Transactions in GIS, 2006, 10(3): 395–415

Research Article

Spatial Prediction of Landslide Hazard Using Logistic Regression and ROC Analysis

Pece V Gorsevski
Department of Forest Resources
University of Idaho

Paul E Gessler
Department of Forest Resources
University of Idaho

Randy B Foltz
USDA Forest Service
Rocky Mountain Research Station
Moscow, Idaho

William J Elliot
USDA Forest Service
Rocky Mountain Research Station
Moscow, Idaho

Abstract
An empirical modeling of road related and non-road related landslide hazard for a large geographical area using logistic regression in tandem with signal detection theory is presented. This modeling was developed using geographic information system (GIS) and remote sensing data, and was implemented on the Clearwater National Forest in central Idaho. The approach is based on explicit and quantitative environmental correlations between observed landslide occurrences, climate, parent material, and environmental attributes while the receiver operating characteristic (ROC) curves are used as a measure of performance of a predictive rule. The modeling results suggest that development of two independent models for road related and non-road related landslide hazard was necessary because spatial prediction and predictor variables were different for these models. The probabilistic models of landslide potential may be used as a decision support tool in forest planning involving the maintenance, obliteration or development of new forest roads in steep mountainous terrain.

1 Introduction

Landslide hazard is defined as the probability of a landslide at a specific location in space and time (Varnes 1984). Landslides result from a complex mixture of geologic,

Address for correspondence: Pece V Gorsevski, Department of Forest Resources, College of Natural Resources, University of Idaho, Moscow, ID 83844-1133, USA. E-mail: peceg@uidaho.edu

geomorphic, and hydrologic conditions causing damage and disruption to people, organizations, industries and the environment (Wu and Sidle 1995, Glade 1997). Modification of natural conditions by human activities, such as road building and forest harvesting accelerates landsliding (Swanson and Dryness 1975, Chung et al. 1995, Montgomery et al. 1998). Human activities in forestlands affect these conditions at a variety of spatio-temporal scales. In individual stands, such activities influence composition, root strength, biomass surcharge, age, and density of the vegetation. At the watershed scale, such activities can influence climate, evapotranspiration, wind, erosion processes, and drainage of surface and ground water. At regional and global scales, landslides contribute to decreased water quality, loss of fish spawning habitat and organic matter, and debris jams that may break during peak flows, thereby scouring channels and destroying riparian vegetation.

While management of forestlands for multiple use is becoming increasingly important, this adds to the complexity of factors influencing landslide processes and further complicates the modeling of landslide hazard. For instance, in managed forests particular locations and the style of road construction can greatly influence landslide hazard (Montgomery 1994, Montgomery et al. 1998). Landslides initiated from roads are defined here as road related (RR) landslides which occur or initiate within the road right-of-way, while non-road related (NRR) landslides are landslides that occur outside the road right-of-way. Road right-of-way includes the clearing width and the roadway with its elements: cut slope, ditch, shoulder, travel way, and fill slope. Public concerns about landslide hazard have created a specific need for developing quantitative hazard models for studying the interaction of landscape processes that incorporate a complex mixture of factors in both roaded and roadless areas. This may enable forestland management to avoid critical areas or provide information that suggests modified practices in areas prone to landslide occurrence. Thus, sites or locations prone to mass failure can be spatially mapped using physical and statistical quantitative modeling approaches.

In recent years, with the development of GIS tools (Burrough and McDonnell 1998, Malczewski 1999), spatially explicit approaches have been applied in various disciplines such as soil-landscape modeling, and more recently in environmental science or wildlife studies (Carrara 1983; Gessler et al. 1995, 2000; Mladenoff et al. 1995; Chamran et al. 2002). Many of these approaches have used digital elevation model (DEM) data for spatial prediction of various landscape properties (Carrara 1983; Moore et al. 1993; Montgomery and Dietrich 1994; Wu and Sidle 1995; Gessler et al. 1995, 2000). DEM's represent the terrain surface and can be used to generate quantitative topographic attributes (Moore et al. 1993; Gallant and Wilson 1996, 2000). Integration of terrain modeling and GIS analysis provides a toolset for rapid spatial prediction of landslide hazard. The relative simplicity of statistical methods (Carrara 1983; Carrara et al. 1995; Chung et al. 1995; Atkinson and Massari 1998; Atkinson et al. 1998; Chung and Fabbri 1999; Massari and Atkinson 1999; Dhakal et al. 2000; Gorsevski et al. 2000, 2001, 2003) is attractive when rapid investigation with minimal cost is required for a large area. For instance, logistic regression has been used for landslide hazard prediction previously with great success by Atkinson et al. (1998) and Atkinson and Massari (1998). This approach explicitly defines the relationship between landslide occurrence and different environmental variables. The relationships defined by the model are then applied for spatial prediction and generation of landslide hazard maps for large areas. However, such empirical models are dependent on the dataset and specific physiographic setting in which they are developed.

This paper presents a robust GIS-based approach that establishes physiographic settings by consolidating *a priori* knowledge about the study area through exploratory data analysis before statistical modeling is implemented (McSweeney et al. 1994). The approach uses logistic regression, following Atkinson et al. (1998) and Atkinson and Massari (1998), but it extends their approach and includes signal detection theory as a first step for spatial prediction of landslide hazard. Different models are constructed for the RR and NRR landslide occurrences and applied to the entire study area separately, to broadly discriminate stable from unstable areas. Receiver operating characteristic (ROC) curves are used to plot predicted probability to understand issues of accuracy, criterion selection, and interpretation (Egan 1975, Swets 1988). The area under the ROC curve (AUC) was used as a convenient measure of overall fit and comparison of modeled predictions of NRR and RR landslide hazard. However, the broad aim of this paper is to improve upon current approaches by enhancing the dichotomous statistical modeling with ROC curves that may provide a measure of accuracy to assist forestland managers in making decisions regarding management in both roaded and roadless areas.

The approach is tested in a case study on the Clearwater National Forest (CNF) for two objectives: (1) to evaluate if the predictor variables from the logistic regressions for RR and NRR landslide hazard models are the same; and (2) to evaluate if the logistic regression models for predicting RR and NRR landslide hazard are different. The first objective is intended to provide pointers to the processes driving landslide occurrence for road versus non-road areas and to provide the basis for understanding landslide processes and hazard in the area. The second objective is to justify whether or not development of two independent models is necessary.

2 Statistical Modeling Approach

2.1 Logistic regression model

Logistic regression (Hosmer and Lemeshow 1989, Allison 1999) is typically used when the predictor variables are not normally distributed and some may be categorical (Johnson 1998, Allison 1999). The spatial prediction is modeled by a dependent variable and a number of independent variables that are available in a spatially continuous fashion across the region. Logistic regression is similar to multiple regression. However, the primary difference is that the dependent variable in the logistic regression is sampled as a binary variable (i.e. presence, absence of landslide). The logistic regression therefore models the probability of presence and absence given observed values of predictor variables. Logistic regression fits a special s-shaped curve by taking a linear regression, that may produce any y-value between $-\infty$ and $+\infty$, and transforming it with the function that produces a probability (p-probability) between 0 (as y approaches minus infinity) and 1 (as y approaches plus infinity).

$$p(y = 1|x) = \frac{\exp(\beta_0 + \beta_1' x)}{1 + \exp(\beta_0 + \beta_1' x)} \tag{1}$$

In the logistic regression model, x is the data vector for a randomly selected experimental unit, β is the coefficient(s) of the independent variable(s), and y is the value of the binary outcome variable. The maximum likelihood method may be used to estimate β prior to prediction. The output of a statistically based model is an equation that can be used for prediction or estimation. However, the accuracy and the reliability of the prediction of these

explicit empirical models are a function of the sampling strategy and choice of explanatory variables that may relate to the basic processes operating at specific measurement scales.

