Research Article

A cloud-based and open-source approach to generate landslide susceptibility maps

Uma abordagem baseada em nuvem e de acesso aberto para gerar mapas de suscetibilidade a deslizamentos de terras

Macleidi Varnier [1]

[1] Federal University of Rio Grande do Sul, Postgraduate Program in Remote Sensing, Porto Alegre, Brazil.
macleidivarnier@gmail.com
ORCID: https://orcid.org/0000-0002-2255-7294

Abstract: Landslides are devastating events with global implications, impacting human lives and infrastructure. Understanding their patterns of occurrence and identifying susceptible regions is crucial to minimizing their impact. Probabilistic susceptibility mapping is a widely used method for studying landslides, providing a detailed spatial overview. By characterizing the terrain attributes of areas where landslides have and have not occurred, and extrapolating these patterns to the entire study area, a landslide probability map is generated. Although these methods are proven to be efficient, they are often implemented using paid software or restricted access programming languages, which makes it difficult for other researchers to reproduce them. This article presents a free and open-access approach to mapping landslide susceptibility using the Google Earth Engine platform, encompassing all the necessary modeling steps. To illustrate how it works, we carried out a case study for the municipalities of São Sebastião and Ilhabela, in southeastern Brazil. The landslide susceptibility map resulting from this application obtained a ROC curve value of 0.931. 5.4% of the study area is highly susceptible to landslides, and these areas are distributed across the municipalities. Ultimately, this method offers a quick, accessible and low-cost computational solution for an initial mapping of landslide susceptibility.

Keywords: Terrain attributes; Google Earth Engine; Random Forest.

Resumo: Deslizamentos de terra são fenômenos devastadores com implicações globais, afetando vidas humanas e infraestruturas. Compreender seus padrões de ocorrência e identificar as regiões mais suscetíveis é crucial para minimizar seus impactos. O mapeamento probabilístico de suscetibilidade é um método amplamente utilizado para estudar os deslizamentos de terra, fornecendo uma visão espacial detalhada. Ao caracterizar os atributos do terreno de áreas onde ocorreram e não ocorreram deslizamentos de terra, e extrapolar estes padrões para toda a área de estudo, é gerado um mapa de probabilidade de deslizamentos. Embora estes métodos sejam comprovadamente eficientes, são frequentemente implementados utilizando softwares pagos ou linguagens de programação de acesso restrito, o que dificulta a sua reprodução por outros investigadores. Este artigo apresenta uma abordagem gratuita e de acesso livre para mapear a suscetibilidade a deslizamentos de terra utilizando a plataforma Google Earth Engine, englobando todas as etapas necessárias para a modelagem. Para ilustrar o seu funcionamento, realizamos um estudo de caso para os municípios de São Sebastião e Ilhabela, no sudeste do Brasil. O mapa de suscetibilidade a deslizamentos de terra resultado desta aplicação obteve um valor de curva ROC de 0.931. 5,4% da área de estudo possui alta suscetibilidade a deslizamentos de terra, estando estes espaços bem distribuídos nos municípios. Em última análise, este método oferece uma solução rápida, acessível e sem custos para um primeiro mapeamento da suscetibilidade a deslizamentos de terra.

Palavras-chave: Atributos de terreno; Google Earth Engine; Random Forest.

1. Introduction

Landslides are events of soil, rock, and rock debris disintegration and transportation downhill under the force of gravity (BIGARELLA, 2003). Events of different natures are called landslides, such as rockfall and rockslide, rock rolling, mudslides, creep and mud, soil, and debris avalanches (VARNES, 1976). In general, these events are primarily reported in areas with high slopes and are induced by intense and prolonged rainfall that overloads the soil (CRUDEN, 1991). The geological context, such as the proximity of geological faults or the presence of specific rocks, can also facilitate the occurrence of these phenomena (KAMP et al., 2008; CLERICI et al., 2010). Although landslides occur naturally, when they affect urban areas, they cause human and material damage. Population growth, the expansion of cities and the existence of vulnerable economic groups are factors that drive the occupation of risk areas, increasing the chances of occurrence and the destructive potential of landslides (DAI et al., 2001; TEIMOURI; GRAEE, 2012). Climate and environmental changes, which have intensified since the second half of the 20th century, have also increased the frequency of extreme events and their negative impacts (CROZIER, 2010). In this context, knowing the characteristics of the occurrence of these phenomena and indicating the locations with the greatest potential for being affected are actions that help to reduce losses (BRITO, 2014).

In the scientific literature, we find different methods for assessing landslides. We can highlight the elaboration of mass movement inventories, representing the location of identified and recorded events; the susceptibility assessment, indicating the probability of landslides occurring in a locality based on its terrain attributes; the hazard assessment, indicating the chances of landslides happening in a locality based on its terrain attributes and the frequency of occurrence of these events; and the risk assessment, indicating the chances of a locality being affected based on its terrain attributes, frequency of events and possible consequences for exposed human and material elements. Of these approaches, susceptibility assessment is the most widely used because its results are easy to interpret and spatially explicit, its preparation is less complex than that required for risk and danger assessment, and it requires the use of a considerably reduced data set (MANTOVANI et al., 1996; GUZZETTI et al., 1999; FELL et al., 2008). In this approach, relevant spatial variables are used to indicate susceptibility to landslides. Remote sensing and geoprocessing data, especially digital elevation models (DEMs) and geological maps, are widely used because they allow terrain attributes to be interpreted quickly and easily, characterizing the space to be assessed (VAN WESTEN et al., 2000).

Two approaches to assessing landslide susceptibility are notable: the heuristic method and the probabilistic method. In the heuristic method, experts are consulted and determine the relative importance of the explanatory variables and stipulate weights for landslide occurrence based on the values of these variables (usually terrain attributes, geology, land use, etc). The weights and the importance defined for each variable are multiplied and the result expresses a landslide susceptibility surface (FELL et al., 2008). For the probabilistic method, it is necessary to define samples of occurrence and non-occurrence of landslides and use a classification method to differentiate the samples. This technique is based on characterizing the terrain attributes of areas where landslides occur and do not occur, and then extrapolating these patterns to the entire study area. The result of this process is a probabilistic surface indicating susceptibility to landslides, allowing the researcher to differentiate areas with a higher and lower degree of susceptibility (ZHANG et al., 2017; TAALAB; CHENG; ZHANG, 2018). Both approaches have positive and negative points. In both methods, the definition of predictive variables can be carried out based on expert knowledge. On the other hand, if there are samples of landslide, the choice of predictive variables can be based on statistical tests. In this case, the subjectivity of the definition is reduced by measuring the ability of the variables to represent the spatial distribution of the phenomenon. The heuristic method does not require an inventory of landslides, but it does require consultation with experts and their results are usually adjusted to a specific area. In the probabilistic method, the construction of a landslide scar inventory is required, but the weights are adjusted by automated computation making it easier to extrapolate to other areas. However, in the probabilistic method, the weights are defined in order to better divide the sample set of occurrences and non-occurrences, and do not indicate probabilities for terrain attributes based on consolidated knowledge. For a good training, it is necessary to guarantee the quality of the data provided. Thus, both methods are useful and should be used considering the data and resources available for the study area. Their results indicate a first approach to identifying locations susceptible to landslides, delimiting priority zones for monitoring and mapping on a large scale of detail (VAN WESTEN et al., 2003; AYALEW et al., 2004; YALCIN, 2008; GEMITZI et al., 2011).

The scientific literature offers a variety of techniques to map susceptibility to landslides. Reichenbach et al. (2018) in research where 565 articles of landslide susceptibility modeling were evaluated from 1986 to 2016,

demonstrating that the 15 most used statistical methods are 1) logistic regression (AYALEW; YAMAGISHI, 2005; LEE, 2005; HEMASINGHE. et al, 2018; WUBALEM; METEN, 2020), 2) data overlay (AWAWDEH; ELMUGHRABI; ATALLAH, 2018; ARUMUGAM et al. 2023), 3) neural networks (OLIVEIRA et al., 2019; WANG; FANG; HONG, 2019), 4) index based (CANTARINO et al., 2018), 5) multi-criteria decision (GIGOVIĆ; DROBNJAK; PAMUČAR, 2019; Khalil et al. 2022), 6) weight of evidence (LEE; CHOI 2004; ARMAŞ, 2011; GETACHEW; METEN, 2021), 7) fuzzy sets (BAHRAMI; HASSANI; MAGHSOUDI, 2020; AGHDA; BAGHERI; RAZIFARD, 2017), 8) probability based (ROODPOSHTI; ARYAL; PRADHAN, 2019), 9) heuristic (SHARMA; MAHAJAN, 2018; OZIOKO; IGWE, 2020), 10) linear regression (ONAGH; KUMRA; RAI, 2012), 11) tree based (ERMINI; CATANI; CASAGLI, 2005; YEON; HAN; RYU, 2010; CATANI et al., 2013; KIN et al., 2017; ARABAMERI et al., 2021); 12) safety factor (RAY; JACOBS; BALLESTERO, 2011; NATH; SENGUPTA; SRIVASTAVA, 2021), 13) discriminant analysis (PHAM; PRAKASH, 2017; WANG; CHEN; CHEN, 2020), 14) support vector machine (POURGHASEMI et al., 2013; HUANG; ZHAO, 2018; LEE; HONG; JUNG, 2017) and 15) bivariate analysis (NOHANI et al., 2019; MERSHA; METEN, 2020). In addition, the authors point out that in the last years of the analysis the methods based on machine Learning have become preferred.

However, although statistical methods are described as efficient techniques (MERGHADI et al., 2020) they often require the use of machines with good processing capacity and paid software, making it difficult for researchers who do not have these resources. It is also observed that in most published works the researchers do not make public the computational codes used, preventing the analysis of the processes and the replication of the methodology.

In this context, we present a probabilistic method for landslide susceptibility mapping using the Google Earth Engine (GEE) platform. GEE is a free online platform that allows remote sensing data to be processed in a cloud environment based on JavaScript language and using Google's computer. Since its launch, it has gained considerable notoriety, with numerous highly relevant studies elaborated using the platform (SOUZA et al., 2020; ZHAO et al., 2021; BROWN et al., 2022). Among its main attractions are free access; the availability of a large catalog of remote sensing data; cloud data processing, not requiring offline processes or use of the user's computer memory; and the possibility of interacting with other researchers, facilitating joint work and the sharing of methodologies (GORELICK et al., 2017).

The aim of developing a probabilistic methodology for assessing landslide susceptibility on GEE is to enable other users to have easy access to an open script made on a free platform. All the processing is carried out on Google's computer infrastructure, which makes it possible for anyone to reproduce the computational methodology independently of the computer used. DEMs are available in the platform's catalog and the creation of morphometric variables relevant to landslide analysis is automated. The algorithm used for modeling was Random Forest because it is available on GEE and has comparable performance to Deep Learning methods that require more processing power (OLIVEIRA et al., 2019), and its use has increased considerably in landslide modeling since 2014 (MERGHADI et al., 2020). Statistical evaluation of the data used in the model and its results, and training validation were also implemented. Thus, we consider that the greatest contribution of this work is the development of a probabilistic method to evaluate the susceptibility to landslides in a free platform with cloud processing. However, the provision of the samples required for the method to perform and the definition of the explanatory variables must be determined taking into account the specific context of each application. The results obtained indicate a first approximation in the assessment of susceptibility, helping to better allocate resources in future studies. In addition, to present the method, an application was carried out for two municipalities on the southeast coast of Brazil.

