Nat. Hazards Earth Syst. Sci., 10, 623–640, 2010
www.nat-hazards-earth-syst-sci.net/10/623/2010/


Natural Hazards and Earth System Sciences

# Landslides and vegetation cover in the 2005 North Pakistan earthquake: a GIS and statistical quantitative approach

**P. Peduzzi**

UNEP/GRID-Europe, Switzerland

Institute of Geomatics and Risk Analysis (IGAR), University of Lausanne, Switzerland

Received: 15 July 2009 – Revised: 26 February 2010 – Accepted: 28 February 2010 – Published: 1 April 2010

**Abstract.** The growing concern for loss of services once provided by natural ecosystems is getting increasing attention. However, the accelerating rate of natural resources destruction calls for rapid and global action. With often very limited budgets, environmental agencies and NGOs need cost-efficient ways to quickly convince decision-makers that sound management of natural resources can help to protect human lives and their welfare. The methodology described in this paper, is based on geospatial and statistical analysis, involving simple Geographical Information System (GIS) and remote sensing algorithms. It is based on free or very low-cost data. It aims to scientifically assess the potential role of vegetation in mitigating landslides triggered by earthquakes by normalising for other factors such as slopes and distance from active fault. The methodology was applied to the 2005 North Pakistan/India earthquake which generated a large number of victims and hundreds of landslides. The study shows that if slopes and proximity from active fault are the main susceptibility factors for post landslides triggered by earthquakes in this area, the results clearly revealed that areas covered by denser vegetation suffered less and smaller landslides than areas with thinner (or devoid of) vegetation cover. Short distance from roads/trails and rivers also proved to be pertinent factors in increasing landslides susceptibility. This project is a component of a wider initiative involving the Global Resource Information Database Europe from the United Nations Environment Programme, the International Union for Conservation of Nature, the Institute of Geomatics and Risk Analysis from the University of Lausanne and the "institut universitaire d'études du développement" from the University of Geneva.

*Correspondence to:* P. Peduzzi
(pascal.peduzzi@grid.unep.ch)

## 1 Introduction

Overexploitation of natural resources and deforestation is one of the main triggers for the observed increase in landslide disasters along with increase in population exposure (Nadim et al., 2006). While timber production, grazing or woodfuel collection are activities supporting livelihoods, their impact on vegetation cover needs to be addressed, as forests are being harvested, converted to crop land or pasture at an accelerating pace. Current deforestation reaches 13 millions ha per year (Fao, 2006). Conversely, restoration of vegetation coverage can be a cost-effective method for risk reduction. Planting mangroves for tropical cyclones protection revealed to be seven fold cheaper than dike maintenance, while also providing secondary benefits for local livelihoods (IFRC, 2005).

The Hyogo Framework for Action (HFA) "*encourages the sustainable use and management of ecosystems*, [for] *reducing the underlying risk*" (UNISDR, 2005). To achieve this goal, both local authorities and international decision-makers need to adopt improved environmental policies. Yet, convincing people to change their practices demands tangible evidence and clear examples of sound environmental management. Post-disaster situations might provide a favourable impetus to bring new concepts and to avoid rebuilding risk during the reconstruction phase. There is a need for a multiplication of scientific evidence and thus for developing simple methods allowing solid scientific assessments. Outputs from this quantitative analysis were used in an interdisciplinary study for disaster risk reduction (Sudmeier-Rieux et al., 2008). It explored the relation between land use factors, such as deforestation, grazing, road building, etc. on the frequency of landslides and coping strategies developed by the population. It was applied and tested on the area affected by the earthquake that hit North Pakistan

Published by Copernicus Publications on behalf of the European Geosciences Union.

and India on 8 October 2005. The epicentre was located at 34.493° N, 73.629° E and had a recorded magnitude of 7.6 $M_w$ on Richter scale (USGS, 2006). It devastated a large stretch of the region, killing between 74 647, injuring 134 622 and leaving 5.15 million homeless and resulted in an economic loss evaluated at 6.2 billion US$ (CRED, 2009). While impressive, these figures fail to capture the level of despair of the surviving population. The heaviest damage arose in the Muzaffarabad area and Kashmir, where entire villages were destroyed. More than 30% of the victims were killed by landslides (Petley et al., 2006). More than 2400 landslides were identified by remote sensing techniques (Sato et al., 2007) following this earthquake. The biggest individual landslide triggered by the Kashmir Earthquake 2005 was the 68 millions $m^3$ Hattian Bala rock avalanche that killed about 1000 people (Dunning et al., 2007).

Understanding why landslides claim such a high death tolls is thus an important task. Not only to identify the potential future slope failure, but also to see if portion of past susceptibility can be attributed to human activities. This is particularly relevant in this context, as five months before the October 2005 earthquake, an IUCN-Pakistan report highlighted the risk "*from a possible human catastrophe due to the growing danger of landslides that was haunting the locals owing to heavy constructions, ruthless deforestation and massive quarrying.*" (IUCN, 2005).

Landslides are complex hazards requesting the collection of many different parameters to produce susceptibility maps: slopes, lithology, identification of recent deforestation, proximity from roads and presence of triggers (such as heavy precipitations or seismic activities) are the most commonly used factors in landslides modelling (Guzzetti et al., 1999; Gorsevski et al., 2001; Vanacker et al., 2003; Ayalew and Yamagishi, 2005).

Guzzetti et al. (1999) describe five main categories of techniques for mapping landslides susceptibility (geomorphological hazard mapping, analysis of landslides inventories, heuristic or index based methods, functional, statistically based models, geotechnical or physically based models). They can be gathered in two broader categories: qualitative and quantitative (Ayalew and Yamagishi, 2005).

The most common types of qualitative methods simply use landslide inventories (Ayalew and Yamagishi, 2005). Landslide inventories attempts to predict future patterns of slope failure by preparing landslide density maps (Guzzetti et al., 1999).

As part of a preliminary study (Sudmeier-Rieux et al., 2007), a first landslide inventory was undertaken by computing the landslides density on different landcover, slope classes and geological formations.

It showed for instance that 57% of landslides occurred in Murree geological formation. Although at first glance it seems that such geological is highly susceptible to landslide, this is less evident when knowing that such formation accounted for 52.3% of the area. Qualitative methods are useful in preliminary tests, but are only of interest if the ratio of percentage of landslides over percentage of coverage of the selected feature is showing significant over (or under) representation. In the case of the Murree geological formation, such ratio is about 1.1 (57/52.3). This is not to say that such geological formation has a neutral role, but only that this is not statistically relevant.

Looking at landcover was more interesting. While forests cover about 45% of the study area, it includes only 17% of total landslides, ratio = 0.36 (17/45), so forest are underrepresented. Deforested and grazing areas covers 42% of the study area, but includes 54.8% of total landslides, ratio = 1.3 (45.8/42), so areas with no or low vegetation density are over-represented. However this can easily be a fluke correlation as one might argue that forested areas may potentially be more located on gentler slopes, or areas further away from fault line or on different geological formation.

Quantitative methods overcome the issue of potential interconnectivity between the susceptibility factors. The most robust method is based on a deterministic approach. This consists of an engineered evaluation of slope instability based on exhaustive collection of relevant data. Such exercises are very time consuming and costly. Deterministic approaches request comprehensive local assessments and are thus usually conducted over small areas (Ayalew and Yamagishi, 2005) and one might add: with high financial values given the resources needed.

For large areas, statistical quantitative approaches are more appropriate. In this category, logistic regression and discriminant analysis are the most frequently chosen models (Guzzetti et al., 1999; Brenning, 2005). This requests first to identify past landslides (usually using remote sensing) and then prepare map layers for each potential susceptibility factors using GIS techniques (Guzzetti et al., 1999; Ayalew and Yamagishi, 2005; Brenning, 2005; Vanacker et al., 2003; Coe et al., 2004). This allows for the identification and the estimation of the relative contribution of the instability factors and the cartography of different hazard degree (Guzzetti et al., 1999).

Such multiple regression analysis associated with extraction of parameters using GIS, was already used for highlighting the role of deforestation in landslides (e.g. Vanacker et al., 2003). In their study, Vanacker et al. extracted slopes, aspect, distance to valley and type of landcover. They also used different sets of aerial photos taken at different years to look at the role of deforestation (and time since deforestation) for landslides susceptibility. They concluded, that in their area of study, the overall susceptibility of slope movement was highly dependent on recent land-use changes. Vegetation can reduce landslide susceptibility (both shallow and deep landslides) by reducing water content in the soil (Popescu, 2002) or may reduce shallow landslides with the mechanical role of roots in anchoring the soil. However, vegetation may also destabilize the forces by adding weight and acting as a

surcharge as well as by wind forces on vegetation exposed (Popescu, 2002).

Remote sensing techniques are useful for estimation of crustal deformation using either passive sensors (Avouac et al., 2006) or radar imagery. A remarkable study (Sato et al., 2007) used Synthetic Aperture Radar (SAR) images from ENVISAT revealing a maximum six-meter uplift north of Muzaffarabad. In the same study, landslide detection performed by comparing a pair of pre- and post-event SPOT-5 images plotted over a Digital Elevation Model (DEM). Sato et al. showed that a majority of landslides (63.3%) occurred on slope steeper than 30 degrees and that gentler slopes were also affected by landslides when closer to active faults. However, they found that this was not systematic, as large-scale slope failures also occurred at slopes less than 30 degrees at a longer distance. Steeper slopes located further away from the active faults were not necessarily more affected. Slope and distance from active faults are thus only part of the story.

In order to highlight the potential role of vegetation in mitigating slope failure, this current study builds on similar methodologies developed for different hazard types (Peduzzi et al., 2002; Chatenoux and Peduzzi, 2007). It uses multiple regressions to normalise geophysical and geographical parameters (such as slopes, distance from active fault, distance from rivers) to highlight other parameters related to human activities (presence of roads, vegetation removal). A logistic regression with stepwise variable selection proved to be adequate for landslide susceptibility modelling (Brenning, 2005). In order to ensure easy reproduction even for low-budget institutions, the research is based on free or low-cost data. It is thus based on published material and free global datasets, with the sole acquisition of a (low-cost) 30 m DEM derived from ASTER satellite sensor. This paper describes how spatial and statistical analysis using remote sensing and GIS techniques were applied (see Table E1 for the list of software used).

