https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021

Natural Hazards and Earth System Sciences
Open Access
Discussions

Assessing the importance of feature selection in Landslide Susceptibility for Belluno province (Veneto Region, NE Italy)

Sansar Raj Meena[1,2] *, Silvia Puliero[1], Kushanav Bhuyan[1,2], Mario Floris[1], Filippo Catani[1]

[1] Department of Geosciences, University of Padova, Padova, Italy.

[2] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, Netherlands.

* Corresponding author Email: sansarraj.meena@unipd.it

Abstract

In the domain of landslide risk science, landslide susceptibility mapping (LSM) is very important as it helps spatially identify potential landslide-prone regions. This study used a statistical ensemble model (Frequency Ratio and Evidence Belief Function) and two machine learning (ML) models (Random Forest and XG-Boost) for LSM in the Belluno province (Veneto Region, NE Italy). The study investigated the importance of the conditioning factors in predicting landslide occurrences using the mentioned models. In this paper, we evaluated the importance of the conditioning factors (features) in the overall prediction capabilities of the statistical and ML algorithms. By the trial-and-error method, we eliminated the least "important" features by using a common threshold. Conclusively, we found that removing the least "important" features does not impact the overall accuracy of the LSM for all three models. Based on the results of our study, the most commonly available features, for example, the topographic features, contributes to comparable results after removing the least "important" ones. This confirms that the requirement for the important factor maps can be assessed based on the physiography of the region. Based on the analysis of the three models, it was observed that most commonly available feature data can be useful for carrying out LSM at regional scale, eliminating the least available ones in most of the use cases due to data scarcity. Identifying LSMs at regional scale has implications for understanding landslide phenomena in the region and post-event relief measures, planning disaster risk reduction, mitigation, and evaluating potentially affected areas.

Commentato [r11]: Do you mean “thematic data”? Feature is confusing

Commentato [r12]: Interesting statement that should be explained better in the text

Commentato [r13]: Not clear

Commentato [r14]: Do you mean recover?

https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021

Natural Hazards and Earth System Sciences
Open Access
EGU
Discussions

1. Introduction

Landslides are one of the most frequently occurring natural disasters that cause significant human casualties and infrastructure destruction. Landslides are triggered by several natural and man-made ~~triggering~~ events such as earthquakes, volcanic eruptions, heavy rains, extreme winds, and unsustainable construction activities such as informal settlement development and cutting of roads along the slopes (Glade et al., 2006; van Westen et al., 2008). Extreme meteorological events such as the Vaia storm of 2018 triggered landslides and debris flow, destroyed critical infrastructures in the northern parts of Italy (Boretto et al., 2021). As reported by ~~(~~Gariano et al., (2021), in the last 50 years between 1969-2018, landslides posed a severe threat to the Italian population. Approximately, 1500 out of the 8100 municipalities in Italy have faced landslides with severe fatalities. Between the years of 1990 and 1999, 263 people were killed by landslides. Studies by ~~(~~Rossi et al.~~,~~ (2019) estimated that approximately 2500 people were killed between 1945-1990. Moreover, predictive modelling of the Italian population at risk to landslides (Rossi et al., 2019) shows massive tendency of risk to the population with data acquired between 1861-2015, emphasizing the necessity of landslide risk studies.

Commentato [r15]: Not sure this is the correct word

Therefore, to assess landslide risk and plan for suitable risk mitigation measures, it is crucial to realize the significance of landslide studies, particularly landslide susceptibility mapping (LSM). LSM ~~is an~~may provide an essential tool that incorporates the potential landslide locations (Senouci et al., 2021). The probability of a landslide occurring in a particular region owing to the effects of several causative factors is referred to as landslide susceptibility. LSM is an essential step towards landslide risk management and helps in effective mapping of the spatial distribution of probable landslide manifestations (Dai et al., 2002). In the past, researchers have used a range of models to assess landslide susceptibility using technologies such as Earth Observation (EO) and Geographic Information Systems (GIS). The ~~extraction~~ recognition and analysis of slope movements have been going on since the early 1970s (Brabb et al., 1972) and is still one of the most important ~~tools~~ component to perform LSM (Castellanos Abella and Van Westen, 2008; Catani et al., 2013; Chacón et al., 2006; Ercanoglu and Gokceoglu, 2002; Floris et al., 2011; Guzzetti et al., 2006; Liu et al., 2021; Pham et al., 2015; Reichenbach et al., 2018; Youssef and Pourghasemi, 2021).

Commentato [r16]: Not sure it is the correct word

Traditional methods such as the expert-based Analytical Hierarchy Process (AHP), multivariate statistics, data-driven Frequency Ratio (FR) have been employed for landslide susceptibility for many years, with satisfactory results (Castellanos Abella and Van Westen, 2008; Komac, 2006; Pradhan, 2010). Examples of such approaches is given in the study area~~,~~ by Floris et al. (2011) which combined traditional LSM methods with an updated online landslide database

