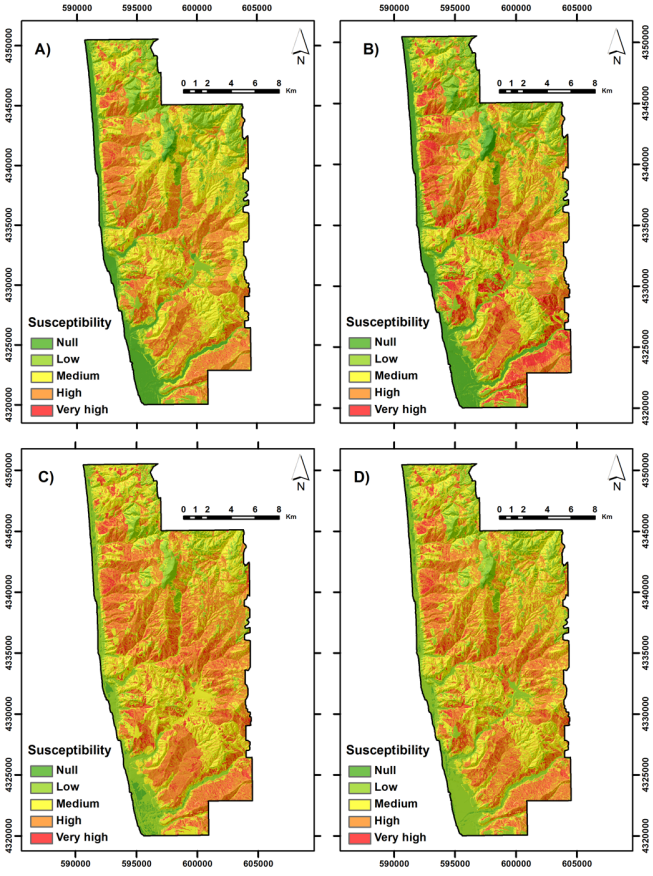

Fig. 4. Susceptibility maps obtained with square cells and grouped variables for tested sampling area: **(A)** All area; **(B)** Transects; **(C)** Buffer 1/1; **(D)** Buffer 1/2.

of the buffer around each landslide was set so that the ratio between unstable and stable cells is 1/1 or 1/2.

By using these four sampling modes as a training set, a series of LR analyses has been performed. Cells matrix and grouped variables, that have been adopted also in previous studies on performance assessment (Greco et al., 2007; Sorriso-Valvo et al., 2009), are also in subsequent assessment sessions.

By employing the training sets resulting from the different sampling modes, the susceptibility to phenomena has been evaluated; this has, in turn, produced four maps that are illustrated in Fig. 4.

The susceptibility maps obtained from the regressions have been validated through the ROC analysis. The results obtained after the validation step are shown in Fig. 5a–d and Table 2a. The results of these LR analyses highlight that the sampling with transects allow more accurate assessment.

4.3 Variables coding

In order to be able to compare the results obtained by means of grouped variables with those obtained through binary variables, we have generated a binary variable for each class of independent variables; we have thus obtained 44 binary dummy variables (Table 1).

Using dummy variables, four different regressions have been performed by considering the four previously illustrated sampling methodologies, which have produced the susceptibility maps illustrated in Fig. 6.

Also in this case, we have validated the results of the regressions by means of the ROC analysis. This second series of regressions confirms a better performance of the sampling with transects and indicates that binary variables provide the best results (Fig. 5e–h, Table 2b).

4.4 Reference land units

Once the best performance with transects sampling and binary variables has been established, the results obtained through a GIS system based on a cells matrix employed in all the above-mentioned regressions, have been compared with the results obtained through the employment of slope units.

Slope units have been generated for the study area by using the 20 m square cells DEM. Thanks to the hydrological functions of Arc-Info (Flowdir, Flowacc, Stream order, Stream