2.2 ROC Curves

ROC curve analysis is a commonly used method for assessing the accuracy of a diagnostic test (Egan 1975, Swets 1988, Williams et al. 1999). ROC curves provide a diagnostic that may be used to distinguish between two classes of events, and to visualize classifier performance (Swets 1988). The curve is a plot of the probability of having a true positive (correctly predicted event response) versus the probability of a false positive (falsely predicted event response) as the cut-off probability varies. For example, a true positive is a prediction of a landslide for a location where a landslide occurred, while a false positive is a prediction of a landslide for a location where a landslide did not occur. An ideal model would have an area equal to 1, because then P (true positive) = 1 and P (false positive) = 0 regardless of the cutoff point (Williams et al. 1999). Each point on the ROC curve may be associated with a specific decision criterion for how much risk the user is willing to take regarding the accuracy of the prediction. This point will vary among observers because their decision criteria may vary even when their ROC curves are the same. However, incorporation of this analysis may allow managers to alter decisions based on knowledge of contextual information (e.g. is the area remote or heavily populated?) (Swets 1988).

In dichotomous statistical modeling such as logistic regression, ROC curves are very useful for evaluating the predictive accuracy of a chosen model. The predicted probabilities generated by the logistic model can be viewed as a continuous indicator to be compared to the observed binary response variable. Although ROC analysis provides an overall accuracy value independent of a cutoff value and occurrence, it does not provide the optimal cutoff value nor does it illustrate how occurrence affects cutoff selection (Mcfall 1999). With this technique, the cutoff is selected so that it provides a specification of a function to be maximized. Thus, the choice of a suitable cutoff can be determined by quantitative assessment of their degree of accuracy (Swets 1988). Valid and precise assessments of quantitative accuracy based on long-term test data could help decision makers to know when to use the system as well as how much faith to put in the system. Such assessments may help managers to determine various criteria of system use considering all possible sets of conditions that could lead to landsliding. For example, if a manager knows that recent cumulative rainfall has left the landscape saturated, the acceptable risk may be different than the acceptable risk when the landscape is unsaturated.

3 Material and Methods

The modeling approach for estimating landslide hazard probability in Idaho's CNF is discussed in the following sections.

3.1 Study Area

The study area is the CNF (115°46'W, 46°07'N, 114°19'W, 47°00'N), located on the western slopes of the Rocky Mountains in north central Idaho covering approximately 7,435 km^2 (Figure 1). The CNF is located west of the Montana border and bounded

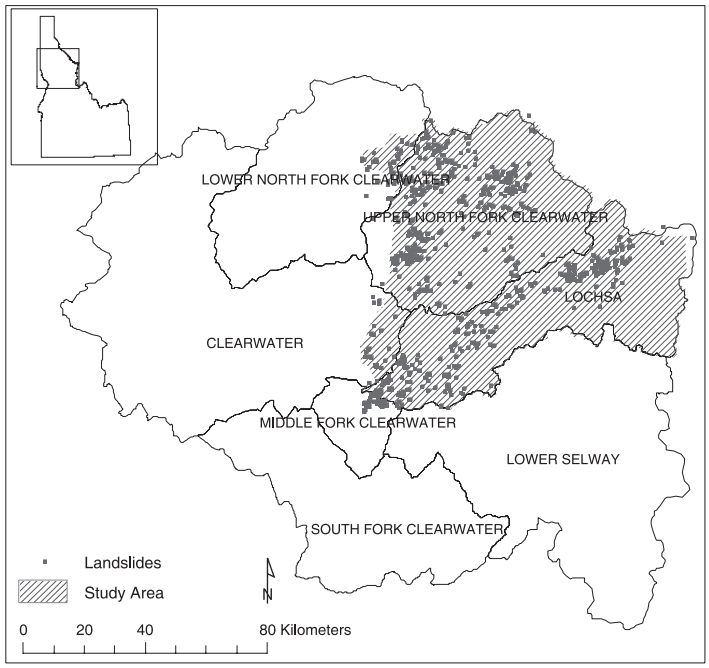


Figure 1 Distribution of landslides over the Clearwater National Forest

on three sides by four other National Forests; the Lolo in Montana; the Bitterroot in Montana and Idaho; the Nez Perce in Idaho; and the Panhandle in Idaho. The highly dissected mountainous topography of the CNF is typical for the headwaters of the Columbia River Basin. Elevation in the CNF ranges from 480 to 2,700 m and slopes vary between 0 and 45°. The climate is characterized by dry and warm summers, and cool wet winters (McClelland et al. 1997). Annual precipitation ranges between 600 mm at low elevations to more than 2,000 mm at high elevations. Much of the annual precipitation falls as snow during winter and spring, while peak stream discharge occurs in late spring and early summer. The highly variable steep soils are well drained and are primarily derived from parent materials such as granitics, metamorphic rocks, quartzites, and basalts or surface erosion and deposition. Vegetation includes Grand fir (*Abies grandis*), Douglas fir (*Pseudotsuga menziesii*), Subalpine fir (*Abies lasiocarpa*), Western redcedar (Thuja plicata), Western white pine (*Pinus Monticola*), and various other shrubs and grasses that have short growing seasons, particularly at the higher elevations.

Historically, the CNF has experienced periodic floods and landslide events. Major floods occurred in 1919, 1933, 1948, 1964, 1968, and 1974. These flooding events were documented through stream flow records (McClelland et al. 1997). Although the frequency of major landslide events is known for the last century, accurate and complete mapping of landslide locations over this period does not exist.

3.2 Data Acquisition and GIS Data Layers

The selection of environmental attributes and GIS layers used here was guided by knowledge of instability factors that were cost-effectively acquired in previous studies

Table 1 Environmental attributes used for modeling landslide hazard

Environmental Attribute		Classes
Terrain	*Elevation*	Continuous
	Slope	Continuous
	Profile Curvature	Continuous
	Plan Curvature	Continuous
	Tangent Curvature	Continuous
	Flow Path Length	Continuous
	Upslope slope	Continuous
	CTI	Continuous
	Solar radiation	Continuous
Parent Material		Alluvium
		Glacial Till
		Granitic
		Revett Quartz
		Quartzite
		Grussic Granitic
		Mica Schist
		Basalts
		Undifferent
Climate (Köppen classification)		Dfb
		Dfc
		Dsb
		Dsc
Landslides		Non-road related
		Road related

and readily available for the selected physiographic setting (Carrara 1983, Varnes 1984, Turner and Schuster 1996). For instance, lithology of the study area existed at a scale of 1:1,000,000 digitized from the original pen and ink mylar map of the general lithologies of the major ground water flow systems in Idaho. We decided not to use lithology in our analysis because of lack of detail. Table 1 shows the explanatory variables used for landslide hazard development. All variables are continuous except parent material and climate (categorical).