https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

in the Veneto Region, Italy, where they used online spatial data from Italian portals for mapping landslide susceptibility at medium and large scales. Afterwards, with the development of new approaches, susceptibility modelling has advanced from traditional approaches. Presently, two approaches: (1) statistical and (2) machine learning, are practised for LSM at investigating the landslide predisposing factors and to map the geographical distribution of landslide processes. Reichenbach et al., (2018) classified landslide susceptibility models into six main groups: (1) classical statistics, (2) index-based, (3) machine learning, (4) multi-criteria analysis, (5) neural networks, and (6) others. Research by ~~(~~Reichenbach et al.~~,~~ (2018) also depicted that before 1995, only five models were used for LSM, but in recent times, an investigation of 19 other models was carried out, which yielded good results. More than 50 per cent of the methods consisting of the first five models mentioned above accounted for landslide susceptibility studies. Recent work of ~~(~~Stanley et al.~~,~~ (2021) emphasized the importance of data-driven methods in global LSM, trained to report landslide spatial occurrences between the periods of 2015-2018. The first version of the Landslide Hazard Assessment for Situational Awareness (LHASA) from their work for NASA, reported landslide occurrences with a decision tree model that first defines the intensity of one week of rainfall. LHASA version 2 used the data-driven model of XG-Boost by adding two dynamically varying factors: snow and soil moisture. However, despite advances in LSM, the advent of feature importance or the importance of the causative factors in the prediction capability of a model is not discussed enough. The need of increasing our control over the model sensitivity to system parameters changes, including those induced by anthropogenic and climate-change dynamics, is becoming a key factor in the implementation of truly efficient LSM for risk mitigation purposes. The VAIA windstorm of 2018, as a typical extreme weather event, may easily escape traditional statistical prediction schemes and represent, therefore, a challenging test for exploring the sensitivity of the various LSM models to changing factors and conditions.

One goal of this research is to look into the relative changes in LSM accuracy when the least "important" conditioning factors are removed. Feature selection in LSM is an approach in reducing landslide conditioning features to improve model performance and reduce computational costs. The purpose of this approach is to find the optimal set of conditioning ~~features~~ factors that will provide the best fit for the model to yield higher accuracy as predictions. Micheletti et al.~~,~~ (2014) emphasized the importance of feature selection in LSM and discussed the use of Machine Learning (ML) models such as Support Vector Machine (SVM), Random Forest (RF), and AdaBoost for LSM, as well as the significance of associated features within the confluence of the ML models for feature importance. However, their study did not consider geological and meteorological features like lithology, land use, and rainfall intensity for

Commentato [r17]: Please rephrase

Commentato [r18]: NOt the correct word

Commentato [r19]: It would be interesting to see where is the area affected by the VAIA windstorm

Commentato [r110]: Why cost?

https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021 c Author(s) 2021. CC BY 4.0 License.

both LSM and feature selection. Studies by Liu et al.~~,~~ (2021) depicted the improvement in the predictive capability of the so-called Feature Selected Machine Learning (FS-ML) model but also remarked on the fact that the same features may contribute differently in different ML models. In this study, we wanted to investigate post-prediction feature selection approach to improve LSM accuracy in contrast to what has been done in literature like Liu et al., (2021), where they perform pre-prediction feature importance using approaches like multi-collinearity analysis, variance inflation factor. The identification of the most ~~crucial~~ relevant features can help in monitoring the effect of extreme events (such as Vaia) on the increase of landslide hazard. This has implications for observation of the influence of extreme events on crucial factors in comprehending the changes in the evolution of hazard can be evaluated.

Commentato [r111]: What do you mean with post-prediction and pre-prediction?

Commentato [r112]: Not clear. Please rephrase

We present a study in the province of Belluno, northern Italy, with the comparison of feature or factor importance of statistical and ML models for LSM before the Vaia storm event. The results from the LSM will be then validated using the IFFI landslide inventory data for testing the various models' prediction capability with/without certain factors. We also investigate whether many of the latter features are crucial for LSM. As in many regions over the world, the same data or factor maps might not be available.

Commentato [r113]: Not clear feature vs factors

Commentato [r114]: This means that your analysis cannot be exported to other test sites?

2. Study area and Data

2.1 Study area

The area of the Belluno Province (Veneto Region, NE Italy) is part of the tectonic unit of the Southern Alps. The territory is 3,672 km² wide, stretching from north to south between the Dolomite Alps and the Venetian Pre-Alps, with elevations ranging from 42 to 3325 m above mean sea level. From a geological point of view, Dolomite Alps comprises the Hercynian crystalline basement consisting of micaschists and phyllites intruded by the Permian ignimbrites (Doglioni, 1990; Schönborn, 1999). These Paleozoic units are mainly outcropping in the NE and central-West sectors. The Middle-Upper Triassic includes carbonate, volcanic and dolomitic formations. In particular, the Upper Triassic Main Dolomite covers 14% of the whole province. Jurassic-Cretaceous limestone and marls are especially located between the Valsugana and Belluno thrusts (Sauro et al., 2013). Moreover, in the Belluno valley and in the southern part of the area, Cenozoic sediments, i.e., flysch and molasse and Quaternary glacial, alluvial and colluvial deposits are largely present. Instead, Venetian Prealps are characterized by Jurassic-Cretaceous sedimentary cover, such as layered limestones and dolomites with cherts (Compagnoni et al., 2005; Corò et al., 2015). Because of its morphological characteristics, the study area is affected by slope instability, which overlay an area of 165 km² corresponding to 6% of the province

https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021 c Author(s) 2021. CC BY 4.0 License.