2. Materials and Methods

Figure 1 shows a summary flowchart of the methodology. The main stages are: 2.2 sampling of the occurrence and non-occurrence of landslides; 2.3 composition of the database; 2.4 training of the model; 2.5 statistical evaluation of sampling and results and 2.6 evaluation of the accuracy of the susceptibility surface.
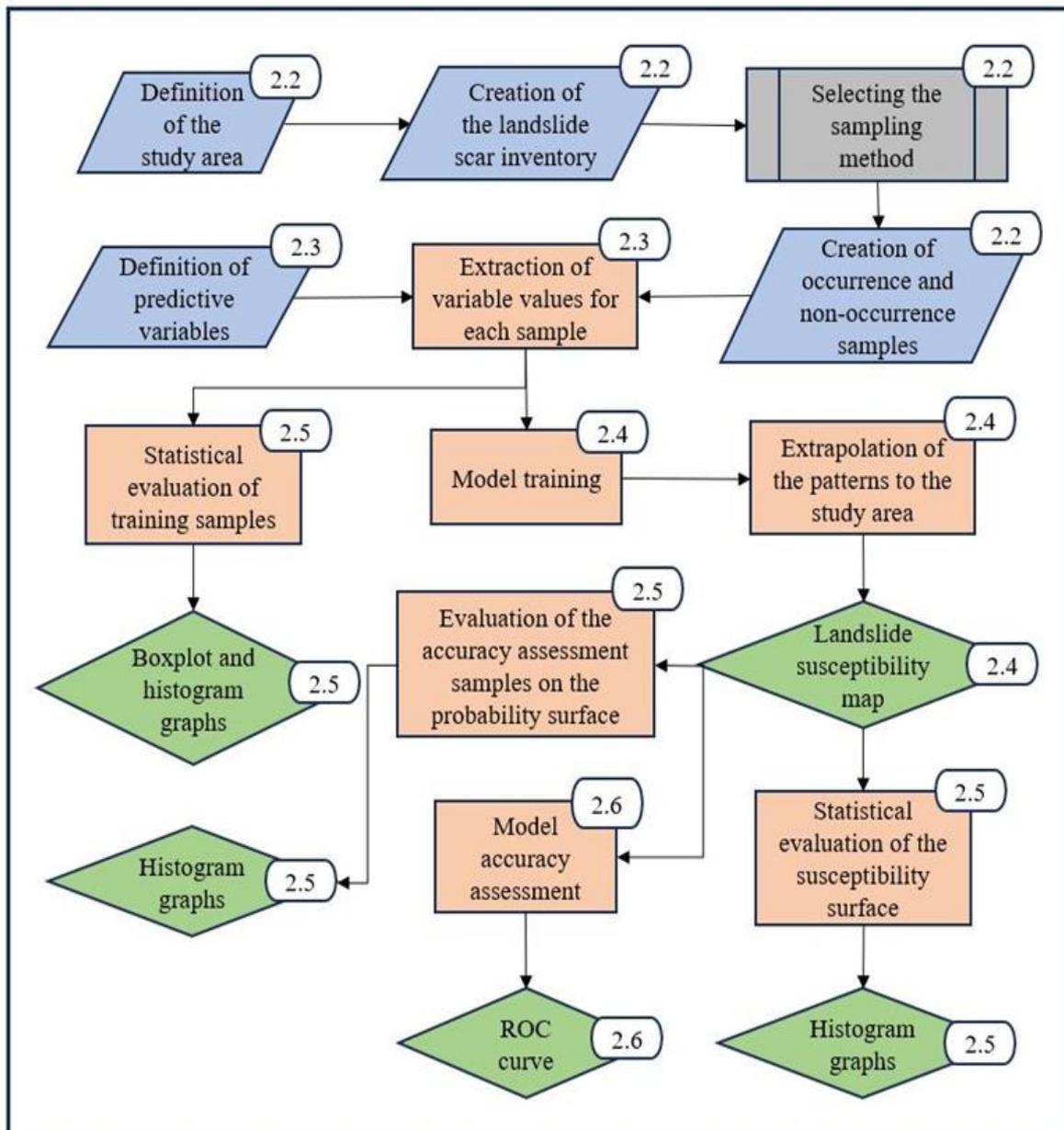
Figure 1. Methodology flowchart.

2.1 *Study area*

The study area for this application is the municipalities of São Sebastião and Ilhabela, located on the north coast of the state of São Paulo, in southeastern Brazil (Figure 1). They are part of the Serra do Mar relief unit, with a highland morphology contoured by the coastal plain (AB'SABER, 2010). Its lithology is made up of exposed rocks from the Brazilian crystalline shield, with altitudes ranging from 0 to 1366 m, with an average of 388 m (CPRM, 2006; NASA, 2023). The pedology is predominantly made up of shallow Haplic Cambisol, covered by dense Atlantic Forest vegetation (ROSSI, 2017). The climate is characterized as tropical humid without drought, with a higher concentration of rainfall in the summer season and less in the winter season (MENDONÇA; DANNI-OLIVEIRA, 2013). The study area is very popular due to its beaches and proximity to large cities such as São Paulo. At the same time as there are high-end condominiums that house the tourists who visit the region, there are also homes of the poor population with precarious infrastructure located in risky places such as steep areas and near hilltops (MARANDOLA JUNIOR et al., 2013; G1, 2023).
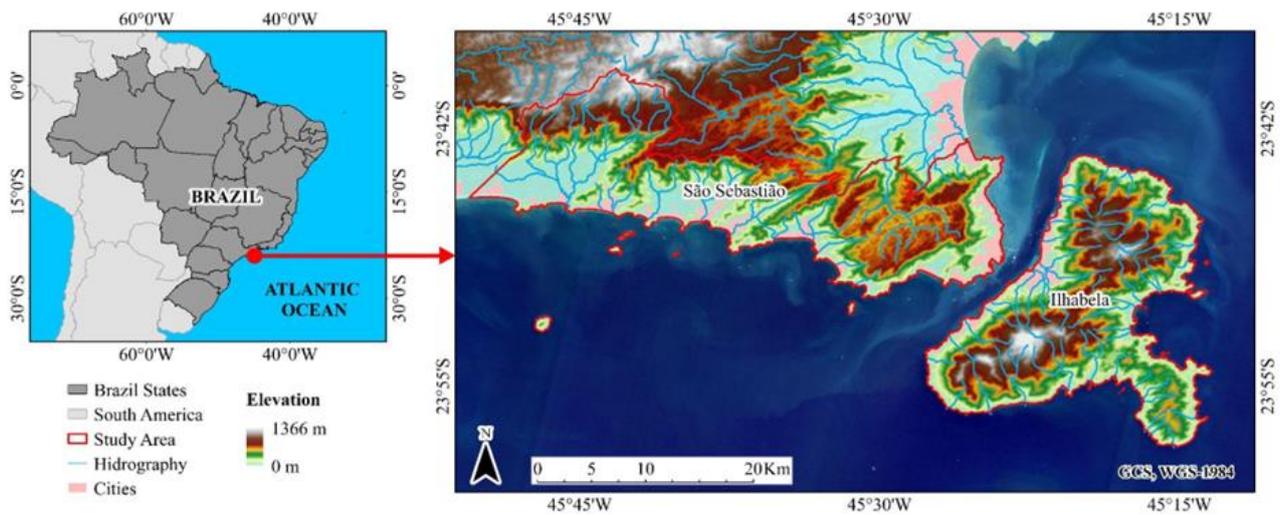
Figure 2. Location of the study area.

On February 19, 2023, a storm hit the northern coast of São Paulo, generating an accumulated rainfall of 683 mm in 24 hours, the highest rainfall recorded in Brazil's history for the same period. This event was the result of a combination of factors: a cold front coming from the south of the continent met an area of low atmospheric pressure in the Atlantic Ocean, making it possible for clouds with a lot of moisture to form. The contact of this cold front with warm winds from the northeast of Brazil, the above-average temperature of the Atlantic Ocean and the orographic effect of the Serra do Mar intensified the phenomenon and kept it parked over the municipalities, providing the extreme rainfall (G1, 2023; BBC, 2023).

These rains contributed to the occurrence of numerous landslides, such as those shown in figure 3, representing the landslide scars after the events. Due to the large amount of rainfall, the soil became saturated with water, resulting in mass movements that transported soil and vegetation from sloping areas to flat areas, where the material was deposited. The areas hit by the landslides were seriously affected, resulting in the deaths of more than 60 people and displacing thousands of residents. The landslide scars resulting from this event were used to assess landslide susceptibility in these municipalities (GUERRA, 1994; G1, 2023).
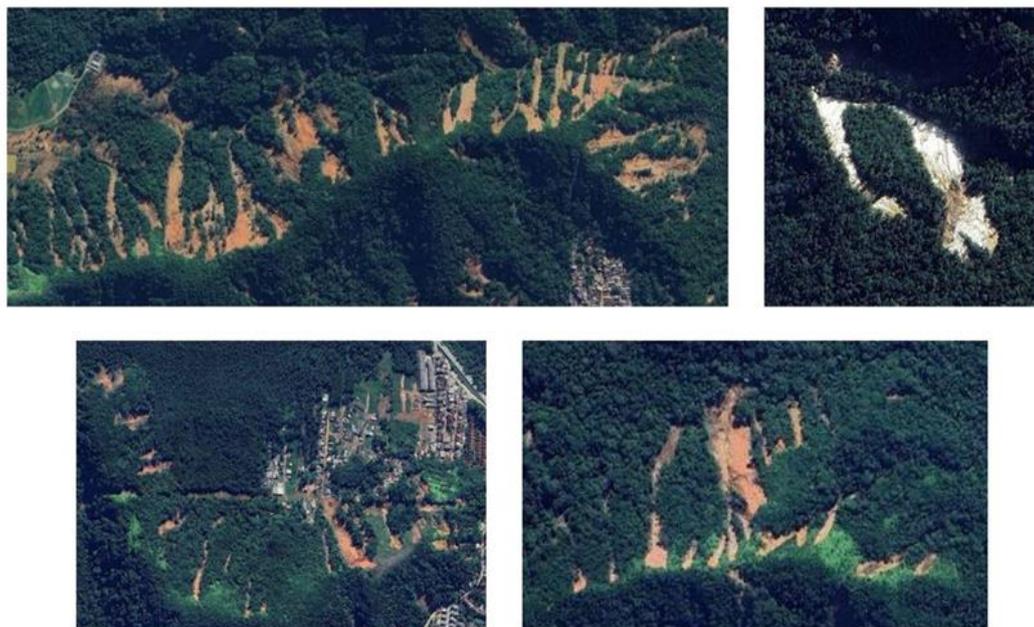
Figure 3. Landslide scars.

2.2 *Sampling for training and accuracy assessment*

Training a probabilistic landslide susceptibility model requires sampling places that represent the occurrence and non-occurrence of this phenomenon. This sample set is necessary to characterize and differentiate between areas subject and not subject to landslides (KALANTAR et al., 2017; GAIDZIK; RAMIREZ-HERRERA, 2021). The inputs required in the script for sampling are:

- roi, delimitation of the study area;
- ls_scar, delimitation of landslide scars;
- digital elevation model (DEM), the DEM used in the study.

The delimitation of the study area refers to the region where it is desired to assess susceptibility to landslides. Commonly in work in this area, the study region is defined as a watershed, relief partition, political boundary, etc (QASIMI et al., 2023). Landslide scars, on the other hand, indicate the occurrence area of this phenomenon in a specific event or in a compilation of events (OLIVEIRA et al., 2019). DEM is the base product used to create the predictive variables (PHAM, BUI; PRAKASH, 2018). At this stage, the DEM is used to define the image's pixel grid.

In this script, the definition of the study area can be done directly in the GEE, by selecting a vector from the platform's data catalog or creating the vector itself; or it can be obtained outside the platform and uploaded. Landslide scars can also be made inside or outside the platform. In the case of landslide scars, the most commonly used method for their delimitation is polygonization by visual identification comparing images from before and after the event/s of interest. There are digital image processing techniques that use optical and/or SAR data to segment the scar area. However, due to limitations such as clouds, shadows and noise inherent in data capture, the segmentation of scars using these techniques tends to be less accurate than those carried out by visual interpretation (GHORBANZADEH et al., 2019; HANDWERGER et al., 2022).

Once we have defined the study area and the landslide scars, we can generate point samples to represent the occurrence and non-occurrence of landslides. For the occurrence class, we use the boundary of the landslide scars to generate a sample for each pixel within the boundary. The pixel grid used in this sampling must be the same as that present in the defined DEM. For the non-occurrence samples, there are studies in the scientific literature that discuss the best spatial arrangement, with a variety of possible arrangements (LUCCHESE; OLIVEIRA; PEDROLLO, 2021; RABBY; LI; HIFALU, 2023). In this script, four different approaches to non-occurrence sampling were automated:

i) considering the boundary of the study area;

ii) considering areas with a slope greater than x° within the study area;

iii) considering a buffer of x meters from the landslide scars;

iiii) considering areas with a slope greater than x° within a buffer of x meters from the landslide scars.

For all four methods, a filter was used to eliminate the samples allocated within a 60 m buffer from the landslide scars. This restriction aims to mitigate possible errors in vectorization, ensuring that the neighboring pixels of the locations delimited as scars, allocated within the buffer, are disregarded from the non-occurrence sampling.

Having chosen the non-occurrence sampling method, we can generate random samples for this area. For this script, a ratio of 1:1 was established for occurrence and non-occurrence samples. However, the user can define other proportions if desired. With the two sample sets created, they are grouped and divided for training and evaluation of training accuracy.

For the case study that we carried out, the study area was obtained from the Brazilian Institute of Geography and Statistics (IBGE) and the landslide scars were delimited by visual interpretation in the Google Earth Pro software and then uploaded to the GEE. The occurrence samples were generated from the delimitation of the landslide scars and allocated to the pixel grid of the NASADEM digital elevation model. The non-occurrence samples were generated using method i. Model examples were not made for all the non-occurrence sampling methods, because this would generate an excessive number of results, and the main objective of this article is to present a probabilistic landslide modeling method and not to evaluate sampling methods. Method i was chosen among the others in order to demonstrate that the program allows the evaluation of landslide susceptibility for the whole of a study area. Often, when performing an initial analysis of a study area, researchers aim to differentiate the most susceptible areas from the others. Even knowing that in some terrain attributes (such as low slope) the probability of occurrence of landslides is very low, it is desired that the map of susceptibility to landslides covers

the entire study area investigated. However, if future users wish to exclude areas with specific terrain characteristics, it is possible to use the other methods of sampling of nonoccurrence.

A total of 1223 samples were defined for each class and 70% of the total was divided up for training and 30% for accuracy assessment. For the accuracy assessment, 366 occurrence samples and 366 non-occurrence samples were used.

2.3 Database composition

In addition to occurrence and non-occurrence samples, modeling requires defining spatial variables that can explain the phenomenon. In applications like this, it is common to use DEMs and by-products because of their ability to represent the shape of the surface. Furthermore, DEMs are available globally, facilitating their use and the replication of methodologies. Other products not related to the topographic surface, such as land use maps, pedology and geology are also frequently used. In this work, we only use variables derived from DEMs because they are globally available on the GEE platform, while data such as pedology and geology are not. However, if future users wish to add variables other than those derived from DEMs, this can be easily done (Park; Lee, 2014; Wubalem, 2021).

