Landslides
DOI 10.1007/s10346-015-0593-2
Received: 30 October 2014
Accepted: 4 June 2015

Alfonso Mariano Ramos-Cañón · Luis Felipe Prada-Sarmiento · Mario Germán Trujillo-Vela · Juan Pablo Macías · Ana Carolina Santos-R

Linear discriminant analysis to describe the relationship between rainfall and landslides in Bogotá, Colombia

Abstract Bogotá is located in the central Andean region of Colombia, which is frequently affected by landslide processes. These processes are mostly triggered during the rainy season in the city. This fact remarks the importance of determining what rain-derived parameters (e.g. intensity, antecedent rain, daily rain) are better related with the occurrence of landslides. For this purpose, the linear discriminant analysis (LDA), a technique derived from multivariate statistics, was used. The application of this type of analysis led to obtain simple mathematical functions that represent the probability of occurrence of landslides in Bogotá. The functions also allow to identify the most relevant variables derived from records of rainfall linked to the generation of landslides. A proof of concept using the proposed methodology was done using historic rainfall data from a 9-km^2 area of homogenous climatology and geomorphology in the south part of Bogotá. Landslides needed to be grouped for the LDA. Each one of these grouping categories represents landslides that occurred in similar geomorphologic conditions. Another set of events with no landslides was generated synthetically. Results of the proof of concept show that rainfall parameters such as normalized rainfall intensity I_{MAP}, normalized daily rainfall R_{MAP} and rainy-days normal RDN have the best statistical correlation with the landslides observed in the zone of analysis.

Keywords Rainfall–landslide relationship for Bogotá (Colombia) · Linear discriminant analysis · Probability of landslide

Introduction

Landslides are recurring phenomena causing damages to private property, public facilities and human lives. In order to reduce these effects, huge efforts have been undertaken in recent years to establish early warning systems capable of predicting the occurrence of landslides (Rosi et al. 2012). These events are triggered by various factors like human intervention on slopes, scour, excavations, root expansion, mining, earthquakes, pipe bursting and, mainly, rainfall. Determining rainfall conditions that trigger landslides can be done with models based on the laws of physics, which generally apply the laws of conservation of mass and momentum involving the infiltration and percolation of water into the slope (Iverson 2000).

The rainfall–landslide relationship has also been analysed from the experience of geotechnical risk assessors, who propose rainfall thresholds such as intensity-duration, antecedent rainfall, daily rainfall and others (Chivatá 2008; Gabet et al. 2004; Glade 1998; Mateos et al. 2012; Pasuto and Silvano 1998; Terlien 1996; Caine 1980; Castellanos and González 1996, 1997), based on their own knowledge and on basic analysis of historical records of slides and simple characteristics of rainy events. These rainfall thresholds are commonly defined and enforced for regions which can often exceed areas of 1000 km^2.

Choosing a rainfall threshold for a region based only on the opinion of experts is not always a reasonable option, given the broad range of mathematical, statistical and computational tools available nowadays to analyse the relationship between landslides and rainfall. Multivariate statistical tools like linear discriminant analysis (LDA) are best suited for determining which of the multiple parameters of rainfall is best related to landslides in a determined area. A linear function that separates and groups information into n categories, referred to as the discriminant function, is obtained as the result of a discriminant analysis of a set of data. The obtained function contains parameters of rainfall with an associated weight factor. Larger weight factors indicate that the associated parameter is of greater importance to explain the occurrence of landslides in a region.

In this research, rainfall and landslides records for the city of Bogotá are characterized with the purpose of finding a suitable mathematical relation between rainfall descriptors and the onset of slides. Linear discriminant analysis is used with the historic series to find which one of the 97 rainfall parameters studied herein can explain better the relationship between rainfall and landslides in Bogotá. This tool proves to be very useful for policy-makers in the city, since the frequency and quantity of landslides increases dramatically during the two rainy seasons of the year. The derived discriminant functions cannot be applied to the whole city, because of the large variability of rainfall patterns and of the geomorphology in the urban area.

Database

A database was assembled using the information available in the official institutions in charge of the management of natural risks in Bogotá like the Instituto Distrital de Gestión del Riesgo y Cambio Climático-IDIGER (City Institute of Risk Management and Climate Change-IDIGER). The 7065 reports of landslides that happened in the city between 1996 and 2013 were initially analysed. These reports are produced with information obtained from visits of experts to the zone of the slide, hours or a few days after the event took place. Sheet forms are filled in by the experts of IDIGER, who have relevant experience in geology and geotechnics. Important information of 2208 landslides was gathered from this set of reports (Fig. 1).

Relevant information like date, time and location of the event, type of movement, volume of slide material, ground slope, geological description of the zone, type of geological sliding material, possible detonating factor of the event and soil coverage was collected for each one of the 2208 landslides. The quality of some of these variables is suspicious, because there is no certainty on many of the analysed reports on the date and time of occurrence of the slide event. Because of the multiple features of geotechnical descriptions of landslides, the information was grouped with some key words or categories as shown in Fig. 2.

Published online: 16 June 2015

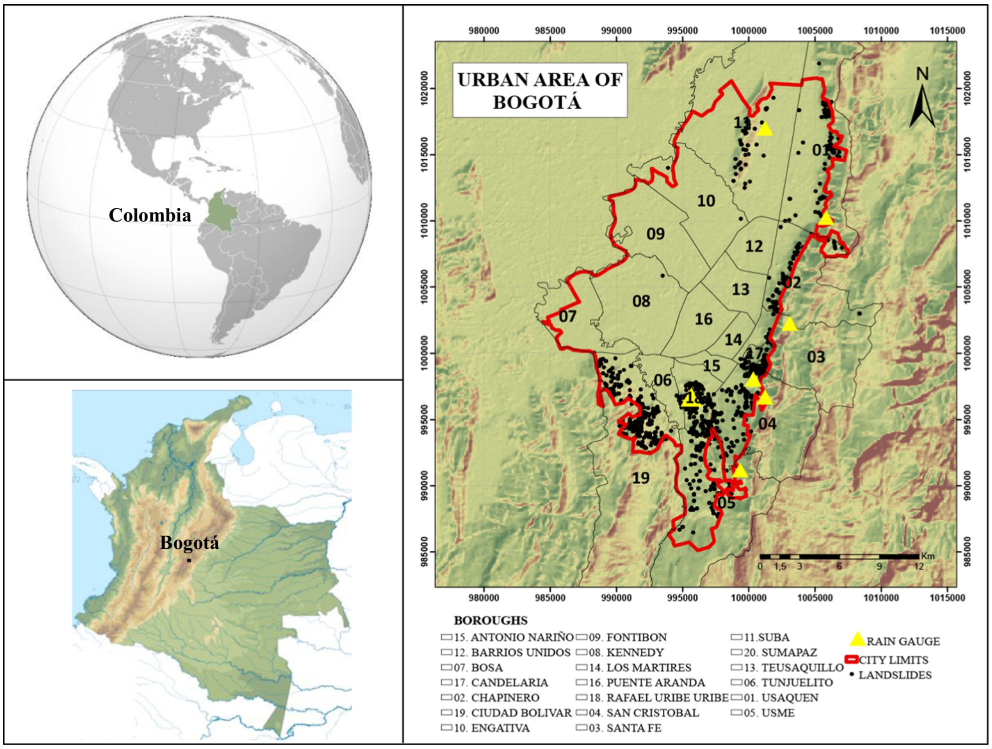

Fig. 1 Map of the city of Bogotá with the location of recorded landslides and of rain gauge stations

Figure 1 shows the location of the city of Bogotá, the capital of Colombia, with an urban area of 307.36 km^2. The same figure depicts the spatial distribution of the 2208 landslides analysed for this research, which cover a time lapse between 1996 and 2013. The map also displays the administrative division of the city into 20 boroughs, namely Antonio Nariño, Barrios Unidos, Bosa, La Candelaria, Ciudad Bolívar, Chapinero, Engativá, Fontibón, Kennedy, Mártires, Rafael Uribe Uribe, Puente Aranda, San Cristóbal, Santa Fe, Suba, Sumapaz, Teusaquillo, Tunjuelo, Usaquén and Usme. Figure 1 also presents the location of nine meteorological gauging stations that cover the urban area of Bogotá.