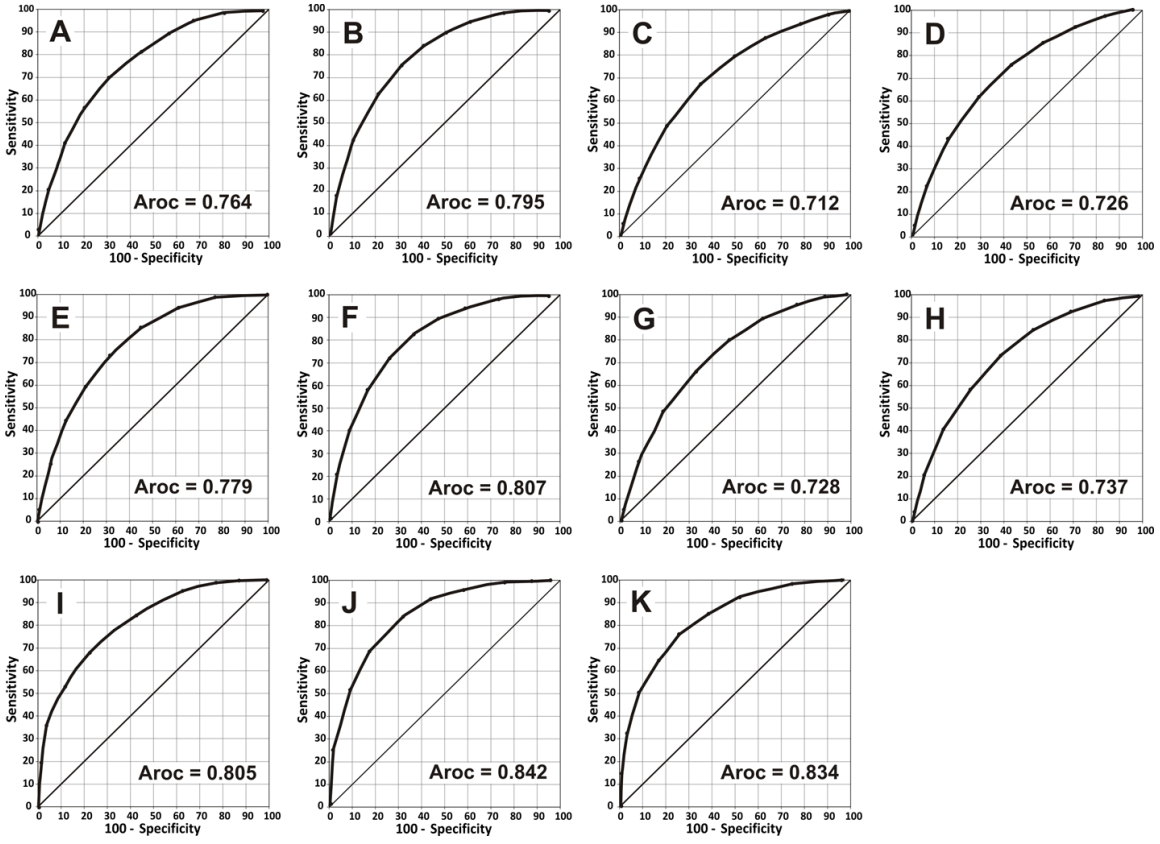

Fig. 5. ROC curves obtained with: square cells and grouped variables for tested sampling area, **(A)** All area; **(B)** Transects; **(C)** Buffer 1/1; **(D)** Buffer 1/2; square cells and dummy variables for tested sampling area, **(E)** All area; **(F)** Transects; **(G)** Buffer 1/1; **(H)** Buffer 1/2; slope units and dummy variables for tested landslide frequency thresholds (Flt), **(I)** Flt > 5 %; **(J)** Flt > 20 %; **(K)** Flt > 40 %.

link, Basin, Watershed) both the Horton ordering of branches of drainage network and the related watershed have been generated. I and II order branches have been excluded since the hydrological functions of Arc-Info generate a drainage network continuously covering the topographic surface, thus only by using branches with order greater than II a network fairly similar to the real one could be produced.

Slope units have been defined by extending the drainage branches up to the watershed. Thus, two slope units have been defined for each basin (Fig. 7). By means of this procedure, and combining among themselves the 769 500 square cells of the GIS system, 1,860 slope units have been obtained (Fig. 8); one of these units is made up by merging coastal plains with alluvial valley floors. This unit has been excluded from the susceptibility assessment test, as its constituent units are not slope units and landslide incidence is nearly null. In the development of a regression analysis employing slope units, it is necessary to define the minimum percentage of surface affected by mass movements for which a unit is considered unstable. In this study, three relevant different thresholds have been heuristically considered, respectively 5 %, 20 % and 40 % (Fig. 9).

Once the landsliding threshold has been defined, we have detected – for each slope units – the class prevailing in the unit itself for the categorical land variables (LTU, LUS, ELEV, ASP) and the average value for the parametric variables (SLO, ACUR, DCUR, TWI, FDIST). Such procedure can be easily applied by means of an overlay analysis between the layer of the slope units and those of the land variables. The prevailing classes and the average values detected have been assigned to the slope units (Table 3).

Through this approach, the layers of the independent land variables have been regenerated, and the independent variables are no longer distributed according to cells matrix, but according to slope units. By the binary recoding of such layers we have obtained 44 binary variables to be employed in this analysis; in the following analyses, binary variables only are used as they turned out to be more performing than grouped variables (see Sect. 4.3). The sample population upon which the regression coefficients had to be calculated has been drawn from 782 sampling slope units, completely or prevalently located in the previously tested transects (Fig. 8).

By employing the slope units and binary variables three regressions for the three different landsliding threshold

Table 2. Synthesis of $P(y)$ values and validation of RL performed: (A) with square cells and grouped variables for all tested sampling area; (B) with square cells and dummy variables for all tested sampling area; (C) with slope units and dummy variables for all tested landslide frequency threshold (Lft): (A) Lft > 5 %; (B) Lft > 20 %; (C) Lft > 40 %.