(Baglioni et al., 2006). Most of the landslides ~~phenomena~~ are located in the NW (Upper basin of Cordevole River) and SE (Alpago district) sectors of the province (Figure 1). The dominant landslide types are slides (47%), rapid flows (20%), slow flows (12%), and shallow soil slips (7%) (Iadanza et al., 2021). The climate of the province of Belluno is continental. The mean annual temperature recorded in the period 1961–1990 is 7°C and the mean precipitation is 1284 mm/year (Desiato et al., 2005) with two peaks distributed in spring and autumn. In the last 27 years, temperature and rainfall intensity in the study area have increased due to climatic changes leading to more frequent meteorological conditions (ARPAV ~~(~~ Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto).

Commentato [r115]: The most important locations should be shown in the map

2.2 Landslide inventory data

The inventory of landslide phenomena in Italy (IFFI) conducted by the Italian Institute for Environmental Protection and Research (ISPRA) and the Regions and Autonomous Provinces was used in this study (Trigila et al., 2010). The IFFI Project was financed in 1997. Since 2005, the catalogue is available online and consists of point features indicating the scarp of the landslides and polygon features delineating the instabilities. The archive stores the main attributes of the landslides, such as morphometry, type of movement, rate, involved material, induced damages and mitigation measures. The inventory currently holds 620,808 landslides collected from historical documents, field surveys and aerial photointerpretation, covering an area of 23,700 km^2, which corresponds to the 7.9% of the Italian territory (Trigila and Iadanza, 2018). In the Belluno province, the IFFI inventory consists of 5934 points of landslides occurred before 2006 (Baglioni et al., 2006).

https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021

Natural Hazards and Earth System Sciences Discussions
Open Access
EGU

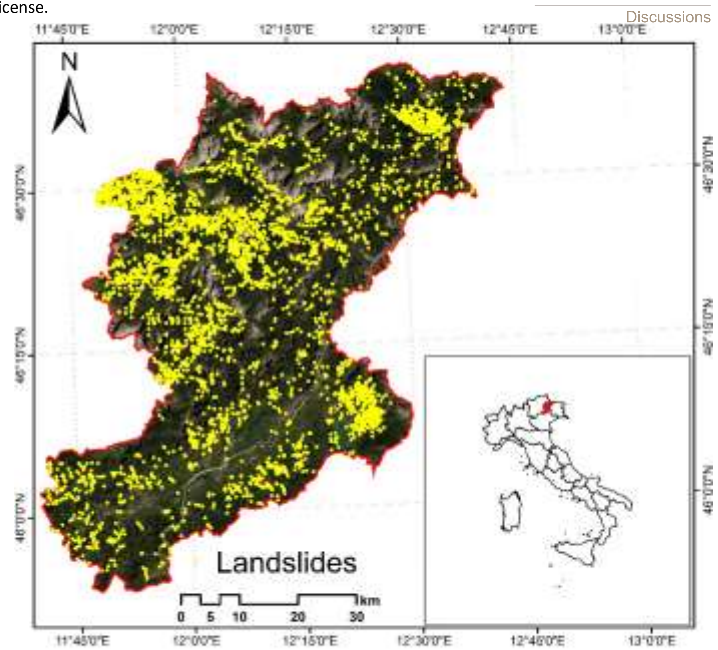


Figure 1: Location of the study area and landslides (yellow points) collected by IFFI (Inventory of Landslide Phenomena in Italy) project.

2.3 Landslide conditioning factors

Based on the ~~regional~~ environmental characteristics of the study area and the scientific literature, fourteen landslide conditioning factors were selected, including: (i) topographical factors such as elevation, slope angle, slope aspect, topographical wetness index (TWI), topographical position index (TPI), topographical roughness index (TRI), profile curvature, and plan curvature; (ii) hydrological factors (i.e., distance to drainage, precipitation); geological factors (lithology); (iii) anthropogenic factors (distance to roads); and (iv) environmental factors like Normalized Difference Vegetation Index (NDVI) and landcover (see figure 2). A freely accessible digital elevation model (DEM) with a spatial resolution of 25 metres was downloaded from the Veneto Region cartographic portal (https://idt2.regione.veneto.it), and was used to derive the topographic~~al~~ layers. Refer to table 1 for a detailed description of the conditioning factors. Land cover, lithology maps, road network and drainage maps were downloaded from the same portal. Rainfall data was

Commentato [r116]: IN the table you should provide information referred to your test area

Commentato [r117]: What is the scale of the maps?

https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021 c Author(s) 2021. CC BY 4.0 License.

downloaded from the Regional Agency for the Environmental Prevention and Protection of Veneto (ARPAV: https://www.arpa.veneto.it/) web site.

Table 1: Description of the conditioning factors for landslide occurrences.

Sl No.	Conditioning Factor	Data Range	Description/Justification
1	Elevation	42 m to 3325 m	The geomorphological and geological processes are affected by elevation (Raja et al., 2017). It has an impact on topographic characteristics, which contribute to spatial differences in many landform processes, as well as the distribution of vegetation.
2	Slope	Flat areas to very high slopes till 86.48°	Slope is a derivative of the DEM which can cause failure of slope (Pham et al., 2018). Landforms having a higher angle of slope are usually more susceptible to collapse, which is closely correlated to landslides.
3	Aspect	North (0 degrees) to North (360 degrees)	Aspect has a correlation with other geo-environmental factors is a crucial factor for LSM that describes the slope direction (Dahal et al., 2008). The slope direction to a degree dictates the frequency of landslides.
4	Topographic wetness index	-2.12 to 20.06	The influence of topography on the location and amount of saturated runoff source areas is an essential conditioning factor (Pourghasemi et al., 2012). TWI measures the amount of accumulated water and distribution of soil moisture at a location. Higher TWI values can relate to higher chances of landslide occurrence.
5	Topographic Position Index	-1143.68 to 243.84	The topographic position index (TPI) shows the difference between the elevation of a point and its surrounding defined by a specified radius. Lower values represents the plausibility of features lower than the surrounding, thus possibly relating to higher odds of landslide occurrence.