3.2.1 Terrain Attributes

The 7.5-minute U.S. Geological Survey DEMs (30 m) were used as input to the TAPES-G software (Gallant and Wilson 1996, 2000) to derive the primary terrain attributes from a DEM (Table 2). Secondary terrain attributes are computed from combinations of two or more primary attributes (Moore et al. 1993, Gallant and Wilson 2000) and include the topographic wetness index or compound topographic index (CTI) (Moore et al. 1993) and solar radiation indices (Fu and Rich 2000) (Table 3). The 30 m DEMs were provided in the Universal Transverse Mercator (UTM) projection using the North

Table 2 Primary terrain attributes derived from a digital elevation model

Attribute	Definition	Significance
Elevation	Height above sea level	Climate, vegetation, potential energy
Slope	Change in elevation divided by horizontal distance	Overland and subsurface flow velocity and runoff rate, precipitation, vegetation, geomorphology, soil water content, land capability class
Aspect	Slope azimuth	Solar insolation, evapotranspiration, flora and fauna distribution and abundance
Profile curvature	Slope profile curvature	Flow acceleration, erosion/deposition rate, geomorphology
Plan curvature	Contour curvature	Converging, diverging flow, soil water content, soil characteristics
Tangent curvature	Curvature of line formed by intersection of surface with plane normal to flow line	Erosion/deposition
Flow path length	Distance from watershed divide to the point of interest	Erosion rates, sediment yield, time of concentration
Upslope area	Area (i.e. square meters)	Runoff velocity, potential energy

Table 3 Secondary terrain attributes derived from a digital elevation model

Attribute	Definition	Significance
Compound topographic index (topographic wetness index)	$CTI = \ln\left(\frac{A_s}{\tan\beta}\right)$	Steady state and uniform soil properties are assumed in this equation in order to predict zones of saturation. A_s is specific catchment area (m^2m^{-1}) and β is the slope angle
Solar radiation	$Global_{tot} = Dir_{tot} + Dif_{tot}$	Solar radiation (insolation) is a primary driver of physical and biological processes. The global radiation is calculated as a sum of direct and diffuse radiation

American Datum of 1927. A final DEM aggregation for the CNF consisted of a total of 80 USGS Level 2 DEM topographic quadrangle maps (7.5-minute).

3.2.2 Parent Material

The Idaho Geological Survey compiled the geologic data used for this research through the Digital Geologic Mapping and GIS Lab at a cartographic scale of 1:100,000. This

is the only digital geologic map currently available for the entire study area. The digital geologic map data model is based on the North American Geologic Map Data Model (NADM) (Johnson et al. 1999). About 40% of the CNF is covered by undifferentiated depositional material, which consists of complex geologic types: usually materials outcropping on steep erosional escarpments. This complex class is used for classification where units relating to individual parent materials cannot be delimited separately at the scale of mapping. It may include colluvium derived from the various genetic materials that rest upon the scarp slope. Glacial till covered about 16%, and granitics 20% of the study area.

3.2.3 Climate

Köppen's climate classification was used to delineate homogeneous climatic zones for the CNF (Godfrey 1999). The raw climate surface data used to produce the classification were obtained from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM), developed at Oregon State University (Daly et al. 1994). They include estimates of mean monthly and annual climatic parameters interpolated from point data and a DEM. The Köppen climate classification used for this study is based predominantly on monthly and annual averages of precipitation and temperature that were derived using multivariate statistical analyses combined within a GIS (Godfrey 1999). The classification has a 4 km spatial resolution.

The Köppen classification divides land areas first into temperature zones, which are further subdivided on the basis of seasonal incidence and amounts of annual rainfall. A total of four climatic types occur on the CNF and include: Dfb, Dfc, Dsb, and Dsc. These climatic zones, according to the Köppen classification, correspond to cold humid climates (*D*). These are defined as cold long winters with several months of frozen ground (mean monthly temperature of 0°C and below) and winter precipitation in the form of snow is characteristic, while summers are dry (mean monthly temperature of 10°C and above). The second letter "*f*" or "*s*" stands for moisture (precipitation) distinction and the distribution of moisture between the wettest and the driest month in different seasons of the year, while the third letter "b" or "c" indicates variation in temperature that may influence rain or snow occurrences.

3.3 Landslide Data Acquisition

Landslides were assessed through aerial photo interpretation (API) in conjunction with field inventory following major landslide events that occurred in November 1995 and February 1996. Aerial photography of the CNF was acquired at a scale of 1:15,840 followed by a photo interpretation phase between October 1996 and February 1997 (McClelland et al. 1997). Landslides interpreted from aerial photos were classified into RR and NRR. A total of 865 landslides were recorded, with 55% road-related and 45% non road-related. The presence or absence of a landslide was represented as a (30 m) grid coverage with values of 1 for presence and 0 for absence. The initiation area of each landslide (i.e. the area where the main scarp of the landslide occurred) was interpreted as the point representing the presence of a landslide.

3.3.1 Data Sampling

In the development of the two models, presence and absence for RR landslides represented the subset of all landslides that occurred in the roaded areas (i.e. spatial locations within

the road right-of-way). The NRR landslides were represented as the remainder of the landslide locations that occurred in non-roaded areas (spatial locations outside road right-of-way). An arbitrary decision was made to sample a sufficient size of a total of 15% (fixed sample proportions) of the non-landslide cells (absence) in a random fashion to provide a non-landslide location dataset. Separate samples were acquired for completion of both the RR and NRR landslide datasets for the entire CNF. Next, a multivariate subset of data was sampled from the digital coverages (i.e. terrain attributes, parent material) for all sample locations in the datasets. The two separate datasets were then used for EDA and subsequent statistical modeling.

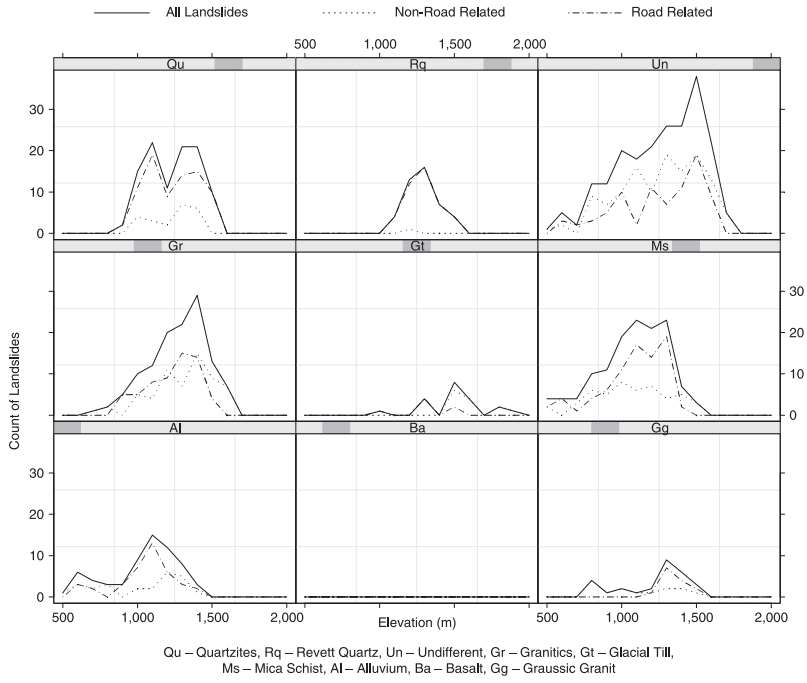
3.4 Exploratory Data Analysis and Statistical Model Development