One of the key inputs consisted of the use of results from two previous assessment made by the *Service Regional de Traitement d'Image et de Télédétection* (SERTIT) and the National Engineering Services of Pakistan (NESPAK) which identified post-disaster landslides using satellite imagery. Both institutes kindly provided the two sets of detected landslides. To study which parameters are potentially linked with landslide susceptibility, a series of potential susceptibility factors were extracted using GIS techniques (slope variation and steepness, vegetation density, and distance from epicentres/active fault, rivers, roads or trails). Satellite imagery and simple remote sensing computation were also used to evaluate vegetation density. The Normalised Difference Vegetation Index (NDVI) is commonly used as a proxy for vegetation density (Tucker, 1979). It was computed and statistical regressions were run to identify potential susceptibility factors associated with observed slope failures. Once identified, the identified factors were introduced into the GIS to provide a landslide susceptibility map.

## 2 Data collection

### 2.1 Selection of the study area

The study area is a 60 by 60 km square (3600 $km^2$) delimited by the choice of the ASTER DEM covering Muzaffarabad and the Neelum valley (the bounding coordinates are: 73.23 E, 34.65 N, 73.86 E, 34.56 N; 73.72 E, 34.02 N; 73.09 E, 34.11 N). It lies in North Pakistan and India with more than half over the disputed territory of Jammu Kashmir (sovereignty status still unsettled). It includes the largest epicentre of 7.6 on Richter scale in the north of the study area, while numerous replicas are just outside in the northeast (see map of the study area in Fig. 1).

The altitudes range between 552 m and 4476 m (average around 1700 m) in this complex pattern featuring a rugged landscape. The rough relief of this mountainous area might be of concern for remote sensing processes, due to the areas in the shadow.

To overlay the different layers of information, the data were all projected in UTM 43 N (datum: WGS 1984). The full list of data sources is provided in Table A1.

### 2.2 Hypothesis and data sources

The dependant variable to be explained is the size of landslides. The original data on detected landslides were obtained through the Humanitarian Information Centre for Pakistan (HIC) but generated by two different offices: (SERTIT) based on 5-m SPOT-4 images and from the National Engineering Services of Pakistan (NESPAK) at a lower resolution. To explain the variation in landslide size, several hypotheses were made. Assuming that distance from active fault is an important factors, the Muzaffarabad and Tanda fault lines were manually digitalized (R. Klaus internship at UNEP/GRID-Europe) from a map extracted from Nakata et al. (1991) at a scale of 1:100 000. From this dataset the distance from the active fault was computed for each pixel. Epicentres were geo-referenced based on latitude/longitude information retrieved from the Advanced National Seismic System (ANSS) composite catalogue, the Northern California Earthquake Data Center (NCEDC); The ANSS composite catalogue. Distances from epicentres (and replicas) were computed for three categories of epicentres: the first one includes epicentres comprised between 5.5 and 7 $M_W$ on Richter scale, the second one for epicentres between 7 and 7.5 $M_W$, the last one consisting of the main epicentre at 7.6 $M_W$.

Another hypothesis was made that the presence of roads and trails could destabilise the slopes by either allowing infiltrations or by destabilising the balance slope of the material. Trails for pedestrians and cattle were distinguished from roads for cars and trucks. The data were provided (and digitalized) by the United Nations Joint Logistics Centre

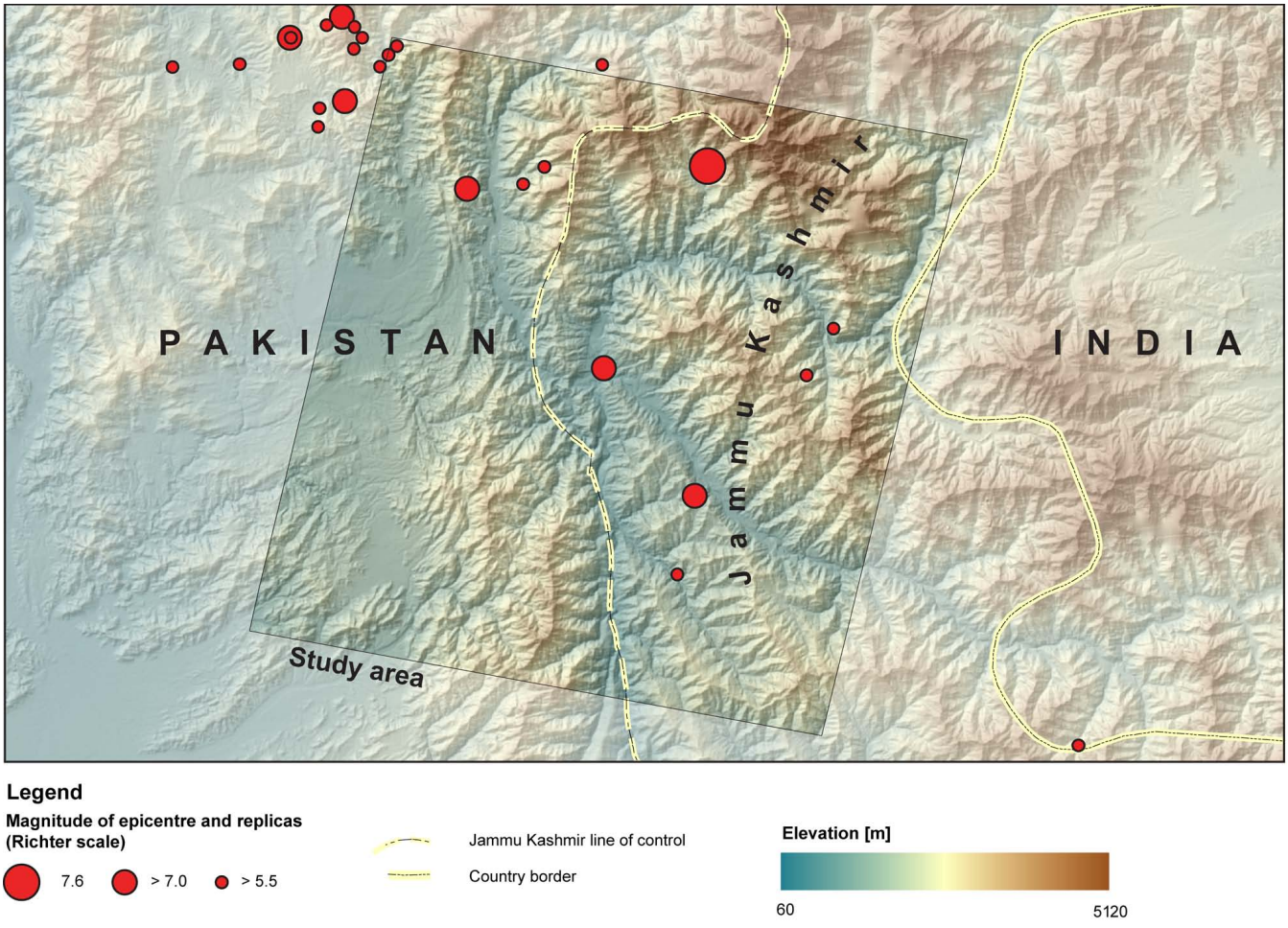



**Fig. 1.** Map of the study area (the boundaries and names shown on this map do not imply official endorsement or acceptance by the United Nations or by the author).

(UNJLC). Based on this dataset, distances from nearest roads and trails were computed for each pixel using GIS.

On the satellite image, numerous landslides appear to be located close to (or touching) rivers. The files for rivers were downloaded from the Data Repository of the Geographic Information Support Team (GIST). Distances from rivers were computed for each pixel. Soil types are also a central element for landslides susceptibility. At this stage, the soil coverage in this region is still a major gap in the analysis that needs to be bridged.

A main hypothesis is that slope is the primary causal factor of landslides (including debris flows, blocks fall and landslides). Two DEM were used. The Shuttle Radar Topography Mission (SRTM) version 3 (at 90 m spatial resolution) was obtained from the CGIAR Consortium for Spatial Information (CGIAR-CSI) and the ASTER (at 30 m spatial resolution) purchased from United States Geological Survey (USGS). These datasets allowed the computation of slopes for each pixel and to derive the maximum, average and standard deviation of slopes within each landslide).

Finally, the aim of the study was to ascertain the role of vegetation in relation with landslides. A pre-event satellite image from Landsat ETM sensor was used (image path: 150, row 36 from 7 October 2002). It was obtained from Landsat.org, Global Observatory for Ecosystem Services, Michigan State University. A normalise difference vegetation Index (NDVI) was computed. This combination of recorded electromagnetic reflectance in near Infra-red (nIR) and red (Red) wavelengths is highly correlated with photosynthesis activity, hence with density of vegetation (Tucker, 1979). Being a complex ratio it also reduces the problem of shadows produced by topographic effects. This was particularly relevant over this mountainous area. The ratio is computed using the equation:

$$\mathrm{NDVI} = \frac{(\mathrm{nIR} - \mathrm{Red})}{(\mathrm{nIR} + \mathrm{Red})} \quad (1)$$

Where: NDVI: Normalized Difference Vegetation Index; nIR: electromagnetic reflectance in Near Infra-Red (not

equal to zero); Red: electromagnetic reflectance in Red (not equal to zero).

## 3 Methodology

### 3.1 Role of vegetation cover and landslides

To understand why some areas led to large landslides while others suffered smaller landslides, some basic tests using correlation matrix between variables and area of landslides, or 3-D surfaces can provide useful hints. However, to better understand the context, multiple regressions analysis should be run. The size of landslide was set as the dependant variable for the magnitude of landslides. Prior to testing whether variations in vegetation cover density have an effect on the size of detected landslides, standardisation of other parameters is needed. A hypothesis was made that slope failures, triggered by earthquakes, were related with slopes, type of soil (not tested due to lack of appropriate data), proximity from active fault or epicentres and proximity from rivers.

Once the geophysical and morphological parameters related to slope failure are identified, proximity from trails (or roads) as well as role of vegetation cover can be introduced to see what are the potential mitigation or enhancing effects of these features.

The dependant variable was set to be the size of landslides. Using the post-event detection of landslides by SERTIT and NESPAK (further corrected by UNEP/GRID-Europe as explained in Sect. 3.2), the area (in $m^2$) of each landslide was computed using GIS. This variable was then transformed by computing the natural logarithm (LN) of the size.

$$\mathrm{LS}_{\mathrm{size}}=\alpha V_1^a \cdot \beta V_2^b \cdot \ldots \cdot \gamma V_n^i \quad (2)$$

#### 3.1.1 Factors to be tested

Other factors were extracted using GIS techniques and associated with each landslide. Examples of variables extracted are provided in Table 1 while the full list of tested variables (including transformed variables) is provided in the Table B1.

