Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Modeling Spatial Landslide Probability in Volcanic Terrains Through Continuous Neighborhood Spatial Analysis and Multiple Logistic Regression in La Ciénega Watershed, Nevado de Toluca, Mexico

Rutilio Castro-Miguel (rutilio.cm@gmail.com)
Universidad Nacional Autonoma de Mexico https://orcid.org/0000-0001-6454-091X

Gabriel Legorreta Paulín
Universidad Nacional Autonoma de Mexico: Instituto de Geografia

Roberto Bonifaz-Alfonzo
Universidad Nacional Autonoma de Mexico: Instituto de Geofisica

José Fernando Aceves Quesada
Universidad Nacional Autonoma de Mexico: Instituto de Geografia

Miguel Ángel Castillo-Santiago
El Colegio De La Frontera Sur - Unidad San Cristóbal: El Colegio de la Frontera Sur

Research Article

Keywords: Spatial models, neighborhood analysis, Multiple Logistic Regression, landslides.

Posted Date: August 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-581705/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Abstract

Little work has been done on the effect of the pixel neighborhood information when modeling landslide susceptibility using Multiple Logistic Regression (MLR). This research uses in situ and neighborhood cartographic information to evaluate how pixel distance of sampling sites affects the precision and accuracy of the MLR landslide susceptibility model. Two landslide susceptibility models are used: MLR-in situ, calibrated and validated by using variables that are collected at the site of the sampling point; and MLR in combination with Continuous Neighborhood Spatial Analysis (CNSA) to incorporate a search radius to extract pixel values for each cartographic variable based on a distance ratio. La Ciénega watershed on the eastern flank of the volcano Nevado de Toluca is selected as a study area. Its climate, topography, geomorphology, and geology predispose it to episodic landslides. The resulting susceptibility maps are validated in terms of the area under the curve (AUC) of the receiver operating characteristic (ROC), and they are compared with an inventory map in a contingency table; the MLR-CNSA model yields the better spatial prediction and representation of landslide susceptibility. The AUC evaluation indicates a predictive capability for the MLR-CNSA model of 0.969.

1. Introduction

Production of spatial models for landslide susceptibility assessment in geographic studies has constantly evolved as a consequence of the development of earth science, as well as of the computing capacity required in the development of new spatial analysis functions and the ability to process a large amount of data (Álvarez et al. 2003; Yang et al. 2019; Zambrano et al. 2017). The need to identify and map areas affected by landslides, and to model and map potential future landslide areas, arises from the fact that these gravitational processes (GPs) represent one of the main natural hazards that cause human and economic losses (Meten et al. 2014; Nourani and Ghaffari 2012; Rodríguez et al. 2006). Cartographic representation of landslide susceptibility allows planned territorial actions to reduce the negative effects that these processes might have on the population and its economic activities.

Areas susceptible to landslides have been identified by analysis of the spatial distribution and frequency of landslides (Hervás and Bobrowsky 2009; Ko and Lo 2018), Multiple Logistic Regression (MLR) (Esper 2013; Legorreta et al. 2014; Mirnazari et al. 2014; Zhao et al. 2019), multicriteria analysis (Dragićević et al. 2015; Legorreta et al. 2016), and fuzzy logic (Dehban 2013; Leonardi et al. 2016, 2020). MLR is widely applied when the outcome is dichotomous.

The model estimates the landslide probability between 0 and 1, and is fairly successful if the model is based on an adequate sample strategy, sufficient samples, and a set of variables that are related to the processes that trigger landslides (Ayalew and Yamagishi 2005; Legorreta et al. 2016; Yang et al. 2019). The model is calibrated by using variables that are collected at the site of the sampling point (herein termed MLR-in situ) (Abdulah and Yulianti 2015; Eskandarj and Chuvieco 2015; Hair et al. 1999; Tayyebi et al. 2010; Xiong and Zuo 2018); this implies an isolation of the point where the information is acquired in relation to its environment, since values from the surrounding area are not considered (Castro 2020; Castro and Legorreta 2019). The isolation of the sites excludes a large amount of spatial data from the study area, resulting in the calibration of the model with a minimal amount of information. To address the above deficiency, MLR is combined with Continuous Neighborhood Spatial Analysis (CNSA), which integrates information from the neighboring areas of the sampling sites in unstable and stable terrains (Castro 2017, 2020; Castro and Legorreta 2019). This model is herein called MLR- CNSA.

Landslide susceptibility studies that apply MLR summarize the error and success of prediction using the AUC of the ROC curve and contingency table values. In landslide susceptibility conducted in other studies, AUC values range between 0.670 and 0.940 and overall accuracy values of contingency table range between 74.2% and 86.1% (Goyes-Peñafiel y Hernandez-Rojas 2021; Smith et al. 2021;Yang et al. 2019; Wubalem y Meten 2020).

In Mexico, several studies have assessed and modeled the probability of landslide occurrence. By combining weighted independent variables with map algebra, and heuristic methods with hierarchical analysis, landslide susceptibility on the Sierra Norte, Chiapas, Mexico has been assessed (Peña 2013). By applying heuristic-geomorphological criteria, volcanic landforms were used to establish landslide frequency, landslide susceptibility, and landslide volumes in watersheds along the southwest flank of the volcano Pico de Orizaba, Mexico (Legorreta et al. 2012, 2015). Also for Pico de Orizaba, three landslide susceptibility models (Stability Index MAPping, MLR, and multicriteria evaluation) were trained with in situ variables, and their results were compared and contrasted (Legorreta et al. 2016). In the study, the MLR model was preferred over the other two because of its close match with the landslide inventory (79.81% producer's accuracy), and greater predictive capability with AUC of 0.831 and overall accuracy of 71.24 % (Legorreta et al. 2016).

In the study area, La Ciénega watershed, Mexico, identification of potential areas for landslides and debris flows by multicriteria analysis has shown a 75% match with the landslide inventory (Aceves et al. 2014a), and landslide frequency per landform has been analyzed (García et al. 2018). MLR with CNSA evaluated the slope values from the in situ focus to a distance of 20 pixels in order to select the size of the neighborhood area that would produce the best probability model (Castro and Legorreta 2019). However, there has been no assessment of the efficacy of using multiple predictive variables in the MLR-CNSA model to predict landslide susceptibility. In this study, landslide susceptibility is derived with reference to 16 cartographic variables in the La Ciénega watershed, obtained from the in situ pixel and, for the CNSA model, from the neighborhood pixels. Validation of the resulting landslide susceptibility map by the Receiver Operating Characteristic (ROC) curve and a contingency table demonstrates that use of the values of neighborhood pixels enhances the performance of the MLR in the prediction of landslides.

2. Study Area

La Ciénaga watershed is a small endorheic watershed of ~ 3 000 ha on the eastern flank of the volcano Nevado de Toluca in central Mexico; this volcano is the fourth most important elevation in Mexico with an approximate altitude of 4 680 m a.s.l. (Macías 2005). The La Ciénega watershed is at 19°7´10.7´´–19°4´21.8´´ N and 99°37´28.7´´–99°44´40.7´´ W. The La Ciénega river is a sub-basin of the Almoloya-Otzolotepec watershed – a tributary of the Lerma-Santiago hydrographic region which flows into the Pacific Ocean (Aceves et al. 2014b; INEGI 2010). The watershed is elongate with a length of 13 km and average width of 2.5 km (Aceves et al. 2014a) (Fig. 1). Pyroclastic flow deposits are widely spread on the eastern flank of the volcano, filling the stream valleys. These streams in the watershed start in old glacial cirque and continue through a long dendritic drainage with an average 100–300 m depth. The slope is 20–35° in

the elevated portions, 6–12° in the piedmont and 2–6° on the plain. In the middle portion of the watershed, lahar deposits cover block and ash flows that form massive units interstratified with surge horizons (Aceves et al. 2014a, b). To the east, the volcanic piedmont continues with active fans composed of pyroclastic materials and alluvial sediment (Capra and Macías 2000; Espinosa-Rodríguez et al. 2014; García-Palomo et al. 2000). The landslide areas frequently create debris flows affecting the town of Santa Cruz Pueblo Nuevo. The town was partially destroyed on 28 June 1940 by a large debris flow. Today, the town is settled in the alluvial fan of an old debris flow deposit. It is highly probable that the town will be affected again by landslides and debris flows (Aparicio and Pérez 2014; Capra et al. 2008; García et al. 2018). In the study area, the mean annual precipitation, the bulk of which falls in summer (May to September), is 1 200 to 1 100 mm/yr at > 4 000 m a.s.l, and 1 100 to 800 mm/yr at elevations of 2 650-4 000 m a.s.l. (García 2004).

3. Materials And Method

A detailed landslide inventory was based on fieldwork and interpretation of satellite images (SPOT-6), and on a compilation of pre-existent landslide inventories. Fieldwork was conducted in year 2015 and 2016 along the main river and some tributary rivers. Field data recorded landslide type, size (length, height, and depth), and location in relation to the river. For the landslide mapping and classification, we followed the landslide hazard zonation protocol (2006) of the Forest Practices Division, Department of Natural Resources, Washington State, Cruden & Varnes (1996), and Wieczorec (1984). Landslides were classified into shallow landslides, debris flow, debris slides, deep-seated landslides, incised meanders, and rock falls; interpretation and mapping used Spot 6 from the year 2015 with spatial resolution of 1.5 m in panchromatic mode and of 6 m in multispectral mode (SIAP-SEDENA 2016). Some landslides were added from the database of Álvarez (2015). With the landslide location mapped, a point coverage was produced in the Arc/Info® program version 10.0. Positioning the points corresponding to landslides over the satellite image, the landslide headscarps were identified, verified, and digitized (Fig. 2).

Two classes of dependent variables were recognized. Landslide headscarp areas and non-landslide areas were rasterized and reclassified in a dichotomous way: landslide areas as 1 (1 299 pixels, 0.97% of the total watershed area) and non-landslide areas as 0 (132 390 pixels, 99.03% of the total watershed area) with a spatial resolution of 15 m per pixel (Fig. 3).

We used stratified random sampling of sites in landslide areas and non-landslide areas to conduct MLR-in situ and MLR-CNSA analysis (McGrew and Monroe 2000; Pardo and Ruíz 2002). Of 512 sites selected, 384 (75%) were used for the calibration of the models and 128 (25%) for model validation. The same number of sample sites was sampled from the landslide areas and non-landslide areas: 192 from each class for the calibration, and 64 from each for the validation.