Some relevant statistics were initially extracted from the set of 2208 events and are summarized in Fig. 3. These statistics were compared with seminal publications in journals that dealt with the assembly of databases used to infer relationships between parameters derived from rainfall and the triggering of different type of landslides. Direct comparisons between the statistics derived from the set of information compiled in this research and information available in the literature cannot be easily done, because the large majority of published articles present information of landslides in Europe and Asia, where geological and hydrological conditions may differ significantly from those registered in Bogotá. However, some references are cited henceforth, where similar trends as those observed for landslides in Bogotá have been reported.

Figure 3a presents as sorting criteria, the type of material that was transported during the landslide. Fifty-two percent of the analysed reports state that the type of material slipped corresponds to soil (disaggregated shallow material), followed by 22 % of reports which do not provide the type of sliding material. Thirteen percent of reports mention rocky materials, and finally, 13 % describe the slide material as detritus. In a similar way, Guzzetti et al. (2008) also present some statistics calculated for databases of landslides induced by rainfall in central Europe, where the lack of knowledge of the lithology of the places where events occurred amounts to 36.4 % of the whole data set.

In Fig. 3b, landslides are grouped depending on the classification of motion registered in the report. Thirty-six percent of the recorded events were generically classified by IDIGER experts as

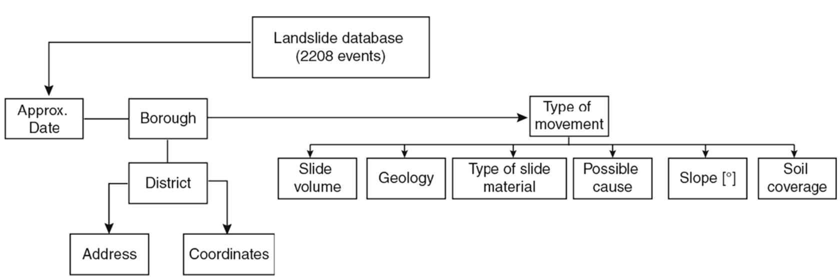

Fig. 2 General structure of the database collected from the information gathered by the experts of IDIGER

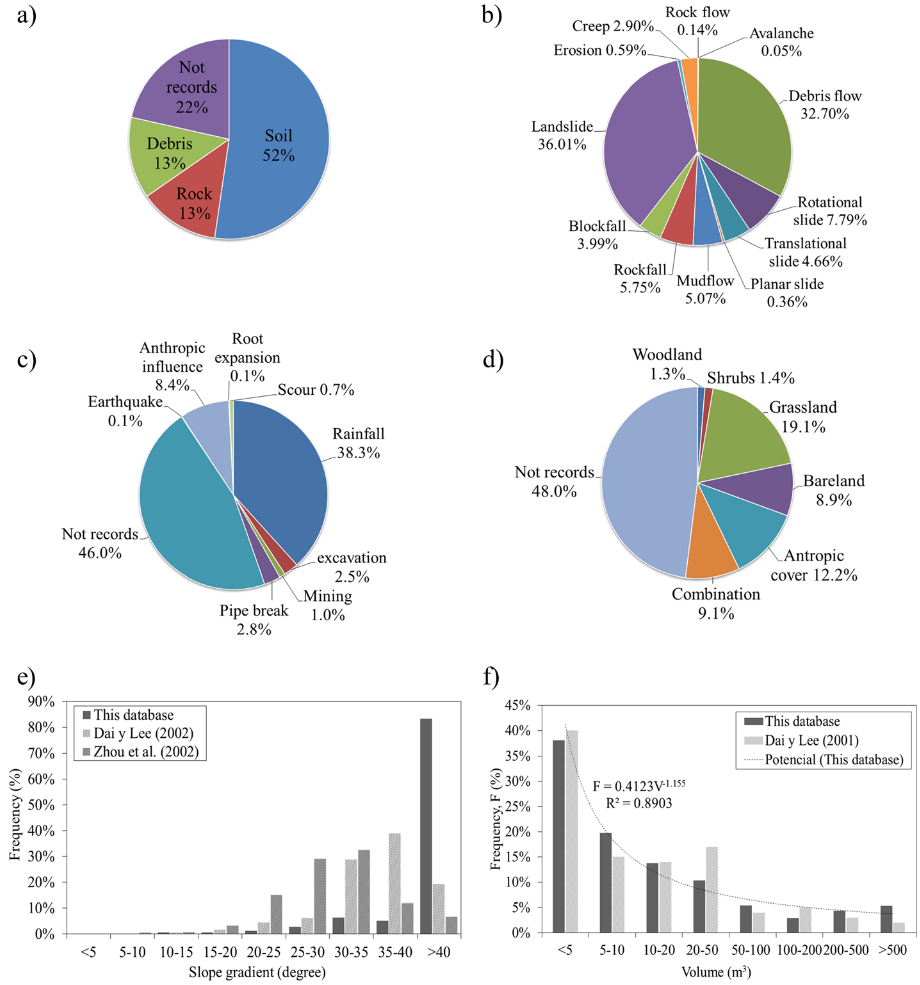

Fig. 3 Statistical description of the database of 2208 landslides registered in Bogotá between 1996 and 2013. Proportion of landslides as function of the type of slide material (**a**). Distribution of landslide types in Bogotá (**b**). Possible triggering factor of the landslides (**c**). Type of soil coverage in the area of landslides (**d**). Angle of slopes as a function of the relative frequency of occurrence of landslides (**e**). Volume of displaced material as a function of the relative frequency of occurrence of landslides (**f**)

landslides, indicating no specificity in the description that allows the events to be classified into the other categories. Thirty-eight percent of the movements that took place in Bogotá happened in the form of flows; 17 % relates to movements where shear bands are evident, and 10 % are classified as block and rock falls. Flow movements are the most common type of slide in Bogotá, similarly to the observations of Guzzetti et al. (2008), who reports 42 % of debris flow in their database. Likewise, 42 % of the disasters caused by landslides in Japan between 1989 and 2007 are attributed to debris flows (Shrestha et al. 2008). Debris flows also represent 32 % of 252 cases of landslides reported in the watershed of the Rules reservoir in Granada, Spain (Jiménez-Perálvarez, et al. 2010).