Commentato [r118]: Some of these justifications are very well and the description are not useful

Commentato [r119]: Which is the radius in your study?

https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021 c Author(s) 2021. CC BY 4.0 License.

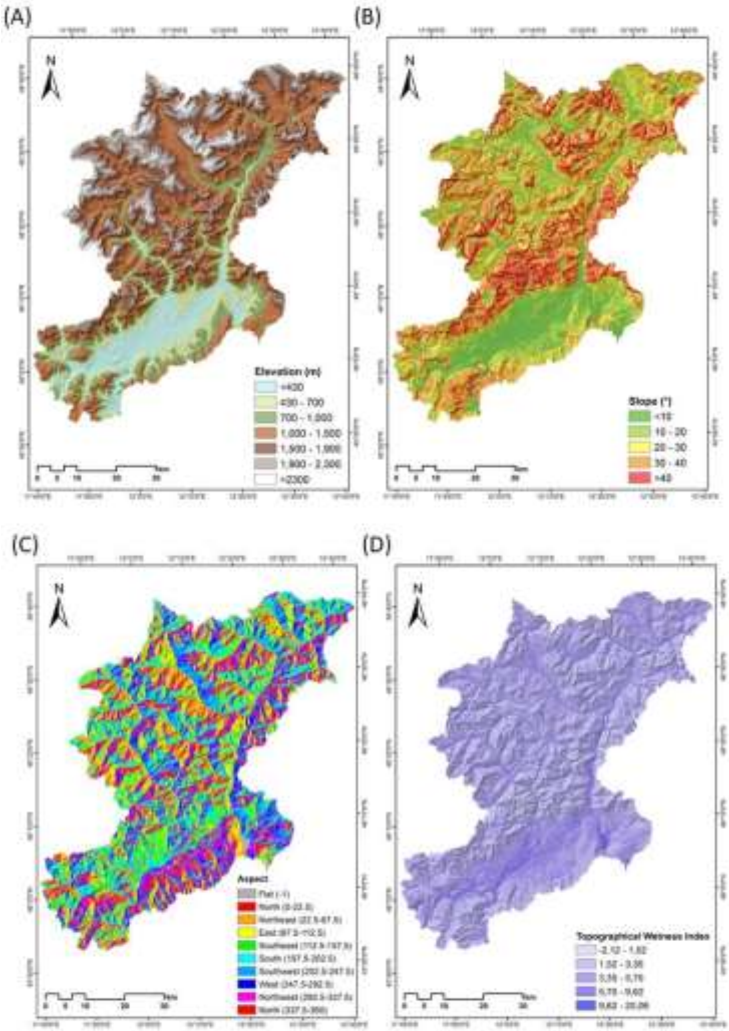
6	Topographic Roughness Index	0 to 1077.30	Topographic Roughness Index (TRI) calculates the difference in elevation between adjacent pixels in a DEM which depicts the terrain fluctuation (Riley et al., 1999). As the slope of a landscape moves, the TRI decreases, relating to slope movement.
7	Profile Curvature	Concave Flat Convex	The driving and resisting forces within a landslide in the slope direction are affected by profile curvature.
8	Plan Curvature	Concave Flat Convex	The direction of landslide movement is controlled by the plan curvature, which regulates the convergence or divergence of landslide material (Dury, 1972; Meten et al., 2015).
9	Drainage	0 to 400	Drainage transports water, which induces material saturation, culminating in landslides in valleys. (Shahabi and Hashim, 2015).
10	Rainfall	84 to 1198.05 (mm/month)	Precipitation characteristics shift by climatic conditions and geographical characteristics, resulting in significant temporal and geographical variations in rainfall quantity and intensity. This can lead to the triggering of landslides across large areas but also for specific smaller areas.
11	Lithology	Volcanites, Pre-Permian, metamorphic, sequence Morainic, Gravels, etc.	The geological strength indices, failure susceptibility, and permeability of lithological units differ (Yalcin and Bulut 2006), where changes in the stress-strain behaviour of the rock strata can be caused by lithological unit variation. Slope failure typically occurs on a slope with low strength and permeability.
12	Distance to Roads	0 to 200	A crucial manmade element impacting the occurrence of landslides is roads because of road clear-cutting and construction activities (Dunning et al., 2009).
13	Landcover	Rock, Forest, Urban cover etc.	Land cover can be utilized to describe the region's vastly dismembered zones and the likelihood of landslide activities.
14	NDVI	-0.66 to 0.66	NDVI is important in realizing the amount of vegetation cover which can be interpreted to understand the strength of the slope and the landslide occurrences. The NDVI reflects the inhibitory effect of landslide occurrence (Huang et al., 2020).

Commentato [r120]: This is not the case if you use average monthly data. What is the resolution of this information?