Thematic maps and second-derivative products at a scale of 1:50 000 were used as the independent variables. The thematic cartography was standardized to a raster format with a spatial resolution of 15 m per pixel. The 16 independent cartographic variables were processed and coded according to the type of data that they represented: six were qualitative (land use, land use change, geomorphology, edaphology, lithology, and soil texture) (Table 1) and the other ten were quantitative (altitude, terrain steepness, down-slope direction, terrain surface curvature, terrain roughness, slope length, distance from faults and fractures, distance from roads, and normalized difference vegetation index) (Table 2). For the MLR-in situ analysis, the qualitative cartographic variables were coded as categorical (string variable) and the quantitative variables by using their raw values (numeric variable). For the MLR-CNSA analysis, the qualitative cartographic variables were coded using dummy variables. Dummy coding is a way of representing groups of a cartographic qualitative variable using only zeros and ones. To do this, we created several variables, one for each cartographic variable class. For instance, the qualitative cartographic variable of land use change had 3 classes: permanence, changes, and deforestation. For each of these classes, a dummy variable with 0 and 1 was created (Table 1, Fig. 4). For the quantitative cartographic variables, the raw values were used (Table 2, Fig. 5).

Table 1
Independent variables of qualitative type

Variable / Class	Variable / Class
Land use 2015 (Castro, 2020)	**Geomorphology** (García, 2017)
Human settlements	Dome 1
Farming	Dome 2
Fragmented forest	Dome 3
Semi-dense forest	Dome 4
Dense forest	Nevado de Toluca Cone
Natural grassland	Upper Hillslope
Induced grassland	Lower Hillslope
Without vegetation	Lava Flow Mesa
Other types of vegetation	Dome 5 partially buried
Land use change 1983–2015 (Castro, 2020)	Lava hillslopes covered with pumice and ash
Permanence	Debris avalanche hillslopes covered with pyroclast
Changes	Pumice hillslopes and fall deposits
Deforestation	Pumice hillslopes, block and ash
Edaphology (INEGI, 2001a)	Lava hillslopes covered with pyroclast
Humic Andosol	Falling pumice hillslopes and pyroclastic flows
Molic Andosol	Pyroclastic hillslopes
Ochric Andosol	Lava flow covered with pumice and ash flows
Haplic Pheozems	Lava flow partially covered with pyroclast
Eutric Fluvisol	Alluvial fan
Leptosol	Floodplain
Eutric Regosol	**Lithology** (INEGI, 2001b)
Soil texture (INEGI, 2001a)	Intermediate extrusive igneous
Gross	Alluvial
Medium	Volcanic breccia
	Tuff
	Andesite
	Basalt

Table 2
Independent variables of quantitative type

Variable / Units	Variable / Units
Altitude m a.s.l. (INEGI, 2013a; ESRI, 2016)	**Slope length** (INEGI, 2013a; FSFI, 2016)
Units meters above sea level.	Units meters.
Terrain steepness (INEGI, 2013a; ESRI, 2016)	**Distance from faults and fractures** (INEGI, 2001b; ESRI, 2016)
In degrees.	Units meters.
Down-slope direction (INEGI, 2013a; ESRI, 2016)	**Distance from rivers** (INEGI, 2010; ESRI, 2016)
Expressed in azimuthal measure; the value − 1 corresponds to flat areas.	Units meters.
Terrain surface curvature (INEGI, 2013a; ESRI, 2016)	**Distance from roads** (INEGI, 2013b; ESRI, 2016)
Positive values for convex surfaces, negative values for concave surfaces, and 0 for flat areas.	Units meters.
Terrain Roughness (INEGI, 2013a; Sappington et al., 2007)	**Normalized Difference Vegetation Index (NDVI)** (SIAP-SEDENA, 2016; ESRI, 2016)
values between 0 and 1, where 0 corresponds to flat areas, 1 corresponds to maximum roughness.	Values between − 1 to 1; values close to -1 correspond to water or snow and values close to 1 indicate vegetation with a high chlorophyll content.

A multicollinearity diagnostic was calculated for the quantitative variables by using the variance inflation factor (VIF). A VIF value greater than 10 is indicative of a serious multicollinearity problem (Field 2013).

MLR-in situ was calculated with the values extracted from the 16 thematic variables for the sample sites. To extract the values, 384 sites corresponding to the calibration sample were coded 1 in areas with landslides and 0 in areas with no landslides. Using the values, an intercept β_0 and β coefficients were calculated by using logistic function (Eq. 1) under SPSS (Pardo and Ruíz 2002).

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}} \qquad [1]$$

where:

Y is the probability that a process will occur.

X_1 to X_k are the independent variables that are part of the model.

β_0 is the constant in the model coefficients.

β_1 to β_k are the independent variables coefficients.

With SPSS, forward MLR examined the variables in the model to see whether any variable should be removed. This method begins the model with only the constant β_0, then selects the independent variable that has the highest simple correlation with the outcome. If this variable significantly improves the ability of the model to predict landslides, then this predictor is retained in the model and the computer searches for a second predictor. The process is repeated until all variables have been evaluated (Spiegel and Stephens 2009).

The MLR-CNSA analysis required, first, use of the Focal Statistics tool in ArcMap for a search radius to extract pixel values for each cartographic variable based on a distance ratio. The distance ratio allowed consideration of the values f the neighborhood pixels that surround a landslide. The specific neighborhood in our case was concentric circles with radii from 1 pixel to 20 pixels. For the qualitative cartographic variables, the values assigned to each variable class was the sum of values within each concentric circle; to do this, we used the Focal Statistic in ArcMap to calculate the Sum statistic for the maps of dummy variables with 0 and 1. For the quantitative cartographic variables, the value assigned to each variable class was the mean of values within each concentric circle; this used the mean function of the Focal Statistics tool in ArcMap. A data matrix was obtained for each neighborhood area analyzed, from a 1-pixel radius to 20 pixels. The matrix thereby generated contained the data corresponding to the sampling site and the number of pixels of each class within the neighborhood area that made up the variable (Tables 3 and 4, and Fig. 6). Hence, a unique ASCII matrix was obtained for each cartographic variable, leading to 320 matrices (16 thematic variables x 20 neighborhood areas).

Table 3
Example of qualitative matrices from the CNSA for the land-use change variable at distances of 1 and 20 pixels

Sum of the CNSA at 1-pixel diameter					Sum of the CNSA at 20-pixels diameter				
Site	Process	Permanence	Changes	Deforestation	Site	Process	Permanence	Changes	Deforestation
230	0	5	0	0	230	0	1 257	0	0
231	0	0	0	5	231	0	941	0	316
232	0	5	0	0	232	0	1 074	0	183
489	1	3	0	2	489	1	941	0	316
490	1	1	0	4	490	1	966	0	291

Table 4
Example of quantitative matrices from the CNSA for the altitude variable at distances of 1 and 20 pixels

Mean of the CNSA at 1-pixel diameter			Mean of the CNSA at 20-pixels diameter		
Site	Process	Altitude	Site	Process	Altitude
230	0	2 888	230	0	2 880
231	0	3 075	231	0	3 066
232	0	3 046	232	0	3 011
489	1	2 981	489	1	3 012
490	1	2 975	490	1	3 008

The next stage in the MLR-CNSA analysis required selection of the neighborhood radius value. MLR was used to evaluate the relationship of the information obtained from the CNSA, and to develop the landslide probability model. Among the statistics of the MLR model obtained with SPSS, the − 2 log likelihood (-2 LL) assesses the level of probability from the model's result by comparing the observed cases with the expected ones (Martín et al. 2008). A smaller value of the − 2 LL indicator signifies a better adjustment for the statistical model, with 0 being the optimal value. The − 2 LL indicator can be compared between models, and the values show the level of adjustment from one model to another (Hair et al. 1999). Therefore, -2 LL can be used to evaluate the data obtained with the CNSA. For each variable, the distance of the neighborhood analysis that reports the best fit of the MLR model with the landslide process was selected, based on the lowest value of -2 LL within the range analyzed (Castro 2020; Castro and Legorreta 2019).

With the selected distances and their neighborhood values, a single matrix was integrated (Fig. 7) for the 16 independent variables. The conditional forward MLR model was applied to this matrix to create the model.

For an objective comparison between the MLR-CNSA and MLR-in situ models, a binary classification scheme was applied (landslide and non-landslide areas). The cut-off point between the two classes for each model was obtained by ROC curve analysis; this maximizes successes in both classification classes (Pardo and Ruíz 2002). Once the cut-off point had been obtained, spatial models of landslide probability were compared to identify possible advantages or disadvantages.

The accuracy in the predictive capability of the MLR-CNSA and MLR-in situ models was evaluated from the ROC curve; the area under the curve corresponded to the proportion of elements correctly classified with the MLR (Pardo and Ruíz 2002). The model maps were evaluated with a contingency table obtained from the overlay between inventory and models; the evaluation was made for the enlarged sample. The contingency table evaluates the classificatory capacity of the discriminant functions such as MLR (Hair et al. 1999). The contingency table yielded the kappa index that expressed the degree of agreement between the real group and predicted group (Pardo and Ruíz 2002). Other statistics calculated were as follows: the overall accuracy, which is the percentage of elements correctly classified (Nemmaoui et al. 2013; Pardo and Ruiz 2002); the producer’s accuracy, which is the ratio of elements correctly classified in a class divided by the total of real elements of the same class (Legorreta et al. 2012); the user’s accuracy, which is the ratio between the elements correctly classified in a class and the total elements classified by the model in that same category (Legorreta et al. 2012); and the model efficiency, which is the ratio of correctly classified elements of a class, minus the incorrectly predicted elements of the same category, divided by the total elements of the real group (Legorreta et al. 2012). In all of these statistics, the MLR-CNSA model had better results, which reflected its more exact spatial representation with respect to the landslides identified in the field.

4. Results

During the assessment of the La Ciénega watershed, a representative sample of 290 landslide features was inventoried (Fig. 9). The landslide area affected 4.8 ha of the watershed. Of the 3 types identified, 250 were debris (86% of the total gravitational processes in the study area, and the predominant mass-wasting feature along the main river and tributaries), 33 (11%) were debris flows, and 7 (3%) were deep-seated landslides.

The assessment of the independent variables with the CNSA identified the radius (in pixels) of the neighborhood area for each variable that best fit the MLR model, as follows: land-use 4 pixels, land-use change 15, geomorphology 20, lithology 20, pedology 4, soil texture 20, altitude 20, terrain steepness 5, down-slope direction 20, terrain surface curvature 6, terrain roughness 2, slope length 4, distance to faults and fractures 1, distance from a river 2, distance from a road 9, and Normalized Difference Vegetation Index (NDVI) 20 (Fig. 10).

The VIF for the 10 quantitative variables showed that the variables can be used in the MLR analysis. These ten variables are strongly related to the probability of distribution for the dependent variable (landslide or non-landslide) but not strongly related to each other. After the assessment of multicollinearity, forward Multiple Logistic Regression showed that of the original 16 variables only 7 contributed to the model (Table 5a). Statistical tests (Table 5) showed the a-intercept and β coefficients to be strong enough to be reliable for the MLR-CNSA model.

