ARTICULO DE INVESTIGACIÓN CIENTÍFICA

Sociedad Colombiana
de la Ciencia del Suelo

APLICACIÓN DE REGRESIÓN LOGÍSTICA Y REDES BAYESIANAS PARA EVALUAR SUSCEPTIBILIDAD A DESLIZAMIENTOS EN MONTAÑAS

María Corina Pineda[1] ✉, Álvaro Viloria[2], Jesús Viloria[1]

[1]Universidad Central de Venezuela, Maracay, Estado Aragua, Venezuela
[2]Universidad Central de Venezuela, Caracas, Venezuela

✉: pinedac@agr.ucv.ve

Palabras claves:

Sistema de Información Geográfica
Factores de inestabilidad
Análisis de máxima verosimilitud
WEKA
Variables morfométricas

RESUMEN

Los movimientos en masa son el resultado de la interacción entre variables intrínsecas y activadoras. La variación espacial de las variables intrínsecas determina la distribución geográfica de la susceptibilidad a deslizamientos en masa. En este estudio, realizado en la cuenca del río Caramacate en la Cordillera de la Costa de Venezuela, se relacionó un mapa de cicatrices de deslizamiento con mapas de variables intrínsecas, por medio de regresión logística (RL) y redes bayesianas (RB). Las variables intrínsecas incluyeron variables geomorfométricas, unidades litogeomorfológicas, distancia a la red de drenaje y la diferencia normalizada del índice de vegetación (NDVI). Los resultados de ambos métodos coinciden en indicar que los atributos más asociados con la susceptibilidad a los deslizamientos en esta cuenca son la forma de la pendiente, la distancia a la red de drenaje, el índice topográfico de humedad, el NDVI, el tipo de relieve y la litología. El modelo de RB mostró más claramente la interacción entre las variables intrínsecas, mientras que los resultados de la RL permitieron representar la distribución espacial de la susceptibilidad a deslizamientos.

APPLICATION OF LOGISTIC REGRESSION AND BAYESIAN NETWORKS TO LANDSLIDE SUSCEPTIBILITY ASSESSMENT IN MOUNTAINS

Keywords:

Geographic Information System
Factors of instability
Maximum likelihood analysis
WEKA Morphometric variables

SUELOS ECUATORIALES
42 (1): 23-27

ISSN 0562-5351

ABSTRACT

Mass movements are the result of the interaction between intrinsic and activator variables. The spatial variation of the intrinsic variables determines the geographical distribution of the susceptibility to landslides. In this study, carried out in the basin of the Caramacate River in the mountain ranges of north-central Venezuela, a map of landslide scars was related to maps of intrinsic variables using logistic regression (LR) and Bayesian networks (RB). The variables included geomorphometric variables, lithogeomorphologic units, distance to the drainage network and the normalized difference vegetation index (NDVI). The results of both methods agree on revealing that the attributes associated with susceptibility to landslides in the studied area are the shape of the slope, the distance to the drainage network, the topographic wetness index, the NDVI, the type of relief and the lithology. The BN model showed more clearly the interaction between the intrinsic variables, while the results of the RL allowed to represent the spatial distribution of the landslide susceptibility.

Recibido: Noviembre 2011
Revisado: Enero 2012
Aceptado: Mayo 2012

INTRODUCCIÓN

Los movimientos en masa son el resultado de la acción conjunta de diferentes factores que pueden agruparse en dos categorías: a) variables intrínsecas o de inestabilidad, tales como las condiciones geológicas, las propiedades del suelo y la pendiente, y b) variables externas o de activación como las precipitaciones, los sismos y las actividades humanas. La variación espacial de las variables intrínsecas determina la distribución geográfica de la susceptibilidad relativa a los movimientos de masas en una región determinada (Huaibin *et al*. 2005). Por lo tanto, este tipo de fenómeno no ocurre al azar en el paisaje, por el contrario, tiende a seguir una distribución geográfica que responde a una combinación particular de factores de control. En consecuencia, la probabilidad de ocurrencia de deslizamientos puede ser estimada por las relaciones estadísticas entre las cicatrices producidas por los deslizamientos de tierra antiguos y un conjunto de datos espaciales representativos de variables de inestabilidad (Can *et al*. 2005).