Figure 3c presents the statistics of landslides in terms of the possible causes that triggered the events. Thirty-nine percent of the reviewed reports show that the possible cause of the movement is rainfall; 8 % determines that the most likely cause is anthropogenic, and only 3 % are caused by other reasons like earthquakes, mining, excavations and pipe bursts. These statistics suggest that given the hydrometeorological regime of the city, rainfall is the most important triggering factor of landslides in the urban perimeter of Bogotá. This can be corroborated in Fig. 4, where the coincidence of the maximum number of events of rainfall and landslides happening in the same two periods of the year is shown.

Figure 3d presents as criterion of classification of the type of soil coverage that was observed and registered in the place where the slide occurred. It is important to contrast the effects of two types of coverage: the first one is the percentage of landslides related to human-made coverage (12 %), where the frequent lack of adequate engineering techniques in both the 'stabilization' of slopes and housing construction have a large influence on the generation of mass movements. The second is the vegetation cover (grasses, shrubs, trees) representing 22 % of the records. It is important to clarify that the soil coverage dominated by trees represents only 1 % of the landslide processes. Similar trends like these were described by Zhou et al. (2002), and it can be inferred that, when the ground is covered with trees, the least amount of slides is generated. Some studies have shown that plant cover, especially of woody type with a set of large and strong roots, helps to improve the stability of the slopes (Gray and Leiser 1982; Greenway 1987; Imaizumi et al. 2008), although further studies to improve the understanding of the role of vegetation cover in the

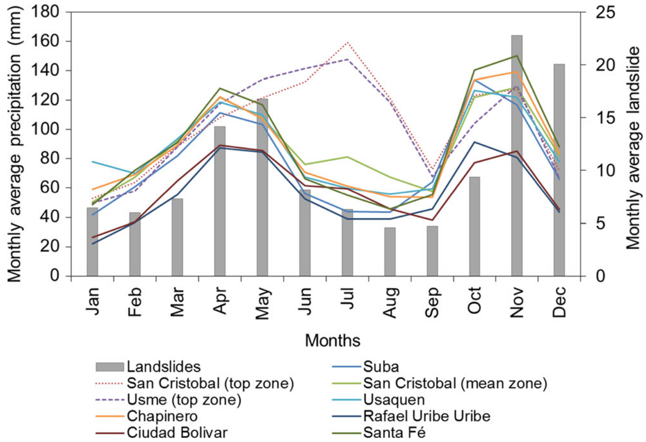

Fig. 4 Multiyear monthly average rainfall between 1996 and 2013 for nine boroughs in Bogotá and landslides multiyear monthly average

occurrence of landslides are needed (Zhou et al. 2002). These statistics suggest that much of the landslides events in Bogotá, not only are controlled by anthropogenic interventions but also are strongly associated with antecedent rainfall events that saturate the ground and not with short rains with large intensities.

Figure 3e shows statistics of landslides according to the angle of the slopes (measured from the horizontal). Most landslide processes in Bogotá occur at angles of inclination greater than 40°. For angles less than 40°, the frequency of landslide occurrence is less than 6 % (Fig. 3e). This could account for the importance of human intervention in high gradient slopes for construction of housing without proper structural design and also without geotechnical works to stabilize the slopes, as it is the case in many impoverished areas of the city, where internal migrants and refugees establish settlements on these highly inclined slopes and build their own houses without proper technical supervision. Imaizumi et al. (2008) reported a similar behaviour to the one described in this work, where the highest frequency of occurrence of landslides per square kilometre per year, is on slopes greater than 30°, for slopes with wood coverage aged between 1 and 10 years. Unlike Dai and Lee (2002) who report a higher frequency of generation of landslide processes depending on the angle of inclination between 30 and 40° for Hong Kong, Zhou et al. (2002) shows a higher frequency of landslides for slopes between 20 and 35° for the Lantau Island in Hong Kong (Fig. 3e).

Figure 3f shows the magnitude of the landslide processes based on the volume of the sliding material. Data presented in this study indicate that 82 % of landslides move volumes less than 50 m^3 of material, while the remaining 18 % are movements that have volumes larger than 50 m^3, including a 1 % of movements reporting volumes larger than 1000 m^3 (Fig. 3f). Moya Sánchez et al. (2013) indicate that the largest number of block topples in the slope Solà d'Andorra in Andorra have volumes between 4 and 30 m^3, reaching some movements volumes up to 150 m^3. Dai and Lee (2001) reported volumes of landslides ranging from 0.1 to 100,000 m^3 and also describe the percentage of landslides based on volume with a power law, similarly to the findings of Brardinoni and Church (2004) and Malamud et al. (2004). These figures show that the highest proportion of landslide events convey small volumes of material, as seen in Fig. 3f.

Rainfall and landslides

The data presented in the previous section already showed in some way the relationship between landslides and rainfall: firstly, because most of the movements can be classified as debris flows (Fig. 3b), indicating the importance of water in the triggering of the event, and secondly, Fig. 3c clearly shows that a large percentage of the movements are possibly caused by rainfall, according to the information registered on the field by the experts of IDIGER who attended the scene of the slides when they happened. Now, a simple analysis of the database of landslides based on rainfall is required, with the aim of elucidating the possible relationship of rainfall and landslide in Bogotá. The original database of 2208 events with detailed information concerning the characteristics of each slide is now filtered in terms of the possible causes of the events. Technical reports filled in by the experts of IDIGER consider nine types of possible causes for the slides, like rainfall, excavations, quarry activities, pipe bursts, earthquakes, scour, etc. Events directly identified by the experts as caused by rainfall amount to 38 %, i.e. 846 events (Fig. 3c). These events will constitute the set of slides used in the linear discriminant analysis and in the subsequent series of statistical analysis.

Figure 4 shows a relationship between rainfall and landslides. On average, the maximum rainfall of Bogotá takes place in the periods of March–May and November–December and coincides with the largest number of slides presented in these same months. Figure 4 shows the multiyear monthly average of landslides between 1996 and 2013 (bars). Multiyear monthly average rainfall for the same period of analysis (solid and dashed lines) recorded in nine of the existing rainfall stations around Bogotá is also plotted.

The occurrence of landslides shows approximately a 1-month lag relative to the multiannual monthly average rainfall, where the greatest amount of landslides is triggered weeks after the highest values of precipitation in the city are recorded (Fig. 4). It is also worth noting that this general trend is not observed in St. Cristobal and Usme rainfall gauging stations (dashed lines), where the hydrological regime differs from the rest of the city by having the highest monthly average values of rainfall in the period June–August and November.

Some differences can be observed between the two rainy seasons in Bogotá. The first season between March and May lasts longer than the second season between November and December. Besides, monthly accumulated rainfall is larger in the second season than the first period between March and May. A 1-month lag between rainfall and landslide occurrence is clearly observed for the second rainy season of the year, which is preceded by the driest period of the year between June and September. Soils need more time to reach a degree of saturation after the dry season in the middle of the year, which could explain the 1-month lag between the peak of rain in November and the peak in the amount of slides in December. This trend is not easily observed during the first rainy season, since the difference of rainfall volumes between this period and the preceding drier months is not as large as the one recorded for the second semester of the year.