Table 5 Variables in the MLR-CNSA model, and its classification levels based on the analysis of the neighborhood areas

a) Variables in the Equation

Variable/Class	β	S.E.	Wald	df	Sig.	Exp(β)	95% C.I. for Exp(β)	
							Lower	Upper
Land use change/Deforestation	0.016	0.007	5.493	1	0.019	1.016	1.003	1.029
Geomorphology/Upper Hillslope	0.002	0.001	12.462	1	0.000	1.002	1.001	1.004
Geomorphology/Debris avalanche hillslopes covered with pyroclast	0.005	0.001	10.792	1	0.001	1.005	1.002	1.008
Geomorphology/Pumice hillslopes and fall deposits	0.001	0.001	6.350	1	0.012	1.001	1.000	1.003
Soil texture/Gross	-0.002	0.001	6.882	1	0.009	0.998	0.997	1.000
Terrain surface curvature	-4.962	1.317	14.189	1	0.000	0.007	0.001	0.093
Distance from a river	-0.051	0.009	32.050	1	0.000	0.951	0.934	0.967
Constant	0.512	0.430	1.419	1	0.234	1.668		

b) Clasification table

Observed		Predicted		
		Process		% correct
		0	1	
Process	0	172	20	89.6
	1	20	172	89.6
Overall %				89.6
The cut-off value is 0.585				

For the MLR-in situ model, only three variables were retained in the model (Table 6): the direction of the slope, distance from faults and fractures, and distance from a river.

Table 6 Variables in the MLR- in situ model, and its classification levels based on the analysis of the neighborhood areas

a) Variables in the Equation

Variable/Class	β	S.E.	Wald	df	Sig.	Exp(β)	95% C.I. for Exp(β)	
							Lower	Upper
Down-slope Direction	0.010	0.003	15.213	1	0.000	1.011	1.005	1.016
Distance from faults and fractures	-0.003	0.001	17.713	1	0.000	0.997	0.996	0.998
Distance from a river	-0.057	0.008	57.746	1	0.000	0.944	0.931	0.958
Constant	1.871	0.376	24.784	1	0.000	6.492		

b) Clasification table

Observed		Predicted		
		Process		% correct
		0	1	
Process	0	165	27	85.9
	1	27	165	85.9
Overall %				85.9
The cut-off value is 0.652				

Based on these results (Tables 5 and 6) the logit function led to maps of landslide probability that could then be used through the ROC and the contingency table to evaluate the models vs inventory (Fig. 11); this entailed a binary classification into landslide and non-landslide, with the cut-off point selected according to the ROC curve: 0.586 for the MLR-CNSA model and 0.652 for MLR-in situ (Tables 5b and 6b).

The maps derived from the models (Figs. 12 and 13) showed 422.28 ha to be affected by landslides according to the MLR-CNSA model and 464.38 ha according to the MLR-in situ model. These models were used together with an enlarged sample area to generate the contingency table from which to calculate the kappa index, overall accuracy, producer's accuracy, user's accuracy, and model efficiency (Table 7).

Table 7a Evaluation of the model based on the CNSA

Clasification Table				
			Predicted Process	
			0	**1**
Observed	Process	**0**	58	6
		1	9	55

Kappa index	0.77
Overall accuracy	88.28%
Producer's accuracy	85.94%
User's accuracy	90.16
Model efficiency	0.77

Table 7b Evaluation of the model with in situ data

Clasification Table				
			Predicted Process	
			0	**1**
Observed	Process	**0**	55	9
		1	11	53

Kappa index	0.69
Overall accuracy	84.38%
Producer's accuracy	82.81%
User's accuracy	85.48
Model efficiency	0.69

The ROC curve indicated that prediction was better with the MLR-CNSA model (96.8%) than with the MLR-in situ model (89.8%). Evaluation in terms of error and success from many locations showed overall accuracy to be higher with the MLR-CNSA model (88.28%) than with the MLR-in situ model (84.38%), a conclusion supported by all the other statistics (kappa index, producer's accuracy, user's accuracy, and model efficiency).

5. Conclusions

Landslide susceptibility is difficult to model not only because of the continuous changes in the topography caused by landslides and environmental conditions in watersheds that are large, remote, and difficult to access, but also because the calibration of a model may consider only local or in situ landslide information. The main goal of this study was to model and assess landslide susceptibility by using MLR-CNSA to integrate information from the neighboring areas; this integration supplied complementary information that enhanced the outcome of the modeling. Use of CNSA identifies the extent of neighborhood areas that will lead to a more significant relationship with the landslide process, and hence to a more accurate probability model. This study is a first step towards a more comprehensive research into landslide susceptibility assessment on one of the highest volcanoes in Mexico.

For the la Ciénega watershed, validation using the kappa index, overall accuracy, accuracies, and model efficiency showed that MLR-CNSA achieved a closer match with the landslide areas than did the MLR-in situ model, a conclusion that was supported by field validation. The MLR-in situ model mapped 9.9% more potential landslide area (464.38 ha) than the MLR-CNSA model (422.28 ha). Field validation showed that in areas with few landslides the in situ model tended to overpredict, whereas the MLR-CNSA model produced a map that indicated the lower potential for landslides. The results point out that the MLR-CNSA model can be used to make more accurate landslide susceptibility assessment. The obtained AUC value with MLR-CNSA model is higher than those reported AUC values that range from 0.670 to 0.940 in recent studies (Goyes-Peñafiel and Hernandez-Rojas 2021; Legorreta et al. 2016; Smith et al. 2021; Wubalem and Meten 2020; Yang et al. 2019). Regarding the spatial validation using contingency table, Yang et al. (2021) and Legorreta et al. (2016) report overall accuracy of 86.1 % and 71.24 % respectively, while the MLR-CNSA model reports value of 88.28. The landslide probability model developed in the present work reports more accurate values in both statistical and spatial validation compared to other studies that use MLR-in situ calibrated data.

The CNSA assessment used a circular distance neighborhood and the sum and average function; as expected, this improved the representation of landslide susceptibility. However, the option of selecting a different geometric distance to collect the spatial information and of using other functions needs to be evaluated. Also, the application of MLR-CNSA is not restricted to gravitational processes but could evaluate deforestation, flood risk, or species distribution, etc. The use of CNSA is also not limited to a MLR model, but can be extended to other spatial statistics.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

6. Declarations

Funding. The authors did not receive support from any organization for the submitted work.

Conflicts of interest/Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material. Not applicable

Code availability. Not applicable

Ethics approval. Not Applicable

Consent to participate. Not applicable

Consent for publication. Not applicable

7. References

1. Abdulah L, Yulianti M (2015) A spatial modeling approach with logistics binary regression for bamboo forest distribution mapping. 10th World Bamboo Congress, Korea. 1-9, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwj5ydGDhO3vAhVEQq0KHZM5Bv8QFjABegQIBRAD&url=https%3A%2F%2Fwww.worldbamboo.net
2. Aceves JF, Legorreta G, Álvarez-Ruíz Y (2014a) Gravitational processes on the eastern flank of the Nevado de Toluca, México. Zeitschrift für Geomorphologie 58:185–200
3. Aceves JF, Legorreta G, Álvarez-Ruíz Y (2014b) Cartografía geomorfológica para el inventario de procesos gravitacionales en la cuenca endorreica del arroyo La Ciénega, flanco oriental del volcán Nevado de Toluca. Boletín de la Sociedad Geológica Mexicana 66 2:329-342
4. Álvarez R, Bonifaz R, Lunetta RS, García G, Gómez G, Castro R et al (2003) Multitemporal land-cover classification of México using Landsat MSS imagery. International Journal Remote Sensing 24:2501-2514
5. Álvarez Y (2015) Relación entre los deslizamientos y cambios en la cubierta de uso de suelo y vegetación entre los años 1983 y 2014 en el arroyo la Ciénega, volcán Nevado de Toluca, México. Dissertation, Universidad Nacional Autónoma de México
6. Aparicio AT, Pérez VV (2014) Representaciones sociales del desastre de 1940 en Santa Cruz Pueblo Nuevo, Estado de México. Investigaciones Geograficas. https://doi.org/10.14350/rig.36918
7. Ayalew L, Yamagishi H (2005) The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko mountains, central Japan. Geomorphology 65:15-31
8. Capra L, Macías JL, (2000) Pleistocene cohesive debris flows at Nevado de Toluca Volcano, central Mexico. Journal of Volcanology and Geothermal Research 102:149–167
9. Capra L, Norini G, Groppelli G, Macías JL, Arce JL (2008) Volcanic hazard zonation of the Nevado de Toluca volcano, México. Journal of Volcanology and Geothermal Research 176 4:469–484
10. Castro R (2017) Modelo espacial de probabilidad a la deforestación en bosques para el estado de Oaxaca. Dissetation, Universidad Nacional Autónoma de México
11. Castro Miguel R, Legorreta Paulín G (2019) Análisis Espacial de Vecindad Continua y regresión logística en el modelado espacial de probabilidad de ocurrencia de deslizamientos. Investigaciones Geográficas. https://doi.org/10.14350/rig.59760
12. Castro R (2020) Aplicación del Análisis Espacial de Vecindad Continua al modelado de probabilidad a deslizamientos en terrenos volcánicos utilizando la Regresión Logística Múltiple. Dissertation, Universidad Nacional Autónoma de México
13. Cruden DM, Varnes DJ (1996) Landslide types and processes, In: Turner, A. K., Schuster, R.L. (ed.) Landslides; Investigation and Mitigation, Transportation Research Board; Special Report 247, National Academy Press, Washington D.C., pp 6-75
14. Dehban AS (2013) Landslide susceptibility mapping for Poulrood Earth Fill dam reservoir (The comparison of two methods) Iran, Guilan Province. Universal Journal of Geoscience 1 2: 69-79
15. Dragićević S, Lai T, Balram S (2015) GIS-based multicriteria evaluation with multiscale analysis to characterize urban landslide susceptibility in data-scarce environments. Habitat international 45:114-125
16. Esper Angillieri, M. Y. (2013) Debris flow susceptibility mapping in a portion of the Andes and Preandes of San Juan, Argentina using frequency ratio and logistic regression models. Earth Sciences Research Journal 17: 159–167
17. Espinosa-Rodríguez LM, Balderas-Plata MA, Cabadas-Báez H V (2014) Caracterización geomorfológica del área natural protegida Nevado de Toluca: complejo de volcanes Nevado de Toluca y San Antonio. CienciaUAT 9 1:6-14
18. ESRI (Environmental Systems Research Institute) (2016) ArcGis desktop. http://desktop.arcgis.com/en/arcmap/
19. Eskandari S, Chuvieco E (2015) Fire danger assessment in Iran based on geospatial information. International Journal of Applied Earth Observation and Geoinformation 42:57–64
20. Field A (2013) Discovering Statistics Using IBM SPSS Statistics. SAGE, London
21. FSFI (Free Software Foundation Inc.), 2016. QGIS 3.6.3-Noosa. ULR: https://qgis.org/en/site/
22. García E (2004) Modificaciones al sistema de clasificación climática de Koppen. Instituto de geografía, Mexico City
23. García S (2017) Susceptibilidad geomorfológica a deslizamientos en arroyos secundarios del sur de la cuenca la Ciénega, volcán Nevado de Toluca. Dissertation, Universidad Nacional Autónoma de México
24. García Reyes S, Legorreta Paulín G, Aceves Quesada F (2018) Susceptibilidad geomorfológica a procesos gravitacionales en arroyos secundarios del sur de la cuenca La Ciénega, volcán Nevado de Toluca. Investigaciones Geográficas. https://doi.org/10.14350/rig.59613
25. García-Palomo A, Macías JL, Garduño VH (2000) Miocene to recent structural evolution of the Nevado de Toluca volcano region, Central Mexico. Tectonophysics 318:281–302.
26. Goyes-Peñafiel P, Hernandez-Rojas A (2021). Landslide susceptibility index based on the integration of logistic regression and weights of evidence: A case study in Popayan, Colombia. Engineering Geology 280(2021):1-9. . https://doi.org/10.1016/j.enggeo.2020.105958
27. Hair J, Anderson E, Tatham R, Black W (1999) Análisis Multivariante, 5th edn. Pearson, Madrid
28. Hervás J, Bobrowsky P (2009) Mapping: Inventories, susceptibility, hazard and risk, In: Sassa K, Canuti P (ed.) Landslides - Disaster Risk Reduction. Springer, Berlin, pp 21-349
29. INEGI (National Institute of Statistics and Geography) (2001a) Carta edafológica en escala 1:50.00. http://www.inegi.org.mx.
30. INEGI (2001b) Carta geológica en escala 1:50,00. http://www.inegi.org.mx.
31. INEGI (2010) Red hidrográfica escala 1:50,000. Documento técnico descriptivo de la red hidrográfica escala *1:50,000*. http://www.inegi.org.mx.