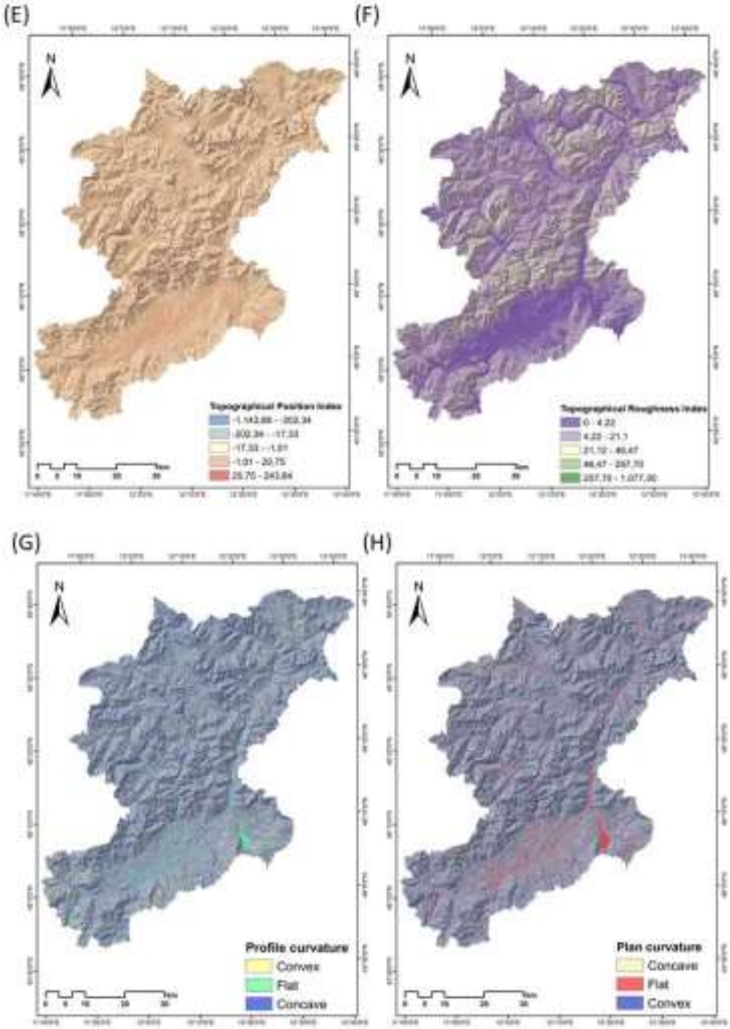
Commentato [r121]: Low permeability is not always true

Commentato [r122]: What do you mean?

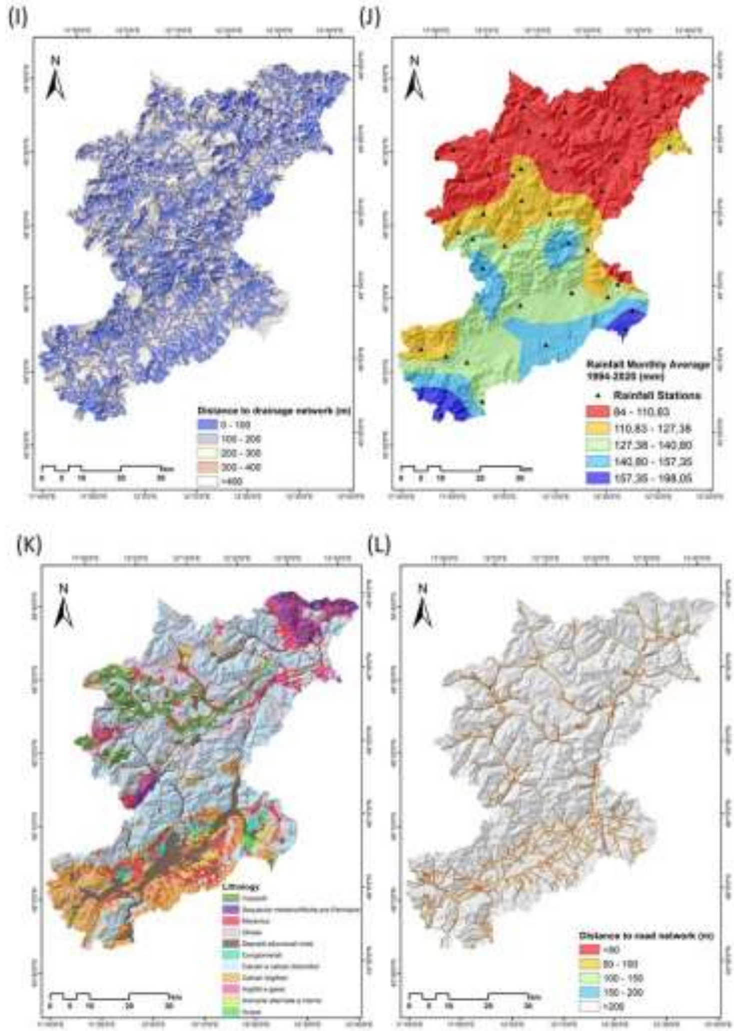
https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021



https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021 c Author(s) 2021. CC BY 4.0 License.