The 1-month lag between rainfall peaks and the initiation of landslides observed in Bogotá supports the question of finding the best set of rainfall variables that can be associated to landslides in the city. The variation of the patterns of rainfall in Bogotá, as shown in Fig. 4, confirms that the rainfall parameters associated with slides cannot be applied to the entire city, as rainfall can change considerably from one gauging station to another separated only by few kilometres. For this reason, analysis should be zoned in the city to associate rainfall events with landslides more accurately.

The trend observed in Bogotá differs from other patterns obtained from cities or countries with yearly seasons like Hong Kong, where an almost perfect coincidence between the series of rainfall and the quantity of landslides has been reported by different researchers (Chau et al. 2004).

Extreme periods of rain and drought are related to the Oceanic Niño Index (ONI) in Colombia. The ONI is a measure of the variation of ocean surface temperature in periods of 3 months compared to historical average values (of a baseline series between January 1950 and December 1999) in a particular region (Niño-3.4 region bounded with the coordinates 120°W–170°W, 5°S–5°N) where El Niño is constantly monitored (NOAA 2014). Negative values of ONI indicate the occurrence of La Niña phenomenon, which, depending on its intensity, can bring volumes of precipitation above the average values observed for some regions of Colombia.

Due to the observed relationship between the rainfall and the rate of occurrence of landslides with a 1-month lag in Bogotá, it is also important to examine if there is a link between ONI and the time series of landslides in the city. The historical series taken from the NOAA ONI (2014) and the time series of landslides between 1996 and 2013 (Fig. 5) were overlapped for comparison purposes. In Fig. 5, a hint of a relationship between the periods of La Niña (extremely rainy seasons depicted with a blue horizontal striped pattern) and the largest number of landslides can be inferred, between the years 1996–1997, 1998–2001 and 2007–2008 and especially between the years 2010 and 2012, where the number of landslides overlaps the index marked as La Niña. Periods between the years 1997–1998, 2002–2003 and 2009–2010 (yellow hatch) are El Niño seasons (drought years) where the amount of landslides reduces in Bogotá compared to La Niña years.

Unlike the results presented in Fig. 5, a statistical analysis done for the Central Andes in the province of Mendoza, Argentina, indicates that there was an increased landslide activity during El Niño phase in the Front Andes mountain range; however, this trend has not been observed in the foothills. Moreover, during the cold phase (La Niña), slope movements decreased particularly in the frontal mountain range, reaching similar numbers to the amount of landslides registered on the foothills (Moreiras 2005). This indicates the importance of the weather variation in different parts of America, according to the ONI, as described by Trauth et al. (2003)) and its implications on the derivation of rainfall indexes that predict the occurrence of landslides in these zones. Besides, climatic variations can also have a strong impact in much smaller areas like in Bogotá as shown in Fig. 4, which confirms the importance of generating rainfall–landslide relationships with an adequate selection of the scale and size of the analysis regions.

Linear discriminant analysis

This work, as many of the references cited beforehand, has highlighted the importance of rainfall as a triggering factor of landslides. A methodology is proposed to answer the question raised about what parameters derived from rainfall can be related to landslides in Bogotá. Some authors like Baeza and Corominas (2001) and Dai and Lee (2001) have used a multivariate statistical method called *linear discriminant analysis* (*LDA*) to find out which parameters derived from rainfall can better describe the occurrence of landslides.

LDA is a method for classifying objects from a set of independent variables, in one or more sets of mutually exclusive categories (Morrison 1969; Anderson 2003; Baecher and Christian 2003). Discriminant analysis produces a linear function which separates (discriminates) and groups the data into n categories, referred to as the discriminant function. In general, the discriminant function for n input variables follows the form $D_i = c_1X_{1i} + c_2X_{2i} + \cdots + c_nX_{ni}$, where X_{ji} is the i^{th} value of the independent variable j; c_j is the discriminant coefficient for the variable j; D_i is the value or discriminant score i. The discriminant coefficients are values that maximize the distance between the vector of mean values for each category (Baecher and Christian 2003), namely in this research, the group of landslide occurrence and the group of non-occurrence of landslides.

This section presents the development of the LDA to determine mainly two aspects: first, the type of variable derived from rainfall that shows a strong relation with the occurrence of landslides; second, once the variables derived from rainfall records that generate landslides are found, it is necessary to establish probabilistic threshold for the discriminant function, which separates cases of probable

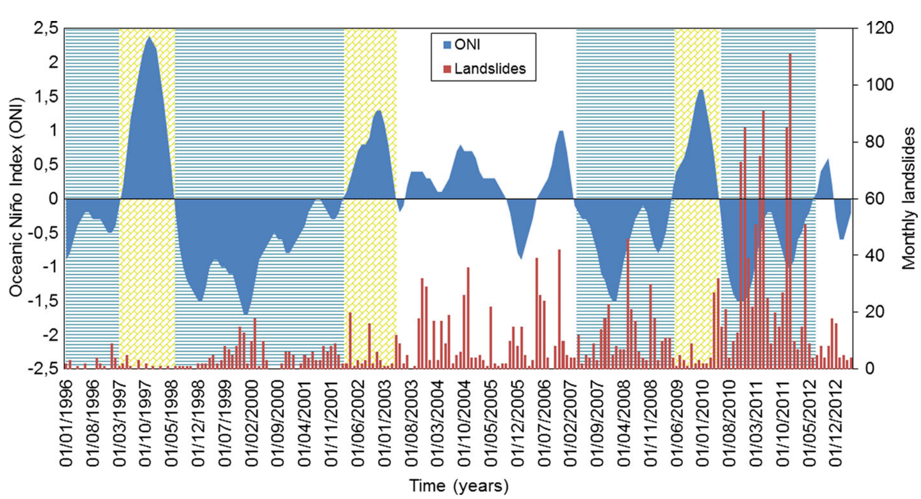

Fig. 5 Evolution of ONI with the occurrence of landslide events in Bogotá

and non-probable occurrence of landslides. The methodology is applied in a small pilot area of 9 km^2 in the area most affected by landslides in Bogotá, defined as the zone with one of the largest densities of slides per area, with homogeneous conditions of climate and geomorphology. A total of 226 events recorded in the pilot zone are used to show the strength of LDA to identify the most relevant rainfall variable that triggers slides in the area.

The LDA requires first that landslides must be grouped into categories, where each class or group reflects similar geomorphological characteristics of the zone where the slides occurred. The grouping criteria in order of importance are the type of slide (TS), angle of inclination of the slope (S), type of geology (TG) linked to the geologic period of the rock outcrop, type of geological material (TM) and the level of landslide hazard (LH) according to the official zoning maps defined by the local authorities of the city. These independent variables were selected, because most of the official reports of post-failure analysis of landslides occurred in Bogotá described these characteristics. The non-numerical categories that describe some of the independent variables (TS, TG, TM, LH) were converted into decimal numbers ranging from zero to one, which could be interpreted as if the grouping index represented a form of quantifying the degree of susceptibility of each slope. In addition, the analyst could define new forms of grouping landslides if other characteristics describing for instance the type and detailed geometry of the slides were available in the technical reports. Table 1 shows an example of the values assigned to each type of slide, material, geology and level of hazard for the landslide events used in this research.

A grouping index (GI) is then calculated for each one of the 226 landslides triggered by rainfall in the pilot zone. This index is proposed as the linear combination of the abovementioned features, linearly weighted as follows:

$$GI = w_1 \cdot TS + w_2 \cdot TG + w_3 \cdot TM + w_4 \cdot S + w_5 \cdot LH$$

where w_i is the weight assigned to each one of the categories based on the judgment and expertise of the analyst.