To compose the database, 11 morphometric indices generated from a DEM or inserted as ready-to-use products by the GEE data catalog were automated. The objective is to process and define variables relevant to the modeling of landslides exclusively in the GEE platform, enabling the replication of the procedure to other locations simply by changing the delimitation of the study area. The inputs required to define the training variables are:

- roi, delimitation of the study area
- mde, digital elevation model
- tagee, (Terrain analysis in GEE) module for terrain analysis in GEE
- fwacc, flow accumulation
- hand, height above the nearest drainage

The DEM used was NASADEM, a product resulting from the reprocessing of Shuttle Radar Topography Mission (SRTM) data with improved accuracy (NASA, 2023). If the user wishes, it is possible to define another DEM as input data for generating the variables. From this product and the terrain functions available in the GEE, it is possible to extract Slope (SLP), representing the slope of the terrain in degrees; Aspect (ASP), the orientation of the terrain in degrees and HillShade (HLSH), a gray scale of the shading of the terrain with adjustable angle of orientation and slope.

From the DEM and the TAGEE package it is possible to generate the variables Northness (NTS), showing the degree of orientation of the terrain to the north; Eastness (ETS), the degree of orientation of the terrain to the east; Horizontal Curvature (HCUR), the tangent curvature in relation to the contour line and Vertical Curvature (VCUR), the tangent curvature in relation to the slope line. The NTS and ETS variables are highly correlated with the ASP variable. The use of correlated variables tends to slow down the modeling process and add little information. However, these data were used in the modeling to illustrate the possibility of using these variables in GEE. (Safanelli et al., 2020).

The variables Height Above the Nearest Drainage (HAND), indicating the height of a pixel above the nearest drainage; and Flow Accumulation (FWACC), the size of the drained area for each pixel; were added as ready-to-use products from the GEE catalog (DONCHYTS et al., 2016; YAMAZAKI et al., 2019).

The Horizontal Distance to the Nearest Drainage (HDND) data, representing the horizontal linear distance from a cell to the nearest drainage; and the Topographic Position Index (TPI), representing the relative elevation of a cell in relation to its surroundings; were derived from other variables by implementing formulas via computer programming. The HDND variable is described by equation 1.

$$d = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2} \tag{1}$$

Where the Euclidean distance (d) of FWACC pixels with a drainage area greater than 0.5 km² (A) to the other pixels in the study area (B) is described by their difference in latitude (y) added to the difference in longitude (x).

Equation 2 was used to formulate the TPI variable.

$$TPI = p - \underline{x} \tag{2}$$

Where TPI is defined by the elevation of a pixel (*p*) subtracted by the average of a 5x5 window in its neighborhood (*x*).

If the user wishes to use variables other than those already automated, they can be added by uploading them to the platform. For our case study, the predictive variables were composed using only the data available on the platform in order to evaluate the performance of a landslide susceptibility model carried out exclusively on GEE. Variables that describe geological characteristics are widely used in this type of modeling, but because they are not available in GEE we decided not to add them to modeling. However, if users of the program wish to add geological data from outside the GEE catalog, this can be done.

After defining the variables, a data cube was created to aggregate them. They were all resampled to a pixel size of 30 m and referenced to the WGS84 coordinate system.

2.4 Random Forest classification

Probabilistic landslide susceptibility modeling is based on a supervised approach that requires the provision of data for training a classifier (REICHENBACH et al., 2018; SUN et al., 2021). To carry out this step, inputs are required:

- training samples, sample set defined for training;
- data cube, set of selected explanatory variables.

From these data, the values of the explanatory variables contained in the data cube are extracted for each sampling point where the pixel coordinates (x and y) match those of the samples. The aim of this stage is to extract information that characterizes the occurrence and non-occurrence samples, providing the classifier with data that makes it possible to differentiate them based on the values obtained from the explanatory variables.

The classifier used in this application is the Random Forest, which is non-parametric and capable of estimating a value or assigning an object to a class. A non-parametric method does not assume the distribution in space or the structure of the classifier. From the samples provided for training (containing in the attribute table: occurrence = 1, non-occurrence = 0 and the value of the explanatory variables) and using regression analysis, the classifier divides the samples into increasingly homogeneous subsets based on the values extracted from the explanatory variables. The parameter for the division is to choose the value of the prediction attribute that minimizes the uncertainty in the division of the sample set. In a decision tree, a sample is classified by the average of the subset where it is allocated in the regression analysis. Its definitive label in a Random Forest classification is the result of the classification in the various regression trees, being the value that has been repeated the most times in these various trees (BREIMAN, 2001).

A random sample subset is used to train each decision tree in order to avoid overfitting and insert variability. The number of samples for each decision tree is defined by the square root of the total number of samples used in training. A sample is usually used in more than one decision tree, generating more than one classification for the same sample point and reducing the overall error. Samples not used in the training of a specific tree are called 'Out of bag' (OOB) and are used to evaluate training performance. The OOB samples pass through the trees that did not see them in training and are classified. This classification is compared with the true label for the sample (knowing that it has been through training in other decision trees) to calculate the error. The average of the errors for all the OOB samples is the out-of-sample error estimate for a Random Forest classification. In short, the higher the generalization capacity of the Random Forest, the lower the error estimate (GOOGLE, 2023; GOOGLE EARTH ENGINE, 2023).

The importance of the variables used to train the model is assessed by the information gain obtained by splitting a subset/node. Using the GINI impurity index, which indicates the degree of mixing of the classes in the sample set, the difference between the value of a node and the value obtained by its child is estimated. This step is important for identifying which variables are most significant for the classification and to evaluate if the set of variables used allows achieving the defined objective. The Gini impurity is given by equation 3.

$$I = 1 - (O^2 + NO^2) \tag{3}$$

Where the Gini index (I) is defined by the sum of the fraction of occurrence (O) and non-occurrence (NO) samples squared and subtracted by 1. Subtracting the Gini index of a node from that obtained by its child reveals

the rate of information gain in this division. The script has been automated to display the degree of importance of the variables in percentages, making it easier to interpret the data (GOOGLE, 2023).

After training the samples, the next step is to extrapolate the identified patterns to the area of interest. Using the same data cube used for training, with information for the entire study area, a classification is made based on the values of the variables on a pixel scale, indicating for each cell its degree of susceptibility to landslides. Equation 4 represents this process

$$Px_{ij} = f(O_{ij}, NO_{ij}) \tag{4}$$

where the probability of landslides (P) occurring in a pixel (x) is determined by the values of the explanatory variables (ij) conditioned on the values of the explanatory variables in pixels defined as occurrence (Oij) and non-occurrence of landslides ($NOij$). The symbol f in the formula represents the method used to train and differentiate the occurrence and non-occurrence samples containing the values extracted from the explanatory variables.

The classification result is a landslide susceptibility surface ranging from 0 (low probability) to 1 (high probability). We recommend that the area used to extrapolate the patterns identified from the training data should be the same as that used to define the non-occurrence samples. However, if the user desires, it is possible to extrapolate to other areas. But, by altering or expanding the extrapolation area, there is no guarantee that the identified patterns will remain valid. Additionally, we have automated the grouping of susceptibility levels into classes, allowing the user to define the desired number of classes and cutoff thresholds.

For our case study, 128 decision trees were defined and the final susceptibility surface was extrapolated to the same area used to create the non-occurrence samples. This number of decision trees was defined based on the performance of the ROC curve. Several rounds were carried out and it was noticed that after 128 trees the ROC curve value increased insignificantly, while the training time increased considerably. The other parameters were not tested, being used the standard established by GEE.

After generating the susceptibility surface, it was downloaded and ArcGIS Pro software was used for cartographic representation.

2.5 Statistical evaluation of training data and results

In addition to developing a probabilistic landslide susceptibility model, it is important to statistically evaluate the samples used in the training and the results provided by the classification (Shano; Raghuvanshi; Meten, 2020; Fleuchaus et al., 2021). The inputs required at this stage are:
- roi, delimitation of the study area;
- landslide scars, delimitation of landslide scars.
- training samples, sample set used for training;
- validation samples, sample set used for accuracy assessment;
- occurrence samples, set of occurrence samples used in training;
- non-occurrence samples, set of non-occurrence samples used in training;
- data cube, explanatory variables listed;
- landslide susceptibility surface, product resulting from the classification.

At this step, the data present in the modeling is examined using four approaches: i) evaluating the distribution of the values that the training samples received from the variables contained in the data cube by BoxPlot; ii) evaluating the histogram of the values that the training samples received from the variables contained in the data cube; iii) evaluating the histogram of the values that the accuracy assessment samples received from the landslide susceptibility surface; and iiii) evaluating the histogram of the landslide susceptibility surface for the study area and for the delineation of landslide scars.

In the first approach (i), the Box Plot graph shows the minimum and maximum values observed; the median; the first quartile, defining the limit that divides ¼ of the samples with the lowest values and the third quartile, representing the limit that divides ¼ of the samples with the highest values. By comparing the Box Plot graph for the occurrence and non-occurrence samples with a specific variable from the data cube, we can analyze if the variable is capable of differentiating the two sample groups. This step is important for evaluating the data used to train the model, helping to remove or add explanatory variables with the aim of improvement and parsimony. In

the way that the script was written, the user needs to indicate which variable in the data cube they want to evaluate using the BoxPlot method.

The second evaluation approach (ii) attempts to provide complementary information to that obtained in the first (i). In this case, a histogram is used to evaluate the frequency distribution of the values that the samples provided for training received for the variables in the data cube. In short, a graph is generated where the 'x' axis shows the values of a variable, and the 'y' axis shows the frequency, the number of samples allocated to a value interval. This approach allows detailed characterization of how the samples were described by the variables used in the training, making it easier to interpret the power of sample differentiation by the variables of interest. It also makes it possible to demonstrate if the sample set is effective for the research objective, and can indicate possible failures and/or outliers in the sampling process.

The third evaluation approach (iii) intends to represent how the samples used to evaluate the model are characterized by the probability surface. In the case of a perfect fit, we would have all the occurrence samples allocated to the highest levels of susceptibility, and all the non-occurrence samples to low levels. In landslide models, it is common to have a mixture between the two sample sets due to the random sampling method. In addition to evaluating the accuracy of the training, this statistical evaluation is useful for indicating the strengths and weaknesses of the model results.

The fourth evaluation approach (iiii) aims to analyze the landslide susceptibility surface. Using a histogram, we evaluated the model's result for the entire study area and for the boundary of the landslide scars. Evaluating the frequency distribution for the whole study area is important for providing information on the amount of area allocated to each degree of susceptibility, indicating the size of the area susceptible to landslides. On the other hand, analyzing how the boundary of the landslide scars is described by the probability surface is a way of evaluating the model's performance. As the areas of the landslide scars are identified as places where the phenomenon occurs, it is expected that this area will be allocated to high levels of susceptibility. If this is not the case, there may be confusion in sampling or in the model's formulation. For this approach, the 'x' axis of the graph will represent the susceptibility to landslides in percentage; while the 'y' axis represents the number of pixels allocated to a susceptibility interval.

For our case study, all four approaches were carried out. After the data was processed in GEE, it was downloaded and Excel software was used for customization and graphical presentation.

2.6 Evaluation of model training accuracy

Evaluating the accuracy of model training is an essential modeling step. Only in this way can we measure the classifier's ability to differentiate the training samples and the amount of information added by the modeling (Fleuchaus et al., 2021). This stage requires the inputs:
- sampling for accuracy assessment, sample set with 30% of total samples;
- surface susceptibility to landslides, the result of the classification.

The evaluation of the model's training serves as a measure of the fit of the samples (containing occurrences and non-occurrences) in relation to the result of the probability surface. The sample set used to evaluate training accuracy is not used to train the model in order to guarantee a bias-free approach, evaluating the model from unknown samples.

It is important to emphasize that this stage is used to assess the accuracy of the training, not the accuracy of the final susceptibility map. In cases such as the one we performed, where the training is performed with samples of occurrence of only one event of landslides, despite dividing 30% of the occurrence samples exclusively for the accuracy assessment, these samples are spatially and temporally correlated with the samples used in the training. Thus, when we assess accuracy in this way, there is a tendency for the fit to be considerably higher because the training and validation samples are correlated. If we consider this result as the ability of the probability surface to indicate future landslide events, we will have a misleading interpretation. In these cases, in order to estimate the probability surface's ability to predict future events, it would be necessary to use uncorrelated samples from another landslide event. Knowing this, we emphasize that the results of the accuracy assessment should be interpreted with care, recognizing the limitations of the proposed method.

To estimate the training performance, we implemented the ROC (Receiver Operating Characteristic) curve method, which is widely used to evaluate binary data samples in relation to a continuous surface. The ROC curve

evaluates the rate of samples classified correctly (true positives) and incorrectly (false positives) by the model for each degree of probability of the susceptibility surface. The true positive rate can be calculated using equation 5, while the false positive rate is measured using equation 6.

$$TPR = TP/(TP + FN) \qquad (5)$$

$$FPR = FP/(FP + TP) \qquad (6)$$

where the true positive rate for each degree of probability (TPR) is obtained by dividing the number of true positives (TP) by the sum of true positives and false negatives (TP + FN). While the false positive rate for each degree of probability (FPR) is obtained by dividing the number of false positives (FP) by the sum of false positives (FP) and true positives (TP). The area under the curve (AUC), commonly used in this evaluation, is a way of simplifying the analysis of the ROC curve by averaging the value of the true positive and false negative rates of all thresholds. An AUC value of 1 indicates a perfect fit, while a value of 0.5 indicates a fit equal to randomness. Values below 0.5 represent unrealistic models.