Multivariate normal probability plot and normal quantile-quantile (normal q-q) plot (Cleveland 1993) graphical methods were used for determining how well the distributions were approximated by a normal distribution. Other techniques such as probability density functions (PDF), cumulative distribution functions (CDF), box plots, and scatter plots were used to explore patterns within the data. First, standardization (z score) was performed on the response variables in order to eliminate the measurement units, so the variables are measured in comparable units (Johnson 1998). Statistical modeling followed using logistic regression with explanatory variables that consist of both continuous and discrete (dummy variables). The maximum likelihood method was used to estimate β and output the model coefficients. A backward elimination procedure that minimized the Akaike Information Criterion (AIC) was used to interactively select the variables useful for predicting landslide hazard. Variables that were not significant at the 0.05 level were eliminated from the regression. Dummy variables were removed only if none of the variables in a group of dummy variables (e.g. parent material) were significant (Johnson 1998). The explanatory variables were checked for multicollinearity prior to other diagnostic methods to ensure stable coefficient fits. A diagnostic method using a bubble plot for identifying outliers and influential points was also used on the predicted values. The plot measures of influence (Pearson chi^2 and Cook's D – bubble size) helped to determine whether individual cases have significant impact on the fitted logistic regression model and the coefficients of individual predictors. Finally, ROC curves were implemented for comparison of various thresholds on the predictor measures; an overall index of accuracy, which accounts for all possible thresholds; the simple identification of the optimal threshold; and the comparison of two or more predictors. ROC curves were utilized for both models (RR and NRR) to further evaluate the objectives and as a measure of performance of a predictive rule.

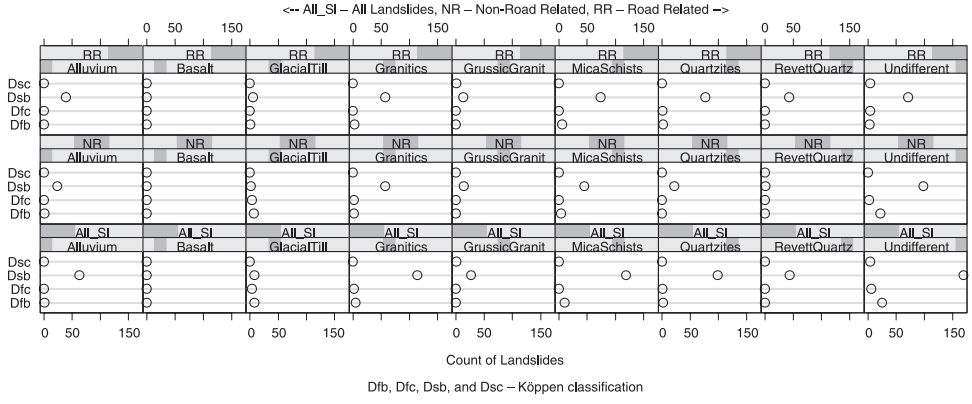
4 Results and Discussion

4.1 Exploratory Data Analysis

Figure 2a shows the count of landslides by parent material and elevation for all landslides along with RR and NRR subsets. A noticeably higher proportion of RR landslides occur on the quartzite and mica-schist parent materials. Figure 2b is a dot plot (data values are represented on a scale using a dot) showing all landslides along with RR and NRR subsets as they interact with climate and parent material. This plot shows that most landslides occur on the Dsb climate type, and again show a greater proportion of



(a)



(b)

Figure 2 (a) Landslides associated with different parent materials; and (b) Dot plot of landslides interacting with climates and parent material

RR landslides on quartzite and mica-schist parent materials. The plots illustrate preliminary relationships between variables.

4.2 Road Related Model

Table 4 shows the statistical model that predicts RR landslide hazard in the CNF, while Table 5 shows the significance of predictor variables for the developed model. The

Table 4 Logistic regression equation for predicting road related landslide hazard

log (p/(1 – p))
where
Road related
p = –3.4037 + 0.7557 * Dsb + 0.5706 * Dsc + 0.272 * Dfb + 0.1061 * Undiffer – 0.2363 * GlacialT + 0.0816 * Alluvium + 0.1998 * MicaS – 2.6598 * Basalts + 0.382 * Quartz + 0.0328 * Granit + 0.432 * RevettQ – 0.6288 * Cti – 0.1394 * Radglb + 0.6588 * Uplog

Table 5 Significance of predictor variables in the logistic regression for the road related landslides

Parameter	Estimate	Error	Chi-Square	Pr > ChiSq	Significance
Intercept	–3.4037	0.2022	283.4598	<0.0001	**
Dsb	0.7557	0.1825	17.1425	<0.0001	**
Dsc	0.5706	0.2366	5.8189	0.0159	
Dfb	0.272	0.205	1.7603	0.1846	
Undiffer	0.1061	0.1014	1.094	0.2956	
Glacial Till	–0.2363	0.1709	1.9132	0.1666	
Alluvium	0.0816	0.1073	0.5792	0.4466	
Mica Schist	0.1998	0.1015	3.8749	0.049	**
Basalts	–2.6598	42.6144	0.0039	0.9502	
Quartz	0.382	0.105	13.2339	0.0003	**
Granitics	0.0328	0.1016	0.1039	0.7472	
Revett Quartz	0.432	0.1149	14.1449	0.0002	**
Cti	–0.6288	0.0931	45.5986	<0.0001	**
Uplog	0.6588	0.0912	52.207	<0.0001	**
Radglb	–0.1394	0.0207	45.538	<0.0001	**

p – probability of landslide hazard.
Dsb, Dsc, Dfb – climate.
Undifferent, Glacial Till, Alluvium, Mica Schist, Basalts, Quartz, Granitics, Revett Quartz – parent material.
Compound Topographic Index (CTI), Logarithm of Upslope Area (Uplog), Solar Radiation (Radglb) – topographic attributes.
** – significant at 0.05 level.

statistical model incorporates topographic, parent material and climate explanatory variables. The *p* values in Table 5 show that some of the parent material categories are not significant, but are included because other parent material categories are significant. A separate analysis of the climate factor (Dsb, Dsc, Dfb, Dfc) was conducted by comparing each factor against the "Dfc" factor. For example, to represent climatic categories using dummy variables (i.e. Dsb, Dsc, Dfb), it takes one less dummy variable than the number of categories. Thus, if the values for all categories are equal to zero, then the excluded dummy variable is referred to as the reference class (i.e. Dfc) and is used for comparison against the other categories. All climate factors showed statistical significance with

positively estimated values suggesting that they have higher probabilities of landslide than "Dfc". A separate analysis of the parent material factors was also conducted. Each parent material factor was compared against the "Grussic granitics" factor. Only the basalts and glacial till factors were statistically significant, and the negative coefficients suggest that landslides in roaded areas are less likely to occur on these parent materials compared to "Grussic granitics".