Transformation and normalisation of variables were needed because the links between landslide area and other contextual parameters are not necessarily linear and statistical linear regressions request variables that follow a normal distribution. In order to see how these variables behave, a visual test using histograms and scatter plots was performed and variables were transformed accordingly (as explained below).

### 3.2 Data preparation

The list of data sources is provided in Table A1. The preparation of data involved:

#### 3.2.1 Improving the recorded landslides data

The recorded landslides from both SERTIT and NESPAK, did not take into account the trans-edge or trans-river effect. In other words, two landslides having the same origin at a mountain edge, or two landslides ending their courses in front of each other at the bottom of a valley, were recorded by SERTIT or NESPAK as one landslide. To correct for these issues, manual transformation of these two datasets were made (thanks to the work of Rafaël Klaus as part of his internship at UNEP/GRID-Europe).

#### 3.2.2 Computation of distances

Distances were computed for the following features: roads, trails, active fault and epicentres. This operation was performed in a GIS using a raster of 30 m×30 m, where each cell includes the minimum distance to one of the selected feature as well as distance between the centre of each landslide and the selected features.

#### 3.2.3 Computation of NDVI

In Fig. 2 one can see that the computation of the NDVI strongly reduced the shadows produced by the relief. In the right image, low NDVI values are displayed in blue and include ice, snow, rivers and lakes, while green reflects the high NDVI values produced by dense vegetation.

#### 3.2.4 Slopes

Slopes were computed using GIS, based on both ASTER (30 m×30 m) and SRTM (90 m×90 m) DEM datasets. The SRTM covers a larger area (in fact the whole world is available), whereas the ASTER DEM purchased, "only" covers 3600 $km^2$. If the 90 m resolution is sufficient, an extrapolation to a larger area using SRTM would be possible.

### 3.3 Data extraction, transformation and integration

#### 3.3.1 Values extraction

By superimposing the detected landslides over the different layers of information, it was possible, for each individual landslide, to compute the minimum distance (or maximum, average…) from a specific feature (such as river, trails, roads, active fault, epicentres) or the maximum slope (average, standard deviation and square of maximum slope were also computed and extracted). The same process was applied to extract the minimum, maximum and average value of NDVI.

#### 3.3.2 Transformation of variables

Prior to developing the statistical analysis, the variables need to be selected and transformed to ensure that they follow a normal distribution. The link between landslide area and

**Table 1.** Examples of variables extracted for each detected landslide.

| Raw data | Derived variables | Type of values recorded for each landslide |
|---|---|---|
| Detected Landslides | Area | Area, maximum width and length. |
| DEM | Slope | Elevation difference, Maximum slope, average slope, standard deviation. |
| Epicentre locations | Distance from epicentres | Minimum distance between either edge of the landslides or centre of the landslide area. |
| active fault | Distance from fault line | Minimum distance between either edge of the landslides or centre of the landslide area. |
| Rivers | Distance from river | Minimum distance between either edge of the landslides or centre of the landslide area. |
| Road and trails | Distance from road and trails | Minimum distance between either edge of the landslides or centre of the landslide area. |
| Landsat ETM+ image | NDVI | Maximum, minimum and average NDVI value. |

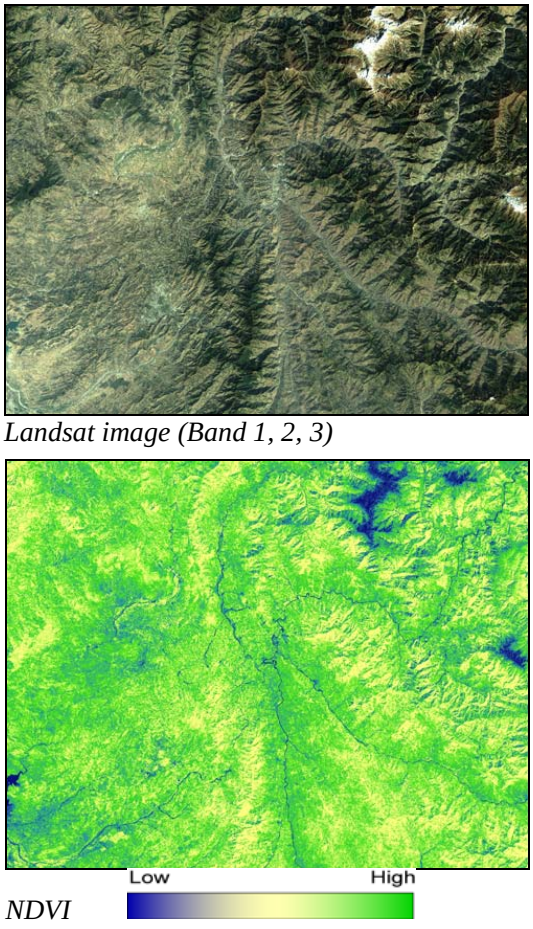



**Fig. 2.** Computation of NDVI using Landsat band 3 and 4.

other factors is not necessarily linear, so a visual interpretation followed to identify whether some functions can be applied to improve the link with landslide area. For example, in a scatter plot *landslide areas* seemed to present a link with the square of *maximum slope*. This function was hence computed for *maximum slope*.

The variables were transformed by taking the natural logarithm (LN) of scalar or, in some cases, the LN of transformed values. Transformations already proved to be efficient in previous studies (Peduzzi et al., 2002) (Chatenoux, 2007). For variables ranging between 0 and 1 (e.g. percentage, or NDVI) Eq. (3) was applied.

$$V_i' = \mathrm{LN}\left(\frac{V_i}{1-V_i}\right) \tag{3}$$

Where $V_i$ is the variable to be transformed and $V_i'$ is the transformed value.

The choice of logarithmic regression was made to reflect the interactivity between the different parameters, given the multiplicative effect on each other (an addition of LN being a multiplication of the exponents). This is believed to be pertinent, given the complexity of sites where one factor can mitigate or enhance another.

All the 36 variables (see Table B1) computed and/or transformed for all the individual landslides were placed in a database and then introduced into statistical software for multiple regression analysis.

### 3.3.3 Groups of independent variables

A correlation matrix (see Table C1) was computed between all the variables and used to discriminate variables that were too correlated to be taken together in regression analysis. Groups of independent variables were generated, each one corresponding to a specific hypothesis, which was tested by running multiple regression analysis. The selection of the most relevant hypothesis was based on relevance (p-level $<0.05$) and maximisation of percentage of variance explained ($R^2$). This process allows the identification of combinations of parameters that best explain the landslide area and thus confirms or rejects the hypothesis on the potential role of the different environmental and geomorphologic features.

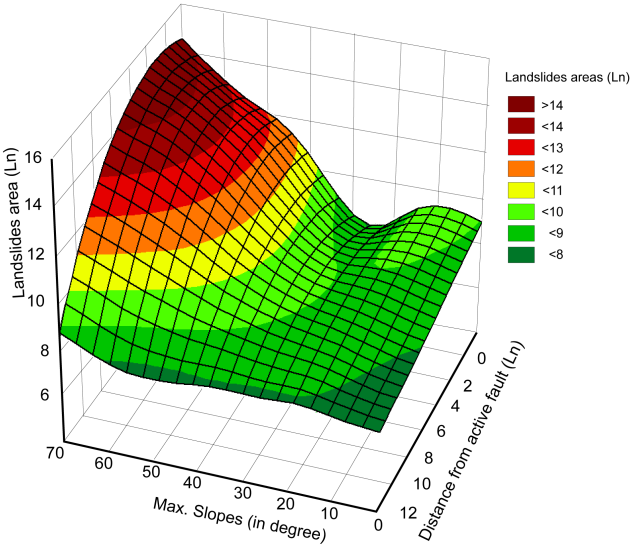



**Fig. 3.** Landslides area versus slopes and distance from active fault.

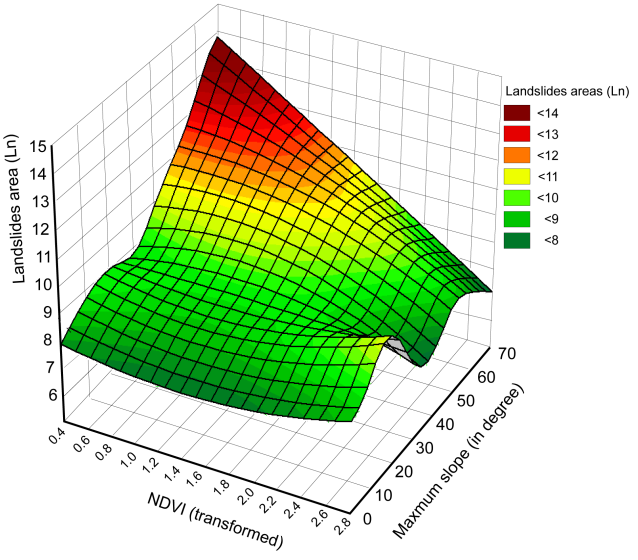



**Fig. 4.** Landslides area versus slopes and vegetation density (NDVI).

## 4 Statistical results

Hypothesis can be quickly tested using correlation matrix between variables and area of landslides. In some cases, plotting observed data in 3-D surfaces can provide useful hints to show correlation of landslides area with two independent variables. For example Fig. 3 shows that the landslides area decreases sharply when slopes decrease (until 30°) or when distance to active fault increase. It also shows that below a slope of 30°, landslides area may increase when located closer to active fault. This is in line with findings provided by Sato et al. (2007).

Similarly, the effect of vegetation density can be quickly tested by looking at 3-D surface between landslides area, maximum slope and a proxy of vegetation density (NDVI). Figure 4 highlights the role of vegetation density, where NDVI is inversely correlated with the Ln of landslides area. However, 3-D plots can only display the link with two explaining variables and some of the modulations viewed in Figs. 3 and 4 suggest that other variables play a role. For example, the increase in landslide area at gentle slope and high vegetation. Could it be along rivers? To really address the weight of each parameters and the potential multiplicative effect of variables, a multiple regression analysis is needed.

### 4.1 General model (all landslides considered – except outliers)

The multiple regressions analysis (Table 2) selected the following variables being associated with the landslides area (see Appendix F for further details on how to read the information provided in the table). The regression coefficients (third column) represent the weight that should be multiplying each independent variable to get the predicted dependent variable.