Diferentes métodos que combinan el análisis estadístico con las herramientas de sistema de información geográfica (SIG) se han aplicado para asociar las variables intrínsecas a los movimientos en masa. Estos métodos incluyen el uso de regresión logística, por ejemplo: Dai y Lee (2002), Martínez-Casasnovas *et al.* (2004), Can *et al*. (2005) y Greco *et al*. (2007) y los modelos bayesianos de probabilidad, por ejemplo: Pistocchi *et al*. (2002), Lee *et al.* (2002), Ermini *et al.* (2005), Demoulin y Chung (2007), Van Den Eeckhaut *et al.* (2009), entre otras técnicas.

La regresión logística binaria o binomial predice la ocurrencia de una variable dependiente cualitativa, dicotómica (en este caso, presencia o ausencia de cicatrices de erosión), a partir de una o más variables explicativas independientes o covariables. Los modelos de regresión logística permiten cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente, y clasificar individuos dentro de las categorías presente/ausente de la variable dependiente (Hosmer y Lemeshow 2000).

Por su parte, una red bayesiana modela un fenómeno por medio de un grafo dirigido acíclico, en el cual los nodos representan variables y los arcos que las unen simbolizan relaciones de dependencia probabilística entre ellas. Con base en este modelo, se puede estimar la probabilidad posterior de las variables no conocidas a partir de las variables conocidas, aplicando el Teorema de Bayes. Las redes bayesianas pueden tener diversas aplicaciones como clasificación, predicción y diagnóstico, entre otras. Además, pueden suministrar información útil sobre la forma cómo se relacionan las variables (Castillo *et al.* 1997). El objetivo de este estudio ha sido comparar los resultados de la aplicación de regresión logística y redes bayesianas como medio para identificar las variables intrínsecas que controlan los deslizamientos e inferir las zonas sensibles, en la cuenca del río Caramacate en la Cordillera de la Costa Central de Venezuela. Esta cuenca forma parte de la cuenca del río Guárico, la cual aporta más del 60 % del agua que consume la capital de este país.

MATERIALES Y MÉTODOS

El área de estudio (6760 ha) se encuentra en una zona de rocas metasedimentarias metavolcánicas y ha sido dividida en cuatro unidades litogeomorfológicas: metatobas de "El Chino-El Caño" (OCSCN), metalavas de "El Carmen" (OCSCA), sedimentos aluviales (OCSCQ) y sedimentos coluvio-aluviales (OCSCC) (Pineda *et al*. 2011). El relieve es montañoso, con una altitud de 334 a 1405 m sobre el nivel del mar y una pendiente media de 40%. La precipitación media anual es de 1100 mm y la temperatura media anual es de 22ºC. La cobertura vegetal dominante es herbácea, sometida a pastoreo extensivo e interrumpida por corredores de bosques de galería y parches de bosque perenne en las tierras más altas. El suelo tiende a ser poco profundo, excepto en las zonas que no han sido afectadas por erosión (Pineda *et al*. 2011).

Para evaluar la susceptibilidad a los deslizamientos se delimitaron 214 cicatrices de deslizamiento por medio de interpretación de fotografías aéreas a escala 1:25000 y validación en campo. También se localizaron 233 puntos de no-deslizamiento seleccionados al azar en lugares situados fuera de un área circular de 50 m de radio, dibujada alrededor de cada cicatriz de deslizamiento (Dai y Lee 2002). Los puntos correspondientes a cicatrices de deslizamiento se identificaron como uno (1) y los puntos de no-deslizamiento se identificaron como cero (0), para crear una variable binaria llamada erosión en masa (EM). Se tomaron 412 valores de esta variable (197 cicatrices de deslizamiento y 215 sitios de no-deslizamiento) para generar modelos de predicción de ocurrencia de deslizamientos. Los 40 puntos restantes de esta variable fueron usados como datos de validación.

Las variables intrínsecas usadas para predecir la ocurrencia de deslizamientos incluyeron, en primer lugar, las variables morfométricas: altitud (m), gradiente de pendiente (%), orientación geográfica (radianes), perfil de curvatura (m/m^2), plano de curvatura (m/m^2), curvagrid o forma de la pendiente (relación entre el plano y perfil de curvatura), área de captación (As- área de drenaje en m^2 que contribuye a cada punto del terreno), y el índice topográfico de

humedad calculado como ln (As/tan β) donde tan β es la pendiente local en grados. Estas variables fueron calculadas con base en un modelo digital de elevación (MDE) en formato de malla, con celdas de 20 m de lado. El MDE fue generado con el algoritmo ANUDEM del comando Topogrid, del programa Arc/Gis 9.2, a partir de un mapa topográfico a escala 1: 25 000, con curvas de nivel cada 20 m de altura.