The most significant predictor variables contributing to the RR landslide model are climate (Dsb), parent material (Mica Schist, Quartz, Revett quartzite), CTI, solar radiation, and upslope drainage area. The negative coefficient for CTI and solar radiation suggests that landslide hazard decreases as the values of these variables increase. Conversely, as the upslope drainage area increases, landslide hazard increases. Therefore, the results suggest, at this scale, the major drivers that influence landsliding in roaded areas are climate, parent material (soil properties), concentration of surface and subsurface flow and biophysical processes (i.e. CTI and solar radiation). Since the Köppen climate classification is based on monthly or annual averages of precipitation and temperature, it may be that the topographic variables provide additional information about water redistribution in the landscape at a finer scale (30 m) than the climate variables (4 km). The importance of the concentration of surface and subsurface flow can be interpreted from the CTI and upslope contributing area. CTI is used to determine potential spatial distribution of water in a landscape, although the model assumes steady-state conditions and invariant conditions for both infiltration and transmissivity (Gallant and Wilson 2000). CTI values increase as the upslope drainage area (specific catchment area) increases and slope gradient decreases. A low CTI value (e.g. 1–3) usually describes steeper slopes, smaller upslope drainage areas, and shallower soils (Gessler et al. 1995, 2002) while larger CTI values (e.g. > 6) typically describe gentler slopes, larger upslope drainage areas, and deeper soils. Therefore, the logistic regression model suggests that locations with equivalent slopes but larger upslope drainage areas (specific catchment area) have increased risk of slope failure. Likewise, high slope gradient and high soil moisture (assumed from larger CTI values), in general, decreases slope stability. As coarser parent materials with low frictional resistance to sliding such as Quartzite and Revett quartzite intersect areas with lower CTI values with roads, the risk of landslide hazard also increases. Finally, the probability of landsliding also increases as solar radiation decreases perhaps because of generally wetter conditions on northerly landscape locations.

4.3 Non-Road Related Model

Table 6 shows the statistical model for prediction of NRR landslide hazard in the CNF, while Table 7 shows the significance of predictor variables for this model. This analysis shows different patterns in the climatic and parent material factors. The climate factors were again compared against the "Dfc" factor. The "Dsc" factor showed statistical significance with a negative relationship suggesting that it is less likely for landsliding compared to "Dfc". The parent material factors were compared against the "Grussic granitics" factor. Only the basalt factor showed statistical significance with a negative relationship meaning that the probability of landslide is lower than the "Grussic granitics" parent materials.

The contributing predictor variables with highest significance to landslide probability for the NRR model were climate, parent material, elevation, CTI, solar radiation,

Table 6 Logistic regression equation for predicting non-road related landslide hazard

log (p/(1 – p))
or
Non-road related
p = –3.714 + 0.4112 * Dsb – 2.6448 * Dsc + 0.3372 * Dfb + 0.087 * Undiffer + 0.1371* Glacial Till + 0.1272 * Alluvium + 0.0291 * Mica Schist – 3.0712 * Basalts + 0.6706 * Quartz + 0.0151 * Granit + 0.4562 * Revett Quartz – 0.1147 * Elevation – 0.2103 * Cti – 0.2145 * Radglb – 0.0466 * Tgcurv + 0.1873 * Uplog

Table 7 Significance of predictor variables in the logistic regression for the non-road related landslides

Parameter	Estimate	Error	Chi-Square	Pr > ChiSq	Significance
Intercept	–3.714	0.1248	885.3652	<0.0001	**
Dsb	0.4112	0.0948	18.808	<0.0001	**
Dsc	–2.6448	49.3237	0.0029	0.9572	
Dfb	0.3372	0.0983	11.7548	0.0006	**
Undiffer	0.087	0.0933	0.8687	0.3513	
Glacial Till	0.1371	0.123	1.2416	0.2652	
Alluvium	0.1272	0.1128	1.272	0.2594	
Mica Schist	0.0291	0.0986	0.0872	0.7678	
Basalts	–3.0712	304.5	0.0001	0.9920	
Quartz	0.6706	0.1092	37.7	<0.0001	**
Granit	0.0151	0.0969	0.0243	0.8760	
Revett Quartz	0.4562	0.1439	10.0559	0.0015	**
Elevation	–0.1147	0.0257	19.9757	<0.0001	**
Cti	–0.2103	0.0824	6.5124	0.0107	**
Tgcurv	–0.0466	0.0171	7.4204	0.0064	**
Uplog	0.1873	0.0807	5.3868	0.0203	**
Radglb	–0.2145	0.0219	96.0433	<0.0001	**

p – probability of landslide hazard.
Dsb, Dsc, Dfb – climate.
Undifferent, Glacial Till, Alluvium, Mica Schist, Basalts, Quartz, Granit, Revett Quartz – parent material.
Elevation, Compounded Topographic Index (CTI), Logarithm of Upslope Area (Uplog), Tangent Curvature (Tgcurv), Solar Radiation (Radglb) – topographic attributes.
** – significant at 0.05 level.

tangential curvature, and upslope drainage area. The negative coefficients for elevation, CTI, tangential curvature, and solar radiation suggest that landslide probability decreases as these variables increase. Conversely, as upslope drainage area increases, the probability of landslide hazard increases, similar to RR areas. These results suggest that NRR landslides may be restricted to a more narrow set of environmental circumstances in comparison to the RR model. Specifically, the inclusion of the elevation variables suggests a more

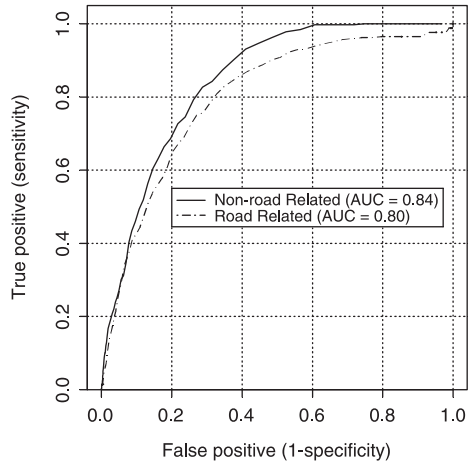


Figure 3 ROC curves for the non-road related and road related landslides

narrow range of elevations where water redistribution is critical to landslide occurrence. For instance, preliminary EDA showing fewer landslides at higher elevations may have resulted because snowpack and cooler temperatures did not contribute to rain-on-snow events. In addition, lower tangential curvature values represent areas of convergent (concave) flow, which parallels findings reported by Montgomery and Dietrich (1994) who showed that topographic convergent areas are at greater risk.

4.4 Model Comparison

The modeled predictions of NRR and RR landslides first were compared using the area under the curve (AUC) from the ROC curves (Figure 3) and evaluated through two objectives. The AUC from the NRR model exhibited higher discrimination power in separating high hazard areas from low hazard areas. The NRR model yields an overall quantitative index of accuracy corresponding to 0.84 (AUC), while the RR model yields an overall quantitative index of accuracy corresponding to 0.80 (AUC). Thus, this suggests that, based on the datasets used, the prediction of NRR landslides can be achieved with higher certainty than the prediction of RR landslides. These findings are also similar to those reported by Montgomery et al. (1998). The ROC curve was also used to compare various cut-off combinations on predicted landslide probabilities with a specific index of accuracy. Table 8 shows an example of three arbitrarily chosen cut-off alternatives (0.048, 0.062, and 0.074) from the ROC curve for NRR landslides that a manager might choose (Figure 4). The table shows the percentage of correctly and incorrectly predicted landslides, as well as the non-landslides. Landslide probability maps corresponding to Table 8 are shown in Figure 6. For example, Figure 6a shows that with a cut-off value of 0.048, a total of 93.4% of the NRR landslides were correctly identified, while only 57.4% of the non-landslide cells were correctly predicted. Figures 6b and c and Figure 7 show the same relation as Figure 6a except that a different cut-off value is used. In addition, Figures 6 and 7 demonstrate that the accuracy of landslide prediction can be increased but only if the user is willing to accept lower accuracies in non-landslide locations. Thus, the uncertainty associated with the accuracy comes with