The expected LN of landslide area from this model would be:

$$\begin{aligned}\text{LN\_LsArea} = {} & -0.125\text{D\_river\_minLn} + 0.206\text{D\_trail\_avLn} \\ & -0.246\text{D\_fault\_minLn} \\ & -0.878\text{NDVIt\_avLn} + 0.263\text{Slope\_max}^2\text{Ln} \\ & +7.913\end{aligned}$$

However, a higher weight do not necessarily implies a higher degree of influence. This is because the units are not the same between the variables as they were not standardized to a mean of 0 and a standard deviation of 1. The magnitude of the "Beta" coefficients provide information on the relative contribution of each independent variable. From the "Beta" coefficient, the percentage of contribution to landslide area can be derived (see column "contribution". *Slopes* and *Distance from active fault* were both explaining most of the variance (about 35% each), the next one *Distance from river* explaining much less (about 12%), *NDVI* and *distance from trail* explain about 9% each.

Except for trails, all the signs are according to the common sense (the steeper the slope, the larger the landslides, the smaller the distances to active fault, river, the larger the landslides and the smaller the NDVI the larger the landslides). The selection shows a high degree of confidence (p-level much smaller than 0.05; the highest p-level (thus worse) is associated with distance from trail with a value of 0.002, hence 99.8% (1–0.002) probability that the selection of distance from trail is not due to random process. It is highlighted has being significant, but calls for improvements.

The percentage of variance explained ($R^2$) by the general model reaches 61.3% (Pearson=0.78), bearing in mind that

**Table 2.** Results from multiple regression analysis.

| Variables | Beta | Coefficient | p-level | Contribution |
|---|---|---|---|---|
| Intercept | | 7.913 | 0.000000 | |
| D_river_minLn | −0.180 | −0.125 | 0.000013 | 12.13% |
| D_trail_avLn | 0.128 | 0.206 | 0.001859 | 8.63% |
| D_fault_minLn | −0.520 | −0.246 | 0.000000 | 35.04% |
| NDVIt_avLn | −0.137 | −0.878 | 0.001007 | 9.23% |
| Slope_max$^2$Ln | 0.519 | 0.263 | 0.000000 | 34.97% |

$r$=0.78, $R^2$=0.613, Adjusted $R^2$=0.601, $N$=246, outliers = 16
Where:
Intercept Intercept value of the regression line
D_river_minLn Logarithm natural of minimum distance between landslides and river
D_trail_avLn Logarithm natural of minimum distance between landslides and trail
D_fault_minLn Logarithm natural of minimum distance between landslides and active fault
NDVI$_t$_avLn Logarithm natural of average transformed value of NDVI
Slope_max$^2$Ln Logarithm natural of maximum slopes as detected from ASTER

the variables have been transformed and expressed here in logarithms. The total number of landslides considered for the analysis is 280, 262 had valid information for the variables studied. 246 cases were considered, excluding 16 outliers (+2.0 sigma).

The correlation matrix (Table 3) between the explicative variables shows no auto-correlations. Maximum slopes and distance from active fault area already strongly correlated with landslides area (0.505 and −0.533, respectively).

The scatter plot featuring predicted versus observed values (Fig. 5) shows a relatively good fit; however it seems that two groups can be identified with a gap between large landslides areas and smaller one. Distance from trails has the opposite sign as expected. This call for an improvement of the model and testing separated processes.

### 4.2 Differencing landslides close and away from rivers

Numerous small landslides were observed along rivers (or close to rivers). Larger landslides seem to be following different rules. Given that distance from river might not be relevant for landslides away from rivers, a hypothesis was made that landslides might be modelled using differentiated regressions. Three different categories of landslides were made: those touching rivers, those close to rivers (but not touching) and those away from rivers (at a distance greater than 100 m), this process was carried out using both Boolean conditions and a (GIS) buffer of 100 m around rivers to intersect with surrounding landslides.

#### 4.2.1 Landslides away from river (minimum distance >100 m)

The number of cases away from rivers is 98 (once excluding 3 outliers, with sigma>2.0). Percentage of variance explained is 54.0% (Pearson=0.73).

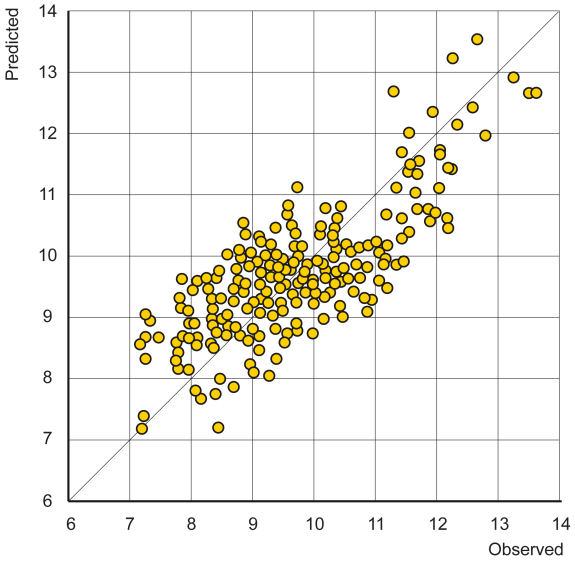



**Fig. 5.** Predicted size of landslides versus observed, scale in Ln[m$^2$].

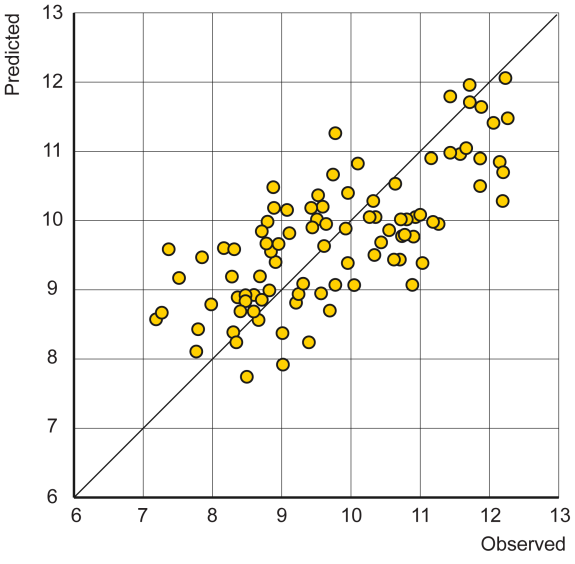



**Fig. 6.** Predicted size of landslides versus observed: Landslides away from rivers (>100 m), scale in Ln(m$^2$).

The variables selected are as follows:

1. Slope (square of Ln maximum slope), positive sign.
2. Minimum distance from active fault (Ln), negative sign.
3. NDVI (Ln of transformed values), negative sign.

The level of significance is very high (all p-level much smaller than 0.05), no auto-correlation suspected, the signs are according to what was expected (see Table 4). According to the Beta coefficient the susceptibility factor most

**Table 3.** Correlation matrix.

| $N$=246 | D_river_minLn | D_trail_avLn | D_fault_minLn | NDVIt_avLn | Slope_max$^2$Ln |
|---|---|---|---|---|---|
| D_trail_avLn | 0.045 | | | | |
| D_fault_minLn | 0.105 | 0.035 | | | |
| NDVIt_avLn | 0.081 | 0.143 | 0.052 | | |
| Slope_max$^2$Ln | −0.020 | 0.100 | 0.018 | 0.153 | |
| Ls_areaLn | −0.251 | 0.134 | −0.533 | −0.081 | 0.505 |

Where: Other factors as above mentioned and Ls_areaLn: Logarithm natural of landslides area

**Table 4.** Results from multiple regression analysis for landslides away from rivers.

| Variables | Beta | Coefficient | p-level | Contribution |
|---|---|---|---|---|
| Intercept | | 7.029 | 0.000000 | |
| D_fault_minLn | −0.511 | −0.219 | 0.000000 | 38.9% |
| NDVIt_avLn | −0.256 | −1.540 | 0.000652 | 19.9% |
| Slope_max$^2$Ln | 0.541 | 0.373 | 0.000000 | 41.2% |

$r$=0.732, $R^2$=0.535, Adj. $R^2$=0.520, $N$=98, outliers = 3

influencing the landslide area is slope (41.2%), followed by distance from active fault (38.9%) and then NDVI (19.9%).

The distance from trails and the distance from rivers are no longer selected by the model, which makes sense for rivers, given the filters applied.

#### 4.2.2 Landslides close to river (at a distance from river <100 m)

The number of valid cases was 178, with 169 considered and 9 outliers. The percentage of variance explained increased to 64.3% (Pearson=0.80).

The variables selected were:

1. Minimum distance from active fault (Ln), negative sign.
2. Maximum slopes (Lnˆ2), positive sign.
3. Distance from trails, negative sign.

The level of significance is very high (all p-level much smaller than 0.05), no auto-correlation suspected (see Table 5). First contributor is slope (42.0%), followed by distance from active fault (36.8%) and then distance from trail (21.2%).

The parameters distance from trail is selected and this time with a negative sign, thus according to what was expected). NDVI is no longer considered by the model.

#### 4.2.3 Landslides touching river

For this category of landslides (touching river), only the slope (52.3% of contribution) and the distance from active fault

**Table 5.** Results from multiple regression analysis for landslides close (but not touching rivers).

| Variables | Beta | Coefficient | p-level | Contribution |
|---|---|---|---|---|
| Intercept | | 9.343 | 0.000000 | |
| D_trail_avLn | −0.249 | −0.172 | 0.000003 | 21.2% |
| D_fault_minLn | −0.433 | −0.230 | 0.000000 | 36.8% |
| Slope_maxLn$^2$ | 0.493 | 0.245 | 0.000000 | 42.0% |

$R$=0.802 $R^2$=0.643, Adj. $R^2$=0.636, $N$=169, outliers = 9
Where:
Dist_trail_av_Ln: Logarithm natural of minimum distance between landslides and trail
Dist_fault_min_Ln: Logarithm natural of minimum distance between landslides and active fault
Slope_max_Ln$^2$: Square of logarithm natural of maximum slope as recorded by ASTER

**Table 6.** Results from multiple regression analysis for landslides touching rivers.

| Variables | Beta | Coefficient | p-level | Contribution |
|---|---|---|---|---|
| Intercept | | 8.630 | 0.000000 | |
| D_fault_minLn | −0.490 | −0.211 | 0.000000 | 47.7% |
| Slope_maxLn$^2$ | 0.537 | 0.200 | 0.000000 | 52.3% |

$r$=0.730, $R^2$=0.532, Adj. $R^2$=0.520, $N$=82, outliers = 5

(47.7%) is relevant according to the statistical model. The explanation value is 53% (Pearson=0.73). The level of significance is very high (all p-level much smaller than 0.05), no auto-correlation suspected (see Table 6). The distance from trails/roads and NDVI is no longer considered.