(A) Sampling	$P(y)$ (%)	ROC (%)
All area	0–85.0	76.4
Transects	0–93.4	79.5
Buffer 1/1	0–87.2	71.2
Buffer 1/2	0–86.7	72.6
(B) Sampling	**$P(y)$ (%)**	**ROC (%)**
All area	0–92.1	77.9
Transects	0–98.3	80.7
Buffer 1/1	0–94.3	72.8
Buffer 1/2	0–98.3	73.7
(C) Lft	**$P(y)$ (%)**	**ROC (%)**
>5 %	0–100	80.5
>20 %	0–100	84.2
>40 %	0–100	83.4

Lft = Landslide frequency threshold.

Table 3. Criteria for assigning values of territorial variables to slope units.

Variable	Criterion
LTU	Prevalence
LUS	Prevalence
ELEV	Prevalence
SLO	Average
ASP	Prevalence
ACUR	Average
DCUR	Average
TWI	Average
FDIST	Average

considered have been performed. The three susceptibility maps obtained are shown in Fig. 10.

Like for the previous regressions, we have validated the results of the regressions by means of the ROC analysis. Figure 5i–k and Table 2c display the results of the validation.

5 Discussion and conclusions

RL is a multivariate statistical analysis widely employed in the assessment of the risk from mass movements based on a set of land variables. In most cases, susceptibility to mass movements is determined by means of different

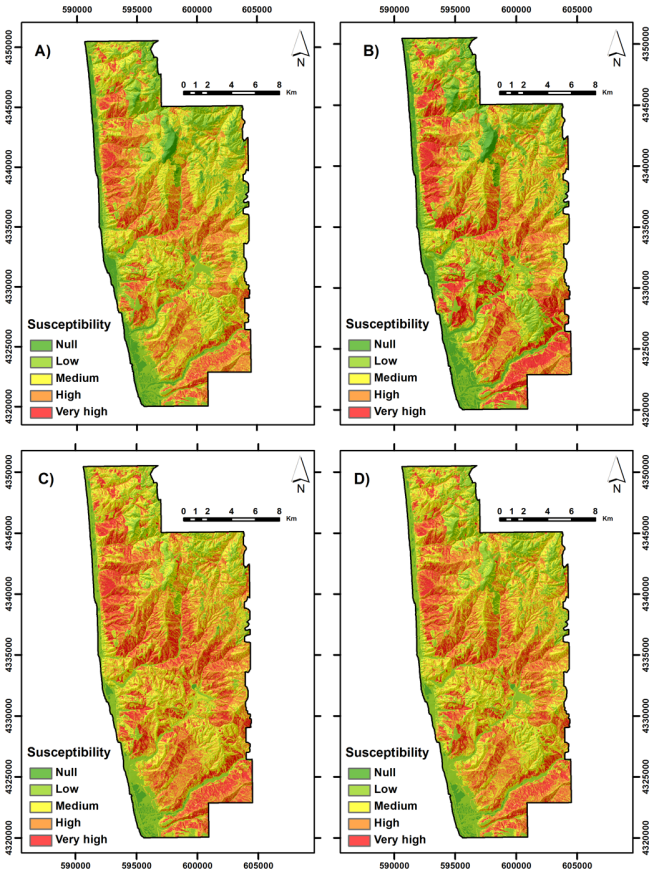

Fig. 6. Susceptibility maps obtained with square cells and dummy variables for tested sampling area: **(A)** All area; **(B)** Transects; **(C)** Buffer 1/1; **(D)** Buffer 1/2.

methodological approaches for those aspects of the procedure, which have not been rigorously standardized. In particular, procedural differences involve different sampling modes and variable management, as well as different types of reference land units.

With the aim to compare the effectiveness of the different methodological approaches, a series of regressions have been performed by using different procedures on a test area located in Calabria (southern Italy Fig. 1). Predictive capacities of the regressions carried out have been validated and compared by means of the ROC analysis (Hosmer and Lemeshow, 1989), and they have also been integrated with the visual analysis of the obtained susceptibility maps (Fig. 11).

A first series of regressions illustrated in Sect. 4.2, has been performed to test four different sampling types: (1) the whole study area (Fig. 3a); (2) transects running parallel to the general slope direction of the study area (Fig. 3b); (3) buffers surrounding the phenomena with a 1/1 ratio between the stable and the unstable area (Fig. 3c); (4) buffers with a 1/2 ratio between the stable and the unstable area (Fig. 3d).

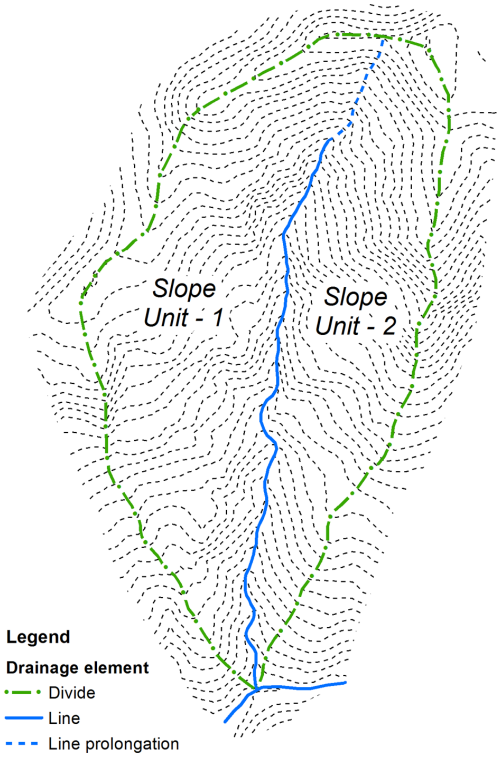

Fig. 7. Scheme of Slope Unit generation.