Table 8 Percentage of correctly and incorrectly predicted landslides and non-landslides for three different arbitrary cut-off values from the ROC curve

	Present	Absent	Present	Absent	Present	Absent
Landslides	93.40	6.60	75.16	24.84	51.26	48.74
Non-Landslides	42.61	57.39	24.97	75.03	12.21	87.79
Cut-off value	0.048		0.062		0.074	

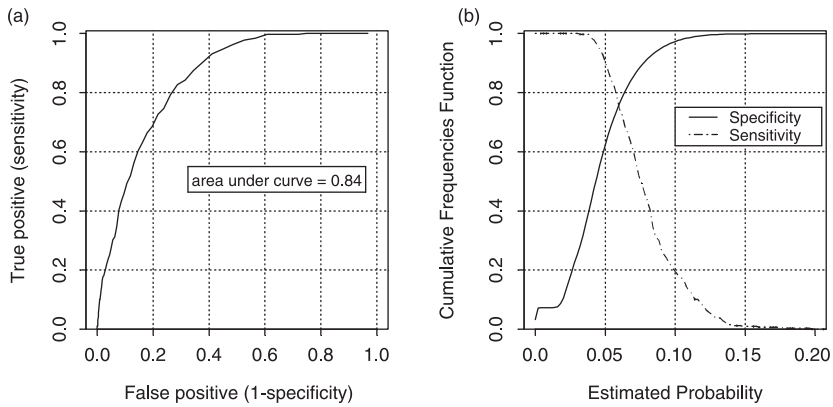


Figure 4 Non-road related model: (a) ROC curve for the logistic regression model; and (b) Cumulative Frequency Function of specificity and sensitivity versus estimated probability

the confidence one might have in making a correct decision. The ROC curve for the RR landslides is shown in Figure 5.

4.4.1 Comparison between predictor variables

The comparison between predictor variables from RR and NRR landslide hazard models showed that the predictor variables were different. The predictor variables with highest significance for each model were interactively selected using a backward elimination procedure employing the Akaike's Information Criterion. The AIC is a statistic that uses the residual deviance penalized by the number of parameters requiring estimation in a model fit. The AIC values for various combinations of predictor variables are shown in Table 9 (NRR model) and 10 (RR model), ordered by the significance of the best predictor variables. From the final models selected by the lowest AIC values, although similarities exist with some of the predictor variables (Dsb, Quartz, Revett Quartize, CTI, upslope contributing area, solar radiation), differences do exist between the predictor variables used for RR and NRR models (Tables 4 and 6). The major difference is that the NRR model includes the Dfb, elevation and tangent curvature variables and the RR model does not. This suggests that the overall processes driving RR and NRR landslides are similar, but that NRR landslides appear to occur in a more tightly constricted set of environments characterized by a specific elevation zone and

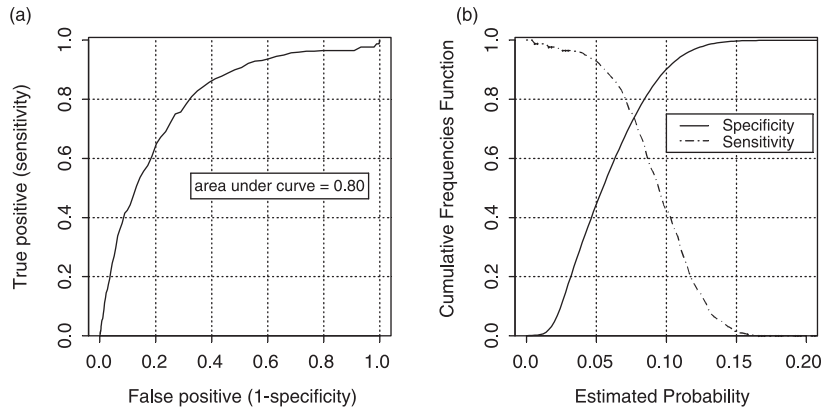


Figure 5 Road related model: (a) ROC curve for the logistic regression model; and (b) Cumulative Frequency Function of specificity and sensitivity versus estimated probability

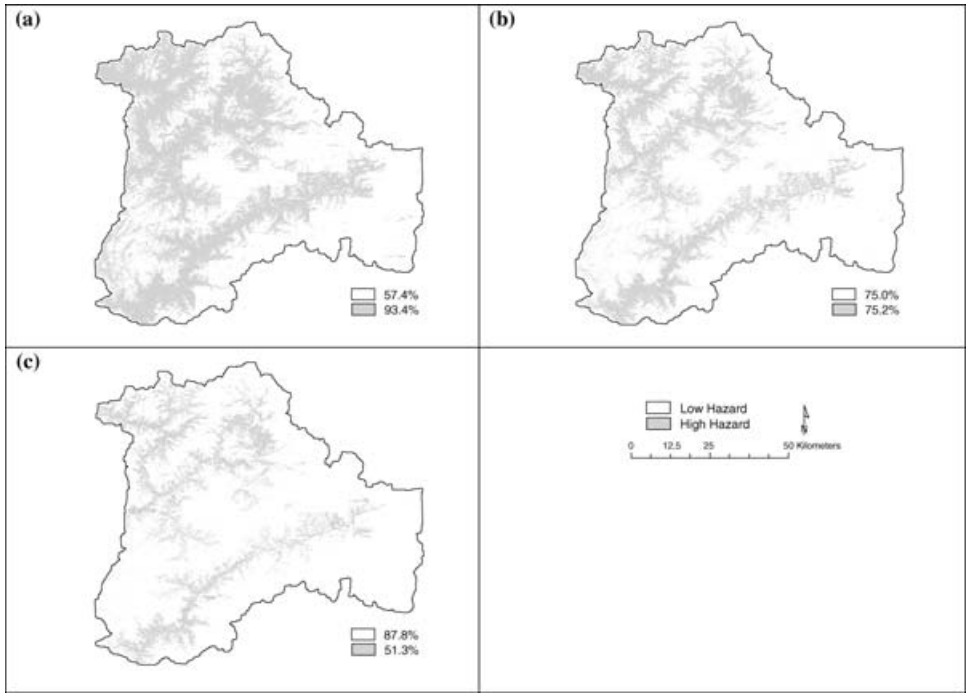


Figure 6 Non-road related landslide hazard maps with different cut-off values derived from ROC curve: (a) cut-off = 0.048; (b) cut-off = 0.062; and (c) cut-off = 0.074

water convergence. It is possible that this could be a result specific to the weather events that caused the major landsliding and hence the landslide dataset used for this analysis. These results suggest that the predictor variables for RR and NRR landslide hazard models are different and two separate logistic regression models should be developed.