## 5 Cartographical results

***Spatial model***

The model was improved by looking at different distance from rivers, although following the theory, three sets of equation should be collected (touching rivers, <100 m from rivers and away from them), given that at such resolution only three pixels account for a distance of 90 m, a simplified model

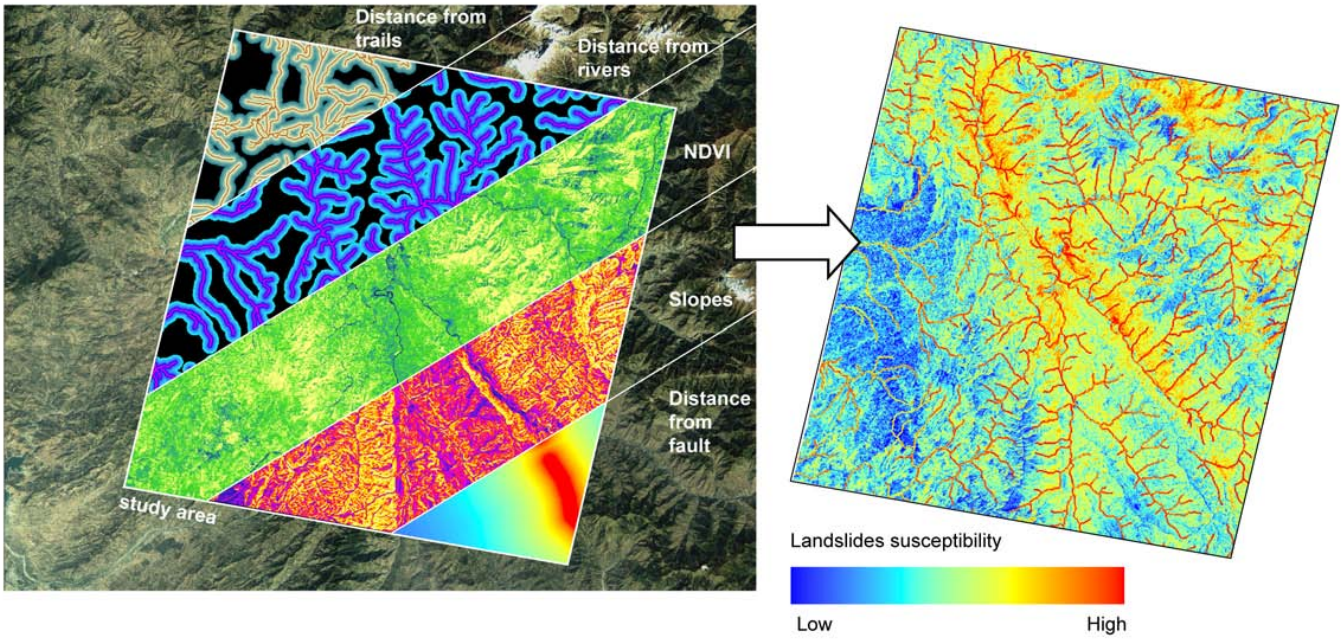



**Fig. 7.** Map of landslides susceptibility as modelled.

based on results provided in Table 4 (away from rivers) and Table 6 (touching rivers) provides the following equations:
If Distance river $>100$, then

$$\text{Ls\_areaLn} = -0.219 \cdot \text{D\_fault\_Ln} - 1.54 \cdot \text{NDVIt\_avLn} + 0.373 \cdot \text{Slope}^2\text{Ln} + 7.029 \quad (4)$$

Else

$$\text{Ls\_areaLn} = -0.211 \cdot \text{D\_fault\_Ln} + 0.2 \cdot \text{Slope\_Ln}^2 + 8.63 \quad (5)$$

Buffers of 100 m were generated on both sides of rivers for discriminating between first and the second case. For each pixel (of 30 m $\times$ 30 m) the distance from active fault, the slope and the NDVI were computed (with the relevant transformations and corresponding weights and exponents). This allows the creation of a susceptibility map as shown in Fig. 7.

## 6 Discussion

The five factors identified as having an influence on landslide area fell in three categories, namely: slope, distance from linear features (active fault; trails or river) and vegetation cover density.

### 6.1 Slopes

Not surprisingly, the main parameter for landslides occurrence is slope. The positive sign associated to the parameters is indicating that the steeper the slope, the higher the landslide susceptibility, which is perfectly logical. Tests were made using the SRTM DEM (at 90 m resolution), however the variance explained dropped significantly, hence extrapolation to larger areas was abandoned. The role of slopes being so predominant, a higher resolution is needed. Higher resolution DEM might improve the model. It may allow computation of concave and convex slopes as this was proven to be an improved explanatory factor in other study (Sato et al., 2007).

### 6.2 Distance from features

The negative sign before the coefficient means that the closer from the active fault, river or trail/road, the larger the value of landslide area. This is consistent with the theory, the maximum energy being closer to the epicentres. Similar links with negative coefficients were found for rivers and trails, although the influences of these features are much more local and the range of influence on instability is not known. Trails could be replaced by roads (although there are fewer of them and thus the variance explained was slightly lower). If distances from trails and roads were highlighted in the general model, it was only selected for landslides near but not touching rivers. Indeed, most of the roads are along main rivers, so while selecting areas touching rivers, it spatially already includes these roads, hence the selection wasn't pertinent anymore. However, their selection in the general model is an indicator that these human infrastructures should be studied with attention.

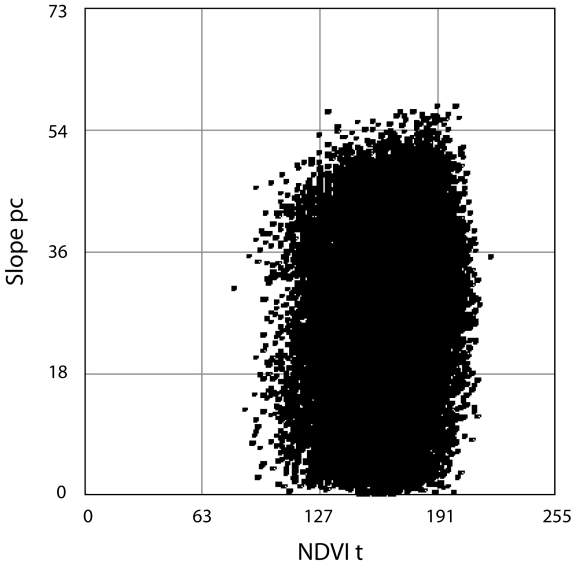



**Fig. 8.** Scatterplot of vegetation density (NDVI t) versus slope (slope pc).

### 6.3 Vegetation density (NDVI)

The use of NDVI proxy appears to be efficient in linking contextual vegetation density with susceptibility of landslides. Although the part of variance explained was not really high, the confidence in the selection (low p-level) clearly indicates that the presence of denser vegetation has a mitigation effect on landslide susceptibility. The close up using the verification model provides some striking examples for the entire map. When displayed to local decision makers, it had a large impact and will hopefully lead to improve environment management. Socio-economical studies on why forests have been cleared should now be conducted in order to see what solutions could be envisaged to reverse the trend. By running the model without the NDVI mitigation effect, the total susceptibility in the study rose by 15.13%. This delineates that vegetation cover is a significant component of risk reduction.

### 6.4 Verification of causality

Correlations between set of parameters do not necessarily imply causality. Providing this link is the usual weakness of statistical regression analysis.

One might argue that because areas on steep slope might be less covered by vegetation. Thus observing less dense vegetation might be an indirect way of looking at slope. To test if slopes and vegetation are correlated, a simple scatterplot of the two susceptibility factors can be computed using the function "scattergram" in the module GRID from ArcINFO workstation (see Fig. 8). This scatterplot shows no correlation between the two variables. Hence, there is no indication that vegetation density is function of slope.

Another test can be run by looking for similar areas, where only one parameter is changing. For example, to test whether vegetation density has mitigation effect on landslide susceptibility, a model can be run without the NDVI component, thus normalising all the features but the vegetation. By plotting forest cover (pre-landslide) and landslides as recorded, landslides should be more frequently observed in areas with similar landslide susceptibility in areas with lower vegetation density as compared with areas covered with dense vegetation.

The Fig. 9 shows three different models, clearly looking at slopes and distance from active fault does not explain for all the landslides on river shores. Thus model "b" is already adding valuable information on susceptibility by adding distance from rivers; however, model "a" and "b" are presenting an area of landslide susceptibility that is much larger than observed impacts. Adding vegetation cover as parameters (model "c"), drastically reduces the area at risk and provides a much better match between observed landslides and the model.

In Fig. 10, the upper panel close up shows a model without a vegetation density component. In a way it shows a theoretical situation if all vegetation were removed. The susceptibility seems to be spread in most of the area, whereas observed landslides areas are featured in bold black. On the right-hand close up, the model includes the vegetation mitigation effect. The areas susceptible to landslides are much more concentrated and fit better with the observed impacts.

These results are encouraging. Although global datasets cannot (and should not) be used for local landuse planning, this method has some great potential as an advocacy tool or to determine where more detailed data should be acquired, allowing saving on – usually – high input costs.

## 7 Conclusion

The study confirmed the hypothesis that landslide occurrence is higher on steep slopes, close to rivers, trails, active fault and that vegetation cover seems to act as stabiliser in this region. The results from this research show that adding the mitigation effect of vegetation cover in the model drastically improves the model as compared with the observed landslide areas. This seems to indicate that, in this region, vegetation seems to play a significant role in decreasing landslide susceptibility. It shows that global available datasets can be used to select layers of information to be gathered and narrow the areas where deeper analysis should be conducted.

Given that this study uses of low costs data and free available data, the resolution of such data (at best 30 m) is not appropriate for local land planning. But given the price of high resolution DEM and satellite sensors (e.g. IKONOS, Quickbird, GeoEye and alike), such study is very useful to identify areas where detailed data should be collected.