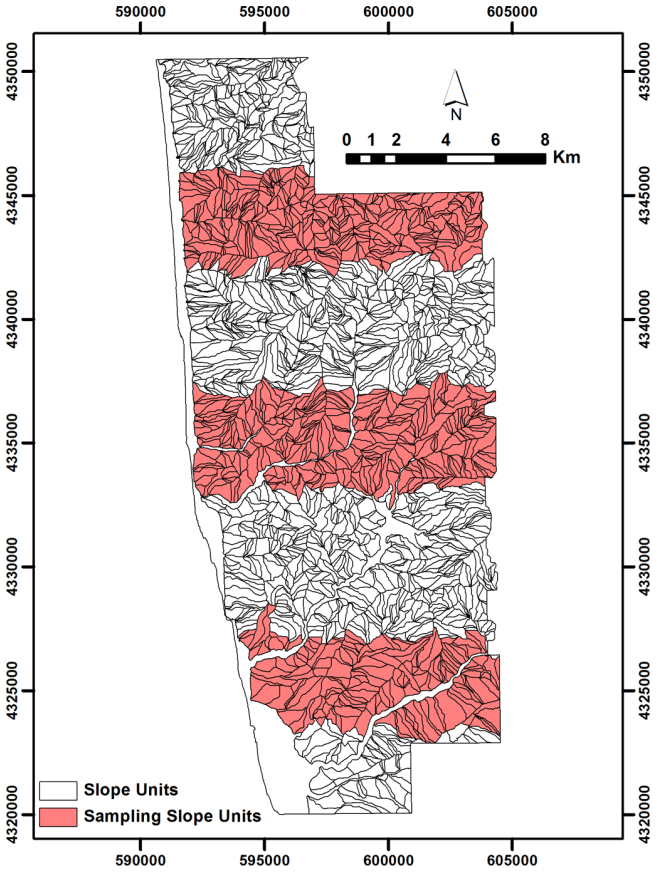

Fig. 8. Slope unit and sampling slope unit map of study area.

By means of ROC analysis the results obtained for the four regressions have been estimated (Fig. 5a–d and Table 2a). As it can be inferred from Table 2a the results obtained, both in terms of range of probability ($P(y)\max > 85\,\%$) and in terms of assessment accuracy (Aroc $> 70\,\%$) must be considered statistically acceptable in all cases. However, the regression employing transects (Fig. 3b) for the selection of the training set, provided the best results ($0 < P(y) \leq 93.4\,\%$; Aroc $= 79.5\,\%$).

The best performance obtained from the sample derived from transects is probably due to a greater representativeness of transects of the whole survey area, as compared to the buffers. On the other hand, transects are one of the sampling modalities making the LR more practical, thereby avoiding to sample the whole study area. Moreover, even the ratio between stable/unstable cells is similar to the values of the whole study area. Finally, the unexpected lower performance obtained through the analysis carried out by means of the sample derived from the whole study area resulted also in others studies (Greco et al., 2007), and it may be attributed to redundancy problems (excessive size of the population). A further reason for this result could be the fact (difficult to demonstrate) that mass movement has not affected some susceptible land units, so far.

The second methodological approach tested, regards the management modalities of independent variables. As for the use of binary variables, four extra regressions have been performed, by reconsidering the four different sample areas and using the 44 variables obtained from the binary reclassification of each category or class of predisposing land factors (Table 1). In this case too the validation (Fig. 5e–h and Table 2b) points out statistically acceptable results for all regressions ($P(y)\max > 92\,\%$; Aroc $> 72\,\%$). The use of the training set obtained by transects (Fig. 3b) provides once again the best results ($0 < P(y) \leq 98.3\,\%$; Aroc $= 80.7\,\%$), definitively confirming that such sampling modality allows a better performance of the LR analysis. The comparison between the sections A and B of Table 2, points out that, notwithstanding the sampling type, the use of binary variables provides, in any case, assessment that is more accurate.

The best performance of the regressions carried out by means of binary variables must be probably attributed to a greater associative capacity of the algorithm of regression between presence/absence of the phenomenon and presence/absence of the independent binary variable as compared to the variables ordered with increasing values according to the frequency of slope instability.