Table 2 Definition of weights used to group the type of slides collected in the database

Variable		Weight
Type of slide	TS	0.15
Geology	TG	0.23
Material	TM	0.15
Slope	S	0.4
Level of hazard	LH	0.07

Table 2 presents the definition of weights w_i assigned in this article to each one of the five variables considered to group the information of recorded slides in Bogotá.

The linear combination of numerical variables with the weight of the category yields grouping indices ranging from zero to one ($0 \le GI \le 1$) for the landslides registered in the database. Once the grouping indices are calculated, it is necessary to define a set of categories of clustering for GIs, based upon the analysis of the histogram of frequencies for the list of grouping indices, see Fig. 6. The levels of clustering for the GIs, obtained from the histogram of the GIs calculated for the database of 286 landslides in the pilot zone of Bogotá, are divided into two classes, denoted as *High* ($GI \ge 0.4$) and *Low* ($GI < 0.4$). The threshold value of 0.4 was selected because it divides the histogram into two areas of similar proportion. The size of each bin, defining each category, is set by the analyst. Sensitivity analysis of the influence of this clustering of grouping indices (GI) on the form of the resulting discriminant functions can be easily done.

Once landslide clusters with similar characteristics are obtained, it is necessary to create rainfall patterns that have not caused landslides as input for the LDA. The number of rainy events that did not cause slides should be the same as the amount of events that triggered slope failures. Generation of non-slide events was done using the nearest rainfall gauging

Table 1 Designation of numerical values to the description of type of slides, materials, type of geology and level of hazard

Type of slide	Value	Type of material	Value
Rock flow	0.7	Soil	0.9
Avalanche	1.0	Rock	0.3
Debris flow	0.7	Debris	0.7
Rotational slide	0.7	Not available	0.5
Traslational slide	0.8	Geologic period	Value
Planar slide	0.8	Cretaceous—K	0.2
Mud flow	0.8	Tertiary—T	0.6
Rock fall	0.4	Quaternary—Q	0.8
Rock topple	0.4	Hazard zoning	Value
Landslide	0.5	Low	0.2
Erosion	0.1	Medium	0.6
Creep	0.1	High	1.0

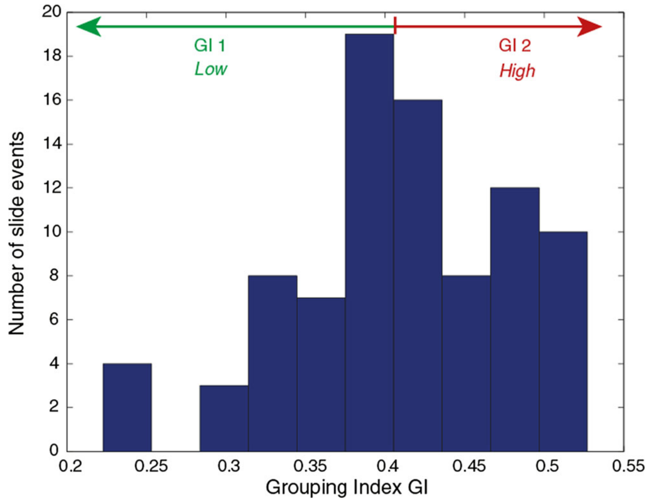

Fig. 6 Histogram of frequency of landslides triggered by rainfall in the pilot zone, clustered in terms of the linear combination of type of slide, type of material, geology, slope and level of hazard of the location of each movement

period is justified because the rainy season in Bogotá typically lasts 3 months (Fig. 4). Most "*non-slide rainfall*" patterns will match the rainless days and must follow an empirical probability density function similar to the function of probability of occurrence of landslides, with the difference that the empirical function should have a 3-month lag.

The discriminant function allows to find the most relevant variables derived from rainfall to explain from a statistical standpoint the generation of landslide categories grouped in terms of GI. Linear discriminant analysis is done separately for each group of indices of actual events that caused landslides, using the free statistical software *R* (R Core Team 2014). Two separate tables of information were prepared for the LDA, where each category of landslides (high, low) is grouped with its corresponding set of "*non-slide rainfall*" synthetic records:

- Grouping index 1 of landslides with "*non-slide rainfall*" records.
- Grouping index 2 of landslides with "*non-slide rainfall*" records.

station to the sliding area. The nearest gauging station must also have rainfall records covering the same time span between the onset of the first and the last landslide. The date of each slide is taken, and 90 days are subtracted thereof to define the beginning of the rainfall time series that did not generate landslides (denoted as "*non-slide rainfall*" henceforth). The 90-day

Parameters derived from rainfall (MAP, R_{MAP}, I, I_{MAP}, RDN, N_{365}, P_1 and A_d) used in the LDA are explained in Table 3. Tables of data needed for the LDA include the calculation of some parameters derived from rainfall like the antecedent rain A_d for every day of the period comprised between the first day and the 90th day previous to the initiation of the landslide. The

Table 3 Definition of rainfall derived parameters used in the linear discriminant analysis

Variable	Description	Units	Reference
A_d	Antecedent rainfall. The total cumulative precipitation measured before the initiation of the landslide, measured during a given period of days [A_d]. For the LDA, this research defined $1 \leq d \leq 90$.	[mm - d]	Terlien (1996), Pasuto and Silvano (1998), Arango-Gartner (2000), Bell and Maud (2000), Guzzetti et al. (2007), Castellanos and González (1996, 1997).
R	Daily rainfall. The total amount of rainfall for the day of the landslide event [R].	[mm]	Terlien (1996), Glade (1998), Pasuto and Silvano (1998), Gabet et al. (2004), Guzzetti et al. (2007), Chivatá (2008), Mateos et al. (2012).
MAP	Mean annual precipitation.	[mm]	Guidicini and Iwasa (1977), Guzzetti et al. (2007).
R_{MAP}	Normalized daily rainfall. Total amount of rainfall for the day of the event divided by the mean annual precipitation [R/MAP].	[mm]	Terlien (1996), Glade (1998), Pasuto and Silvano (1998), Gabet et al. (2004), Guzzetti et al. (2007), Chivatá (2008), Mateos et al. (2012).
I	Rainfall intensity.	[mm/h]	Caine (1980), Au (1998), Ahmad (2003), Cannon and Gartner (2005), Guzzetti et al. (2007), Brunetti et al. (2010), Akcali et al. (2010).
I_{MAP}	Normalized rainfall intensity. Rainfall intensity divided by mean annual precipitation [I_{MAP}=I/MAP].	[1/h]	Paronuzzi et al. (1998), Bacchini et al. (2003), Aleotti (2004), Guzzetti et al. (2007), Dahal and Hasegawa (2008), Guzzetti et al. (2008), Saito et al. (2010), Zhou and Tang (2013).
RDN	Rainy days normal. Mean annual precipitation divided by mean number of rainy days in a year [RDN=MAP/RD_y].	[mm/h]	Guzzetti et al. (2007), Guzzetti et al. (2008), Mathew et al. (2013)
N_{365}	Average number of rainy days in a year. Multiyear average of days with rain in a region or gauging station.	[d]	Wilson and Jayko (1997)
P_1	Precipitation of the antecedent day. Total rainfall measured the day before the landslide occurred.	[mm]	Arango-Gartner (2000)

h hours, *d* days, *mm* millimetres

input table has as many rows as the number of landslide events (226 for the pilot zone) and 97 columns with the calculation of the parameters derived from rainfall.