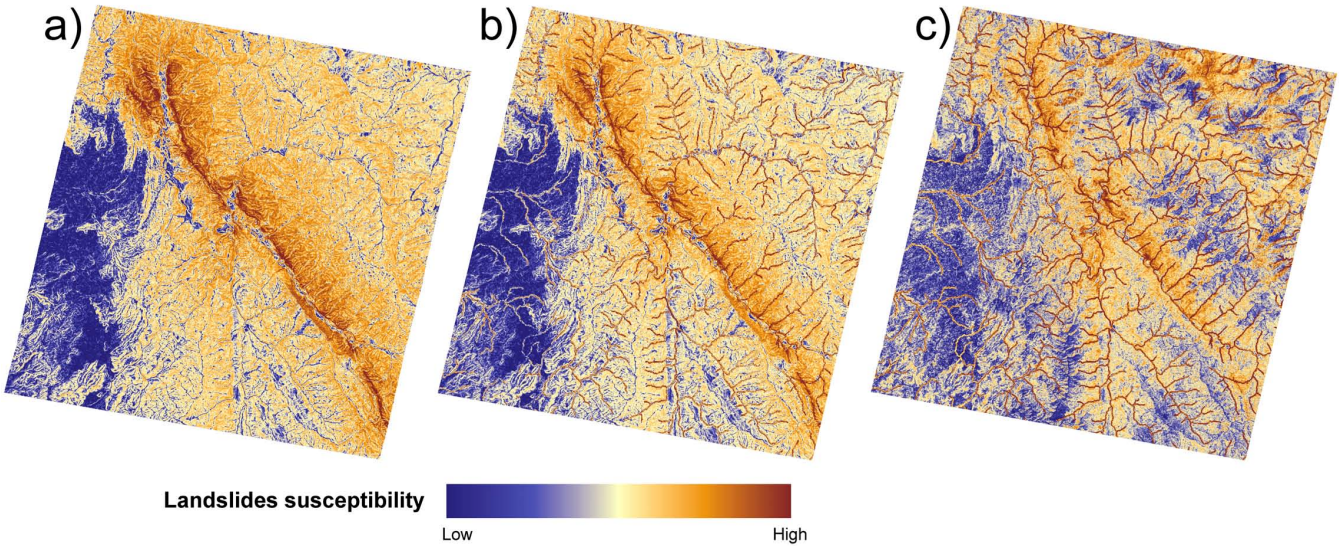



**Fig. 9.** Different models of landslides susceptibility (**a** including slope and active faults, **b** adding distance from rivers, **c** adding presence of vegetation).

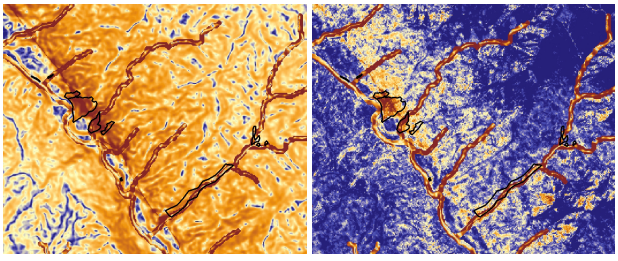


**Fig. 10.** Close up without vegetation mitigation effect (model "b"), left and with vegetation mitigation effect (model "C"), right.

The applied method proved to be successful in providing statistical links between contextual parameters, vegetation density and size of landslides. The extrapolation of the model to the area provides a general quick look of the areas of potential high risk of landslide occurrence. The spatial precision of the DEM was the main reason for the success of this analysis; hence the extrapolation using lower spatial distribution (such as SRTM) was not possible without significant decrease in accuracy. New studies from University of Lausanne are now been implemented with very high resolution satellite data (Quickbird, 0.6 m resolution) to zoom in the Neelum Valley and bring more in depth analysis.

The role of small tracks in inducing risk of soil instability was highlighted. This parameter (along with vegetation cover, slope, distance from rivers and proximity from active fault) should be considered for landuse planning. Whereas for timber exploitation, pasture or access, such high risk areas cannot be managed without improved information in landslide occurrence. The larger gap to be bridged is the lack of data on soil types, which would most probably, increase the level of accuracy. It would, however, introduce a difficulty, as type of soil is not a continuous variable. It would request either to run several models (one for each type of soil), or to use expert judgment on the susceptibility of each soil type.

The methodology is quite simple, based on global datasets and/or easily accessible data. This was applied in the case of landslides triggered by earthquakes, but should now be tested on other areas with landslides triggered by heavy precipitation in deforested areas. It can be adapted to allow gathering evidence in different areas of the planet where heavy deforestation has been recorded, thus hopefully address the message that healthy ecosystems can help reduce disaster risk.

This model was presented to local authorities. The simple message it carries was well-received. It should be emphasized that reforestation cannot suppress all the susceptibility factors (such as slope, rivers and distance from active fault). Landuse planning in areas where such magnitude of earthquakes can take place is not an easy task. Planting trees and increasing vegetation density cannot be the only solution and should not be done blindly. Methodologies for reforesting using local useful species is one of the recommendations, however, some cities in the region are located straight on the active fault (or potentially active fault), on Quaternary rocks, where recorded vertical deformation of land (uplift) was between 1 and 6 m. Clearly in these locations, forest cover will not be of sufficient protection. If delocalisation of population is to be carried out, it is hoped that the new settlements will be chosen with care not to recreate risk and that it will be done with more consideration for environmental features.

**Table A1.** Raw data sources.

| Features | Raw data providers | URL link (if relevant) |
|---|---|---|
| Post event landslides | Service Régional de traitement d'image et de télédétection (SERTIT) | http://sertit.u-strasbg.fr/documents/asie/asia_en.html |
| Post event landslides | National Engineering Services of Pakistan (NESPAK), both obtained through the Humanitarian Information Centre For Pakistan (HIC) | |
| active fault | Manually digitalised from a map published by Nakata et al. (1991) | |
| Roads/trails | United Nations Joint Logistics Centre (UNJLC) | |
| Rivers | Data Repository of the Geographic Information Support Team (GIST) | https://gist.itos.uga.edu/index.asp?body=repository |
| Epicentres | Advanced National Seismic System (ANSS) | http://quake.geo.berkeley.edu/anss/ |
| Vector country borders | NIMA Vmap level 0, UN Cartographic Section | www.mapability.com/info/vmap0_intro.html |
| Digital Elevation Model (DEM), 90 m | Consortium for Spatial Information (CGIAR-CSI) SRTM version 3. | http://srtm.csi.cgiar.org |
| DEM, 30 m | From ASTER data purchased at USGS | http://lpdaac.usgs.gov/aster/ast14dem.asp |
| Satellite imagery | Landsat 7 ETM+, path/row: 150/36, 7 Oct 2001, original data sources: Landsat ETM from 7 Oct 2001, path/row 150/36 obtained through the Global Observatory for Ecosystem Services, Michigan State University | http://landsat.org |

**Table B1.** Set of independent variables extracted.

| ID | Variable name | Description |
|---|---|---|
| V1 | Geol_class | Classified lithology |
| V2 | TYPES | Landslides shape (horizontal, vertical and large) |
| V3 | EPI55_Ln | Log Natural (Ln) of Distance Epicentre>5.5 and centre of landslides |
| V4 | EPI7_Ln | Ln of Distance Epicentre>7 and centre of landslides |
| V5 | EPI8_Ln | Ln of distance between epicentre = 7.6 and centre of landslides |
| V6 | DFPC2001 | Dense forests % in 2001 (transformed and Ln) |
| V7 | AFPC2001 | All forest % in 2001 (transformed and Ln) |
| V8 | DFPC1992 | Dense forests % in 1992 (transformed and Ln) |
| V9 | AFPC1992 | All forest % in 1992 (transformed and Ln) |
| V10 | DFPC1979 | Dense forests % in 1979 (transformed and Ln) |
| V11 | AFPC1979 | All forest % in 1979 (transformed and Ln) |
| V12 | D_AF_N | Deforestation 2001-1979 all forest, normalised et Ln |
| V13 | D_DF_N | Deforestation 2001-1979 dense forest, normalised and Ln |
| V14 | DEM_D_Ln | Difference in elevation en Ln |
| V15 | SL_MAXPC | Slope max in % |
| V16 | SLMEANPC | Slope min in % |
| V17 | SLOPESTD | Slope standard dev. |
| V18 | SL_MEDPC | Slope median % |
| V19 | FAU_E_Ln | Ln distance fault (edge) |
| V20 | FAU_C_Ln | Ln distance faille (centre) |
| V21 | RIV_E_Ln | Ln distance to river (edge) |
| V22 | RIV_C_Ln | Ln distance to river (centre) |
| V23 | ROAD_ELn | Ln distance route (edge) |
| V24 | ROAD_CLn | Ln distance route (centre) |
| V25 | TRAILELn | Ln distance trail (edge) |
| V26 | TRAILCLn | Ln distance trail (centre) |
| V27 | LN__AS_D | Ln DEM Aster Difference DEM |
| V28 | AS_SLMAX | Maximum slope max from ASTER DEM |
| V29 | AS_SLAV | Slope average from ASTER DEM |
| V30 | AS_STD | Slope standard deviation from ASTER DEM |
| V31 | minR_Tc | Distance minimum between road and trail (centre) |
| V32 | minR_Te | Distance minimum between road and trail (edge) |
| V33 | MinNDVIt_Ln | Minimum NDVI transformed and Ln |
| V34 | MaxNDVIt_Ln | Maximum NDVI transformed and Ln |
| V35 | AVNDVIt_Ln | Average NDVI transformed and Ln |
| V36 | Slopemax$^2$_Ln | Ln of the square of the maximum slope |
| V37 | Slopemax_Ln$^2$ | Square of the Ln of maximum slope |

**Table C1.** Example of a selection of non-correlated features using correlation matrix.

| Variables | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| V1 | – | | | | | | | | | | |
| V2 | **−0.611** | – | | | | | | | | | |
| V3 | 0.030 | 0.064 | – | | | | | | | | |
| V3 | 0.151 | −0.021 | 0.681 | – | | | | | | | |
| V5 | −0.157 | 0.150 | **0.511** | 0.408 | – | | | | | | |
| V6 | −0.306 | 0.263 | 0.397 | 0.291 | **0.661** | – | | | | | |
| V7 | −0.190 | 0.160 | −0.054 | 0.010 | **0.557** | 0.379 | – | | | | |
| V8 | **−0.574** | **0.509** | −0.018 | −0.059 | 0.132 | 0.194 | 0.127 | – | | | |
| V9 | −0.027 | −0.047 | 0.073 | 0.026 | 0.086 | 0.137 | 0.039 | −0.110 | – | | |
| V10 | 0.246 | −0.180 | 0.164 | 0.240 | 0.115 | −0.034 | −0.002 | −0.191 | 0.006 | – | |
| V11 | −0.084 | 0.070 | 0.013 | 0.000 | 0.069 | 0.244 | 0.064 | 0.038 | −0.030 | −0.035 | – |
| V12 | 0.322 | −0.329 | 0.023 | 0.069 | −0.111 | 0.004 | −0.078 | −0.358 | −0.218 | 0.085 | 0.546 |

In bold the variables that cannot be placed in the same group of analysis.