Finally, as regards the influence of the type of land units on the results, we have compared the results obtained by means of the employment of the slope units with those obtained by means of the square cells. By using the slope units

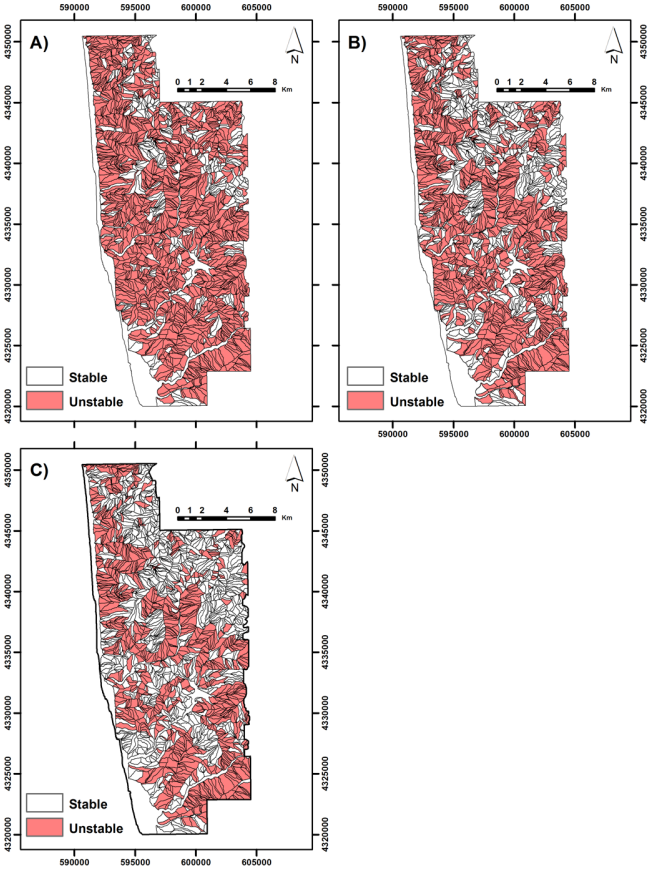

Fig. 9. Stable and unstable slope unit maps related to tested landslide frequency threshold (Lft): **(A)** Lft > 5 %; **(B)** Lft > 20 %; **(C)** Lft > 40 %.

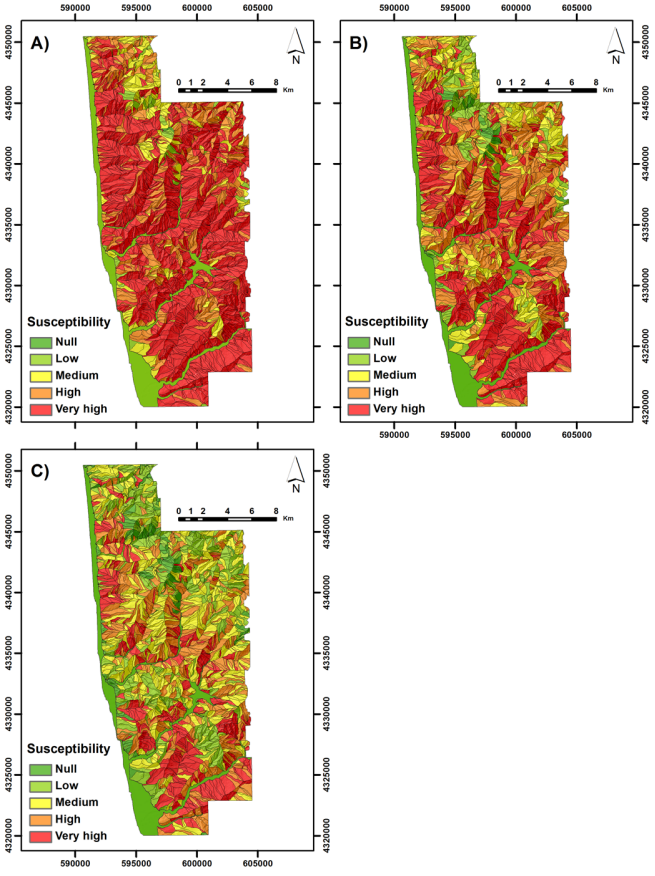

Fig. 10. Susceptibility maps obtained with slope units and dummy variables for tested landslide frequency threshold (Lft): **(A)** Lft > 5 %; **(B)** Lft > 20 %; **(C)** Lft > 40 %.

(Fig. 8), three regressions for different landsliding threshold (Lft) (Fig. 9) have been performed. As for sampling and transformation modalities of independent variables for these regressions, we have made recourse to the choices that had produced the best results in previous tests, i.e. transects (Fig. 8) and binary variables. The validation step (Fig. 5i–k and Table 2c) has shown more than acceptable results for the three threshold considered ($P(y)\max = 100\,\%$; Aroc > 80 %). However, the landsliding threshold of 20 % for the slope units $P(y) = 1$, provides a realistic probability range $0 < P(y) \leq 100\,\%$, and high performance (Aroc = 84.2 %).

If we compare the ROC curves of all illustrated regressions (Fig. 5a–k), it is evident that those related to the regressions based on slope units (Fig. 5i–k) display a subtended area, which is always higher than those based on cells matrix.