Two main characteristics inherent in the formulation of probabilistic landslide susceptibility models can influence the performance of the ROC curve. i) Landslides are events that tend to be concentrated in relatively small areas of space. When assessing landslide susceptibility for a watershed, municipality or buffer around landslide scars, the area of occurrence will be considerably reduced, possibly less than 1% of the total study area. ii) When we generate a landslide susceptibility model for an area such as a watershed and do not use any type of restriction to allocate the non-occurrence samples (as in method i presented in section 2.2), many of these will be allocated to areas that are obviously not susceptible to landslides (such as low slopes), while another considerable portion will be allocated to areas that are potentially susceptible to landslides (such as high slope areas). This is because in these cases the non-occurrence samples are distributed randomly.

For the first situation listed (i), the occurrence samples used for training and evaluating the model will be little distributed in the study area analyzed and allocated to a considerably reduced space. This arrangement can introduce a bias into the analysis, because the occurrence of a landslide event can be induced by variables other than just terrain attributes, such as rainfall concentrated over a short period of time in specific spaces of the study area. It is important to investigate the characteristics of the event/s used to generate the model, understanding that the result of the probability surface will be relative to an event/s that occurred under specific conditions. It will be useful for predicting a future event with the same conditions as the event modeled. In the same way, despite the known difference in size between the occurrence and non-occurrence classes in the study area, the number of occurrence and non-occurrence samples has equal proportions in the accuracy assessment. A much larger number of non-occurrence samples located in areas previously known not to be susceptible to landslides would increase the accuracy assessment in a potentially misleading way.

For the second situation highlighted (ii), in non-occurrence sampling, points that are allocated to areas that are obviously not susceptible to landslides will increase the AUC value in a way that may be misleading, because a model that indicates that areas with low slopes are not susceptible to landslides does not add new information to what is already known[1]. On the other hand, non-occurrence samples that are allocated to areas susceptible to landslides will decrease the AUC in an equally misleading way. This is because the model result will consider these areas as susceptible due to their attributes being similar to those of the occurrence areas, and because the sampling is random, these areas are sampled as non-occurrence.

Pontius and Parmentier (2014) evaluated the use of the ROC curve in models with Boolean variables and indicated some recommendations for its use. Among the main suggestions, they point out that in addition to the AUC value, it is also important to analyze the shape of the ROC curve. As different curve shapes can result in

---

[1] This is an ambiguous situation. Although it is known that areas with low slopes are not susceptible to landslides, in most studies researchers aim to develop a model for their entire study area. In these cases, non-occurrence samples at low slopes are necessary for training the model. Also, although low slopes do not cause landslides, the material resulting from this phenomenon tends to be deposited in flat areas. Therefore, by using the difference image technique to delimit the landslide scars, the areas where material is deposited are sampled and inserted into the model. This adds greater variability to the occurrence data values and makes it difficult to divide the two sample groups. When evaluating the model's performance, it is important to take into account the characteristics of the phenomenon being studied and the modeling method used. Understanding these variables helps to better understand the results of the accuracy assessment.

identical AUC values, presenting the shape of the curve makes it clearer for the reader to see the model's performance. The authors point out that for applications to rare and spatially restricted phenomena (such as landslides), it is important to analyze the fit of the ROC curve at its lower limit. This is because the lower limit of the curve represents the rate of true and false positives for low susceptibility indices; a poor fit in this region would indicate gross errors. On the other hand, in landslide susceptibility models, a worse fit of the ROC curve at its upper limit is more common and acceptable. Due to the fact that non-occurrence sampling is carried out randomly, part of the sample points are allocated to high susceptibility indices, increasing the rate of false positives. Another recommendation highlighted by the researchers is to map the density of sample presence within each compartment of the ROC curve. This analysis is implemented in the code by approach (iii) in section 2.5 and should be interpreted in conjunction with the ROC curve.

In this sense, knowing the details of each user's particular application is important in order to interpret the results of the training accuracy assessment critically, enhancing communication to readers and potential users of the data. Likewise, it is also important to understand how the methods used for the evaluation work, making it possible to communicate the results transparently.

3. Results

The result of the spatial modeling for the study area is shown in figure 4. The degree of susceptibility to landslides for each pixel is represented, making it possible to differentiate between spaces considered more and less susceptible. In green are the areas with the lowest susceptibility to landslides, while in red are the areas with the highest susceptibility. In yellow are the areas considered to be of intermediate susceptibility. We can see that the locations considered to be less susceptible by the model are mainly concentrated in flat areas close to the coast, in valley bottoms and locations close to the hydrographic network, while the most susceptible areas have been allocated to hilltops, in areas with greater horizontal and vertical distance to the hydrographic network.
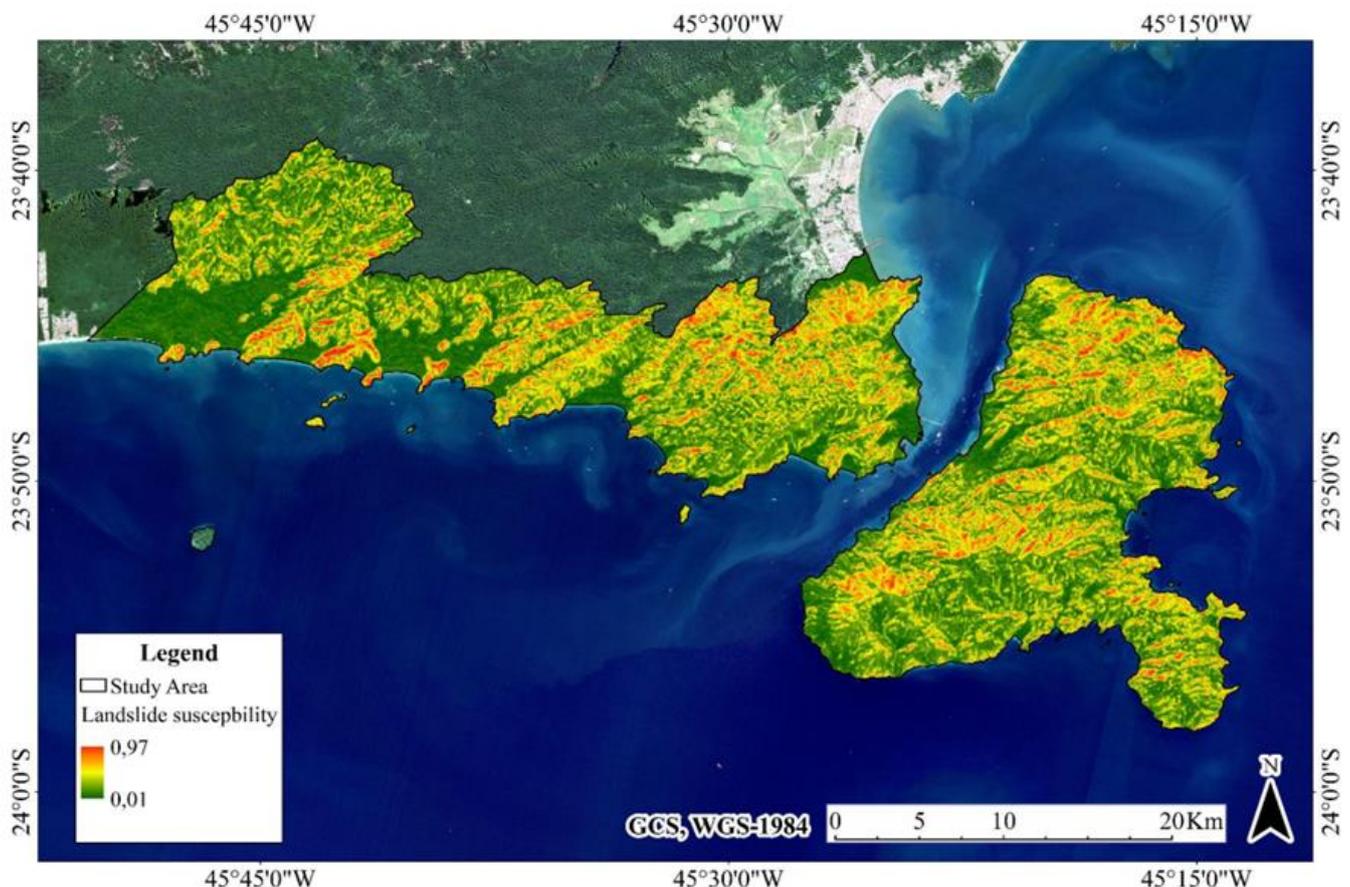
Figure 4. Landslide susceptibility map.

Figure 5 (A) shows that most of the pixels in the study area were considered to have a low susceptibility to landslides by the model result. 74% of the pixels were considered to have a susceptibility equal to or less than 0.4. While 5.4% of the pixels were considered to have a susceptibility equal to or greater than 0.7.

When we analyze figure 5 (B), representing the result of the susceptibility surface for the boundary of the landslide scars, we see that most of the pixels are allocated to the highest levels of susceptibility. 72% of the pixels are allocated to susceptibility levels equal to or greater than 0.7. While 3.4% are allocated to levels of 0.4 or less.
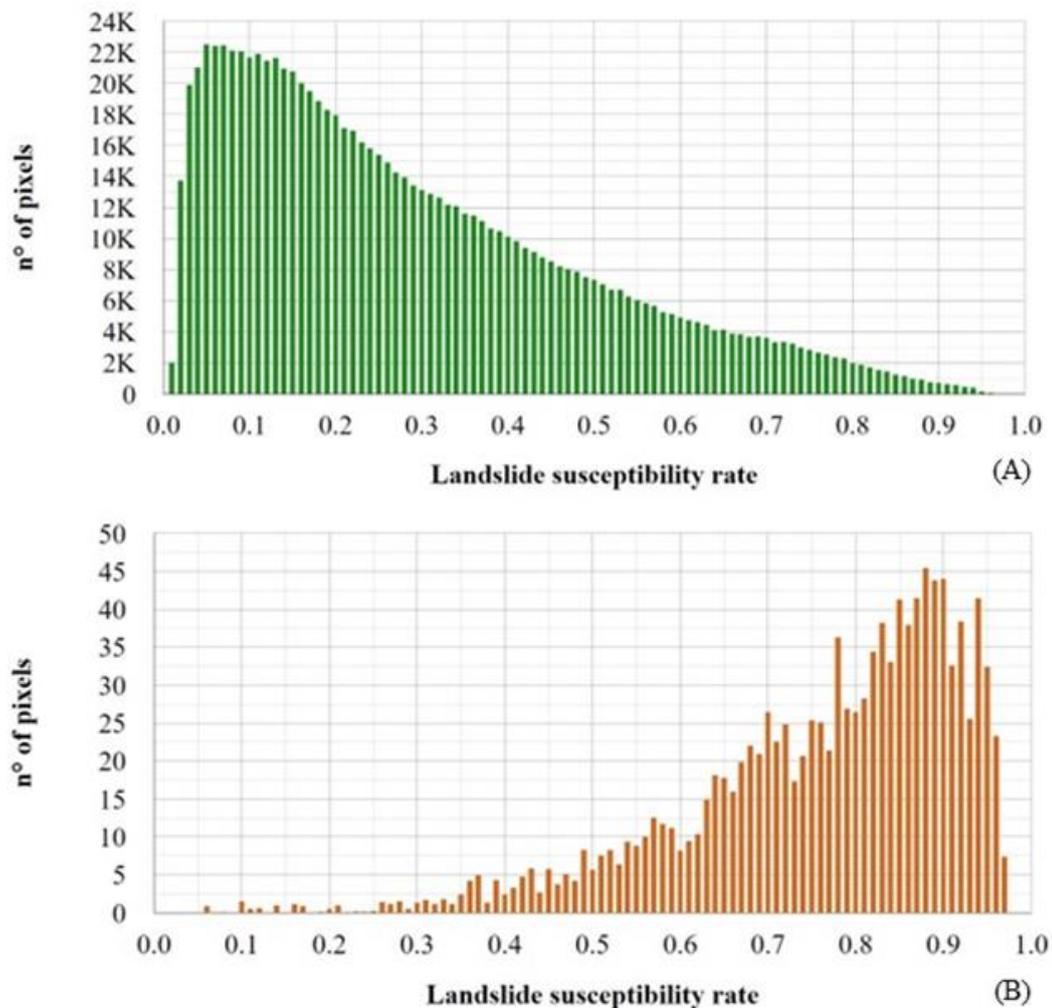
Figure 5. A) Histogram of the landslide susceptibility map for the entire study area. B) Histogram of the landslide susceptibility map for the boundary of the landslide scars.