With the LDA, discriminant function coefficients (c_i) for each parameter of rainfall are obtained, as shown below in a reduced form:

$$D = c_1P_1 + c_2\mathrm{MAP} + c_3R_{\mathrm{MAP}} + c_4I + c_5I_{MAP} + c_6N_{365} + c_7\mathrm{RDN} + c_8A_1 + c_9A_2 + c_{10}A_3 + \ldots + c_{97}A_{90}$$

To reduce the number of involved variables in the previous equation, the analyst proceeds to leave only the relevant variables in the discriminant equation. This “cleansing” was based on the ordering of the discriminant coefficients from highest to lowest. When the discriminant coefficient of a hydrological variable is small (i.e. $c_i \leq 10^{-1}$), this variable is of less importance in the discriminant function, as the significance of the independent variables is directly associated with the order of magnitude of the discriminant coefficients c_i. If the coefficient value $c_i \rightarrow 0$, the independent variable i has practically no effect on the separation of the categories (Morrison 1969; Anderson 2003; Baecher and Christian 2003). The highest values of the coefficients c_i indicate that the independent variable i has a higher importance to discriminate the categories of output (i.e. occurrence of landslide vs no occurrence).

The variables derived from rainfall with the largest absolute value of discriminant coefficients were finally selected for the pilot zone. The type of variables (R_{MAP}, I_{MAP}, RDN) coincided in both categories of landslide clusters (high, low) as shown in Table 4.

Discriminant function for the first group of landslides shows that normalized daily rainfall has a stronger contribution in the occurrence of this type of events, whereas normalized rainfall intensity has a greater contribution triggering the second group of slides.

Linear discriminant functions presented in Table 4 have the ability to group, according to the input variables R_{MAP}, I_{MAP} and RDN, the probability of occurrence or non-occurrence of certain type of landslides described by the categories of the grouping index GI. Hence, the linear discriminant functions need then to be expressed in probabilistic terms. With the mean and variance of the discriminant function, it is possible to adjust a normal probability density function to estimate the probability of an occurrence of a landslide given as input of the value of the discriminant function D for the rainfall data of interest. Mean value μ_{Dj} and standard deviation σ_{D_j} of discriminant function D_j are calculated with the following expression using discriminant coefficients a_i:

$$\mu_{Dj} = \sum_{i=1}^{n} a_i\mu_{xi} \text{ and } \sigma^2{}_{Dj} = \sum_{i=1}^{n} a_i^2\sigma_{x_i}^2$$

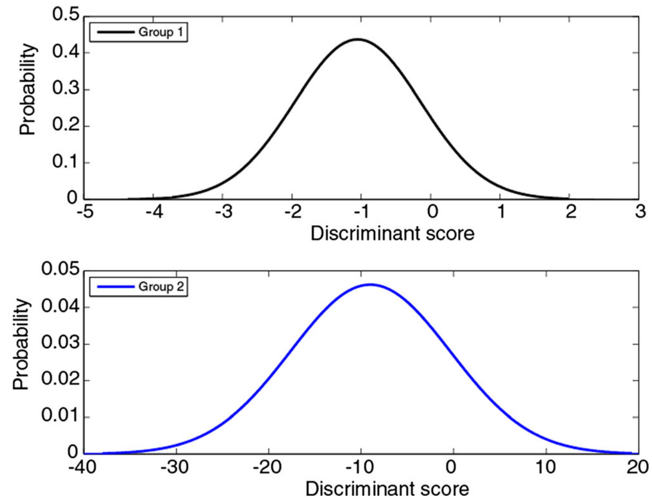

Fig. 7 Probability distribution functions of the discriminant scores of two categories of clustering of GI for the pilot zone

By way of example, mean and standard deviation are calculated for landslides grouped in category 1 as the following:

$$\mu_{D1} = -84.5\,\mu_{R_{\mathrm{MAP}}} + 18.1\mu_{I_{\mathrm{MAP}}} - 4.6\,\mu_{\mathrm{RDN}} = -1.05$$

$$\sigma_{D_1} = \sqrt{(-84.5)^2\sigma^2_{R_{\mathrm{MAP}}} + (18.1)^2\sigma^2_{I_{\mathrm{MAP}}} + (-4.6)^2\sigma^2_{\mathrm{RDN}}} = 0.91$$

Once the mean and standard deviations of each discriminant function, representing both groups of landslides, are calculated, probability density function of both clustering of grouping indexes for the 9-km^2 pilot zone in the south of Bogotá are presented in Fig. 6.

Weights of the discriminant function have an effect on the mean and standard deviation of the probability density function of occurrence of landslides and hence in the shape of the distribution bell. When the rainfall parameters have a large weight (calculated from the linear discriminant analysis), the standard deviation and therefore the width of the PDF curve will significantly increase, as shown in Fig. 7 for the second grouping of slides.

To test the goodness of the linear discriminant analysis to identify variables derived from rainfall that can trigger landslides is to calculate the probability of occurrence of the events that did and did not cause slides for each grouping and to plot the histogram showing the distribution of frequencies. Figure 8 shows the histogram of probability of occurrence for rainfall events for the second grouping (GI 2). It can be seen from the histogram that the probabilistic discriminant

Table 4 Results of the LDA and thresholds of the linear discriminant function

Linear discriminant functions (LDF)	Mean values of LDF	Variance of LDF
$D_1 = -84.5\ R_{\mathrm{MAP}} + 18.1\ I_{\mathrm{MAP}} - 4.6$ RDN	$\mu_{D1} = -1.05$	$\sigma_{D_1} = 0.91$
$D_2 = -63.9\ R_{\mathrm{MAP}} - 807.7\ I_{\mathrm{MAP}} - 5.1$ RDN	$\mu_{D2} = -9.01$	$\sigma_{D_2} = 8.62$

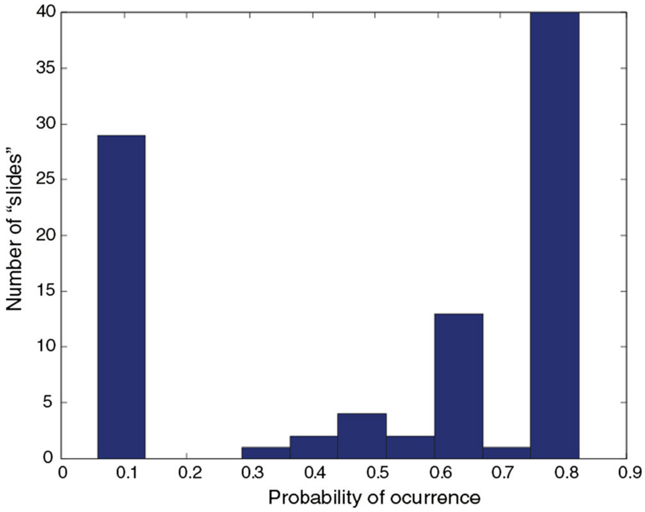

Fig. 8 Histogram of probability of occurrence of landslides, using the discriminant function for events grouped in the second category (GI 2) for the pilot zone

function can separate rainfall events that caused slides (probability of occurrence larger than 50 %) from those which did not triggered landslides (probability of occurrence less than 10 %).