It is therefore evident that the features of the slope units, derived from the prevalence or the average of the values of the cells composing the unit themselves, are those that best represent the associations between characteristic elements of the areas affected by mass movement phenomena, while the estimate of $P(y)$ based on the cells, cannot physically represent the features of areas much wider than the cell itself, since this $P(y)$ estimate is expressed for each single cell.

Once a "positive" land unit has been determined by LR, this is then considered as being completely affected by the mass movement, while actually it is affected for surface rates higher than 50 % in the case of cells, and for a percent greater than the selected threshold in the case of the slope units. The contrary occurs for "negative" units. In any case, uncertainty results on the exact spatial location of the predicted event within the land unit. As the slope units on average contain 380 cells, it is evident that the topographic accuracy of the cells is greater than that of the slope units, while the accuracy of the prediction (ROC analysis) is greater by using slope units. To conclude, in the application of the LR the choice lies in what type of territorial units has to be adopted for the GIS based on the purposes of the survey.

The visual comparison among the best maps obtained by means of each methodological choice and the actual situation in situ (Fig. 11), points out that the map obtained by slope units and binary variables is the one being more similar to the actual situation, also in terms of sizes of unstable areas, above all in the central and southern part of the study

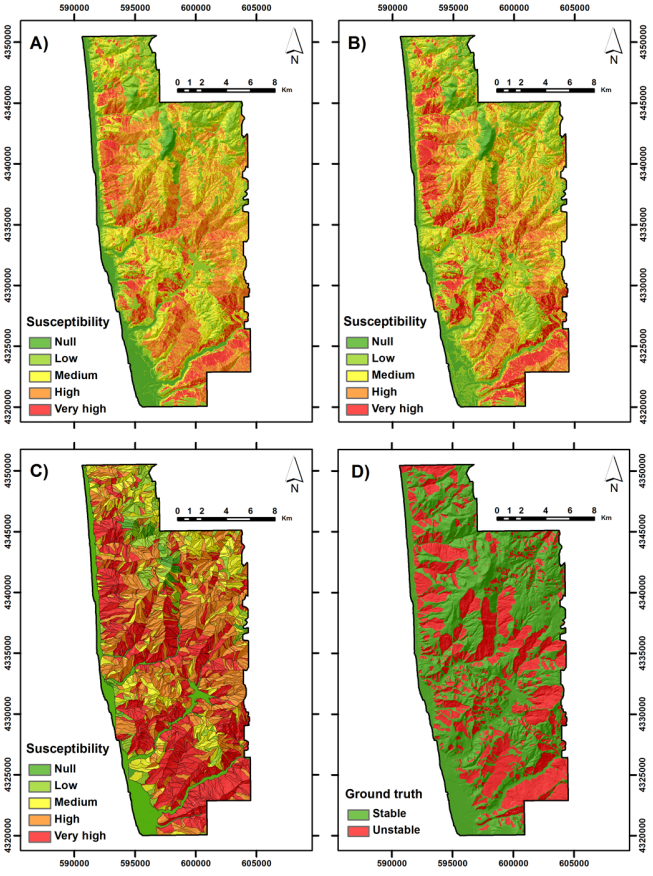

Fig. 11. Visual comparison between best susceptibility maps obtained with: **(A)** square cells and grouped variables, **(B)** square cells and dummy variables, **(C)** slope units and dummy variables; and ground truth **(D)**.

area., The comparison between observed landslide frequency in true positive land units of the best susceptibility maps (Table 4), provides a further confirmation of the best predictive ability of the analysis performed using dummy variables and slope units. The correspondence is greater in areas of estimated high susceptibility and effectively unstable areas. Such correspondence, which is implicit in the high percentage of correctly estimated cases, adds a direct check, which is immediately useful for the transmission of the results to users who may not be familiar with statistic land analyses.

To conclude, it is particularly important to ascertain that all LR analyses carried out have shown at least adequate predictive capacities (Table 2), with Aroc values in any case higher than 70 %, and with a maximum of 84.2 %. Such values are in line with the results obtained by other authors who have dealt with the susceptibility analysis trough LR in different land contexts (Nandi et al., 2009; Chauhan et al., 2010; Erener et al., 2010; Rossi et al., 2010). Such aspect suggests that Logistic Regression is a robust analytical method, which maintains high predictive capacities also when the features of input data are modified.

Table 4. Observed landslide frequency (%) in true positive land units of the best susceptibility maps obtained: (A) with square cells and grouped variables; (B) with square cells and dummy variables; (C) with slope units and dummy variables.

Susceptibility	A	B	C
Null	3.3	2.7	7.2
Low	19.3	17.7	11.7
Medium	34.8	35.9	41.3
High	56.5	57.3	61.6
Very high	73.4	79.8	85.5

Finally, considered the results of the analyses as a whole, we can state that the choice made in terms of sampling modalities, variables transformation and reference land units providing more accurate estimates of susceptibility, are based on transects sampling, binary reclassification of variables and the using slope units as GIS land units.

The values obtained are useful for land management on a medium-high territorial scale (whole municipalities or larger areas) but not for detailed actions such as the planning of the municipal housing sector, projects for lifelines, single buildings etc. This is particularly true if slope units are used, given the uncertainty of what parts are truly involved in slope instability phenomena by predicted phenomena.