https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021



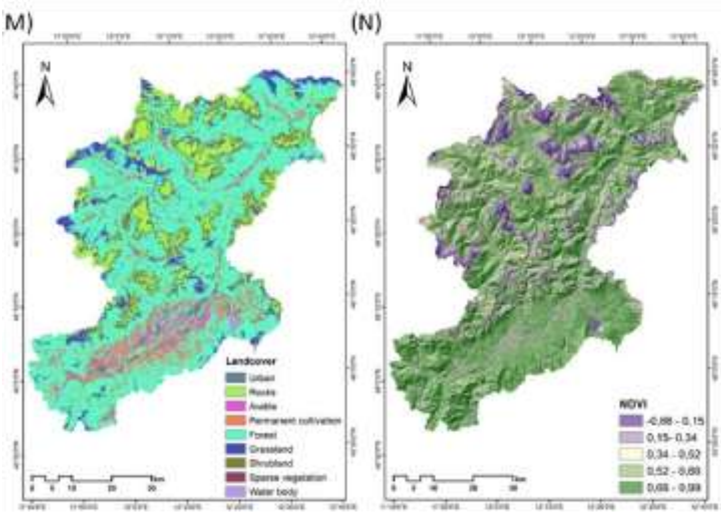


Figure 2: Maps of the conditioning factors used in this study: (A) Elevation, (B) Slope, (C) Aspect, (D) Topographical wetness index, (E) Topographical position index, (F) Topographical roughness index, (G) Profile curvature, (H) Plane curvature, (I) Distance to drainage networks, (J) Rainfall monthly average (1994-2020) mm, (K) Lithology, (L) Distance to road network (M) Landcover, (N) NDVI

Commentato [r123]: The dimension of these maps can be reduced

Commentato [r124]: The legend is in Italian

3. Methodology

Commentato [r125]: The entire chapter should be better organized. You should start with the description of the flowchart and then describe the single models.

We propose an approach that helps understand the intrinsic relationship between the features and the output post-prediction, which can be then refined by removing the less "important" features throughout the statistical and ML models. As stated previously, the study attempts the application of sensitivity analysis to understand relative feature importance as a preliminary step towards the modelling of a space-time changing parameter in LSM methods. The apparent reality is not as simple as using a certain model that gives the highest LSM accuracy and using said derived outputs maps for disaster risk management and mitigation measures. Therefore, It is important to test the effects of the features and it's relative importance in LSM. The successive sub-sections address the definitions of the statistical and ML models for LSM.

Commentato [r126]: Explain better what do you mean

Commentato [r127]: Not clear. Why space-time?

Commentato [r128]: Rephrase because it's not clear.

3.1 Statistical approach

3.1.1 Ensemble Frequency Ratio - Evidence Belief Function

In landslide susceptibility studies, the frequency ratio (FR) model is often applied. This is a ~~straightforward~~ evaluation ~~tool~~ method which calculates the likelihood of landslide occurrence and non-occurrence for each conditioning factor.

https://doi.org/10.5194/nhess-2021-299 Preprint. Discussion started: 18 November 2021

(Lee, 2013; Mondal and Maiti, 2013; Shahabi et al., 2014). For each landslide, the FR is a probabilistic model based on observed correlations between landslide distribution and related parameters (Lea Tien Tay 2014). The model depicts the relationship between spatial locations and the factors that determine the occurrence of landslides in a specific area. Spatial phenomenon and factor classes correlation can be found through FR and is very helpful for geospatial analysis (Mahalingam et al. 2016; Meena et al. 2019b). Figure 3 gives an overview of the methodology employed in this study.

Commentato [r129]: Do you mean "for each factor"?!!

Commentato [r130]: So far, you didn't explain the methodology but the FR. The flowchart should go later

The proportion of landslide inventory points for all classes within each factor can be used to compute FR weights. The area ratio for each of the factor classes in relation to the total area of the study region was calculated by overlapping the landslide inventory points with the conditioning factors. The FR weights are calculated by dividing the landslide occurrence ratio in a class by the entire area in that class (Demir et al. 2012).

Commentato [r131]: A bit confused

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

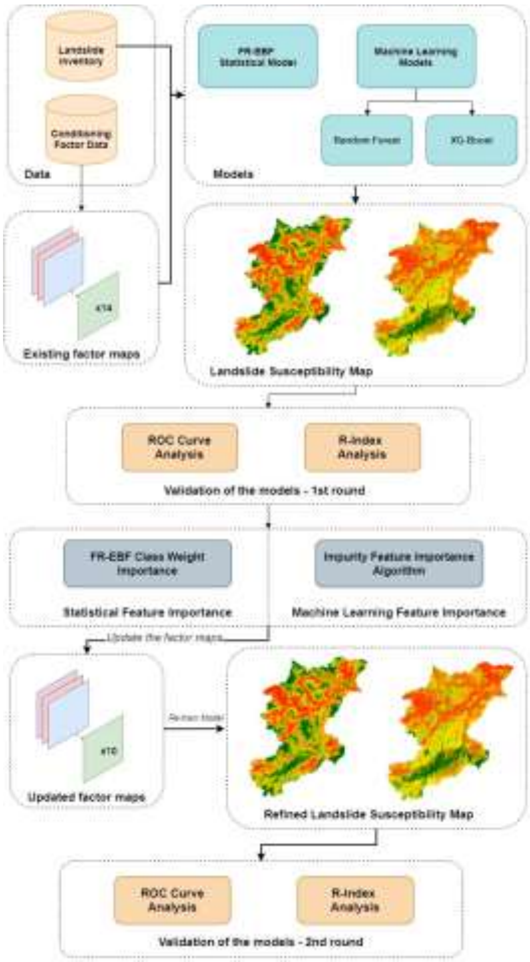


Figure 3: Overview of the conceptual workflow of methodology for landslide susceptibility assessment.