Figure 6 shows the areas indicated by the modeling with landslide susceptibility greater than or equal to 0.7. When seeking to mitigate the human and material impacts that future landslides could cause, it is important to evaluate the areas classified by the model with high susceptibility indices, as these areas are likely to be the regions that will be affected. Thus, thinking about the organization of space, housing and infrastructure construction should be avoided in these areas in order to mitigate future damage. At the same time, knowing that landslides generate a mass flow that will only be deposited in flat areas, construction should also be avoided in areas that are likely to be depositories for the displaced sediment.
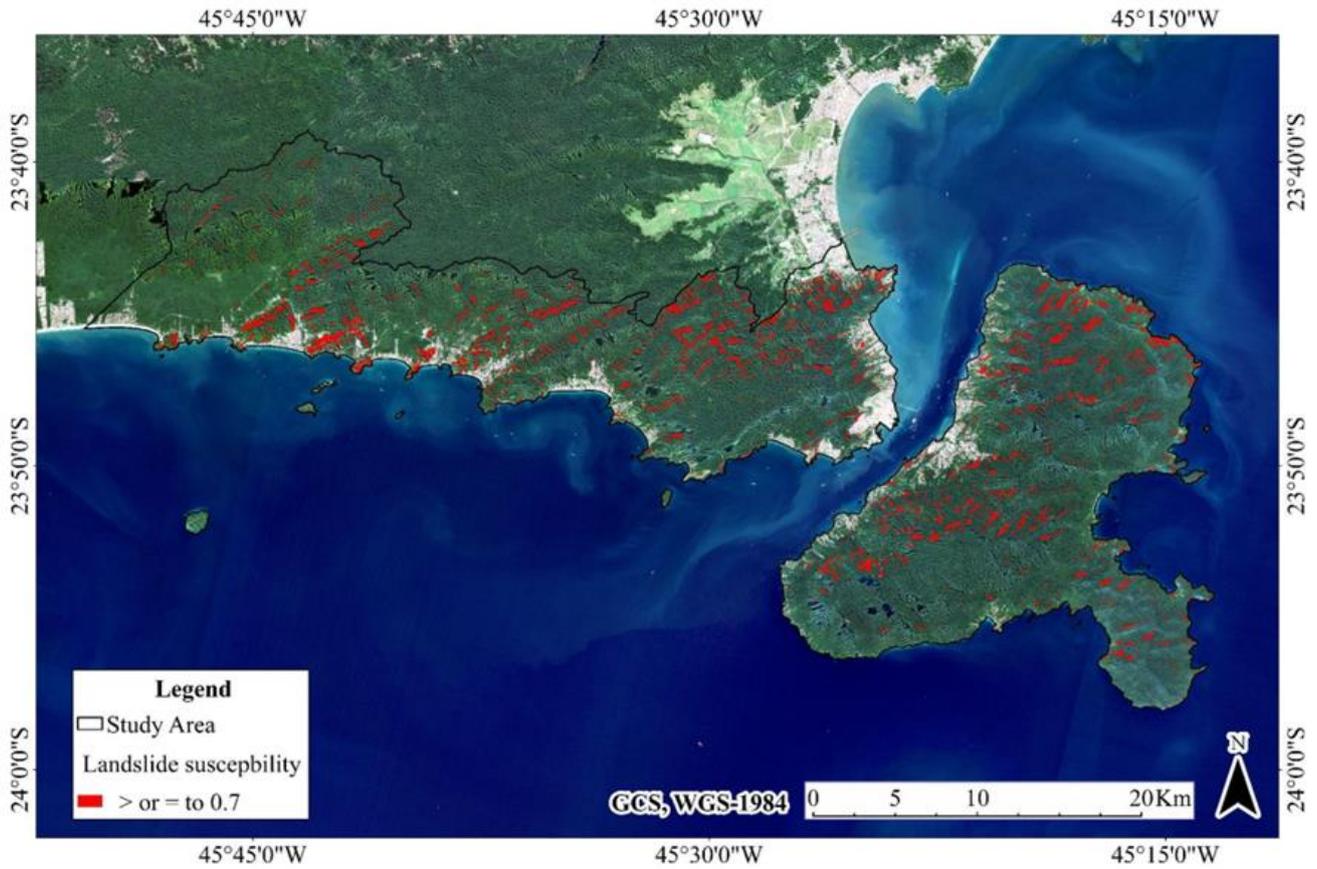
Figure 6. Landslide susceptibility map greater than or equal to 0.7.

The most important variables in training the model are shown in figure 7. FWACC, NTS and HAND were the most important, while ETS, ASP and HLSH were the least important.
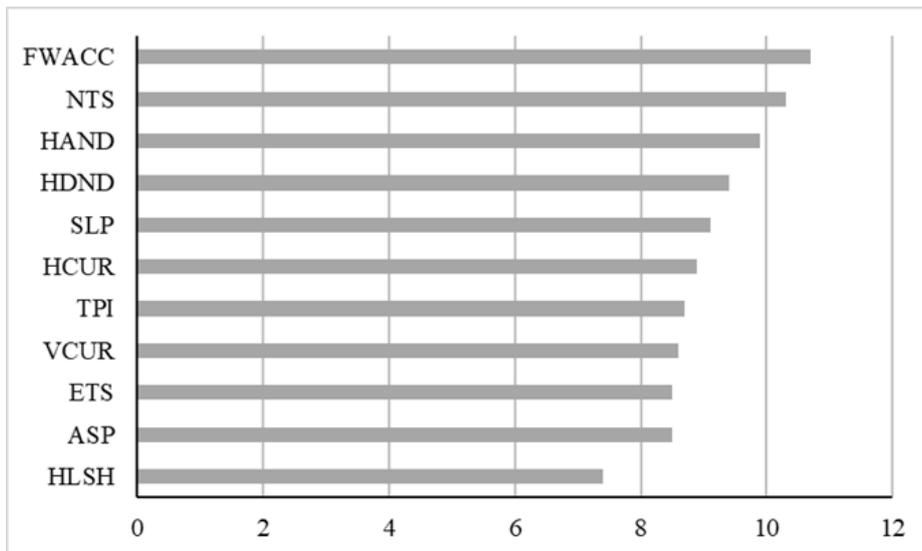
Figure 7. Importance of variables in classification.

In figure 8 we observed through Boxplot the distribution of the values that the samples used in the training received from the predictive variables. From this approach, we can evaluate the differentiation capacity of the sample set by these variables. We noticed small differences between the occurrence and non-occurrence samples for most variables. Among the most important variables in training, only NTS showed considerable differences between occurrence and non-occurrence. On the other hand, the ASP variable, which was the second least

important in training, showed significant differences between the samples of occurrence and non-occurrence. The differences that can be observed between occurrence and non-occurrence samples occur mainly in the amplitude and concentration of the data.
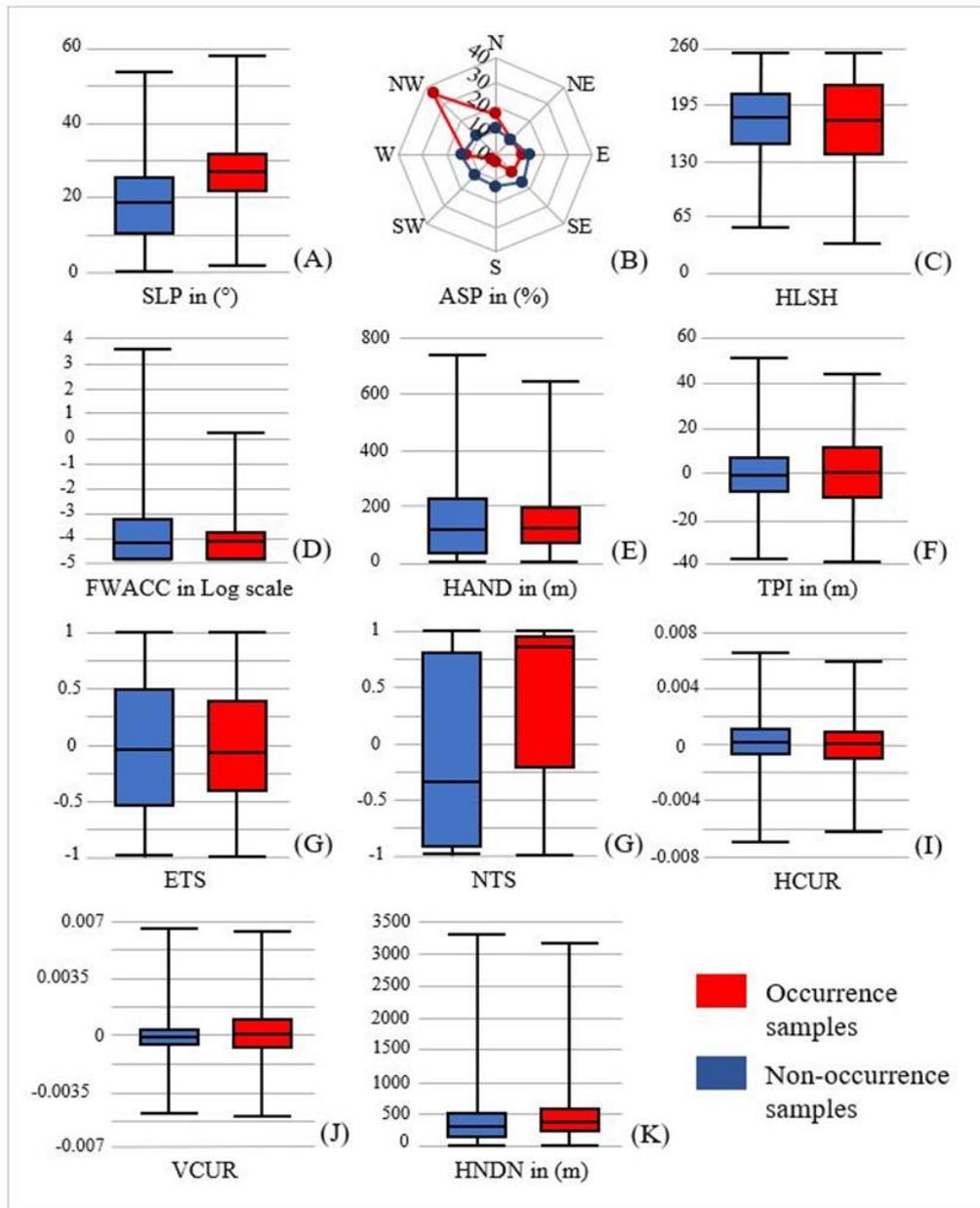
Figure 8. Boxplot of samples used to train the model.

In addition to the Boxplot approach, figures 9 and 10 present a histogram analysis for the values that the samples used in the training received from the predictive variables. This approach allows a detailed characterization of the distribution of these values, presenting information that can be difficult when we only analyze the Boxplot charts.
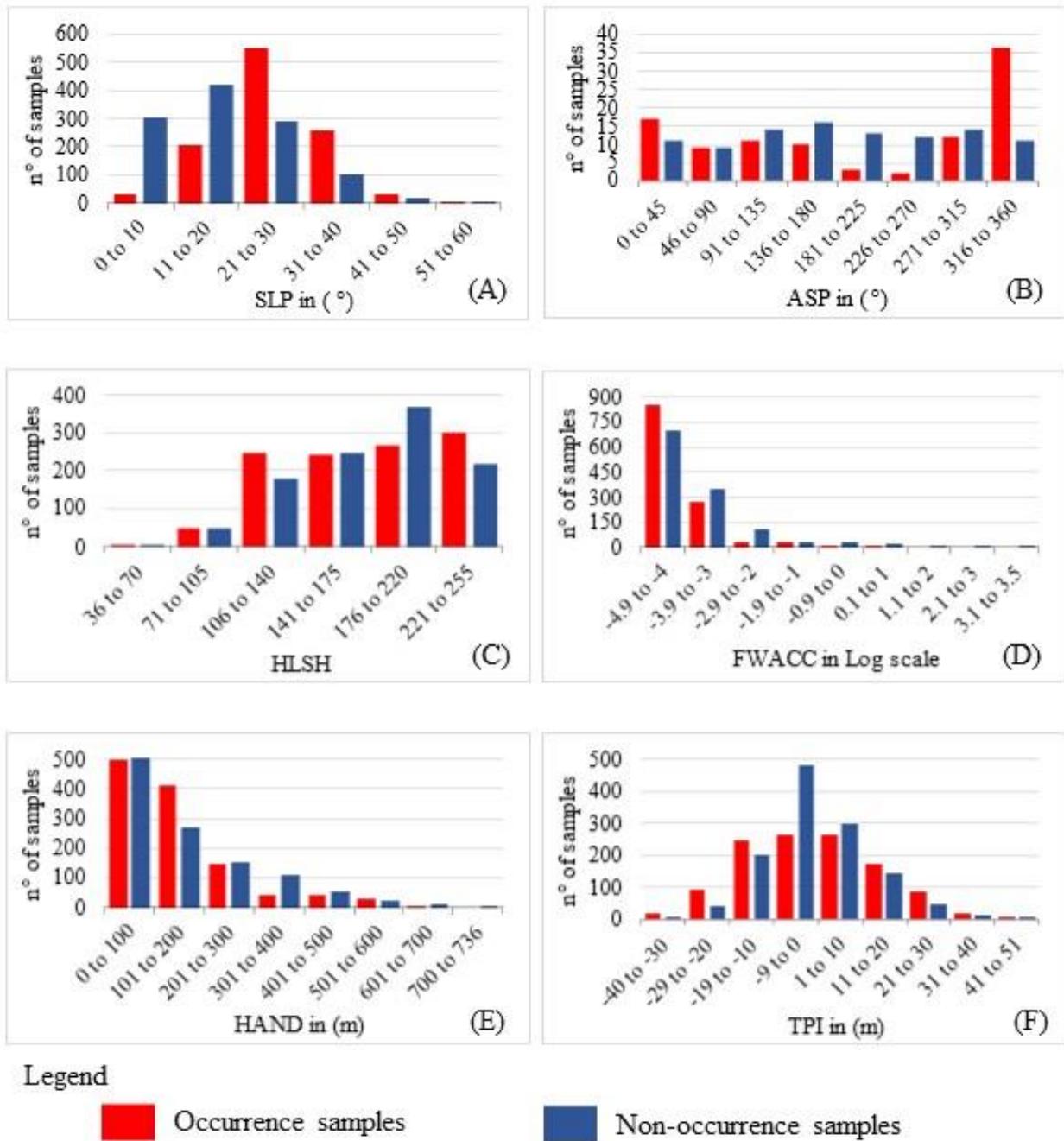
Figure 9. Histogram of samples used to train the model.