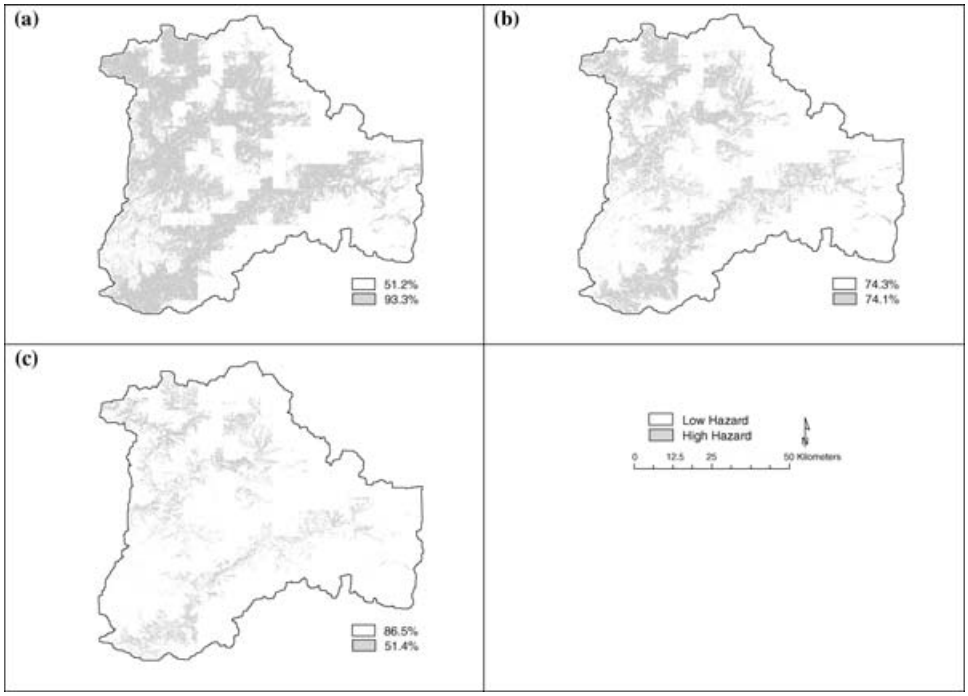


Figure 7 Road related landslide hazard maps with different cut-off values derived from ROC curve: (a) cut-off = 0.052; (b) cut-off = 0.076; and (c) cut-off = 0.094

Table 9 Akaike's information criterion (AIC) for the non-road related model development where best predictor variables are arranged by *p*-value significance

AIC	Variable (by importance)
4,598.479	Solar Radiation, Elevation, Dsb, Quartz, Dfb, Revett Quartz, Tangent Curvature, Cti, Upslope Area, Alluvium, Glacial Till, Undifferent, Mica Schist, Granit, Dsc, Basalts
4,611.682	Solar Radiation, Elevation, Quartz, Dsb, Revett Quartz, Tangent Curvature, Dfb, Slope, Glacial Till, Alluvium, Undifferent, Mica Schist, Granit, Dsc, Basalts
4,614.264	Solar Radiation, Elevation, Quartz, Dsb, Dfb, Revett Quartz, Tangent Curvature, Glacial Till, Alluvium, Undifferent, Mica Schist, Granit, Dsc, Basalts
4,615.329	Solar Radiation, Elevation, Quartz, Dsb, Dfb, Revett Quartz, Tangent Curvature, Glacial Till, Alluvium, Undifferent, Mica Schist, Granit, Dsc, Basalts
4,656.707	Solar Radiation, Elevation, Quartz, Revett Quartz, Tangent Curvature, Plcurv, Prcurv, Alluvium, Glacial Till, Undifferent, Granit, Mica Schist, Basalts
4,674.792	Solar Radiation, Elevation, Dsb, Dfb, Slope, Dsc
4,721.143	Solar Radiation, Elevation, Tangent Curvature, Slope

Table 10 Akaike's information criterion (AIC) for the road related model development where best predictor variables are arranged by *p*-value significance

AIC	Variable (by importance)
3,899.414	Solar Radiation, Cti, Upslope Area, Dsb, Revett Quartz, Quartz, Mica Schist, Dsc, Glacial Till, Dfb, Undifferent, Alluvium, Granit, Basalts
3,916.042	Solar Radiation, Slope, Revett Quartz, Quartz, Cti, Mica Schist, Tangent Curvature, Undifferent, Alluvium, Granit, Dsb, Dsc, Dfb, Glacial Till, Basalts
3,943.885	Solar Radiation, Revett Quartz, Quartz, Mica Schist, Tangent Curvature, Undifferent, Granit, Alluvium, Dsb, Dsc, Dfb, Glacial Till, Basalts
3,944.111	Solar Radiation, Revett Quartz, Quartz, Mica Schist, Upslope Area, Undifferent, Granit, Alluvium, Dsb, Dsc, Dfb, Glacial Till, Basalts
3,962.398	Solar Radiation, Cti, Upslope Area, Elevation, Mica Schist, Undifferent, Granit, Alluvium, Glacial Till, Basalts
3,969.766	Solar Radiation, Slope, Tangent Curvature, Cti, Dsb, Dsc, Dfb
4,043.2	Solar Radiation, Cti, Upslope Area, Elevation, Tangent Curvature

4.4.2 Comparison between predictive models

The paired *t*-test verified that differences exist between the predictive models for RR and NRR landslides. The paired *t*-test was performed on the number of correctly predicted landslide cells associated with a specific percent ROC cut-off value (i.e. 95, 90, 85, 80, 75, 70, 65, 60, 55, 50, 45, 40, 35, 30, 25, 20, 15, 10, 5) derived from both predictions. The probability of the two-tailed *t*-test was $p = 0.2361$ for the 95% confidence interval (lower = –0.9395037; upper = 3.5710826) with 18 degrees of freedom and $t = 1.2257$. The two landslide hazard models yielded different patterns of landslide probabilities (Figures 6 and 7). The difference in patterns is particularly apparent in Figures 7a and b through the "blocky" appearance that is likely an influence of the coarser spatial resolution of the Köppen climate classification (4 km). Therefore, for the second objective, we verified that predictive models of landslide probability using logistic regression for RR and NRR landslide subsets are different suggesting the need for development of individual models.

5 Conclusions

Accurate prediction of landslide hazard is difficult because of the complex landslide processes and the human activities that constantly reshape the Earth's surface. This paper illustrates the application of a logistic regression modeling technique that incorporates ROC curve analysis for modeling landslide probability with datasets limited to a specific set of circumstances. The derived models can be applied to a large geographical area using GIS and statistical modeling with environmental variables mostly derived from a DEM. The landslide dataset was further subdivided into NRR and RR landslides, and two separate models were developed and compared. The identification of climate, parent material, elevation, CTI, solar radiation, tangential curvature, and upslope drainage area for NRR landslides, and the identification of climate, parent material,

CTI, solar radiation, and upslope drainage area for RR landslides, yielded different patterns of landslide probabilities.

The RR and NRR models may be considered a first approximation of a tool that can be used in managed forest areas, roadless, or wilderness areas within the region. This research also incorporated ROC curves as an additional tool that may help decision makers who are often required to alter the level of risk they are prepared to accept depending on circumstances. Although the quantitative method is capable of evaluating the predictive accuracy of a chosen model, the method does not provide the optimal cutoff value to be used. Thus, the choice of an appropriate cutoff value can be refined as our understanding of landslide processes improves and perhaps by integrating other models (Hammond et al. 1992, Montgomery and Dietrich 1994, Wu and Sidle 1995, Gorsevski 2002).

This explicit quantitative modeling approach may provide a first approximation of a landslide hazard surface associated with NRR and RR landslides. The models could be redeveloped as a larger landslide occurrence database is gathered through time to include other events. In addition, this empirical modeling technique could be used to initiate and calibrate process-based models that may enable the simulation and understanding of landslide hazard with changes in environmental variables such as climate, root strength, and vegetation. Therefore, this procedure could be useful as a first step in landslide hazard development and to perhaps indicate stable areas where the focus of landslide hazard rating may not be required.

Acknowledgements

This work was supported through data provided by the Clearwater National Forest and the Rocky Mountain Research Station in Moscow, Idaho.