En adición a las variables señaladas, se incluyeron las siguientes variables intrínsecas: primero, la distancia a la red de drenaje obtenida por medio de curvas de contorno trazadas cada 50 metros desde las líneas de drenaje (Dai y Lee 2002). Segundo, la diferencia normalizada del índice de vegetación (NDVI) determinada a partir de una imagen de satélite SPOT-4. Tercero, mapas de unidades litogeomorfológicas y de tipos de relieve (cresta y viga o laderas) derivados de un mapa geomorfológico a escala 1:25 000 del área de estudio.

Por medio de un análisis de componentes principales se identificaron y eliminaron variables redundantes. En consecuencia, sólo las siguientes variables fueron consideradas en el análisis: curvagrid (CD), gradiente de pendiente (GP), la distancia a la red de drenaje (DR), el índice topográfico de humedad (IH), NDVI, unidad litogeomorfológica (N5) y el tipo de relieve (FT). Se realizó un análisis de regresión logística entre la erosión en masa (EM), como variable dependiente, y las otras variables temáticas como variables independientes, utilizando el programa SPSS versión 12 (SPSS Inc., Chicago, IL, EE.UU.). Este programa creó variables ficticias para las unidades litogeomorfológicas (N5) y los tipos de relieve (FT), las cuales fueron codificadas como se muestra en las tablas 1 y 2, respectivamente. El modelo de regresión fue generado por el análisis de máxima verosimilitud (estadístico de Wald). La probabilidad de deslizamientos fue determinada por medio de la ecuación ($P = 1 / 1 + e^{-\hat{g}}$), donde P es la probabilidad y $\hat{g}$ es la ecuación de regresión logística generada por el modelo. Debido a que el factor tiempo no se toma en cuenta, esta probabilidad debe ser interpretada como una susceptibilidad a los deslizamientos y no como una probabilidad de ocurrencia de deslizamientos (Can *et al*. 2005).

Adicionalmente, se generó un modelo de red bayesiana para representar las relaciones probabilísticas entre los deslizamientos y las variables intrínsecas. Para este fin, se utilizaron el algoritmo de búsqueda K2 y el estimador de parámetros simples implementados en el software "Waikato Environment for Knowledge Analysis" (WEKA) (Hall *et al*. 2009). Dado que estos algoritmos utilizan variables discretas, las variables seleccionadas fueron clasificadas, como se muestra en la Tabla 3. Para cada variable fija como una clase se estimó una red bayesiana diferente y la red con mayor poder de predicción fue seleccionada. El poder predictivo de ambos modelos (regresión logística y la red bayesiana) se evaluó comparando los valores observados y predichos usando los datos de validación.

RESULTADOS Y DISCUSIÓN

La regresión logística produjo la siguiente ecuación con un poder de predicción 80.8%:

$$\hat{g}(EM) = -17{,}788 - 0{,}238\,(CD) - 0{,}015(RD) + 0{,}15(IH) - 13{,}353(NDVI) - 2{,}367(FT)_{(1)}$$

$$- 21.857(FT_{(2)}) + 21{,}39(N5_{(1)}) + 19{,}958\,(N5_{(2)}) + 21{,}768(N5_{(3)})$$

Donde,

CD = curvagrid o forma de la pendiente
RD = distancia a la red de drenaje
IH = índice topográfico de humedad
NDVI = índice de vegetación de diferencia normalizada

$FT_{(1)}$, $FT_{(2)}$ y $N5_{(1)}$, $N5_{(2)}$ y $N5_{(3)}$ corresponden a los códigos de las variables categóricas N5 (unidad litogeomorfológica) y FT (tipo de relieve) como se muestra en las tablas 2 y 3.

De acuerdo a esa ecuación la ocurrencia de deslizamientos disminuye a medida que los valores de CD, RD, y NDVI aumentan y el valor de IH disminuye. El efecto de las variables de CD, RD y NDVI coincide con los resultados reportados por D'Amato *et al*. (2004) y Federici *et al* (2006). Sin embargo, según Ohlmacher (2007), el efecto de la variable CD puede variar en función del tipo de erosión en masa considerado. Para las variables discretas el modelo de regresión indica que la incidencia de los deslizamientos es más alta en la unidad litogeomorfológica "Metatobas de El Chino-El Caño" (N54) y en las laderas (FT3).