Looking at the histogram graphs, we can see that the values of some variables are very concentrated in specific ranges. This is the case of the variables FWACC and HAND, concentrated in the lowest values; NTS, concentrated in both extremes and TPI, HCUR and VCUR, concentrated in intermediate values. It is also interesting to note the specific behavior of the FWACC variable, which obtained the greatest importance in classification. Although its values for the samples of occurrence and not occurrence are very concentrated in the shortest intervals, we noticed that in the highest levels of the series there are no samples of occurrence. In this sense, despite the high concentration of samples in this variable, the classifier interprets that pixels with high values of FWACC are not susceptible to sliding. This is coherent, because the highest values of FWACC are allocated in the hydrographic network or near it, being in most cases flat areas.
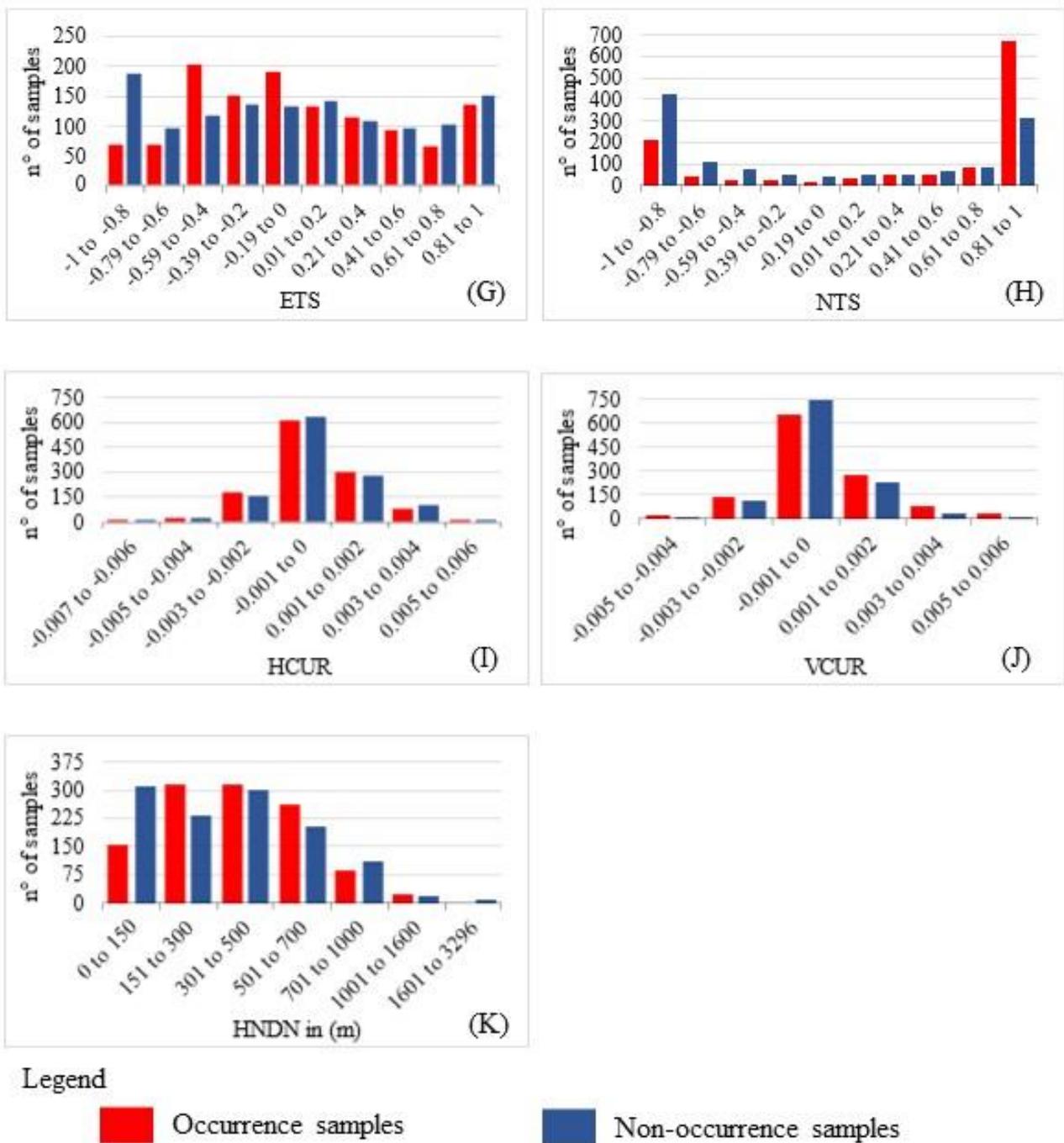
Figure 10. Histogram of samples used to train the model.

The evaluation of the model's training accuracy received an AUC value of 0.931. Figure 11 shows the shape of the ROC curve for this approach. We noticed that a good fit was obtained for the lower limit of the curve. From the susceptibility values of 0 to 0.6, the rate of true positives was close to 1 and the rate of false positives close to 0. From the level of 0.6 we observed a considerable increase in the rate of false positives and a decrease in the rate of true positives. This behavior is a result of what was exposed in section 2.6, being associated with the model formulation.
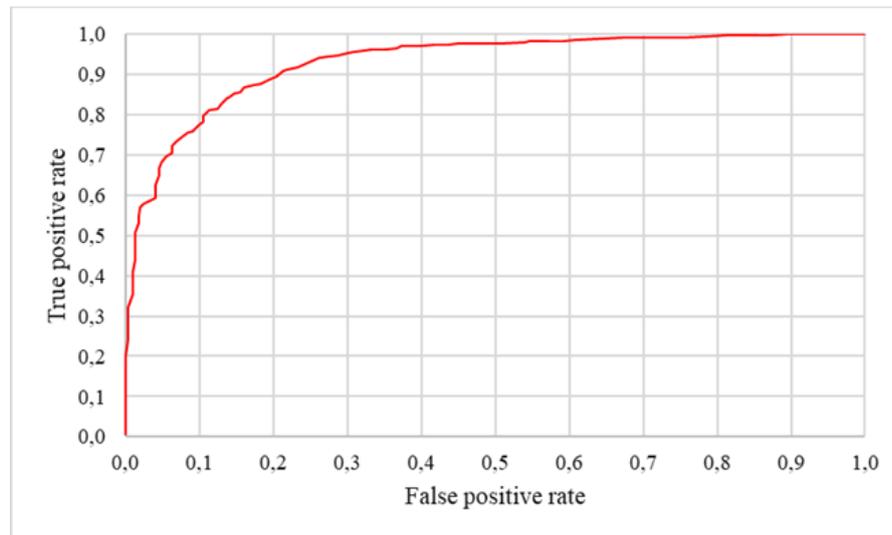
Figure 11. ROC curve for the landslide susceptibility map.

Figure 12 shows the distribution of the samples used to assess the training accuracy, we notice that the samples of occurrence are concentrated in the highest levels, while those of non occurrence in the lowest. For occurrence, 61% of the samples are allocated at susceptibility levels equal to or greater than 0.7; while for non occurrence, 76% of the samples are allocated at levels equal to or lower than 0.4. This result is as expected and helps to understand the behavior of the ROC curve. We can see that the greatest degree of mixing between occurrence and non-occurrence occurs for the intermediate values of the susceptibility surface, which is exactly the area of the two-dimensional space of the ROC curve with the worst performance.
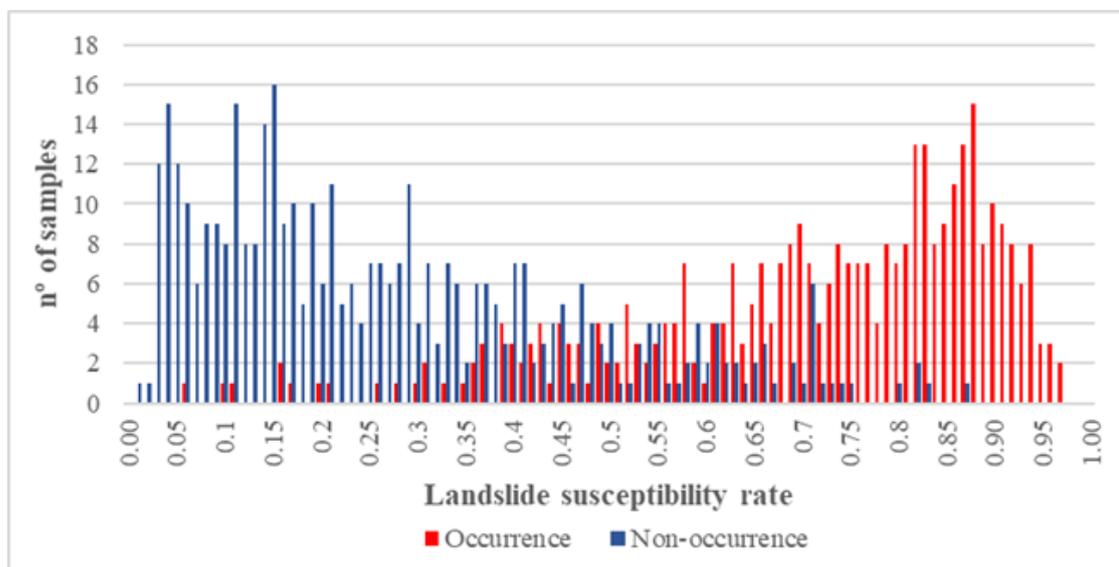
Figure 12. Histogram of the accuracy assessment samples for the landslide susceptibility map.

4. Discussion

The field of research in Remote Sensing is constantly evolving. The launch of new sensors, the availability of open data and the creation of innovative methodologies are examples of scientific advances in this area. In this context of constant progress, the GEE platform has gained considerable notability by being free and allowing the cloud processing of a wide variety of geospatial data through computational programming. The possibility of users accessing and developing their methodologies on the platform allows democratizing the act of doing science, not requiring a computer with great processing capacity or the use of paid software for conducting research (GORELICK et al., 2017).

In the scientific literature, there are not many works on modeling landslide susceptibility using the GEE platform, only two were found. The oldest available is the work by Ilmy, Darminto and Widodo (2021), where the authors assessed landslide susceptibility in southern Indonesia. And the work by Wu et al. (2022), which assessed an area in southern China. Both studies used a previously prepared landslide inventory, predictive variables with spatial information and machine learning algorithms. In our application, GEE demonstrated considerable adaptability to perform the modeling. All the necessary steps for the model formulation could be performed using exclusively the functions available on the platform with a reduced processing time. For the case study presented, all stages of the program and the resulting images were processed in less than five minutes. If we compare this method to others that require the formulation of samples and training variables in a GIS software and script processing using the memory of an offline computer, we will see very different processing times. In addition, when we compare the method presented with other approaches that assess landslide susceptibility using GEE, we observe that our approach differs from others due to the automation of sampling and the creation of training variables on the platform itself; and the availability of the code used. These features make modeling faster and facilitate access to the model by other researchers.

The formulation of a methodology for assessing susceptibility to landslides using GEE was successful. The intention in proposing and describing in detail this application is to enable researchers to have easy access to a free methodology, fast processing and low computational cost, allowing anyone to access a computer modeling method for assessing susceptibility to landslides. From the platform it was possible to create i) samples of occurrence and non occurrence of landslides; ii) morphometric indexes based on a DEM and its grouping in a data cube; iii) the training of a probabilistic model for susceptibility to landslides using Random Forest; iiii) the statistical evaluation of the samples used in the training and the results of the model and iiiii) the evaluation of the training accuracy. All these steps are widely described in the scientific literature and necessary to perform the modeling. Its open availability can facilitate the work of modelers.

On the other hand, some limitations can be highlighted. The GEE interface does not allow the editing and creation of a cartographic layout for the model's resulting image. The same is true for graphics, because the GEE interface allows relatively limited style adjustments. Therefore, if users want a better presentation of the graphs and maps resulting from the model, we recommend downloading the data and processing it in traditional software's. However, this is only a style question, because data visualization is available on the platform itself. In addition, although the platform's data catalog contains many products, it may happen that the user wishes to add a variable that is not present in the GEE database, especially for geological data. In these cases, it will be necessary to have this data for the study area of interest and the user will have to load the desired variable and insert it into the code.

For the application carried out, we consider the model's performance to be acceptable, with an ROC curve value comparable to other studies in the scientific literature. However, for a careful comparison of this methodology with others that are already established, it would be necessary to carry out case studies under controlled conditions. Furthermore, it should be emphasized that the result of the accuracy assessment refers to the training of the model. In this way, the probability surface should be used with caution, as it represents a first approximation for the study area.