References

Atkinson P M, Jiskoot H, Massari R, and Murray T 1998 Generalized linear modelling in geomorphology. *Earth Surface Processes and Landforms* 23: 1185–96

Burrough P A and McDonnell R A 1998 *Principles of Geographic Information Systems*. Oxford, Oxford University Press

Carrara A 1983 Multivariate models for landslide hazard evaluation. *Mathematical Geology* 15: 403–26

Carrara A, Cardinali M, Guzzetti F, and Reichenbach P 1995 GIS technology in mapping landslide hazard. In Carrara A and Guzzetti F (eds) *Geographical Information Systems in Assessing Natural Hazards*. Dordrecht, Kluwer: 135–75

Chamran F, Gessler P E, and Chadwick O A 2002 A spatially explicit treatment of soil-water dynamics along a semiarid catena. *Soil Science Society of America Journal* 66: 1571–83

Chung C F, Fabbri A G, and van Westen C J 1995 Multivariate regression analysis for landslide hazard zonation. In Carrara A and Guzzetti F (eds) *Geographical Information Systems in Assessing Natural Hazards*. Dordrecht, Kluwer: 107–33

Chung C F and Fabbri A G 1999 Probabilistic prediction model for landslide hazard mapping. *Photogrammetric Engineering and Remote Sensing* 65: 1389–99

Cleveland W S 1993 *Visualizing Data*. Summit, NJ, Hobart Press

Daly C, Neilson R P, and Phillips D L 1994 A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology* 33: 140–58

Dhakal A S, Amada T, and Aniya M 2000 Landslide hazard mapping and its evaluation using GIS: An investigation of sampling schemes for a grid-cell based quantitative method. *Photogrammetric Engineering and Remote Sensing* 66: 981–9

Egan J P 1975 *Signal Detection Theory and ROC Analysis*. New York, Academic Press

Fu P and Rich P M 2000 The Solar Analyst Users Manual. WWW document, http://www.hemisoft.com/doc/sa_manual/solarext.htm

Gallant J C and Wilson J P 1996 TAPES-G: A grid-based terrain analysis program for environmental sciences. *Computers and Geosciences* 22: 713–22

Gallant J C and Wilson J P 2000 Primary topographic attributes. In Wilson J P and Gallant J C (eds) *Terrain Analysis Principles and Applications*. New York, John Wiley and Sons: 51–85

Gessler P E, Moore I D, McKenzie N J, and Ryan P J 1995 Soil-landscape modeling and spatial prediction of soil attributes. *International Journal of Geographic Information Systems* 9: 421–32

Gessler P E, Chadwick O A, Chamran F, Althouse L, and Holmes K 2000 Modeling soil-landscape and ecosystem properties using terrain attributes. *Soil Science Society of America Journal* 64: 2046–56

Glade T 1998 Establishing the frequency and magnitude of landslide-triggering rainstorm events in New Zealand. *Environmental Geology* 35: 160–74

Godfrey B R 1999 Delineation of Agroclimate Zones in Idaho. Unpublished MS Thesis, University of Idaho

Gorsevski P V 2002 Landslide Hazard Modeling Using GIS. Unpublished PhD Dissertation, University of Idaho

Gorsevski P V, Gessler P E, and Foltz R B 2000 Spatial prediction of landslide hazard using logistic regression and GIS. In *Proceedings of the Fourth International Conference on Integrating GIS and Environmental Modeling (GIS/EM4): Problems, Prospects and Research Needs*, Banff, Alberta [CD-ROM]

Gorsevski P V, Foltz R B, Gessler P E, and Cundy T W 2001 Statistical modeling of landslide hazard using GIS. In *Proceedings of the Seventh Federal Interagency Sedimentation Conference*, Reno, Nevada, 2: 103–9

Gorsevski P V, Gessler P E, and Jankowski P 2003 Integrating a fuzzy k-means classification and a Bayesian approach for spatial prediction of landslide hazard. *Journal of Geographical Systems* 5: 223–51

Hammond C, Hall D, Miller S, and Swetik P 1992 *Level I Stability Analyses (LISA) Documentation for Version 2.0*. Ogden, UT, U.S. Department of Agriculture, Forest Service, Intermountain Research Station General Technical Report No INT-285

Hosmer D W and Lemeshow S 1989 *Applied Logistic Regression*. New York, John Wiley and Sons

Johnson B R, Brodaric B, Raines G L, Hastings J T, and Wahl R 1999 Digital Geologic Map Data Model, Version 4.3a. WWW document, http://geology.usgs.gov/dm/

Johnson D E 1998 *Applied Multivariate Methods for Data Analysis*. Belmont, CA, Duxbury Press

Malczewski J 1999 *GIS and Multicriteria Decision Analysis*. New York, John Wiley and Sons

Massari R and Atkinson P M 1999 Modelling susceptibility to landsliding: An approach based on individual landslide type. *Transactions of the Japanese Geomorphological Union* 20: 151–68

Mcfall R M 1999 Quantifying the information value of clinical assessments with signal detection theory. WWW document, http://www.findarticles.com/cf_0/m0961/ 1999_Annual/54442299/print.jhtml

McClelland D E, Foltz R B, Wilson W D, Cundy T W, Heinemann R, Saurbier J A, and Schuster R L 1997 Assessment of the 1995 and 1996 floods and landslides on the Clearwater National Forest: Part I, Landslide assessment. Unpublished report (to Regional Forester, Northern Region, U.S. Forest Service)

McSweeney K, Gessler P E, Slater B, Hammer R D, Bell J, and Petersen G W 1994 Towards a new framework for modeling the soil-landscape continuum. In Amundson R (ed) *Factors of Soil Formation: A Fiftieth Anniversary Retrospective*. Madison, WI, Soil Science Society of America Special Publication No 33: 127–45

Mladenoff D J, Sickley T A, Haight R G, and Wydeven A P 1995 A regional landscape analysis and prediction of favorable gray wolf habitat in the northern Great Lakes regions. *Conservation Biology* 9: 279–94

Montgomery D R 1994 Road surface drainage, channel initiation, and slope stability. *Water Resources Research* 30: 1925–32

Montgomery D R and Dietrich W E 1994 A physically based model for the topographic control on shallow landsliding. *Water Resources Research* 30: 1153–71

Montgomery D R, Sullivan K, and Greenberg H M 1998 Regional test of a model for shallow landsliding. *Hydrological Processes* 12: 943–55

Moore I D, Gessler P E, Nielsen G A, and Peterson G A 1993 Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* 57: 443–52

Swanson F J and Dryness C T 1975 Impact of clear-cutting and road construction of soil erosion by landslides in the western Cascade Range, Oregon. *Geology* 3: 393–6

Swets J A 1988 Measuring the accuracy of diagnostic systems. *Science* 240: 1285–93

Turner A K and Schuster R L 1996 *Landslides: Investigation and Mitigation*. Washington, DC, Transportation Research Board Special Report No 247

Varnes D J 1984 *Landslide Hazard Zonation: A Review of Principles and Practice*. Paris, UNESCO Press

Williams C J, Lee S S, Fisher R A, and Dickerman L H 1999 A comparison of statistical methods for prenatal screening for Down syndrome. *Applied Stochastic Models and Data Analysis* 15: 89–101

Wu W and Sidle R C 1995 A distributed slope stability model for steep forested basins. *Water Resources Research* 31: 2097–110