Commentato [r132]: The flowchart should show the steps described below. The "Feature selection algorithms" is for example missing

Using the equation, the landslide susceptibility index (LSI) was computed by summing the values of each factor ratio (Lee, 2013):

$$LSI = \sum FR \quad \text{(Eq.2)}$$

Commentato [r133]: Equation 1 is missing
Where is this step in the flowchart?
What is the mapping unit of your analysis?

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Natural Hazards and Earth System Sciences Discussions Open Access EGU

LSI= (DEM)+(slope)+(aspect)+(Topographic Wetness Index)+(Topographic Roughness Index)+(Topographic Position Index)+(Distance to road)+(Distance to drainage)+(Land Cover)+(Lithology)+(NDVI)+(Rainfall)+(Profile Curvature)+(Plain Curvature)

Commentato [r134]: Do you sum weights?

Where the landslide susceptibility index is the LSI, and the frequency ratio of each factor type is the FR. An FR value of 1 in the relationship analysis implies that the density of landslides in a specific class is proportionate to the size of the class in the map; an LSI value of 1 is an average value. Higher LSI values suggest a stronger correlation, whereas lower LSI values imply a weaker correlation. In a nutshell, a greater LSI value represents higher landslide susceptibility and the vice-versa. We integrated the LSI results with evidence belief functions (EBF) derived predictor values. The EBF uses the conditioning factors defined by FR as the input data. Eq. (3) was applied to the rating of every spatial factor with the training dataset.

Commentato [r135]: What do you mean?

Commentato [r136]: This means that the average value of LSI is 1?

Commentato [r137]: Something wrong… If LSI is the landslide susceptibility Index what do you mean with correlation? Correlation between what?

Commentato [r138]: ?????

Commentato [r139]: You didn't mention the training data set before. How did you define it?

$$PR = \frac{SAmax - SAmin}{SAmax - SAmin} min \quad \text{(Eq.3)}$$

where SA is the indicator of spatial association (Bel) between spatial variables and landslides and PR is the prediction rate. The lowest absolute difference of all variables is divided by the computed absolute difference between the maximum and least SA values (Table 2). The eigenvectors of the matrix were calculated by normalising each column's pairwise result. The eigenvalue was calculated by dividing each pairwise importance rate in a column by the total of the pairwise importance rates in that column. The fractional predictor is obtained by averaging the eigenvectors across a row of matrices. Pairwise comparison of the PR values of the slope failure predictors yielded the pairwise rating matrix of the predictor rating.

Commentato [r140]: Not clear. Rephrase

3.2 Machine learning models

3.2.1 Random Forest model

Random Forest (RF) is based on the ~~fundamental~~ concept of the "wisdom of crowds" where multiple decision trees, introduced by (Breiman, 2001), has been utilized in a number of remote sensing research for a variety of applications.(Melville et al., 2018). RF creates many deep

decision trees using the training data and it can overcome the overfitting problem mostly resulting from complex datasets better than other decision trees. Each RF decision tree gives a prediction, which is then weighted according to the value created from votes from each tree (see figure 4). Since the RF has shown an impressive performance for classification purposes, it is regarded as one of the most efficient non-parametric ensembles models (Chen et al., 2017).

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

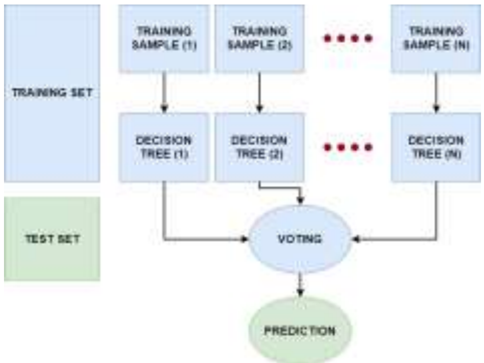


Figure 4: Conceptual diagram of the Random Forest model.

Commentato [r141]: Explain better how the model works in the case of LSM

3.2.2 XG-Boost model

Extreme gradient boosting or commonly known as the XG-Boost ML model is an optimized gradient boosting algorithm that is designed for optimum speed and performance and boosting ensembles are used to generate a prediction model. (Sahin, 2020). The core idea of a boosting algorithm is to combine the weaker learners to improve accuracy (Can et al., 2021). The model is known for its fast-training speed for classification tasks. In the study, we use training parameters to adjust the XG-Boost algorithm like learning rate, subsample ratio, maximum depth of the tree and others. It uses boosting techniques to reduce overfitting problems to improve accuracy results (figure 5).

Commentato [r142]: Explain what is the meaning in case of LSM

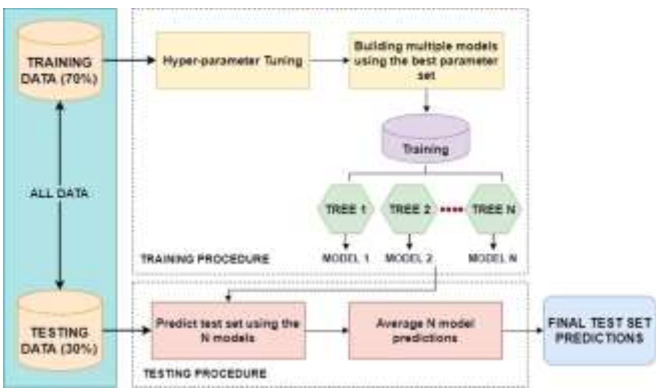


Figure 5: Training and testing procedure of the XG-Boost model.