**Table D1.** Descriptions of acronyms used in this paper.

| Acronyms | Description |
|---|---|
| ANSS | Advanced National Seismic System |
| ASTER | Advanced Spaceborne Thermal Emission and Reflection Radiometer |
| CGIAR-CSI | Consortium for Spatial Information |
| CNSS | Northern California Earthquake Data Center |
| CRED | Centre for Research on Epidemiology of Disasters |
| DEM | Digital Elevation Model |
| GIS | Geographical Information System |
| GIST | Geographic Information Support Team |
| HFA | Hyogo Framework for Action |
| HIC | Humanitarian Information Centre For Pakistan |
| IFRC | International Federation of Red Cross and Red Croissant Societies |
| ISDR | International Strategy for Disaster Reduction |
| IUCN | International Union for Conservation of Nature |
| IUED | Institut Universitaire d'Etude du Développement |
| NDVI | Normalised Difference Vegetation Index |
| NESPAK | National Engineering Services of Pakistan |
| NIR | Near Infra-Red |
| SAR | Synthetic Aperture Radar |
| SERTIT | Service Régional de traitement d'image et de télédétection |
| SRTM | Shuttle Radar Topography Mission |
| UNEP | United Nations Environment Programme |
| UNEP/GRID | United Nation Environment Programme, Global Resource Information Database |
| UNJLC | United Nations Joint Logistics Centre |
| UTM | Universal Transverse Mercator |

**Table E1.** List of software used for the analysis.

| Tasks | Software |
|---|---|
| GIS | ArcGIS 9.2; ArcINFO workstation 9.2 |
| Remote sensing | ERDAS IMAGINE 8.4 |
| Statistics | Statistica 8, Minitab 15.1.30.0. |
| Cartography, graphs | Adobe Illustrators CS3, Photoshop CS3 |

## Appendix F

### Statistical concepts

For readers who are not familiar with some of the statistical concepts used in this paper, here is a small summary. This section is adapted from the on-line help of StatSoft Electronic Statistics Textbooks (http://www.statsoft.com/textbook/statistics-glossary/).

### F1 Multiple regression analysis

When addressing the potential link between one variable (e.g. slope) and a dependant variable (e.g. landslide areas) simple scatter plots provide useful information (see Fig. F1).

Some variables can directly be linked with landslide areas (e.g. slope), however, one variable is usually not enough to model the behaviour of a dependent variable. Variables can have a multiplicative effect when associated one to another.

To understand what the best combinations of susceptibility factors are and how they contribute to landslide area, a *multiple regression analysis* can be made. Such statistical process aims to highlight the relationship between a dependent variable (e.g. landslide area) and several independent variables (potential susceptibility factors, e.g. slopes, distance from active fault, presence of vegetation,...).

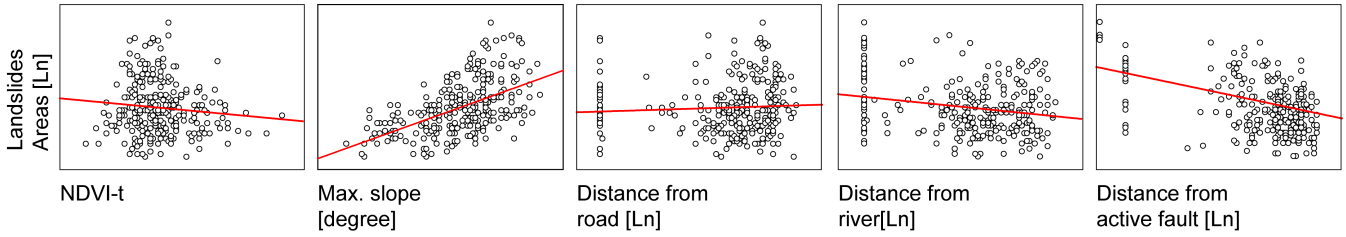



**Fig. F1.** Mattrix plots of correlations.

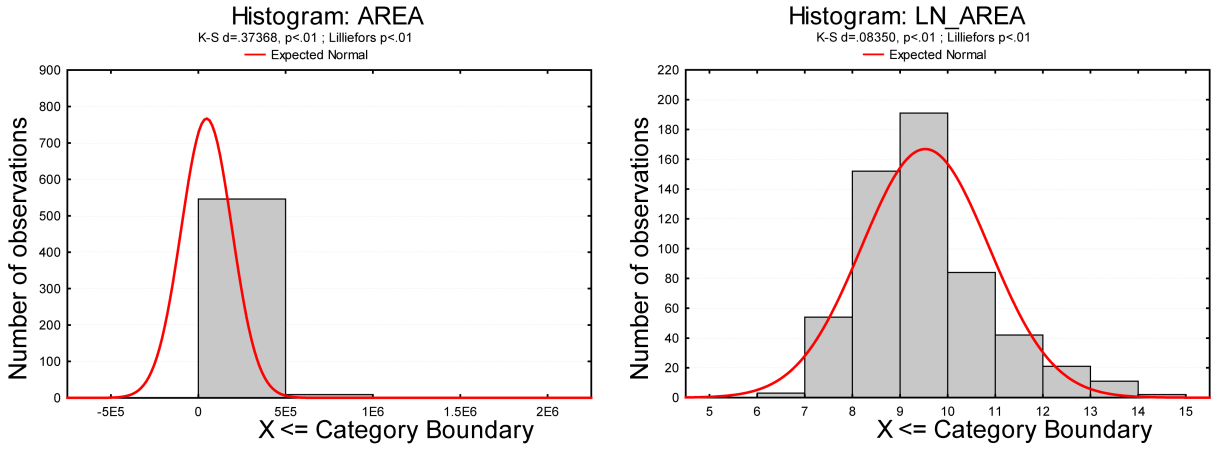



**Fig. F2.** Normality.

The aim is to obtain an equation such as:

$$\mathrm{LA} = \alpha X_1 + \beta \cdot X_2 + \ldots + \theta \cdot X_n + I$$

Where LA = Landslide area; $X_1$ = first susceptibility factor (e.g. slope); $X_2$ = second suscpeptibility factor (e.g. distance from active fault); $X_n$ = last susceptibility factors (e.g. vegetation density); $\alpha$, $\beta$ and $\theta$ = weights which multiplies the factors.

The purpose is double, first it allows a better understanding of the underlying processes leading to landslides, secondly, it provides the weights associated with each susceptibility factors, allowing creating maps of landslides susceptibility. The main limitation being that although it shows potential link, it cannot ensure causality (see causality below).

## F2 Pearson coefficient, *r*

The independent variables in the model should not have influence between them. To produce group of independent variables a correlation matrix is computed and variables that are too correlated should not be tested in the same hypothesis. Thus group of uncorrelated variables should be created (see Table C1). The $r$ is the pearson coefficient, it is computed as follows:

$$r = \frac{\sum(x-\overline{x}) \cdot \sum(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \cdot \sum(y-\overline{y})^2}}$$

where $\overline{x}$ is the average for a observed dependant variable; $\overline{y}$ is the average for the modelled variable.

In this study, two independent variables could be placed in the same group if $|r| < 0.5$.

## F3 $R^2$ and adjusted $R^2$

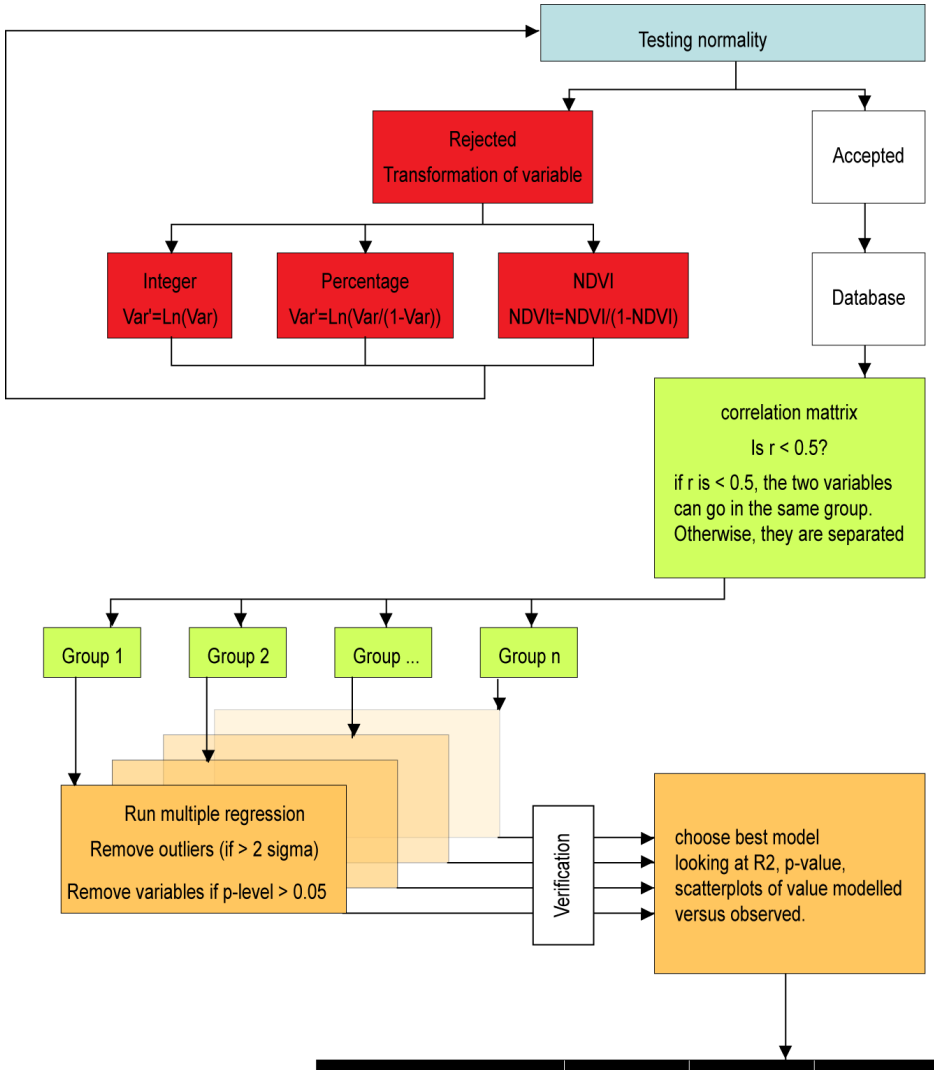
$R^2$ is the square of $r$, it provides an indication of the percentage of variance explained.