32. INEGI (2013a) Continuo de Elevación Mexicano 3.0 (CEM 3.0). http://www.inegi.org.mx.
33. INEGI (2013b) Conjunto de datos vectoriales de información topográfica en escala 1:50,00. Serie III. http://www.inegi.org.mx.
34. Ko FW, Lo FL (2018) From landslide susceptibility to landslide frequency: A territory-wide study in Hong Kong. Engineering geology 242:12-22
35. Legorreta G, Bursik M, Aceves JF, Paredes LM, Lugo J, Alcántara I, López J (2012) Evaluación práctica y teórica de los modelos de regresión logística múltiple y cartográfica-hidrológica para la predicción de deslizamientos de tierra, In: Quintero JA (Coor.) Temas selectos de Geomática. Instituto de Geografía, Mexico City, pp 47-76
36. Legorreta G, Bursik M, Orózco JJ, García JE (2015) Landslide susceptibility of volcanic landforms in the Río El Estado watershed, Pico de Orizaba volcano, Mexico. Natural Hazards 77:559-574
37. Legorreta G, Pouget S, Bursik M, Quesada JF, Contreras T (2016) Comparing landslide susceptibility models in the Río El Estado watershed on the SW flank of Pico de Orizaba volcano, Mexico. Natural Hazards 80:127-139. https://doi.org/10.1007/s11069-015-1960-y
38. Leonardi G, Palamara R, Cirianni F (2016) Landslide susceptibility mapping using a fuzzy approach. Procedia engineering 161:380-387
39. Leonardi G, Palamara R, Suraci F (2020) A fuzzy methodology to evaluate the landslide risk in road lifelines. Transportation research procedia 45:32-739
40. Macías JL (2005) Geología e historia eruptiva de algunos de los grandes volcanes activos de México. Boletín de La Sociedad Geológica Mexicana 57:379–424
41. Martín Q, Cabero MT, Paz YR (2008) Tratamiento estadístico de datos con SPSS. Paraninfo, Madrid.
42. Meten M, PrakashBhandary N, Yatabe R (2014) The application of weights of evidence modeling for landslide susceptibility mapping of Dejan-GohaTsiyon transect in the Blue Nile Gorge, central Ethiopia. International symposium. Goehazard: Science, engineering and management, Kathmandu, Nepal, 62-76
43. McGrew JC, Monroe CB (2000) An Introduction to Statistical Problem Solving in Geography, 2nd edn. McGraw Hill, United States of America
44. Mirnazari J, Ahmad B, Mojaradi B, Sattari F (2014) Using frequency ratio for spatial landslide prediction. Research Journal of Applied Sciences, Engineering and Technology 7 15:3174-3180
45. Nemmaoui A, García-Lorca AM, Aguilar FJ, Aguilar MA (2013) Series temporales de vegetación para un modelo forestal destruido: El caso de Tadla Azilal (Marruecos). Agrociencia 47: 267-280
46. Nourani V, Ghaffari H (2012) Landslide susceptibility mapping with analytical hierarchical process and GIS techniques (case study: Zonouz, East Azerbaijan, IRAN). 10th International congress on advances in civil engineering, Middle east technical university, Turkey. 1-10. https://www.researchgate.net/publication/269411195_Landslide_susceptibility_mapping_with_analytical_hierarchical_process_and_GIS_techniques_case
47. Pardo A, Ruiz MA (2002) SPSS 11. Mc Graw Hill, Madrid
48. Peña IC (2013) Modelo cuantitativo para inestabilidad de laderas en un ambiente SIG, aplicado en la Sierra Norte de Chiapas. Dissertation, Universidad Nacional Autónoma de México
49. Rodríguez S, Mora-González I, Murrieta H (2006) Flujos de baja concentración asociados con lluvias de intensidad extraordinaria en el flanco sur del volcán Pico de Orizaba (Citlaltepetl), Mexico. Boletín de la Sociedad Geológica Mexicana 2:223-236
50. Sappington JM, Longshore KM, Thomson DB (2007) Quantifying landscape ruggedness for animal habitat analysis: A case study using Bighorn sheep in the Mojave Desert. Journal of Wildlife Management 7 5:1419-1426
51. SIAP-SEDENA (Servicio de Información Agroalimentaria y Pesquera – Secretaria de la Defensa Nacional) (2016) Spot6_MS and Spot6_P 2015 image
52. Smith HG, Spiekermann R, Betts H, Neverman AJ (2021). Comparing methods of landslide data acquisition and susceptibility modelling: Examples from New Zealand. Geomorphology 381: 1-17. https://doi.org/10.1016/j.geomorph.2021.107660
53. Spiegel MR, Stephens LJ (2009) Estadística, 4th edn. McGraw-Hill/Interamericana, México
54. Tayyebi A, Reza M, Yazdanpanah M, Pijanowsky B, Saeedi S, Hossein A (2010) A spatial logistic regression model for simulating land use patterns: A case study of the Shiraz Metropolitan Area of Iran, In: Chuvieco, E., Jonathan, L. y Xiaojun, Y. (ed.), Advances in Earth Observation of Global Change, Springer
55. Wieczorek GF (1984) Preparing a detailed landslide inventory map for hazard evaluation and reduction. Bulletin of the Association of Engineering Geologists XXI 3:337–342
56. WSDNR (Washington State Department of Natural Resources) (2006) Landslide Hazard Zonation Project Protocol version 2.1. Washington State Department of Natural Resources, Washington, U.S.A. https://www.dnr.wa.gov/publication/fp_lhz_protocol_v2_1_final.pdf?ee5umyf
57. Wubalem A, Meten M. (2020). Landslide susceptibility mapping using information value and logistic regression models in Goncha Siso Eneses area, northwestern Ethiopia. SN Applied Sciences 2(5):1–19. https://doi.org/10.1007/s42452-020-2563-0
58. Xiong Y, Zuo R (2018) GIS-based rare events logistic regression for mineral prospectivity mapping. Computers and Geosciences 111:18–25
59. Yang J, Song C, Yang Y, Xu C, Guo F, Xie L (2019) New method for landslide susceptibility mapping supported by spatial logistic regression and GeoDetector: A case study of Duwen Highway basin, Sichuan province, China. Geomorphology 324:62–71. https://doi.org/10.1016/j.geomorph.2018.09.019
60. Zambrano L, Pacheco-Muñoz R, Fernández T (2017) A spatial model for evaluating the vulnerability of water management in Mexico City, Sao Paulo and Buenos Aires considering climate change. Anthropocene 17:1–12
61. Zhao Y, Wang R, Jiang Y, Liu H, Wei Z (2019) GIS-based logistic regression for rainfall-induced landslide susceptibility mapping under different grid sizes in Yueqing, Southeastern China. Engineering Geology 259:1051-47

Figures

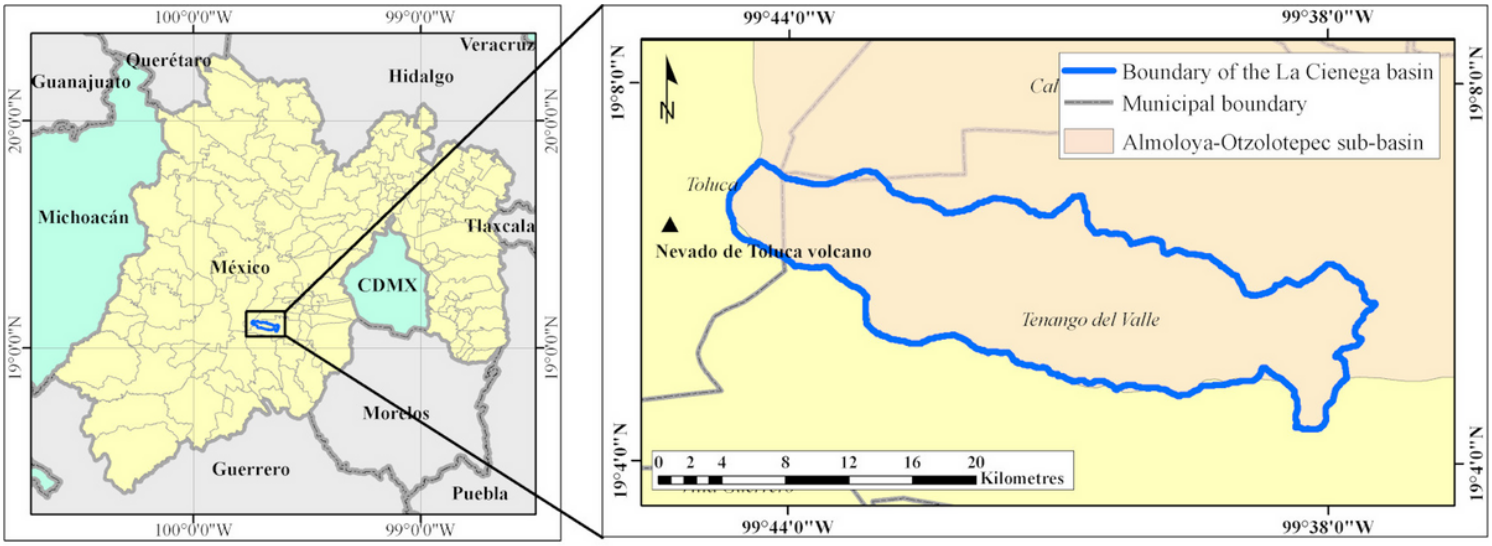


Fig. 1 Study area in central Mexico

Figure 1

See image above for figure legend.