5. Conclusions

As has been demonstrated, modeling landslide susceptibility requires several steps. Traditionally, researchers need to use several programs for modeling, including paid software. This often hinders the work, making it more time consuming and sometimes preventing the fulfillment of some tasks. In this work, we present a methodology for assessing landslide susceptibility on the GEE platform. The main contribution of this method is to perform all necessary steps of landslide susceptibility modeling in a single open access computation environment. Automating the workflow in a cloud environment allows any researcher to have access to a computational method that allows an initial investigation of landslide susceptibility without requiring the processing capacity of their machine. The script developed is available at the end of this article and can be reproduced by other researchers for their specific purposes.

The results of the case study presented are a first approximation to indicate the places most susceptible to landslides. It is observed that 5.4% of the study area was classified with an index of susceptibility to landslides equal to or above 0.7, indicating a large coverage and distribution of susceptible areas. These results may help in

future studies that aim to characterize in detail the locations most susceptible to landslides, making it possible to allocate resources more appropriately.

References

1.   AB'SABER, A. **Os domínios de natureza no Brasil: potencialidades paisagísticas**. 7ª Ed. Cotia-SP, Ateliê Editorial, 2010. 158p.

2.   AGHDA, F.; BAGHERI, V.; RAZIFARD, M. Landslide Susceptibility Mapping Using Fuzzy Logic System and Its Influences on Mainlines in Lashgarak Region, Tehran, Iran. Geotechnical And Geological Engineering, [S.L.], 2017. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10706-017-0365-y.

3.   ALEOTTI, P.; CHOWDHURY, R. Landslide hazard assessment: summary review and new perspectives. **Bulletin Of Engineering Geology And The Environment,** v. 58, n. 1, p. 21-44, 1999. DOI: http://dx.doi.org/10.1007/s100640050066.

4.   ARMAŞ, I. Weights of evidence method for landslide susceptibility mapping. Prahova Subcarpathians, Romania. **Natural Hazards**, v. 60, n. 3, p. 937-950, 2011. DOI: http://dx.doi.org/10.1007/s11069-011-9879-4.

5.   ARABAMERI, A.; PAL, S.REZAIE, F.; CHAKRABORTTY, R.; SAHA, A.; BLASCHKE, T.; NAPOLI, M.; GHORBANZADEH, O.; NGO, P. Decision tree based ensemble machine learning approaches for landslide susceptibility mapping. Geocarto International, [S.L.], v. 37, n. 16, p. 4594-4627, 2021. Informa UK Limited. http://dx.doi.org/10.1080/10106049.2021.1892210.

6.   ARUMUGAM, T.; KINATTINKARA, S.; VELUSAMY, S.; SHANMUGAMOORTHY, M.; MURUGAN, S. GIS based landslide susceptibility mapping and assessment using weighted overlay method in Wayanad: a part of western ghats, kerala. Urban Climate, [S.L.], v. 49, 2023. Elsevier BV. http://dx.doi.org/10.1016/j.uclim.2023.101508.

7.   AYALEW, L.; YAMAGISHI, H. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. **Geomorphology**, v. 65, n. 1-2, p. 15-31, 2005. DOI: http://dx.doi.org/10.1016/j.geomorph.2004.06.010.

8.   AYALEW, L.; YAMAGISHI, H.; UGAWA, N. Landslide susceptibility mapping using GIS-based weighted linear combination, the case in Tsugawa area of Agano River, Niigata Prefecture, Japan. **Landslides**, v. 1, n. 1, p. 73-81, 2004. DOI: http://dx.doi.org/10.1007/s10346-003-0006-9.

9.   AWAWDEH, M.; ELMUGHRABI, M.; ATALLAH, M. Landslide susceptibility mapping using GIS and weighted overlay method: a case study from north jordan. Environmental Earth Sciences, [S.L.], v. 77, n. 21, p. 1-15, 30, 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s12665-018-7910-8.

10.   BAHRAMI, Y.; HASSANI, H.; MAGHSOUDI, A. Landslide susceptibility mapping using AHP and fuzzy methods in the Gilan province, Iran. Geojournal, [S.L.], v. 86, n. 4, p. 1797-1816, 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10708-020-10162-y.

11.   BBC. **Litoral de SP foi atingido por 'evento extremo', com recorde de chuvas e elevação do mar**. Available in: <https://www.bbc.com/portuguese/articles/c3gj49n6jwno>. Access in: 17/09/2023.

12.   BIGARELLA, J. **Estrutura e origem das paisagens tropicais e subtropicais**. Florianópolis: UFSC. 2003. 426p.

13.   BRABB, E. Innovative approaches to landslide hazard and risk mapping. **International landslide symposium proceedings,** p. 17-22. Toronto, Canada. 1985.

14.   BREIMAN, L. "Random Forests". **Machine Learning**. v. 45, p. 5–32, 2001.

15.   BRITO, M. **Geoprocessamento aplicado ao mapeamento de suscetibilidade a escorregamentos no município de Porto Alegre, RS**. Dissertação. UFRGS, Porto Alegre, 2014.

16. BROWN, C.; BRUMBY, S.; GUZDER-WILLIAMS, B.; BIRCH, T.; HYDE, S.; MAZZARIELLO, J.; CZERWINSKI, W.; PASQUARELLA, V.; HAERTEL, R.; ILYUSHCHENKO, S. Dynamic World, Near real-time global 10 m land use land cover mapping. **Scientific Data**, v. 9, n. 1, p. 1-17, 2022. DOI: http://dx.doi.org/10.1038/s41597-022-01307-4.

17. BUI, D.; PRADHAN, B.; LOFMAN, O.; REVHAUG, I. Landslide Susceptibility Assessment in Vietnam Using Support Vector Machines, Decision Tree, and Naïve Bayes Models. **Mathematical Problems In Engineering** v. 2012, P. 1-26, 2012. DOI: http://dx.doi.org/10.1155/2012/974638.

18. CANTARINO, I.; CARRION, M.; GOERLICH, F.; IBAÑEZ, V. A ROC analysis-based classification method for landslide susceptibility maps. Landslides, [S.L.], v. 16, n. 2, p. 265-282, 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10346-018-1063-4.

19. CATANI, F.; LAGOMARSINO, D.; SEGONI, S.; TOFANI, V. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. **Natural Hazards And Earth System Sciences**, V. 13, n.11, p. 2815-2831, 2013. DOI: http://dx.doi.org/10.5194/nhess-13-2815-2013.

20. CLERICI, A.; PEREGO, S.; TELLINI, C.; VESCOVI, P. Landslide failure and runout susceptibility in the upper T. Ceno valley (Northen Apemmimes, Italy). **Natural Hazards**, v. 52, n. 1, p. 1-29, 2010.

21. CPRM – Serviço Geológico do Brasil. **Geologia e recursos minerais do Estado de São Paulo**. São Paulo, CPRM. 195p.

22. CROZIER, M. Deciphering the effect of climate change on landslide activity: a review. **Geomorphology**, v. 124, n. 3-4, p. 260-267, 2010. DOI: http://dx.doi.org/10.1016/j.geomorph.2010.04.009.

23. CRUDEN, D. A simple definition of a landslide. **Bulletin Of The International Association Of Engineering Geology**, v. 43. n. 1, p. 27-2, 1991. doi: http://dx.doi.org/10.1007/bf02590167.

24. DONCHYTS, G.; WINSEMIUS, H.; SCHELLEKENS, J.; ERICKSON, T.; GAO, H.; SAVENIJE, H.; GIESEN, N. Global 30m Height Above the Nearest Drainage. **Geophysical Research Abstracts**, v. 18, p. 1-1. DOI: http://dx.doi.org/10.13140/RG.2.1.3956.8880.

25. ERMINI, L.; CATANI, F.; CASAGLI, N. Artificial Neural Networks applied to landslide susceptibility assessment. **Geomorphology**, v. 66, n. 1-4, p. 327-343, 2005. DOI: http://dx.doi.org/10.1016/j.geomorph.2004.09.025.

26. FELL, R.; COROMINAS, J.; BONNARD, C.; CASCINI, L.; LEROI, E.; SAVAGE, W. Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. **Engineering Geology**, v. 102, n.3, p. 85-98, 2008. DOI: http://dx.doi.org/10.1016/j.enggeo.2008.03.022.

27. FLEUCHAUS, P.; BLUM, P.; WILDE, M.; TERHORST, B.; BUTSCHER, C. Retrospective evaluation of landslide susceptibility maps and review of validation practice. **Environmental Earth Sciences**, v. 80, n. 15, p. 1-15, 2021. DOI: http://dx.doi.org/10.1007/s12665-021-09770-9.

28. G1. **Exclusivo: Inspeção do MP identificou risco de deslizamento na Vila Sahy em 2020 e apontou crescimento na ocupação irregular do local**. Available in: <https://g1.globo.com/sp/vale-do-paraiba-regiao/noticia/2023/02/22/inspecao-do-mp-identificou-risco-de-deslizamento-na-vila-sahy-em-2020-e-apontou-crescimento-na-ocupacao-irregular-do-local.ghtml.>. Access: 17/09/2023.

29. G1. Temporal devastador no Litoral Norte de SP completa uma semana: veja resumo da tragédia. Available in: <https://g1.globo.com/sp/vale-do-paraiba-regiao/noticia/2023/02/26/temporal-devastador-no-litoral-norte-de-sp-completa-uma-semana-veja-resumo-da-tragedia.ghtml>. Access in: 17/09/2023.

30. GAIDZIK, K.; RAMÍREZ-HERRERA, M. The importance of input data on landslide susceptibility mapping. **Scientific Reports**, v. 11, n. 1, p. 1-14, 2021. DOI: http://dx.doi.org/10.1038/s41598-021-98830-y.

31. Gemitzi, A.; Falalakis, G.; Eskioglou, P.; Petalas, C. Evaluating landslide susceptibility using environmental factors, fuzzy membership functions and GIS. **Global NEST Journal**, v. 13, n. 1, p. 28-40, 2021.

32. GETACHEW, N.; METEN, M. Weights of evidence modeling for landslide susceptibility mapping of Kabi-Gebro locality, Gundomeskel area, Central Ethiopia. Geoenvironmental Disasters, [S.L.], v. 8, n. 1, 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s40677-021-00177-z.

33. GOOGLE. **Machine learning glossary**. Available in: <https://developers.google.com/machine-learning/glossary.>. Access: 17/09/2023.

34. GOOGLE EARTH ENGINE. **Smile Random Forest**. Available in: <https://developers.google.com/earth-engine/apidocs/ee-classifier-smilerandomforest>. Access: 17/09/2023.

35. GIGOVIć, L.; DROBNJAK, S.; PAMUčAR, D. The Application of the Hybrid GIS Spatial Multi-Criteria Decision Analysis Best–Worst Methodology for Landslide Susceptibility Mapping. Isprs International Journal Of Geo-Information, [S.L.], v. 8, n. 2, p. 79, 2019. MDPI AG. http://dx.doi.org/10.3390/ijgi8020079.

36. GORELICK, N.; HANCHER, M.; DIXON, M.; ILYUSHCHENKO, S.; THAU, D.; MOORE, R. Google Earth Engine: planetary-scale geospatial analysis for everyone. **Remote Sensing Of Environment**, v. 202, p. 18-27, 2017. DOI: http://dx.doi.org/10.1016/j.rse.2017.06.031.

37. GUERRA, A. **Geomorfologia: Uma atualização de bases e conceitos**. 16ª Ed. Rio de Janeiro: Bertrand do Brasil, 1994. 474p.

38. GUZZETTI, F.; CARRARA, A.; CARDINALI, M.; REICHENBACH, P. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, central italy. **Geomorphology**, v. 31, n, 1-4, p. 181-216, 1999. DOI: http://dx.doi.org/10.1016/s0169-555x(99)00078-1.

39. HANDWERGER, A.; HUANG, M.; JONES, S.; AMATYA, P.; KERNER, H.; KIRSCHBAUM, D. Generating landslide density heatmaps for rapid detection using open-access satellite radar data in Google Earth Engine. **Natural Hazards And Earth System Sciences**, v. 22, n. 3, p. 753-773, 2022. doi: http://dx.doi.org/10.5194/nhess-22-753-2022.

40. HEMASINGHE, H.; RANGALI, R.; DESHAPRIYA, N.; SAMARAKOON, L. Landslide susceptibility mapping using logistic regression model (a case study in Badulla District, Sri Lanka). Procedia Engineering, v. 212, p. 1046-1053, 2018. Elsevier BV. http://dx.doi.org/10.1016/j.proeng.2018.01.135.

41. HUANG, Y.; ZHAO, L. Review on landslide susceptibility mapping using support vector machines. **Catena**, v**.** 165, p. 520-529, 2018. DOI: http://dx.doi.org/10.1016/j.catena.2018.03.003.

42. ILMY, H., DARMINTO, M., WIDODO, A. 2021. "Application of Machine Learning on Google Earth Engine to Produce Landslide Susceptibility Mapping (Case Study: pacitan)". Iop Conference Series: Earth and Environmental Science 731 (1): 12028. doi: http://dx.doi.org/10.1088/1755-1315/731/1/012028.

43. KALANTAR, B.; PRADHAN, B.; NAGHIBI, S.; MOTEVALLI, A.; MANSOR, S. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (svm), logistic regression (lr) and artificial neural networks (ann). **Geomatics, Natural Hazards And Risk**. v. 9, n. 1, p. 49-69, 2017. DOI: http://dx.doi.org/10.1080/19475705.2017.1407368.