Another way to analyse the goodness of the prediction is to calculate a confusion matrix that reports the number of correct and wrong guesses. This matrix is obtained by applying the discriminant equation (Table 4) for each landslide of the corresponding grouping index, using the relevant rainfall variables R_{MAP}, I_{MAP} and RDN of each event. The number of false positives, false negatives, true positives (rainfall data that caused an actual slide is classified as a triggering event of slides) and true negatives (rainfall data that did not cause an actual slide is classified as an event that cannot generate slides) is counted to produce the confusion matrix. An example of the confusion matrix for slides analysed for the second grouping index is shown in Table 5. The set of events that constitute the second grouping comprises 52 events that triggered slides and 40 events that did not result in landslides. From this series of episodes, 70 % of events were used to run the linear discriminant analysis (calibration series). The remaining 30 % of events (both with occurrence and non-occurrence of slides) were used to test the goodness of prediction (validation series). Table 5 presents the results of the confusion matrix for the calibration and validation series.

Table 5 Confusion matrix for the linear discriminant function of grouping index 2 using 70 % of events for calibration and 30 % of events for validation

Actual event	Prediction–grouping index 2–calibration		
		Yes-slide	No-slide
	Yes-slide	15	0
	No-slide	22	17
Actual event	Prediction–grouping index 2–validation		
		Yes-slide	No-slide
	Yes-slide	10	4
	No-slide	4	10

The model calibrated with 70 % of events misclassifies 41 % of events as false positives. In the validation series, the amount of events misclassified reduces to 29 %. This fact implies that the thresholds and weights used in the definition of the grouping indices for the real landslides need to be fine-tuned. However, the proportion of correct guesses is important (70 % approximately), moreover, taken into consideration the uncertainties in the quality of the description of the slides collected by IDIGER like the time and date of occurrence of the slide, possible triggering factor, correct location of the events, and the definition of non-slide events.

Conclusions

The occurrence of landslides in Bogotá shows an approximate 3-month lag with respect to the multiyear mean monthly precipitation, where the largest amount of landslides initiate time after the recording of the highest values of precipitation in the city. Hence, the need for using statistical methods like the linear discriminant analysis to determine which rainfall parameters relate better with actual landslides, instead of resorting to the traditional approach in Colombia of producing empirical rainfall thresholds based on the experience of geotechnical experts.

The spatial variation of patterns of rain in Bogotá evidences that the analysis of rainfall events that can trigger landslides should not be applied to the entire urban area, as rainfall can change considerably from one gauging station to another separated just by few kilometres. Therefore, it is recommended to apply the linear discriminant analysis methodology presented in this research to smaller homogeneous zones of the city (in terms of geomorphology and climate), to relate more accurately rainfall events with landslides.

The multiyear monthly average amount of landslides and of rainfall shows similar patterns in most of the boroughs of Bogotá.

The analysis of the time series of landslides gives some hints of a direct relation with the Oceanic Niño Index (ONI). Some periods of "La Niña" show an increase in the occurrence of landslides, as this period is related to more rainfall for Bogotá, while some periods of "El Niño" show a decrease in the occurrence of landslides, as this period is related to dryer conditions.

This research has shown the potentiality of using linear discriminant analysis to find the best parameters that relate variables derived from rainfall with the occurrence of landslides, through the selection of weighting coefficients associated with each one of the rainfall parameters. The discriminant function and its coefficients can be updated when new records of landslides and rainfall are available for the zone of analysis. The methodology presented in this paper yields the probability of determining whether a landslide may or may not occur for an event of rainfall.

The application of the proposed LDA methodology to the pilot zone, where homogenous conditions of historic rainfall and geomorphology are found, showed that rainfall parameters such as normalized rainfall intensity I_{MAP}, normalized daily rainfall R_{MAP} and rainy-days normal RDN can be used to explain the occurrence of landslides observed in the zone of analysis. Performance of the probabilistic discriminant function for both clusters of landslides defined in the pilot zone can be easily assessed by means of the confusion matrix. The methodology presented herein recommends to use a large fraction of landslides to determine the relevant coefficients and variables of the discriminant function and to use the remaining percentage of landslides to validate the ability of the function to classify rainfall events as a trigger to new slides.

Results of discriminant coefficients are sensitive to the definition of groups of slides. Further analysis should be undertaken to improve the prediction of the model. One alternative to improve the model could be to separate rainfall events that happened in the first semester from those occurred in the second semester of the year, since the dry spells are different between both seasons.

Many uncertainties remain implicit in the quality of the database used for this analysis. Experts who are in charge of recording the description of slides as soon as they take place should register the information in well-designed forms, where basic information such as the exact date and time of occurrence of the landslide are properly recorded.

Acknowledgments

The authors wish to thank IDIGER (Instituto Distrital de Gestión del Riesgo y Cambio Climático) for the access and use of the database of technical reports of landslides in Bogotá, registered between 1996 and 2013. In addition, the authors acknowledge the partial financial support for the development of this work from Pontificia Universidad Javeriana by grant number 6143 'Determination of onset of rainfall that triggered landslides at Bogotá - Colombia'.

References

Akcali E, Arman H, Firat S, Saltabas L, Gunduz Z (2010) Rainfall threshold for the initiation of landslides in Trabzon province of Turkey. Int J Eng Appl Sci (IJEAC) 2(4):14–26

Ahmad R (2003) Developing early warning systems in Jamaica: rainfall thresholds for hydrological hazards. National Disaster Management Conference, Ocho Rios. St Ann, Jamaica 9–10

Aleotti P (2004) A warning system for rainfall-induced shallow failures. Eng Geol 73(3):247–265

Anderson TW (2003). An introduction to multivariate statistical analysis. John Wiley, 3rd Edition. 721 p

Arango-Gartner J (2000). Relaciones lluvias–deslizamientos y zonificación geotécnica en la comuna dos de la ciudad de Manizales. Master's thesis, Universidad Nacional de Colombia-Sede Manizales

Au S (1998) Rain-induced slope instability in Hong Kong. Eng Geol 51(1):1–36

Baecher G, Christian J (2003) Reliability and statistics in geotechnical engineering, 1st edn. John Wiley, Chichster

Baeza C, Corominas J (2001) Assessment of shallow landslide susceptibility by means of multivariate statistical techniques. Earth Surf Process Landf 26(12):1251–1263

Bacchini M, Zannoni A et al (2003) Relations between rainfall and triggering of debris-flow: case study of Cancia (Dolomites, Northeastern Italy). Natur Hazards Earth Syst Sci 3(1/2):71–79

Bell F, Maud R (2000) Landslides associated with the colluvial soils overlying the Natal Group in the greater Durban region of Natal, South Africa. Environ Geol 39(9):1029–1038

Brunetti M, Peruccacci S, Rossi M, Luciani S, Valigi D, Guzzetti F (2010) Rainfall thresholds for the possible occurrence of landslides in Italy. Natur Hazards Earth Syst Sci 10(3):447–458

Brardinoni F, Church M (2004) Representing the landslide magnitude–frequency relation: Capilano River basin, British Columbia. Earth Surf Process Landf 29(1):115–124