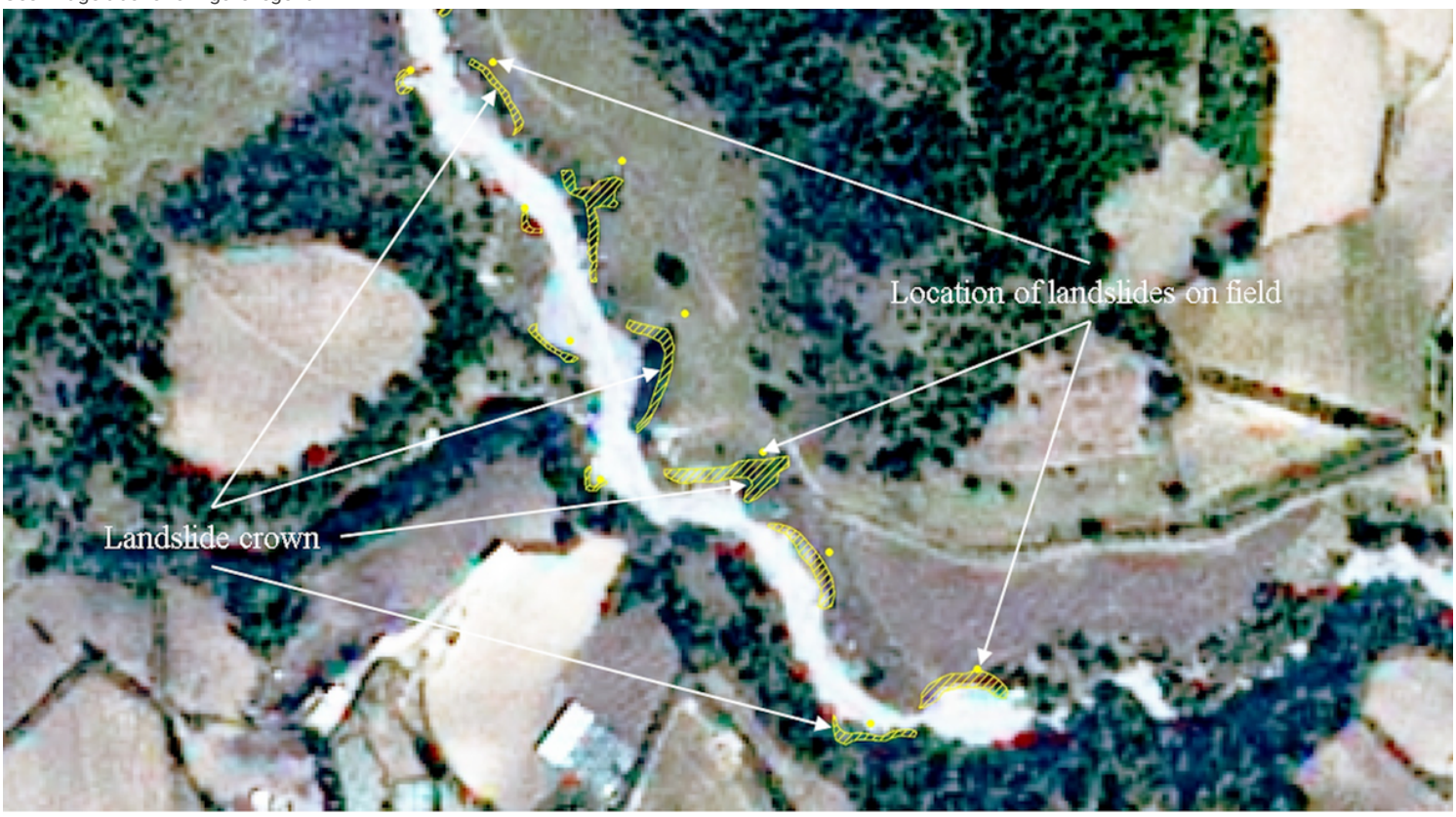


Fig. 2 Digitization of the landslides headscarps

Figure 2

See image above for figure legend.

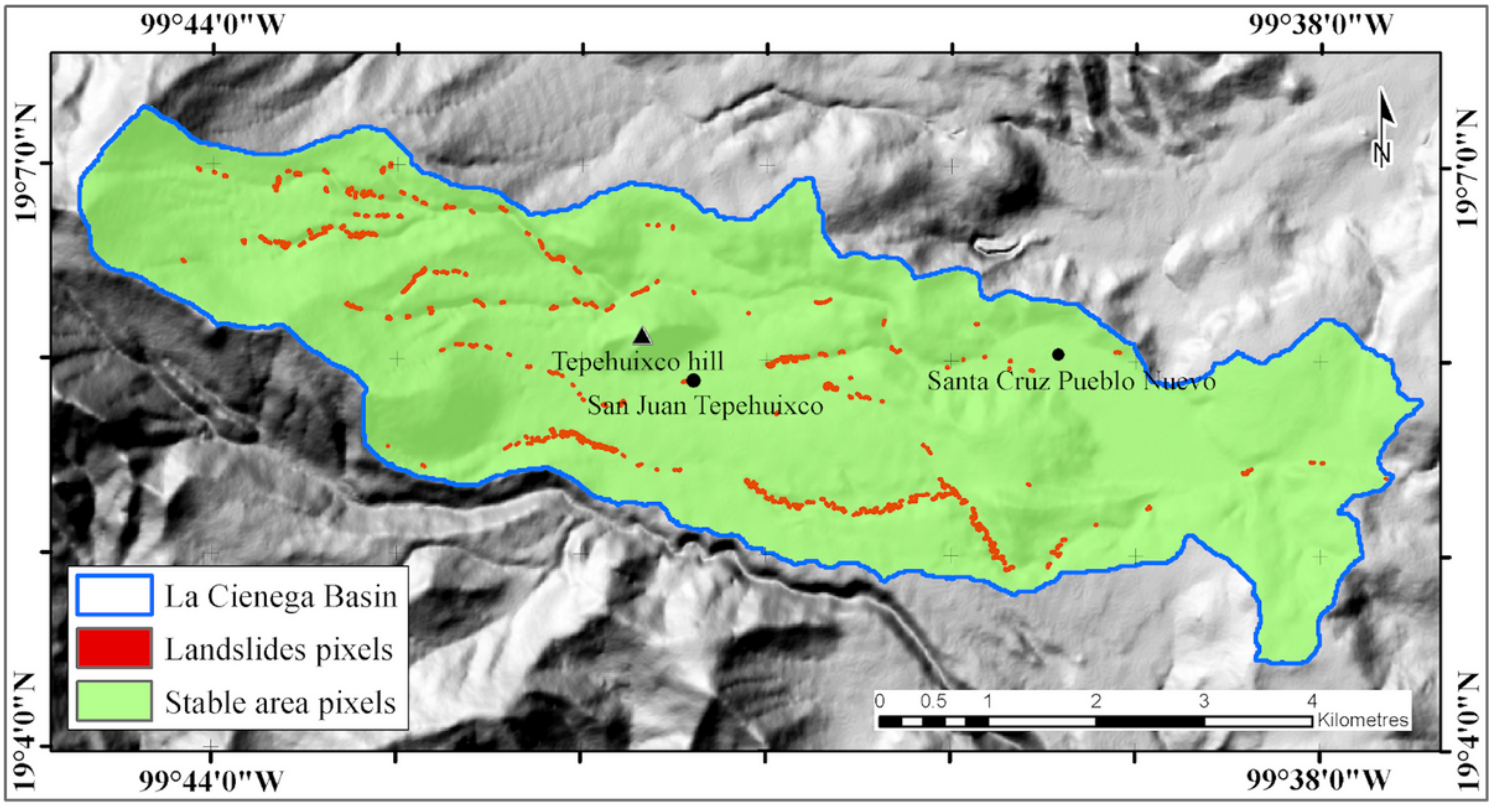


Fig. 3 Digital map of the dependent variable. Red, landslides areas; green, stables areas

Figure 3

See image above for figure legend.

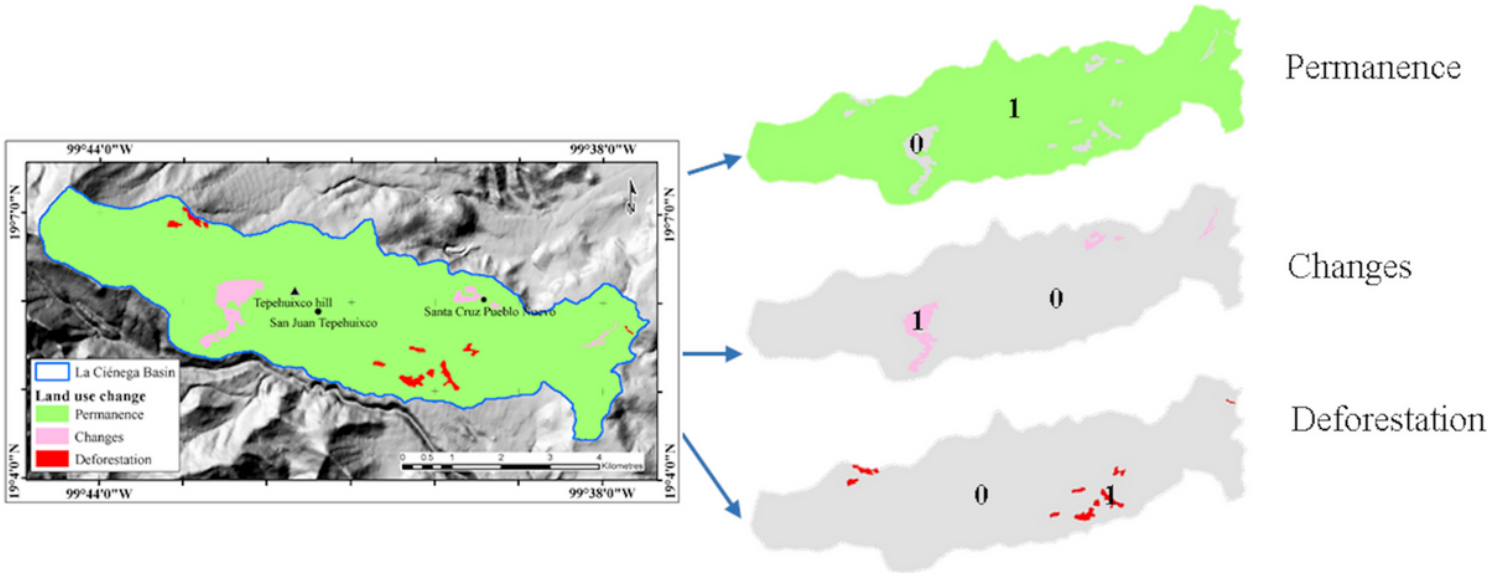


Fig. 4 Qualitative dummy variables. For each qualitative class in a thematic map, a dummy variable with 0 and 1 was created to be used in the MLR-CNSA analysis

Figure 4

See image above for figure legend.