Adjusted $R^2$ is a modification of $R^2$ that adjusts for the number of explanatory terms in a model. Meaning that by adding more explanatory variables you might increase the $R^2$, but it could also be by chance (over fitting models). The Adjusted $R^2$ is particularly useful in the selection of potential susceptibility factors as it takes into account the number of explanatory variables and only increase if the added explanatory variable explains more than as a result of a coincidence.

$$\mathrm{adj}.R^2 = 1 - (1-R^2) \cdot \frac{n-1}{n-p-1}$$

Where $n$ is the sample size, $p$ is the number of independent variables in the model.

In general terms, the more explanatory variables you have the less the $R^2$, because by introducing more independent variables, you increase the risk that the results is obtained by random. This is often called "overfitting".




| Variables | Beta | B | p-level | Contribution |
|---|---|---|---|---|
| Intercept | | 7.03 | 0.000000 | |
| D_fault_minLn | -0.51 | -0.22 | 0.000000 | 38.93 % |
| NDVIt_avLN | -0.26 | -1.54 | 0.000652 | 19.85 % |
| Slope_max2Ln | 0.54 | 0.37 | 0.000000 | 41.22 % |

Model
LN(Landslide area)= -0.22D_fault_minLn -1.54NDVIt_avLN + 0.37Slope_max2Ln +7.03

**Fig. F3.** Simplified version of the statistical process used in this study.

## F4 Normal distribution

The variables should follow a normal distribution. This can be done by looking at histograms and by applying some statistical normality tests (e.g. Normal expected frequencies, Kolmogorov-Smirnov & Lilliefors test for normality, Shapiro-Wilk's W test.). If a variable do not follow a normal distribution, it needs to be transformed so that it does (e.g. computing the Ln, or using transformation formula). The figure below shows the distribution of landslide areas. Taking the Ln greatly improve the normality.

Other functions can be used as specified in the article.

## F5 Outliers

Outliers are cases that do not follow the general assumption. In the real environment, it is difficult to take all the parameters reflecting the complexity of the situations. Some isolated

cases, might have specific settings, and they don't follow the general trends. These outliers are easy to identify as they are distant to the rest of the data. They should be identified and removed so that the general rule can be better identified. However, an analysis of these outliers should be performed to ensure that they don't follow another rule. If this is the case, then the dataset might need to be split so that two (or more) models can be generated. For example, we differentiated landslides close to rivers and landslides away from rivers as it seems that these two groups follow different rules.

### F6 Causality

Correlation between two variables (e.g. $A$ & $B$) do not imply that variation of A is the origin of the variation of $B$.

If multiple regression can shows potential link, it cannot ensure causality.

What we aim to do is to say that factor $A$ (e.g. vegetation density) influence $B$ (landslide area). Now having a correlation between factors $A$ & $B$ can have several origins:

$A$ is indeed having an influence on $B$ or

$B$ is influencing $A$ or

$C$ is influencing $A$ & $B$.

The p-level provide good insight on the probability that the link is due to a coincidence. However, addressing causality is always the main challenge (see discussion under "verification of causality").

In the final table (in blue) provided, the coefficient "Beta" provides information on the relative contribution of each susceptibility factor to landslides area. Slopes is the main one (0.54) and has a positive sign, hence the steeper the slope the bigger the landslide area; distance from active fault is the second contributor ($-0.51$) and has a negative sign, hence the further away from active fault line, the smaller the landslide areas and finally NDVI has a negative sign, hence denser vegetation (high NDVI) relates to smaller landslide areas.

One can even compute the percentage of contribution from each factor. For this we need to sum the absolute value of the Beta coefficient (0.51+0.26+0.54=1.31), then the ratio of each absolute value of the Beta coefficient provides the percentage of contribution for each variable: slope contributes for 41.22% (0.54/1.31), distance from active fault for 38.93% and NDVI for 19.85%. These percentage are provided in the last column (contribution).

The coefficient $B$ (third column) provide the weights which can be used to model the landslide area (see equation of the model below the table).

The p-level indicates the probability that the variable was selected by coincidence. For example a p-level of 0.05 indicates that there is 5% of chance that the selected variable is a "fluke". This level is customarily treated as a "border-line acceptable" error level. So the lowest the p-level the highest the confidence in the selection. In this study the highest p-level was 0.0018, meaning the all the selected variables have less than 0.18% chance of being selected by coincidence.

*Acknowledgements.* Thank you to Raphaël Klaus for his valuable contribution on digitalisation of active faults and improvement of the recorded landslides dataset, to Karen Sudmeier-Rieux (University of Lausanne/IUCN) for successful collaboration with us on this subject and her useful editing suggestions, finally to Michel Jaboyedoff (University of Lausanne) for his review.

Edited by: T. Glade
Reviewed by: M. W. Stirling and another anonymous referee

## References

ANSS: Advanced National Seismic System composite catalogue from the Northern California Earthquake Data Center (NCEDC), http://quake.geo.berkeley.edu/anss/, 2007

Avouac, J. P., Ayoub, F., Leprince, S., Konca, O., and Helmberger, D. V.: The 2005, Mw 7.6 Kashmir earthquake: Sub-pixel correlation of ASTER images and seismic waveforms analysis, Earth Planet. Sci. Lett., 249, 514–528, 2006.

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, 2005.

Brenning, A.: Spatial prediction models for landslide hazards: review, comparison and evaluation, Nat. Hazards Earth Syst. Sci., 5, 853–862, 2005, http://www.nat-hazards-earth-syst-sci.net/5/853/2005/.

Chatenoux, B. and Peduzzi, P.: Impacts from the 2004 Indian Ocean tsunami: analysing the potential protecting role of environmental features, Natural Hazards, 40, 289–304, 2007.

Coe, J. A., Michael, J. A., Crovelli, R. A., Savage, W. Z., Laprade, W. T., and Nashem, W. D.: Probabilistic assessment of precipitation-triggered landslides using historical records of landslide occurrence, Seattle, Washington, Environ. Eng. Geosci., 10, 103–122, 2004.

Dunning, S. A., Mitchell, W. A., Rosser, N. J., and Petley, D. N.: The Hattian Bala rock avalanche and associated landslides triggered by the Kashmir Earthquake of 8 October 2005, Eng. Geology, 93, 130–144, 2007.

EM-DAT: Emergency events Database, Centre for Research on the Epidemiology of Disasters (CRED), Université Catholique de Louvain, http://www.emdat.be, 2007.

FAO, U. N.: Global Forest Resources Assessment 2005. Progress Towards Sustainable Forest Management, Forestry Paper, 147, 737–750, 2006.

Gorsevski, P. V., Foltz, R. B., Gessler, P. E., and Cundy, T. W.: Statistical modeling of landslide hazard using GIS, Links, 2001.

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.

IFRC: Mangrove planting saves lives in Vietnam, Press release issued by the International Federation of the Red Cross Red Crescent Societies in June 2002, the text is an excerpt from the IFRC's World Disasters Report 2002 – Focus on reducing risk republished by UNEP in: Environment and Poverty Times No. 3, Special Edition for the World Conference on Disaster Reduction, Kobe, Japan, available at: http://www.grida.no/publications/et/ep3/page/2610.aspx, 18–22 January 2005.

IUCN: Rapid Environmental Appraisal of Developments in and Around Murree Hills, 16 pp., available at: http://www.waterinfo.net.pk/pdf/new_murre_report.pdf, 2005.

Nadim, F., Kjekstad, O., Peduzzi, P., Herold, C., and Jaedicke, C.: Global landslide and avalanche hotspots, Landslides, 3, 159–173, 2006.

Nakata, T., Tsutsumi, H., Khan, S. H., and Lawrence, R. D.: Active faults of Pakistan: map sheets and inventories, Research Center for Regional Geography, Hiroshima University, Hiroshima, 1991.

Peduzzi, P., Dao, H., Herold, C., and Mouton, F.: Global risk and vulnerability index trends per year (GRAVITY), phase II: development, analysis and results, on-line technical document, UNDP, UNEP, http://www.grid.unep.ch/product/publication/download/ew_gravity2.pdf, 2002.

Petley, D. N., Dunning, S. A., Rosser, N. J., and Kausar, A. B.: Incipient landslides in the Jhelum Valley, Pakistan following the 8th October 2005 earthquake, Disaster Mitigation of Debris Flows, Slope Failures and Landslides, Frontiers of Science Series, 47, 47–56, 2006.

Popescu, M. E.: Landslide causal factors and landslide remediatial options, on-line technical document, Illinois Institute of Technology, Chicago, USA, p. 1–21, http://www.geoengineer.org/Lanslides-Popescu.pdf, 2002.

Sato, H. P., Hasegawa, H., Fujiwara, S., Tobita, M., Koarai, M., Une, H., and Iwahashi, J.: Interpretation of landslide distribution triggered by the 2005 Northern Pakistan earthquake using SPOT 5 imagery, Landslides, 4, 113–122, 2007.

SERTIT: Service Régional de traitement d'image et de télédétection (SERTIT), post landslides detection http://sertit.u-strasbg.fr/SITE_RMS/2005/13_pakistan_2005/pakistan_2005.html, first access date: July 2006, verification date: 17 March 2010.

Sudmeier-Rieux, K., Qureshi, R. A., Peduzzi, P., Jaboyedoff, M. J., Breguet, A., Dubois, J., Jaubert, R., and Cheema, M. A.: An interdisciplinary approach to understanding landslides and risk management: a case study from earthquake-affected Kashmir., Mountain Forum, Mountain GIS e-Conference, 16 pp., 2007.

StatSoft Electronic Statistics Textbooks: http://www.statsoft.com/textbook/statistics-glossary/, access date: 27 February 2010.

Sudmeier-Rieux, K., Jaboyedoff, M., Breguet, A., Dubois, J., Peduzzi, P., Qureshi, R., and Jaubert, R.: Strengthening Decision-Making Tools for Disaster Risk Reduction: An Example of an Integrative Approach from Northern Pakistan, Regions: Laboratories for Adaptation, 74–77 pp., 2008.

Tucker, C. J.: Red and photographic infrared linear combinations for monitoring vegetation., Rem. Sens. Environ., 8, 127–150, 1979.

USGS: Earthquake Hazards Program, http://earthquake.usgs.gov/eqcenter/eqinthenews/2005/usdyae/, 2006.

USGS ASTER DEM: ASTER Digital Elevation Model, http://lpdaac.usgs.gov/aster/ast14dem.asp, access date: August 2006.

UNISDR: Hyogo Framework for Action 2005–2015: Building the Resilience of Nations and Communities to Disasters, World Conference on Disaster Reduction, Kobe, Hyogo, Japan, 25 pp., 2005.

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.