La Figura 1 muestra el modelo de red bayesiana con el más alto poder de predicción el cual utiliza la variable de FT fija en la raíz de la red. La Tabla 4 muestra la interacción entre las variables intrínsecas (nodos de la red bayesiana). La probabilidad de ocurrencia de deslizamientos es mayor cerca de las líneas de drenaje (<50 m) en las laderas de la unidad litogeomorfológica "Metatobas de El Chino-El Caño". La regresión logística clasificó correctamente 80% de los valores observados en los datos de validación, mientras que la red bayesiana clasificó correctamente el 87.5% de estos valores.

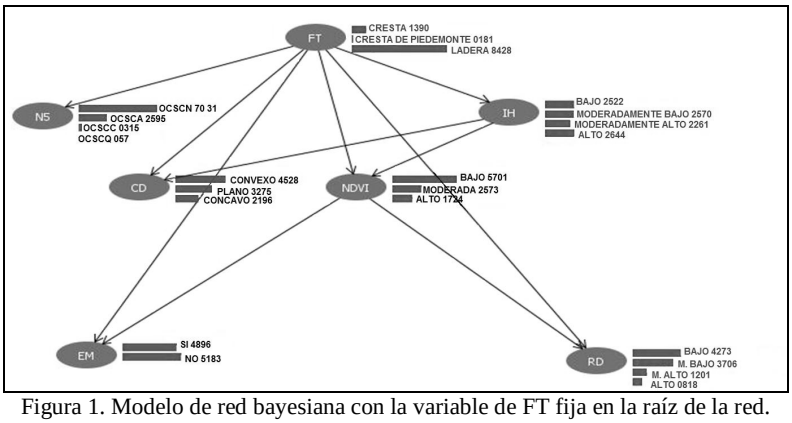

Figura 1. Modelo de red bayesiana con la variable de FT fija en la raíz de la red.

Tabla 1. Codificación de la variable categórica unidad litogeomorfológica (N5).

			Código			
		Descripción	$N5_{(1)}$	$N5_{(2)}$	$N5_{(3)}$	Frecuencia
51	OCSCA	Metalavas de “El Carmen”	1	0	0	107
N52	OCSCC	Sedimentos coluvio-aluviales	0	1	0	12
N53	OCSCN	Metatobas de “El Chino-El Caño”	0	0	1	292
N54	OCSCQ	Sedimentos aluviales	0	0	0	1

Tabla 2. Codificación de la variable categórica tipo de relieve (FT).

		Código		
FT	Descripción	FT(1)	FT(2)	Frecuencia
FT1	Cresta y viga	1	0	57
FT2	Cresta de piedemonte	0	1	7
FT3	Ladera	0	0	348

Tabla 3. Categorización de las variables incluidas en la red Bayesiana.

CD	GP (%)	IH	N5	RD (m)	FT	NDVI
Cóncava (-4.64 a -0.49)	29.33	Bajo 2.9-4.3	OCSCA	Bajo 50-99	CV	Bajo 0.01-0.19
Plano (-0.50 a 0.49)	43.63	Moderadamente bajo 4.4-4.8	OCSCC	Moderadamente bajo 100-149	CP	Moderado 0.20-0.39
Convexa (0.50 a 7.58)	56.28	Moderadamente alto 4.9-5.5	OCSCN	Moderadamente alto 150 -199	L	Alto 0.40-0.63
_	2.01	Alto 5.6-13.4	OCSCQ	Alto 200 -250	_	-

CD= Forma de la pendiente; GP=gradiente; IH= índice topográfico de humedad N5= unidad litogeomorfológica RD= distancia a la red de drenaje; FT= tipo de relieve (cresta y vigas=CV; CP= Cresta de piedemonte; L = Laderas).

Tabla 4. Resumen de la distribución de probabilidades de la red Bayesiana.

		Probabilidad Combinada							
		Clases de N5				RD			
FT	PI para FT	N51	N52	N53	N54	1	2	3	4
Cresta y viga	0.13	0.36	0.01	0.62	0.01	0.35	0.46	0.09	0.10
Cresta de piedemonte	0.02	0.17	0.28	0.50	0.06	0.28	0.61	0.06	0.06
Laderas	0.85	0.24	0.03	0.73	0.00	0.45	0.34	0.13	0.08

PI= Probabilidad individual

CONCLUSIONES

Tanto el modelo de regresión logística como la red bayesiana mostraron que en el área de estudio, los atributos que más se asocian con la ocurrencia de deslizamientos son: la forma del terreno o curvagrid, la distancia a la red de drenaje, el índice topográfico de humedad, el NDVI (como indicador del vigor de la cobertura vegetal), el tipo de relieve y la unidad litogeomorfológica. Los resultados obtenidos de la aplicación de ambas técnicas son complementarios.