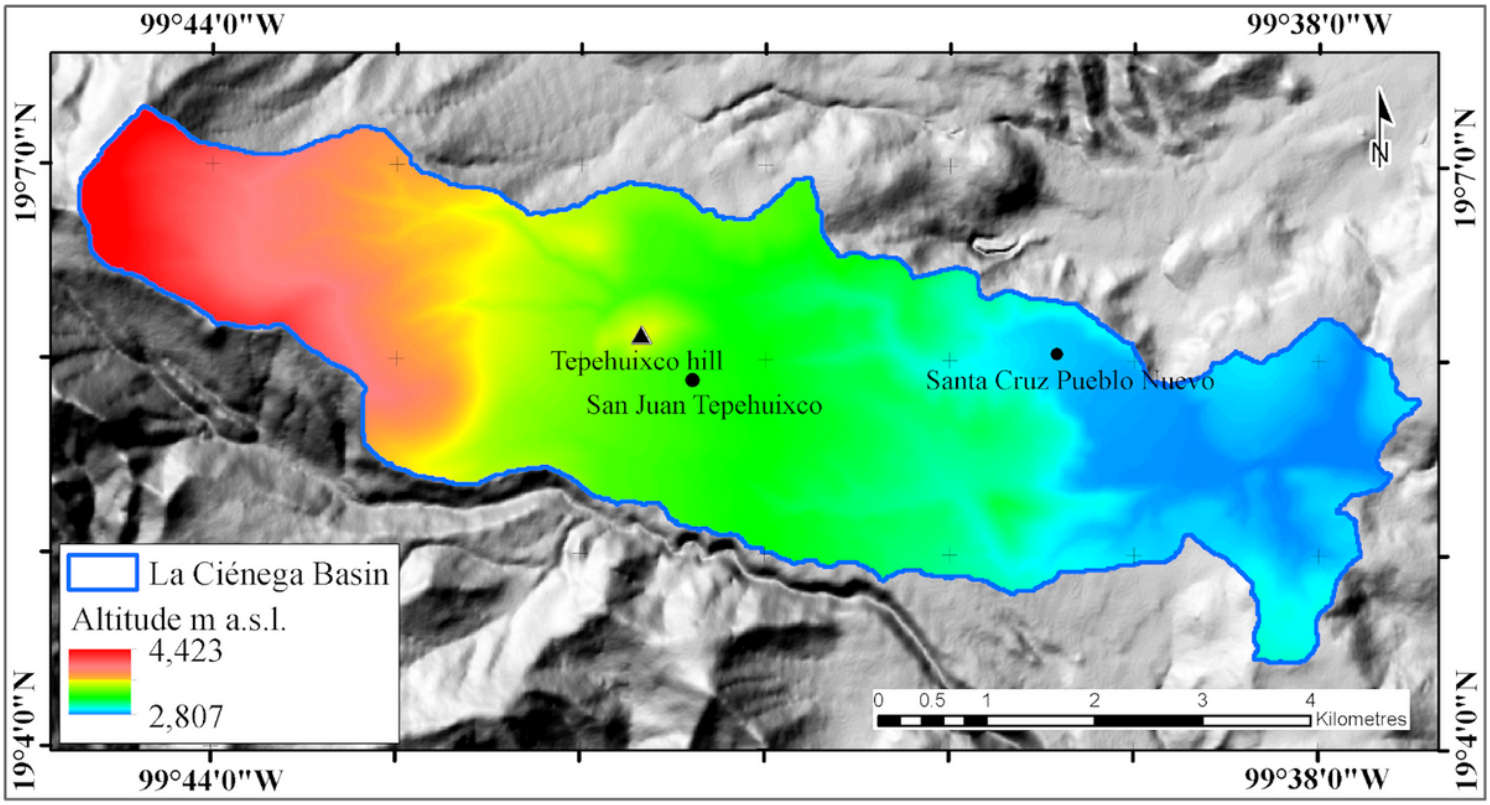
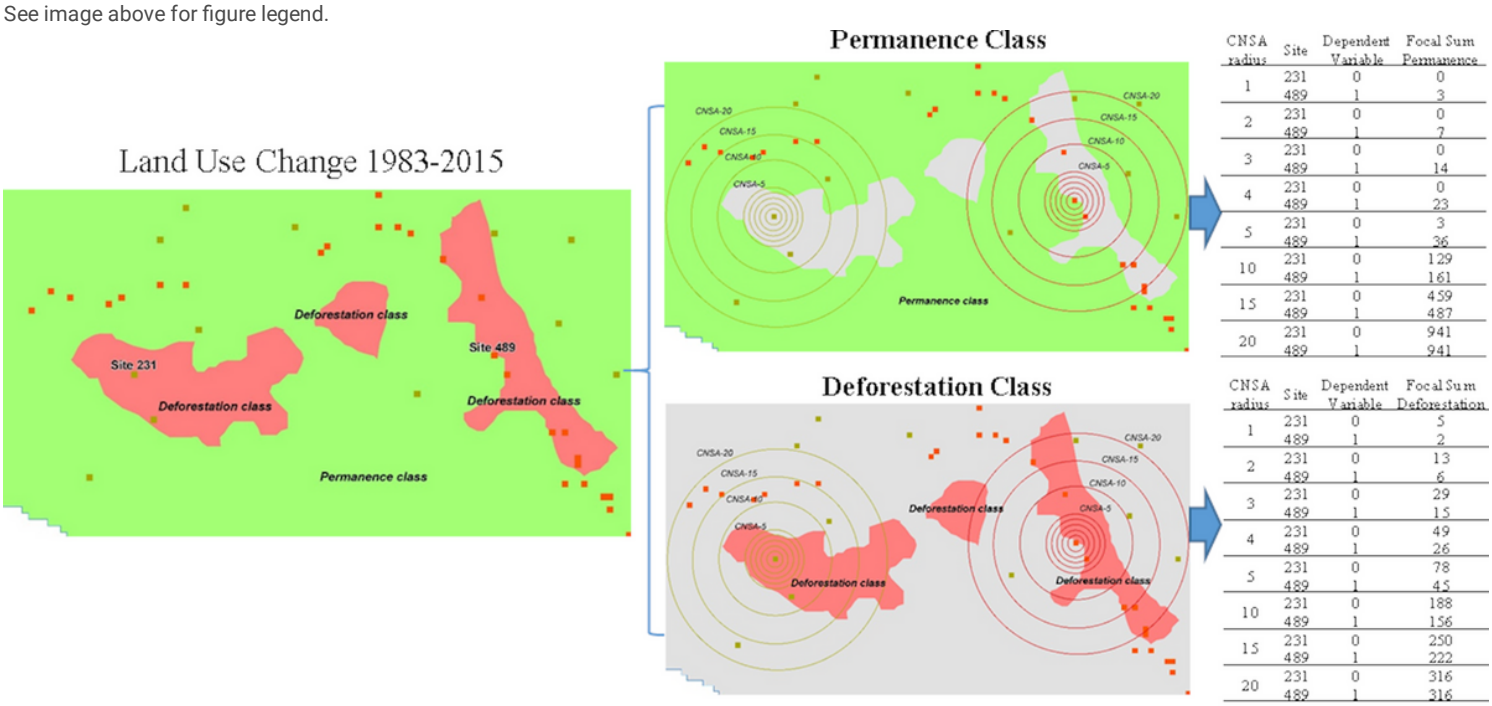


Fig. 5 Quantitative variables with raw values are used directly in the MLR-CNSA analysis

Figure 5

See image above for figure legend.



CNSA radius	Site	Dependent Variable	Focal Sum Permanence
1	231	0	0
	489	1	3
2	231	0	0
	489	1	7
3	231	0	0
	489	1	14
4	231	0	0
	489	1	23
5	231	0	3
	489	1	36
10	231	0	129
	489	1	161
15	231	0	459
	489	1	487
20	231	0	941
	489	1	941

CNSA radius	Site	Dependent Variable	Focal Sum Deforestation
1	231	0	5
	489	1	2
2	231	0	13
	489	1	6
3	231	0	29
	489	1	15
4	231	0	49
	489	1	26
5	231	0	78
	489	1	45
10	231	0	188
	489	1	156
15	231	0	250
	489	1	222
20	231	0	316
	489	1	316

Fig. 6 Matrix for land use variables, indicating the sampling site and the number of pixel of each class with the neighborhood area

Figure 6

See image above for figure legend.