Acknowledgements. Use of dataset by kind permission from Autorità di Bacino della Regione Calabria, POR Calabria 2000–2006, Azione 1.4c, Lotto 1, appointed actuator CNR-IRPI, Scientific Responsible G. Gullà, 2010. Authors wish to thank M. Komac and one anonymous referee for helpful advice.

Edited by: T. Glade
Reviewed by: M. Komac and one anonymous referee

References

AA.VV.: Climate Change: Global Risks, Challenges and Decisions, Synthesis Report, International Conference, available at: www.climatecongress.ku.dk (last access: 14 June 2012), 2009.

Aleotti, P. amd Chowdhury, R.: Landslide hazard assessment: summary review and new perspectives, Bull. Eng. Geol. Environ., 58, 21–44, 1999.

Ayalew, L. and Yamagishi, H.: The application GIS-based logistic regression for Landslide susceptibility mapping in the Kakuda-Yahiko Muuntains, central Japan, Geomorphology, 65, 15–31, 2005.

Baeza, C. and Corominas, J.: Assesment of shallow landslide susceptibility by means of multivariate statistical techniques, Earth Surf. Proc. Landf., 26, 1251–1263, 2001.

Bernknopf, R. L., Brookshire, D. S., and Shapiro, C. D.: A probabilistic approach to landslide hazard mapping in Cincinnati, Ohio, with applications for economic evaluation, Bull. Assoc. Eng. Geol., 24, 39–56, 1988.

Burton, A. N.: Carta Geologica della Calabria (1:25,000) – Relazione Generale. Cassa per Opere Straordinarie di Pubblico Interesse nell'Italia Meridionale (Cassa per il Mezzogiorno), 120 pp., 1971.

Can, T., Nefeslioglu, H. A., Gokceoglu, C., Sonmez, H., and Duman, T.: Susceptibility assessment of shallow eathflows tiggered by heavy rainfall at three catchments by logistic regression analyses, Geomorphology, 72, 250–271, 2005.

Carrara, A., Cardinali, M., Detti, R., Guzzetti, F., Pasqui, V., and Reichenbach, P.: GIS Techniques and statistical models in evaluating landslide hazard, Earth Surf. Proc. Landf., 16, 427–445, 1991.

Carrara, A., Cardinali, M., Guzzetti, F., and Reichenbach, P.: GIS technology in mapping landslide hazard, Kluwer Academic Publishers, Dordrecht, 135–175, 1995.

Carrara, A., Crosta, G., and Frattini, P.: Comparing models of debris-flow susceptibility in the alpine environment, Geomorphology, 94, 353–378, 2008.

Chau, K. and Chan, J. E.: Regional bias of landslide data in generating susceptibility maps using logistic regression: Case of Hong Kong Island, Landslides, 2, 280–290, 2005.

Chauhan, S., Sharma, M., and Arora, M. K.: Landslide susceptibility zonation of the Chamoli region, Garhwal Himalayas, using logistic regression model, Landslides, 7, 411–423, 2010.

Chen, Z. and Wang, J.: Landslide hazard mapping using logistic regression model in Mackenzie Valley, Canada, Nat. Hazard, 42, 75–89, 2007.

Choi, J., Oh, H. J., Lee, H. J., Lee, C., and Lee, S.: Combining landslide susceptibility maps obtained from frequency ratio, logistic regression, and artificial neural network models using ASTER images and GIS, Eng. Geol., 124, 12–23, 2012.

Chung, C.-J. F. and Fabbri, A. G.: Probabilistic prediction models for landslide hazard mapping, Photogram. Eng. Remote Sens., 65, 1389–1399, 1999.

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., Lee, C. F., Tham, L. G., Ng, K. C., and Shum, W. L.: Logistic regression modelling of storm-induced shallow landsliding in time and space on natural terrain of Lantau Island, Hong Kong, Bull. Eng. Geol. Environ., 63, 315–327, 2004.

Das, I., Sahoo, S., Van Westen, C., Stein, A., and Hack, R.: Landslide susceptibility assessment using logistic regression and its comparison with a rock mass classificatiuon system, along a road section in the northern Himalayans (India), Geomorphology, 114, 627–637, 2010.

Erener, A., Sebnem, H., and Duzgun, B.: Improvement of statistical landslide susceptibility mapping by using spatial and global regression methods in the case of More and Romsdal (Norway), Landslides, 7, 55–68, 2010.

Falaschi, F., Giacomelli, F., Federici, P. R., Puccinelli, A., D'Amato Avanzi, G., Pochini, A., and Ribollini, A.: Logistic regression versus artificial neural networks: landslide susceptibility evaluation in a sample area of the Serchio River valley, Italy, Nat. Hazard, 50, 551–569, 2009.

García-Rodríguez, M. J., Malpica, J. A., Benito, B., and Díaz, M.: Susceptibility assessment of earthquake-triggered landslides in El Salvador using logistic regression, Geomorphology, 95, 172–191, 2008.