Por un lado, el modelo de red bayesiana mostró más claramente el efecto de la interacción entre las variables relacionadas con los deslizamientos y, por el otro, el modelo de regresión logística muestra la distribución espacial de la susceptibilidad a los deslizamientos. En general, el área de estudio presenta una sensibilidad alta o muy alta a los deslizamientos.

Ambos métodos, la regresión logística y la red bayesiana, no consideraron al gradiente de pendiente como una variable predictora de la susceptibilidad a deslizamientos. Esto obedece a que la casi totalidad de los puntos considerados en el estudio (90%) se encuentran en el mismo rango de pendiente (20 a 30%).

REFERENCIAS

CAN T, NEFESLIOGLU H, GOKCEOGLU C, SONMEZ H, DUMAN TY (2005) Susceptibility assessments of shallow earthflows triggered by heavy rainfall at three catchments by logistic regression analyses. *Geomorphology* 72:250–271.

CASTILLO E, GUTIÉRREZ JM, HADI AS (1997) Sistemas Expertos y Modelos de Redes Probabilísticas. Monografías de la Academia de Ingeniería. España. 627p.

DAI FC, LEE CF (2002) Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. *Geomorphology* 42:213– 228.

D'AMATO AVANZI G, GIANNECCHINI R, PUCCINELLI A (2004). The influence of the geological and geomorphological settings on shallow landslides. An example in a temperate climate environment: the June 19, 1.996 event in northwestern Tuscany (Italia). *Engineering Geology*. 73:215–228.

DEMOULIN A, CHUNG C (2007). Mapping landslide susceptibility from small datasets: A case study in the Pays de Herve (E Belgium). *Geomorphology* 89:391-404.

ERMINI L, CATANI F, CASAGLI N (2005). Artificial Neural Networks applied to landslide susceptibility assessment. *Geomorphology* 66:327–343.

FEDERICI PR, PUCCINELLI RA, CANTARELLI E, CASAROSA N, D'AMATO AVANZI G, FALASCHI F, GIANNECCHINI R, POCHINI A, RIBOLINI A, BOTTAI M, SALVATI N, TESTI C (2006) Multidisciplinary investigations in evaluating landslide susceptibility. An example in the Serchio River valley (Italia). *Quaternary Internacional*. 171-172:52-63.

GRECO R, SORRISO-VALVO M, CATALANO E (2007) Logistic Regression analysis in the evaluation of mass movements susceptibility: The Aspromonte case study, Calabria, Italy. *Engineering Geology* 89: 47–66.

HALL M, FRANK E, HOLMES G, PFAHRINGER B, REUTEMANN P, WITTEN IH (2009) The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11:1.

HOSMER DW, LEMESHOW S (2000) Applied logistic regression. 2nd ed. New York: John Wiley & Sons. 373p.

HUABIN W, GANGJUN L, WEIYA X, GONGHUI W (2005) GIS-based landslide hazard assessment: an overview. Progress in Physical Geography 29(4), 548–567.

LEE S, CHOI J, MIN K (2002) Landslide susceptibility analysis and verification using the Bayesian probability model. *Environmental Geology* 43:120–131.

MARTÍNEZ-CASSASNOVAS JA, RAMOS MC, POESEN J (2004) Assessment of sidewall erosion in large gullies using multi-temporal dems and logistic regression analysis. Geomorfhology. 58:305-321.

OHLMACHER GC (2007) Plan curvature and landslide probability in regions dominated by earth flows and earth slides. *Engineering Geology* 91:2-4:117-134.

PINEDA MC, ELIZALDE G, VILORIA J (2011) Relación suelo-paisaje en un sector de la cuenca del Río Caramacate, Aragua, Venezuela, Revista de la Facultad de Agronomía. UCV. 37(1):27-37.

PISTOCCHI A, LUZI L, NAPOLITANO P (2002) The use of predictive modeling techniques for optimal exploitation of spatial databases: a case study in landslide hazard mapping with expert system-like methods *Environmental Geology* 41:765–775.

VAN DEN EECKHAUT M, MOEYERSONS J, NYSSEN J, ABRAHA A, POESEN J, HAILE M, DECKERS J (2009) Spatial patterns of old, deep-seated landslides: A case-study in the northern Ethiopian highlands. *Geomorphology* 105:239–252