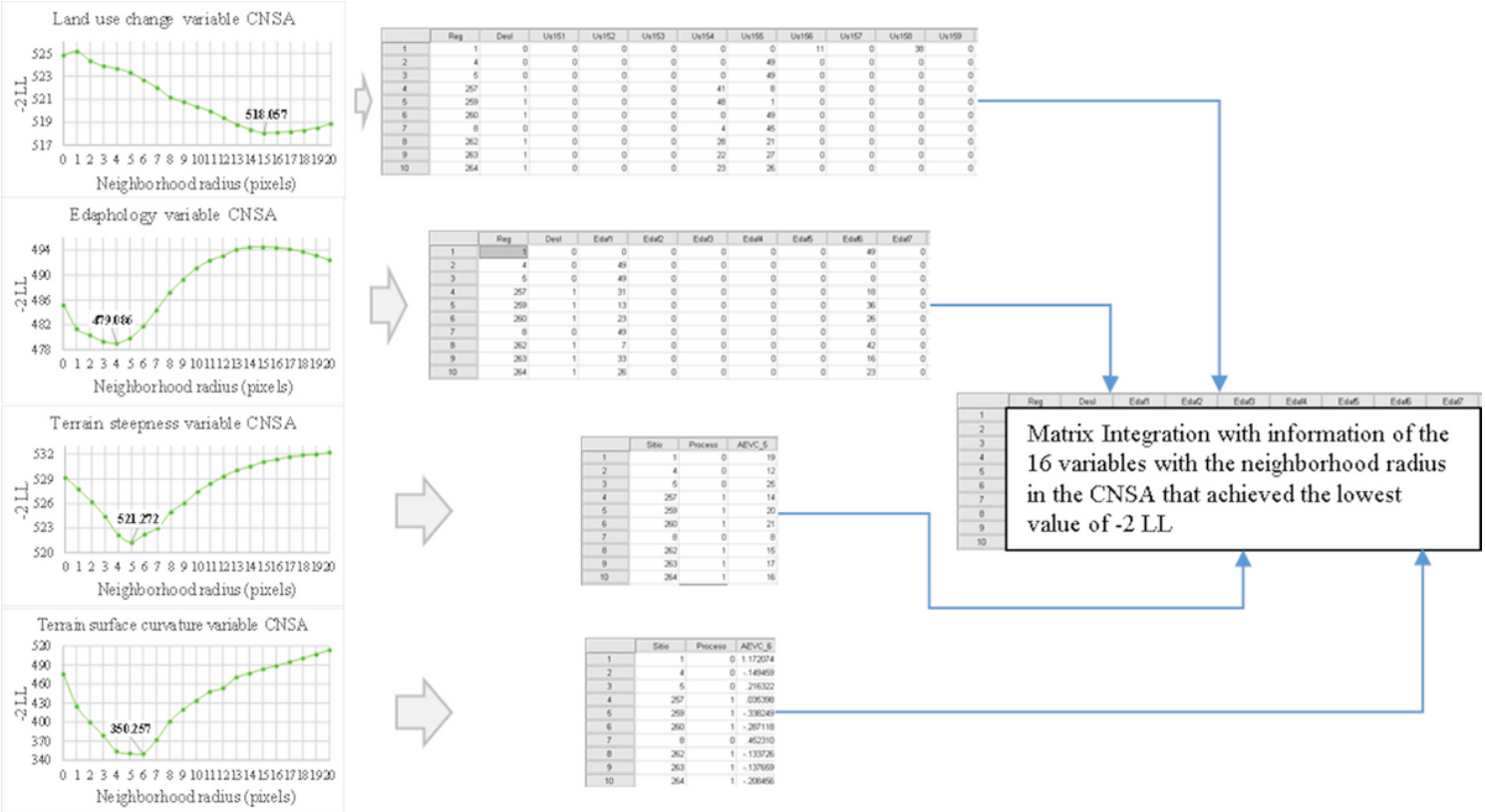


Fig. 7 Integration of the data matrix with information regarding the 16 independent variables selected by applying the CNSA by lowest values of -2 LL

Figure 7

See image above for figure legend.

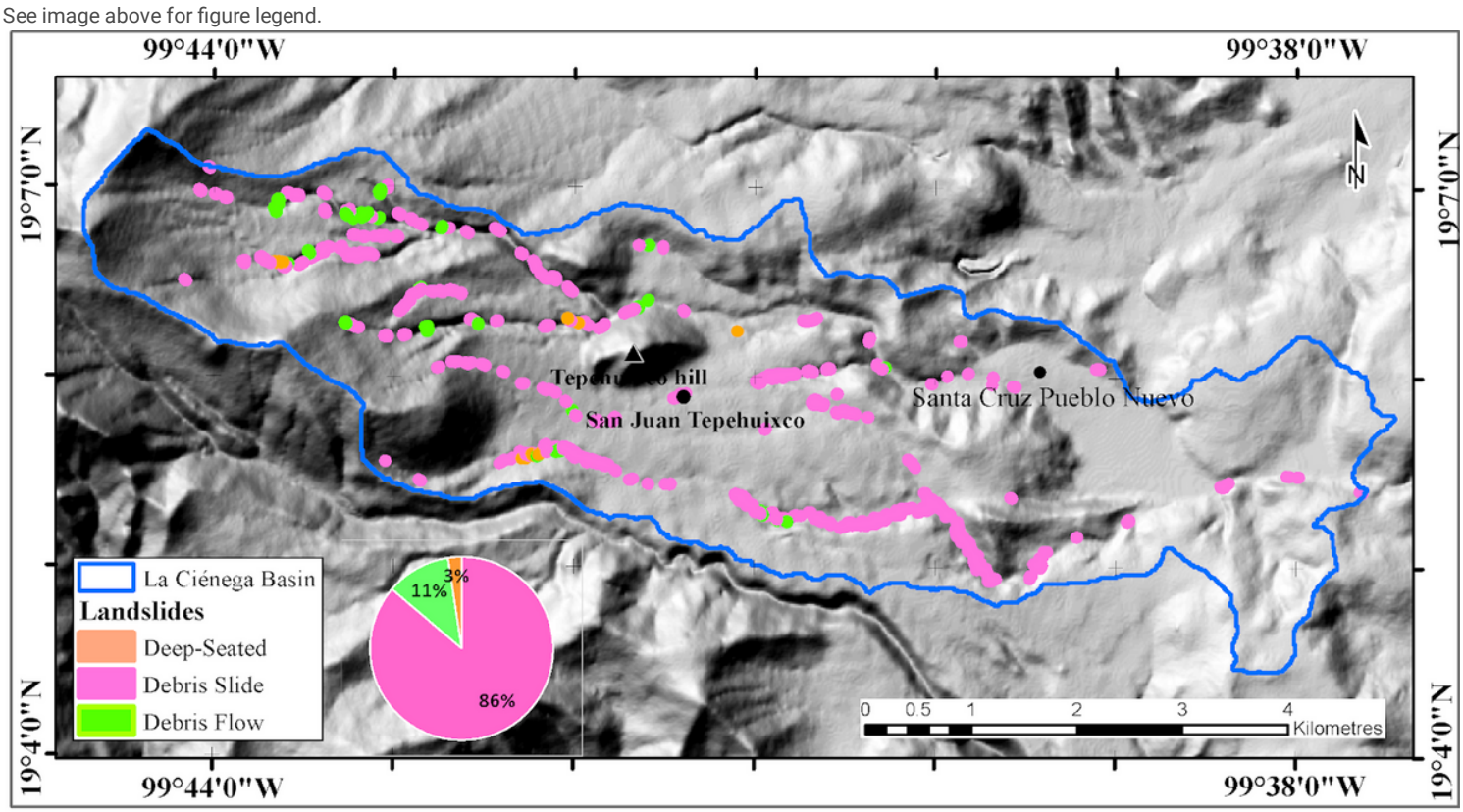


Fig. 8 Landslide inventory distribution by type along the La Ciénega watershed

Figure 8

See image above for figure legend.

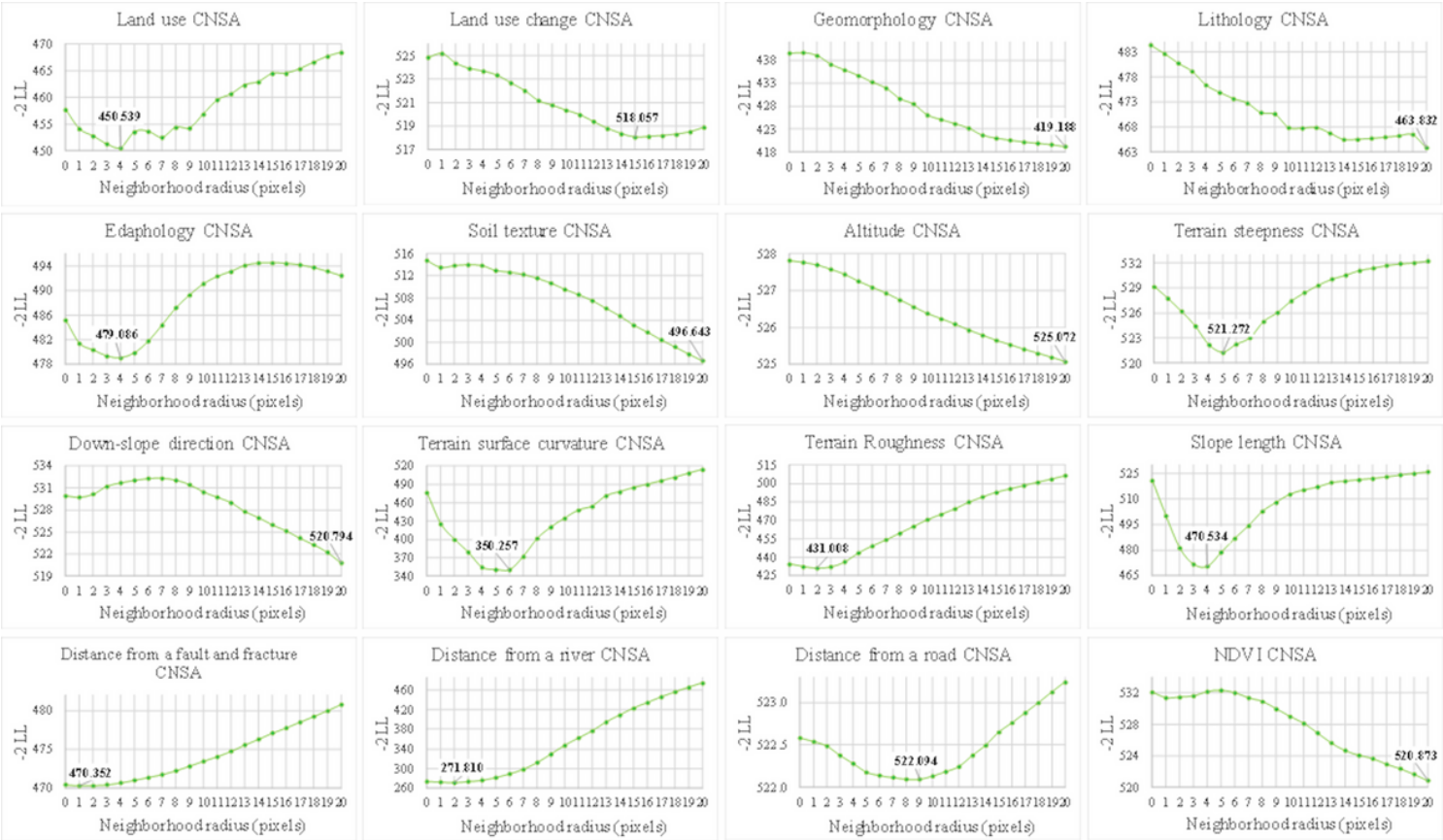
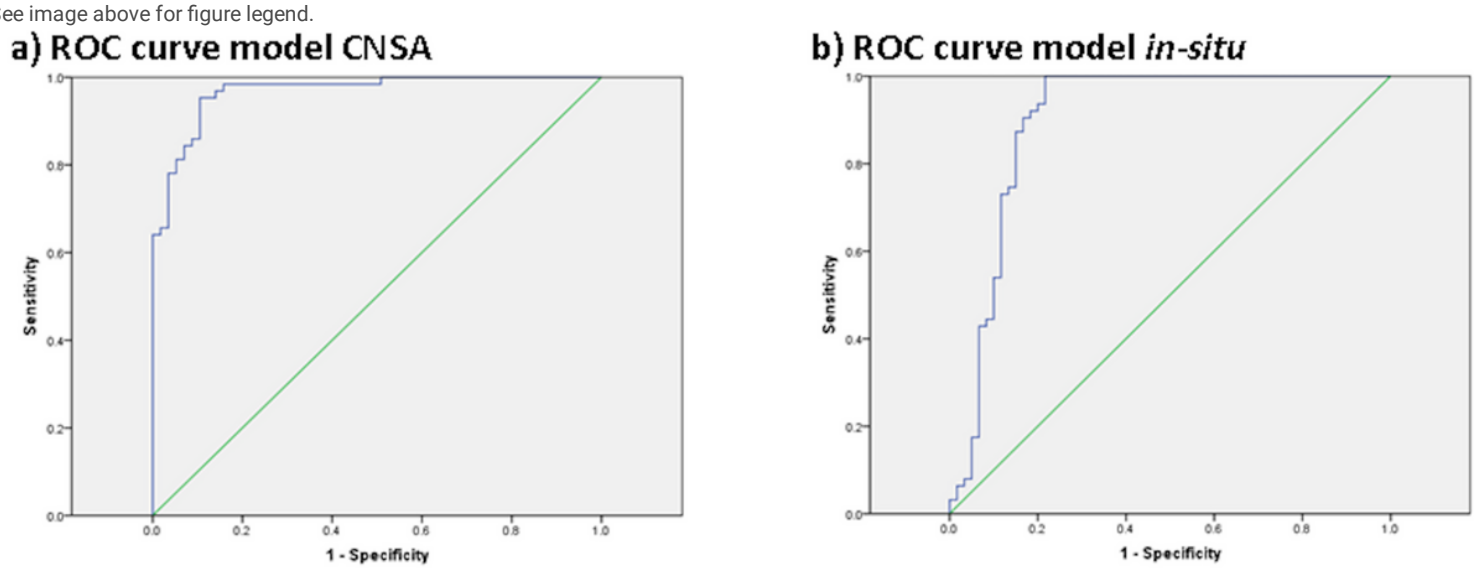