44. KHALIL, U.; IMTIAZ, I.; ASLAM, B.; ULLAH, I.; TARIQ, A.;QIN, S. Comparative analysis of machine learning and multi-criteria decision making techniques for landslide susceptibility mapping of Muzaffarabad district. Frontiers In Environmental Science, [S.L.], v. 10, 2022. Frontiers Media SA. http://dx.doi.org/10.3389/fenvs.2022.1028373.

45. KAMP, U.; GROWLEY, B.; KHATTAK, G.; OWEN, L. GIS-based landslide susceptibility mapping for the 2005 Kashmir earthquake region. **Geomorphology**, v. 85, p. 247-366, 2006.

46. KIM, J.; LEE, S.; JUNG, H.; LEE, S. Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. **Geocarto International**, v. 33, n. 9, p. 1000-1015, 2017. DOI: http://dx.doi.org/10.1080/10106049.2017.1323964.

47. LEE, S. Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. **International Journal Of Remote Sensing**, v. 26, n. 7, p. 1477-1491, 2005. DOI: http://dx.doi.org/10.1080/01431160412331331012.

48. LEE, S.; LEE, M.; JUNG, H.; LEE, S. Landslide susceptibility mapping using Naïve Bayes and Bayesian network models in Umyeonsan, Korea. **Geocarto International**, v. 35, n. 15, p. 1665-1679, 2019. DOI: http://dx.doi.org/10.1080/10106049.2019.1585482.

49. LEE, S.; CHOI, J. Landslide susceptibility mapping using GIS and the weight-of-evidence model". **International Journal Of Geographical Information Science**, v. 18, n. 8, p. 789-814, 2004. DOI: http://dx.doi.org/10.1080/13658810410001702003.

50. LEE, S.; HONG, S.; JUNG, H. A Support Vector Machine for Landslide Susceptibility Mapping in Gangwon Province, Korea. Sustainability, [S.L.], v. 9, n. 1, p. 48, 2017. MDPI AG. http://dx.doi.org/10.3390/su9010048.

51. LUCCHESE, L.; OLIVEIRA, G.; PEDROLLO, O. Investigation of the influence of nonoccurrence sampling on landslide susceptibility assessment using Artificial Neural Networks. **Catena**, v. 198, p. 105067, 2021. DOI: http://dx.doi.org/10.1016/j.catena.2020.105067.

52. MANTOVANI, F.; SOETERS, R.; VAN WESTEN, C. Remote sensing techniques for landslide studies and hazard zonation in Europe. **Geomorphology**, v. 15, n. 3-4, p. 213-225, 1996. DOI: http://dx.doi.org/10.1016/0169-555x(95)00071-c.

53. MARANDOLA JUNIOR, E.; MARQUES, C.; PAULA, L.; CASSANELI, L. Crescimento urbano e áreas de risco no litoral norte de São Paulo. **Revista Brasileira de Estudos de População**, v. 30, n. 1, p. 35-56, 2013. DOI: http://dx.doi.org/10.1590/s0102-30982013000100003.

54. MENDONÇA, F.; DANNI-OLIVEIRA, I. **Climatologia: noções básicas e climas do Brasil**. 1°ed. São Paulo: Oficina de Textos. 2013. 208p.

55. MERGHADI, A.; YUNUS, A.; DOU, J.; WHITELEY, J.; THAIPHAM, B.; BUI, D.; AVTAR, R.; ABDERRAHMANE, B. Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance. **Earth-Science Reviews**, v. 207, p. 103225, 2020. doi: http://dx.doi.org/10.1016/j.earscirev.2020.103225.

56. MERSHA, T.; METEN, M. GIS-based landslide susceptibility mapping and assessment using bivariate statistical methods in Simada area, northwestern Ethiopia. Geoenvironmental Disasters, [S.L.], v. 7, n. 1, 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s40677-020-00155-x.

57. NASA. 2023. **NASADEM: Creating a New NASA Digital Elevation Model and Associated Products**. Available in: <https://www.earthdata.nasa.gov/esds/competitive-programs/measures/nasadem>. Access: 17/09/2023.

58. NATH, S.; SENGUPTA, A.; SRIVASTAVA, A. Remote sensing GIS-based landslide susceptibility & risk modeling in Darjeeling–Sikkim Himalaya together with FEM-based slope stability analysis of the terrain. Natural Hazards, [S.L.], v. 108, n. 3, p. 3271-3304, 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11069-021-04823-5.

59. NOHANI; MOHARRAMI; SHARAFI; KHOSRAVI; PRADHAN; PHAM; LEE; MELESSE. Landslide Susceptibility Mapping Using Different GIS-Based Bivariate Models. Water, [S.L.], v. 11, n. 7, 2019. MDPI AG. http://dx.doi.org/10.3390/w11071402.

60. OLIVEIRA, G.; RUIZ, L.; GUASSELLI, L.; HAETINGER, C. Random forest and artificial neural networks in landslide susceptibility modeling: a case study of the fão river basin, southern Brazil. **Natural Hazards**, v. 99, n.2, p. 1049-1073, 2019. DOI: http://dx.doi.org/10.1007/s11069-019-03795-x.

61. ONAGH, M.; KUMRA, V.; RAI, P. Landslide susceptibility mapping in a part of Uttarkashi district (India) by multiple linear regression method. International Journal of Geology, Earth and Environmental Sciences, v. 2, n. 2, p. 102-120, 2012.

62. OZIOKO, O.; IGWE, O. GIS-based landslide susceptibility mapping using heuristic and bivariate statistical methods for Iva Valley and environs Southeast Nigeria. Environmental Monitoring And Assessment, [S.L.], v. 192, n. 2, 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10661-019-7951-9.

63. PARK, I.; LEE, S. Spatial prediction of landslide susceptibility using a decision tree approach: a case study of the pyeongchang area, korea. **International Journal Of Remote Sensing**, v. 35, n. 16, p. 6089-6112, 2014. DOI: http://dx.doi.org/10.1080/01431161.2014.943326.

64. PHAM, B.; BUI, D.; PRAKASH, I. Landslide susceptibility modelling using different advanced decision trees methods. **Civil Engineering And Environmental Systems**, v. 35, n. 1-4, p. 139-157, 2018 DOI: http://dx.doi.org/10.1080/10286608.2019.1568418.

65. PHAM, B; PRAKASH, I. Evaluation and comparison of LogitBoost Ensemble, Fisher's Linear Discriminant Analysis, logistic regression and support vector machines methods for landslide susceptibility mapping. Geocarto International, [S.L.], v. 34, n. 3, p. 316-333, 2017. Informa UK Limited. http://dx.doi.org/10.1080/10106049.2017.1404141.

66. PONTIUS, R.; PARMENTIER, B. Recommendations for using the relative operating characteristic (ROC). **Landscape Ecology**, v. 29, n. 3, p. 367-382, 2014. DOI: http://dx.doi.org/10.1007/s10980-013-9984-8.

67. QASIMI, A.; ISAZADE, V.; ENAYAT, E.; NADRY, Z.; MAJIDI, A. Landslide susceptibility mapping in Badakhshan province, Afghanistan: a comparative study of machine learning algorithms. **Geocarto International**, v. 38, n. 1, p. 1-22, 2023. doi: http://dx.doi.org/10.1080/10106049.2023.2248082.

68. RABBY, Y.; LI, Y.; HILAFU, H. An objective absence data sampling method for landslide susceptibility mapping. **Scientific Reports**, v. 13, n. 1, p. 1-15, 2023. doi: http://dx.doi.org/10.1038/s41598-023-28991-5.

69. RAY, R.; JACOBS, J.; BALLESTERO, T. Regional landslide susceptibility: spatiotemporal variations under dynamic soil moisture conditions. Natural Hazards, [S.L.], v. 59, n. 3, p. 1317-1337, 2011. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s11069-011-9834-4.

70. REICHENBACH, P.; ROSSI, M.; MALAMUD, B.; MIHIR, M.; GUZZETTI, F. A review of statistically-based landslide susceptibility models. **Earth-Science Reviews**, v. 180, p. 60-91, 2018. doi: http://dx.doi.org/10.1016/j.earscirev.2018.03.001.

71. ROODPOSHTI, M.; ARYAL, J.; PRADHAN, B. A Novel Rule-Based Approach in Mapping Landslide Susceptibility. Sensors, [S.L.], v. 19, n. 10, p. 2274, 2019. MDPI AG. http://dx.doi.org/10.3390/s19102274.

72. ROSSI, M. 2017. **MAPA PEDOLÓGICO DO ESTADO DE SÃO PAULO: REVISADO E AMPLIADO.** 1° ed. São Paulo: Instituto Florestal, 2017. 118p.

73. SAFANELLI, J.; POPPIEL, R.; RUIZ, L.; BONFATTI, B.; MELLO, F.; RIZZO, R.; DEMATTê, J. Terrain Analysis in Google Earth Engine: a method adapted for high-performance global-scale analysis. **Isprs International Journal Of Geo-Information**, v. 9, n. 6, p. 400, 2020. DOI: http://dx.doi.org/10.3390/ijgi9060400.

74. SHARMA, S.; MAHAJAN, A. Comparative evaluation of GIS-based landslide susceptibility mapping using statistical and heuristic approach for Dharamshala region of Kangra Valley, India. Geoenvironmental Disasters, [S.L.], v. 5, n. 1, p. 1-12, 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s40677-018-0097-1.

75. SOUZA, C et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. **Remote Sensing**, v. 12, n. 17, p. 2735, 2020. DOI: http://dx.doi.org/10.3390/rs12172735.

76. SUN, D.; XU, J.; WEN, H.; WANG, D. Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: a comparison between logistic regression and random forest. **Engineering Geology**, v. 281, p. 105972, 2021. DOI: http://dx.doi.org/10.1016/j.enggeo.2020.105972.

77. TAALAB, K.; CHENG, T.; ZHANG, Y. Mapping landslide susceptibility and types using Random Forest. **Big Earth Data**, v. 2, n. 2, p. 159-178, 2018. DOI: http://dx.doi.org/10.1080/20964471.2018.1472392.

78. TEIMOURI, M.; GRAEE, P. Evaluation of AHP and frequency ratio methods in landslide hazard zoning (case study: Bojnord urban watershed, Iran). **International Research of Applied and Basic Sciences**, 3, n. 9, p. 1978-1984, 2012.

79. ZHANG, K.; WU, X.; NIU, R.; YANG, K.; ZHAO, L. The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China. **Environmental Earth Sciences**, v. 76, n. 11, p. 1-20, 2017. DOI: http://dx.doi.org/10.1007/s12665-017-6731-5.

80. WANG, G; CHEN, X; CHEN, W. Spatial Prediction of Landslide Susceptibility Based on GIS and Discriminant Functions. Isprs International Journal Of Geo-Information, [S.L.], v. 9, n. 3, p. 144, 2020. MDPI AG. http://dx.doi.org/10.3390/ijgi9030144.

81. WANG, Y.; FANG, Z.; HONG, H. Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. Science Of The Total Environment, [S.L.], v. 666, p. 975-993, 2019. Elsevier BV. http://dx.doi.org/10.1016/j.scitotenv.2019.02.263.

82. WU, W.; ZHANG, Q.; SINGH, V.; WANG, G.; ZHAO, J.; SHEN, Z.; SUN, S. A Data-Driven Model on Google Earth Engine for Landslide Susceptibility Assessment in the Hengduan Mountains, the Qinghai–Tibetan Plateau. **Remote Sensing**, v. 14, n. 18, p. 4662, 2022. doi: http://dx.doi.org/10.3390/rs14184662.

83. WUBALEM, A. Landslide susceptibility mapping using statistical methods in Uatzau catchment area, northwestern Ethiopia. **Geoenvironmental Disasters**, v. 8, n. 1, p. 1-21, 2021. DOI: http://dx.doi.org/10.1186/s40677-020-00170-y.

84. WUBALEM, A.; METEN, M. Landslide susceptibility mapping using information value and logistic regression models in Goncha Siso Eneses area, northwestern Ethiopia. Sn Applied Sciences, [S.L.], v. 2, n. 5, p. 1-19, 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s42452-020-2563-0.

85. YALCIN, A. GIS-based landslide susceptibility mapping using analytical hierarchy process and bivariate statistics in Ardesen (Turkey): comparisons of results and confirmations. **Catena**, v. 72, n. 1. p. 1-12, 2008. DOI: http://dx.doi.org/10.1016/j.catena.2007.01.003.

86. YAMAZAKI, D.; IKESHIMA, D.; SOSA, J.; BATES, P.;ALLEN, G.; PAVELSKY, T. MERIT Hydro: a high resolution global hydrography map based on latest topography dataset. **Water Resources Research**, v. 55, n. 6, p. 5053-5073, 2019. DOI: http://dx.doi.org/10.1029/2019wr024873.

87. YEON, Y.; HAN, J.; RYU, K. Landslide susceptibility mapping in Injae, Korea, using a decision tree. Engineering Geology, [S.L.], v. 116, n. 3-4, p. 274-283, 2010. Elsevier BV. http://dx.doi.org/10.1016/j.enggeo.2010.09.009.

88. VAN WESTEN, C.; SOETERS, R.; SIJMONS, K. Digital geomorphological landslide hazard mapping of the Alpago area, Italy. **International Journal Of Applied Earth Observation And Geoinformation**, v. 2, n. 1, p. 51-60, 2000. DOI: http://dx.doi.org/10.1016/s0303-2434(00)85026-6.

89. VARNES, D. Slope movement and types and process. In: SCHUSTER. R., KRIZEK. R. 1° ed. **Landslides: Analysis and control**. Washington, National academy press. 1978. 224p.