Gorsevski, P. V., Gessler, P., and Foltz, R. B.: Spatial prediction of landslide hazard using logistic regression and GIS, Proc. 4th International Conference on Integrating GIS and Environmental Modelling (GIS/EM$): Problems, Prospect and Research Needs, Banff, Alberta, Canada, 2000.

Greco, R., Sorriso-Valvo, M., and Catalano, E.: Logistic Regression analysis in the evaluation of mass-movement susceptibility: the Aspromonte case study, Calabria, Italy, Eng. Geol., 89, 47–66, 2007.

Guarascio, G., Silvano, S., and Sorriso-Valvo, M.: Evaluation of susceptibility to landslides applying Logistic Regression in the "Fiumara Laverde" basin (Aspromonte, Calabria, Italy). Unpublished stage report, University of Padova, CNR-IRPI, 2005.

Guha-Sapir, D., Vos, F., Below, R., and Ponserre, S.: Annual Disaster Statistical Review 2010: The Numbers and Trends, Brussels, CRED, 50 pp., 2011.

Gullà, G., Antronico, L., Brunetti, M., Coscarelli, R., Critelli, S., Dramis, F., Iovine, G., Mattei, M., Molin, P., Muto, F., Nanni, T., Petrucci, O., Robustelli, G., Sorriso-Valvo, M., Terranova, O., and Versace, P.: Sviluppo e applicazione di metodi per la valutazione della pericolosità dei fenomeni di dissesto dei versanti. Relazione Finale, POR Calabria 2000–2006, Asse 1, Misura 1.4, Azione 1.4.c, Lotto 1, 350 pp., 2010.

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

Guzzetti, F., Reichenbach, P., Cardinali, M., Galli, M., and Ardizzone, F.: Probabilistic landslide hazard assessment at the basin scale, Geomorphology, 72, 272–299, 2005.

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., and Galli, M.: Estimating the quality of landslide susceptibility model. Geomorphology, 81, 166–184, 2006.

Hosmer, D. W. and Lemeshow, S. (Eds.): Applied Regression Analysis, Wiley, New York, 307 pp., 1989.

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

Mathew, J., Jha, V. K., and Rawat, G. S.: Landslide susceptibility zonation mapping and its validation in part of Garhwal Lesser Himalaya, India, using binary logistic regression analysis and receiver operating characteristic curve method, Landslides, 6, 17–26, 2009.

Moore, I. D., Grayson, R. B., and Ladson, A. R.: Digital terrain modeling: a review of hydrological, geomorphological, and biological application, Hydrol. Process., 5, 3–30, 1991.

Nandi, A. and Shakoor, A.: A GIS based landslide susceptibility evaluation using bivariate and multivariate statistical analyses, Eng. Geol., 110, 11–20, 2009.

Nefeslioglu, H. A., Gokceoglu, C., and Sonmez, H.: An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps, Eng. Geol., 97, 171–191, 2008.

Ohlmacher, G. C. and Davis, J. C.: Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA, Eng. Geol., 69, 331–343, 2003.

Pampel, F. C. (Ed): Logistic Regression: A Primer.,Sage Universitu Paper Series on Quantitative Application in The Social Sciences, 07-132, Thousand Oaks, CA, Sage, 2000.

Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A. C., and Perucacci, S.: Optimal landslide susceptibility zonation based on multiple forecast, Geomorphology, 114, 129–142, 2010.

Scheuren, J.-M., le Polain de Waroux, O., Below, R., Guha-Sapir, D., and Ponserre, S.: Annual Disaster Statistical Review 2007: The Numbers and Trends. Brussels: CRED; 64 pp., 2008.

Sorriso-Valvo, M., Greco, G., and Catalano, E.: Spatial prediction of regional scale mass movement using the Logistic Regression analysis and GIS – Calabria, Italy, Israel J. Earth Sci., 57, 263–280, 2009.

Thiery, Y., Malet, J.-P., Sterlacchini, S., Puissant, A., and Maquaire, O.: Landslide susceptibility assessment by bivariate methods at large scales: Application to a complex mountainous environment, Geomorphology, 92, 38–59, 2007.

Van Den Eeckhaut, M., Vanwallerghem, T., Poesen, J., Govers, G., Verstraten, 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, 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.

Van Den Eeckhaut, M., Marre, A., and Poesen J.: Comparison of two landslide susceptibility assessments in the Champagne-Ardenne region (France), Geomorphology, 115, 141–155, 2010.

Van Westen, C. J., Castellanos, E., and Kuriakose, S. L.: Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview, Eng. Geol., 102, 112–131, 2008.

Vos, F., Rodriguez, J., Below, R., and Guha-Sapir, D.: Annual Disaster Statistical Review 2009: The Numbers and Trends, Brussels, CRED, 46 pp., 2010.

Yalcin, A., Reis, S., Aydinoglu, A. C., and Yomralioglu, T.: A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods for landslide susceptibility mapping in Trabzon, NE Turkey, Catena, 85, 274–287, 2011.

Yesilnacar, E. and Topal, T.: Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey), Eng. Geol., 79, 251–266, 2005.