Fig. 9 Value of -2 LL for each variable analyzed . The lowest value reached is shown in bold

Figure 9

See image above for figure legend.



Area Under the Curve

Test result model CNSA

Area	Std. error	Asymptotic Sig.	Asymptotic 95 % Confidence Interval	
			Lower bound	Upper bound
.968	.014	.000	.941	.995

Area Under the Curve

Test result model *in-situ*

Area	Std. error	Asymptotic Sig.	Asymptotic 95 % Confidence Interval	
			Lower bound	Upper bound
.898	.033	.000	.833	.963

Fig. 10 Evaluation of landslide probability models based on the ROC curve

Figure 10

See image above for figure legend.

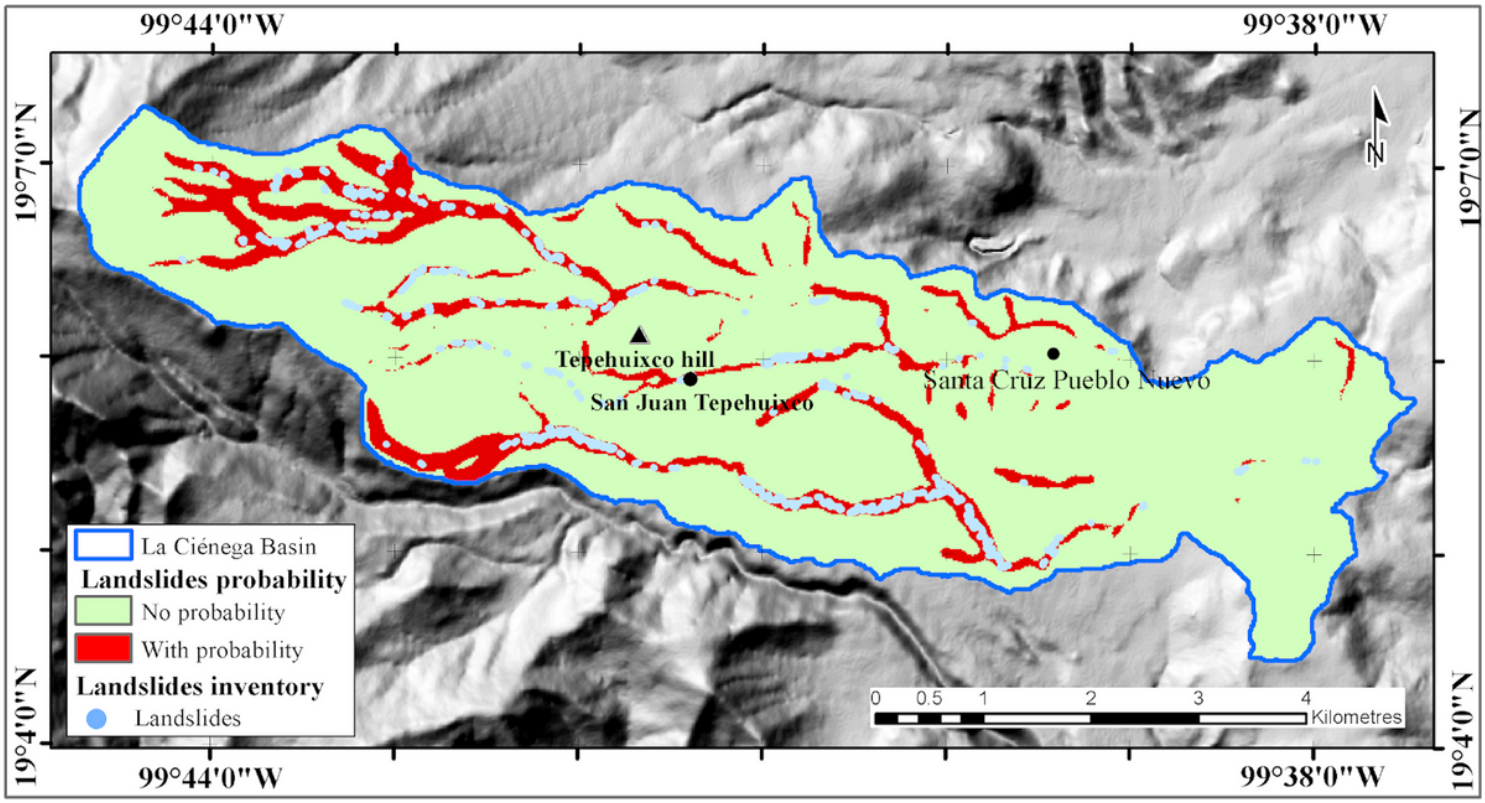


Fig. 11 Landslide probability model based on CNSA

Figure 11

See image above for figure legend.

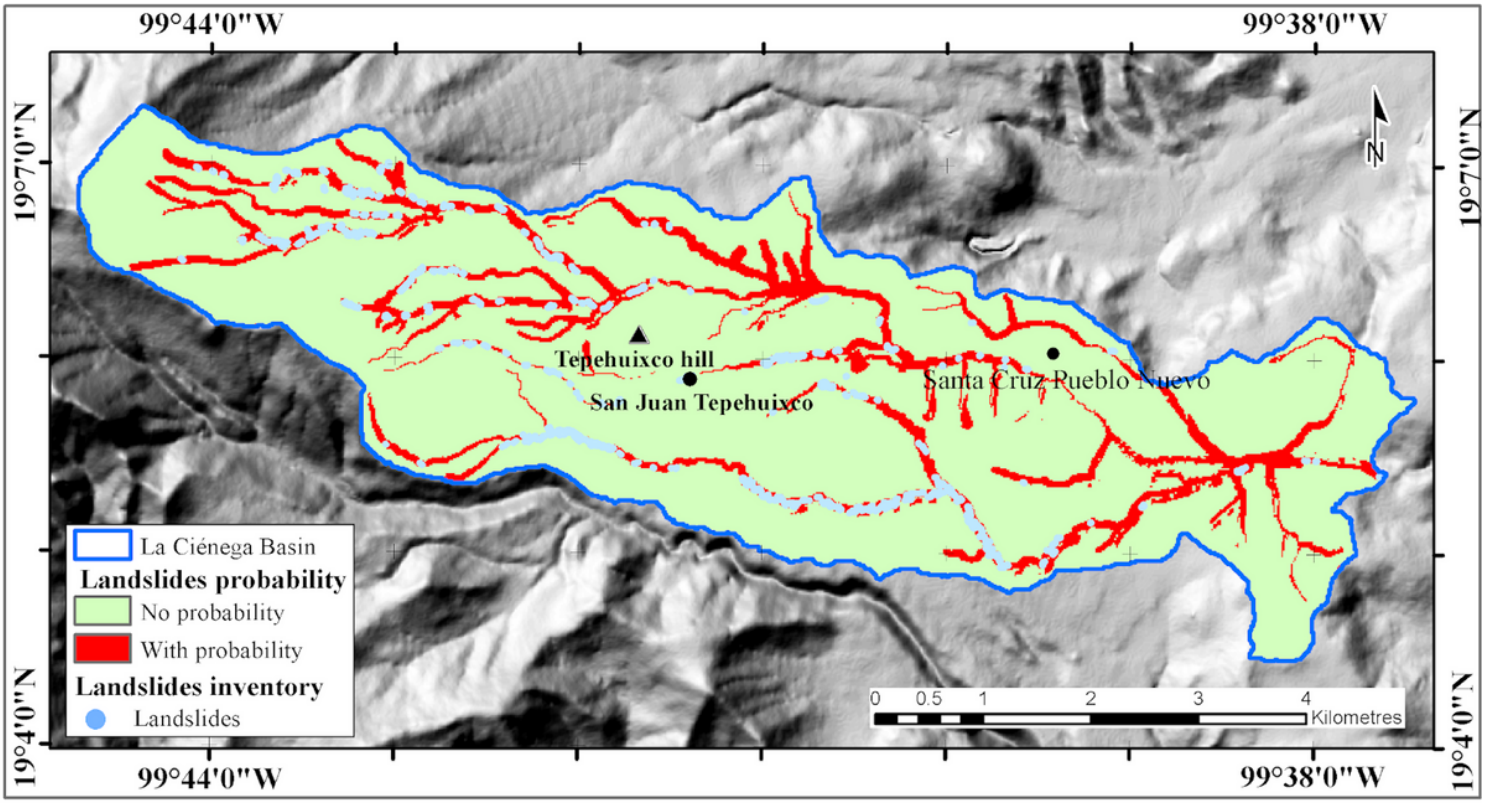


Fig. 12 Landslide probability model with *in-situ* data

Figure 12

See image above for figure legend.

Image not available with this version

Figure 13