Caine N (1980) The rainfall intensity: duration control of shallow landslides and debris flows. Geografiska Annaler. Series A. Physical Geography 23–27

Cannon SH, Gartner JE (2005) Wildfire-related debris flow from a hazards perspective. In Debris-flow Hazards and Related Phenomena, pages 363–385. Springer

Castellanos R, González A (1996) Relaciones entre la lluvia anual y la lluvia crítica que dispara movimientos en masa. IX Jornadas Geotécnicas de la Ingeniería Colombiana. Santafé de Bogotá, pp 4.62–4.70

Castellanos R, González A (1997) Algunas relaciones de precipitación crítica-duración de lluvias que disparan movimientos en masa en Colombia. 2nd Panamerican Symposium on Landslides - ABMS, vol 2. Río de Janeiro, pp 863–878

Chau KT, Sze YL, Fung MK, Wong WY, Fong EL, Chan LCP (2004) Landslide hazard analysis for Hong Kong using landslide inventory and GIS. Comput Geosci 30(4):429–443

Chivatá Cárdenas I (2008) Using multinomial and imprecise probability for non-parametric modelling of rainfall in Manizales (Colombia). Ing e Investigación 28(2):22–29

Dai FC, Lee CF (2001) Frequency–volume relation and prediction of rainfall-induced landslides. Eng Geol 59(3):253–266

Dai FC, Lee CF (2002) Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. Geomorphology 42(3):213–228

Dahal RK, Hasegawa S (2008) Representative rainfall thresholds for landslides in the Nepal Himalaya. Geomorphology 100(3):429–443

Gabet EJ, Burbank DW, Putkonen JK, Pratt-Sitaula BA, Ojha T (2004) Rainfall thresholds for landsliding in the Himalayas of Nepal. Geomorphology 63(3):131–143

Glade T (1998) Establishing the frequency and magnitude of landslide-triggering rainstorm events in New Zealand. Environ Geol 35(2–3):160–174

Gray D H, Leiser AT (1982) Biotechnical slope protection and erosion control. Van Nostrand Reinhold Company Inc

Greenway DR (1987) Vegetation and slope stability. Slope stability: geotechnical engineering and geomorphology/edited by MG Anderson and KS Richards

Guidicini G, Iwasa O (1977) Tentative correlation between rainfall and landslides in a humid tropical environment. Bull Int Assoc Eng Geol Bull de l'Assoc Int Géol l'Ingénieur 16(1):13–20

Guzzetti F, Peruccacci S, Rossi M, Stark CP (2007) Rainfall thresholds for the initiation of landslides in central and southern Europe. Meteorog Atmos Phys 98(3–4):239–267

Guzzetti F, Peruccacci S, Rossi M, Stark CP (2008) The rainfall intensity–duration control of shallow landslides and debris flows: an update. Landslides 5(1):3–17

Imaizumi F, Sidle RC, Kamei R (2008) Effects of forest harvesting on the occurrence of landslides and debris flows in steep terrain of central Japan. Earth Surf Process Landf 33(6):827–840

Iverson RM (2000) Landslide triggering by rain infiltration. Water Resour Res 36:1897–1910

Jiménez-Perálvarez JD, Irigaray C, El Hamdouni R, Chacón J (2010) Landslide susceptibility mapping in a semi-arid mountain environment: an example from the southern slopes of Sierra Nevada (Granada, Spain). Bull Eng Geol Environ 70(2):265–277

Malamud BD, Turcotte DL, Guzzetti F, Reichenbach P (2004) Landslide inventories and their statistical properties. Earth Surf Process Landf 29:687–711

Mateos RM, García-Moreno I, Azañón JM (2012) Freeze–thaw cycles and rainfall as triggering factors of mass movements in a warm Mediterranean region: the case of the Tramuntana Range (Majorca, Spain). Landslides 9(3):417–432

Mathew J, Babu DG, Kundu S, Kumar KV, Pant C (2013) Integrating intensity–duration-based rainfall threshold and antecedent rainfall-based probability estimate towards generating early warning for rainfall-induced landslides in parts of the Garhwal Himalaya, India. Landslides 1–14

Moreiras SM (2005) Climatic effect of ENSO associated with landslide occurrence in the Central Andes, Mendoza province, Argentina. Landslides 2(1):53–59

Morrison D (1969) On the interpretation of discriminant analysis. J Mark Res 6:156–163

Moya Sánchez J, Corominas Dulcet J, Mavrouli OC, Copons Llorens R (2013) Aproximación probabilística al número y tamaño de bloques en desprendimientos con fragmentación. VIII Simposio Nacional sobre Taludes y Laderas Inestables. Palma de Mallorca: 2013, pages 1107–1118

NOAA (2014) Changes to Oceanic Niño Index (ONI). Consultado el 29 de Agosto de 2014. http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml

Paronuzzi P, Coccolo A, Garlatti G (1998) Eventi meteorici critici e debris flows nei bacini Montani del Friuli. La Acqua, Sezione I-Memorie, pages 39–50

Pasuto A, Silvano S (1998) Rainfall as a trigger of shallow mass movements. A case study in the Dolomites, Italy. Environ Geol 35(2–3):184–189

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/
Rosi A, Segoni S, Catani F, Casagli N (2012) Statistical and environmental analyses for the definition of a regional rainfall threshold system for landslide triggering in Tuscany (Italy). J Geogr Sci 22(4):617–629
Saito H, Nakayama D, Matsuyama H (2010) Relationship between the initiation of a shallow landslide and rainfall intensity-duration thresholds in Japan. Geomorphology 118(1):167–175
Shrestha BB, Nakagawa H, Kawaike K, Baba Y (2008) Numerical simulation on debris-flow deposition and erosion processes upstream of a check dam with experimental verification. Ann Disaster Prev Res Inst 51:613–624
Terlien MTJ (1996) Modelling spatial and temporal variations in rainfall-triggered landslides: the integration of hydrologic models, slope stability models and geographic information systems for the hazard zonation of rainfall-triggered landslides with examples from Manizales (Colombia). International Institute for Aerial Survey and Earth Sciences (ITC)
Trauth MH, Bookhagen B, Marwan N, Strecker MR (2003) Multiple landslide clusters record Quaternary climate changes in the northwestern Argentine Andes. Palaeogeogr Palaeoclimatol Palaeoecol 194(1):109–121
Wilson RC, Jayko AS (1997) Preliminary maps showing rainfall thresholds for debris-flow activity, San Francisco Bay Region, California. US Geological Survey Open-File Report 97–745F
Zhou CH, Lee CF, Li J, Xu ZW (2002) On the spatial relationship between landslides and causative factors on Lantau Island, Hong Kong. Geomorphology 43(3):197–207
Zhou W, Tang C (2013) Rainfall thresholds for debris flow initiation in the Wenchuan earthquake-stricken area, southwestern China. Landslides 1–11

A. M. Ramos-Cañón (✉) · L. F. Prada-Sarmiento · M. G. Trujillo-Vela (✉)
Pontificia Universidad Javeriana,
Carrera 7 No. 40-62, Bogotá, Colombia
e-mail: a-ramos@javeriana.edu.co
e-mail: mario.trujillo@javeriana.edu.co

J. P. Macías · A. C. Santos-R
INGFOCOL Ltda.,
Cra. 19 A No. 79-18, Bogotá, Colombia