Commentato [r143]: Explain better how the model works in the case of LSM

Natural Hazards and Earth System Sciences
Open Access
EGU
Discussions

3.3 Feature selection algorithms

The goal of feature selection is to aid in the discovery of acceptable conditions for training the models and to increase generalisability in landslide prediction. This selection help eliminates the irrelevant (less important) conditioning factors to obtain optimal prediction accuracy (Micheletti et al., 2014). For the statistical model, we used class weights obtained from frequency ratio and used them as input for generating predictor rate from FR-EBF model which gives the final weights of the conditioning factors. So, we used the predictor rate weights to select the suitable features.

In terms of the feature importance for selecting the right set of features (or factors in this case) for both RF and XG-Boost, we use the in-built impurity feature importance algorithm which is performed on the training set. Based on the results as ranks of features sorted in a descending order, the most important features will be selected to investigate the improvement of model performance in terms of the accuracy obtained. Thus, we can comment on whether certain factors are impactful in performing LSM with ML models. Besides, the comparison of the resulting important features of the different models can be interpreted to highlight the respective strengths of the models and allow drawing better conclusions towards the robustness of the relevant features for landslide predictions.

Commentato [r144]: Where is this step in the flowchart?

Commentato [r145]: Too general... what do you mean?

Commentato [r146]: Not shown in the flowchart

Commentato [r147]: Not clear

4. Results

4.1 Statistical model

The class weights were derived from data driven FR model and the final weights of the factors were derived by using predictor rate from evidence belief function given in Table 2. The class and factor weights were calculated using equations 1 and 2. The final weights of landslide conditioning factors were calculated using an ensemble of FR-EBF, and then utilised to create the final LSM. Because there is no common approach for identifying landslide susceptibility classes in the final LSM, we normalised the findings to 0 to 100 for uniformity and comparability. Using a quantile classification, which separates the values into groups with an equal number of values, the resultant LSM was classified into five classes: very low, low, moderate, high, and very high, as shown in figure 7. (Chung and Fabbri, 2003). This method of classification gives a better distribution of values in each class than common approaches such as natural breaks, which can result in certain classes having limited or excessive data. In terms of the feature importance that we observe in figure 6 and Table 2 (normalized weights), based on the trial-and-error approach, factors (or features) under the threshold of 0.3 were discarded as they did not make much of a difference in terms of predicting landslide occurrences in the study area. Therefore, five conditioning factors having coefficient values lower than 0.30 were

Commentato [r148]: Equation 1 is missing

Commentato [r149]: How did you utilize them?

Commentato [r150]: This should be figure 6

Commentato [r151]: Why 0.30?

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

dropped and overall, the area under the curve (AUC) accuracy still remained similar to the original accuracy with the 14 factors.

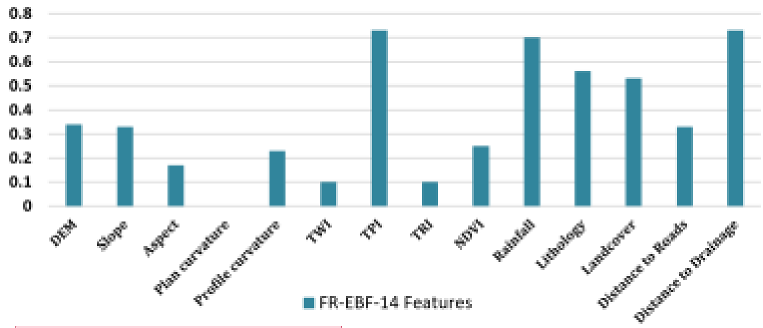


Figure 6: Feature importance of the statistical model

Commentato [r152]: What is the value of the y axes?

Table 2: Frequency ratio values for spatial factors class weighting and EBF coefficients for predictor rates (PR) based on degrees of spatial associations.

Commentato [r153]: What is?

Factors and classes	*Bel*	Min	Max	[Max-Min]	Predictor Rate	FR Weights	Normalized weights
Elevation		0.07	0.24	0.17	0.73		
<430	0.07					0.50	0.06
430 - 700	0.15					1.13	0.20
700 - 1000	0.13					0.96	0.19
1000 - 1500	0.12					0.86	0.15
1500 - 1900	0.11					0.81	0.12
1900 - 2300	0.24					1.72	0.17
>2300	0.18					1.31	0.12
Profile Curvature		0.00	0.53	0.53	2.30		
Concave	0.53					1.05	0.40
Flat	0.00					0.00	0.30
Convex	0.47					0.95	0.30
Plan Curvature		0.00	0.52	0.52	2.26		

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Concave	0.52					1.03	0.35
Flat	0.00					0.00	0.33
Convex	0.48					0.97	0.32
Slope		0.14	0.25	0.11	0.48		
<10	0.14					0.70	0.14
10 - 20	0.23					1.11	0.22
20 - 30	0.25					1.25	0.27
30 - 40	0.20					0.99	0.20
>40	0.17					0.86	0.17
Distance from drainage		0.02	0.36	0.34	1.49		
0 - 100	0.36					1.15	0.28
100 - 200	0.30					0.97	0.19
200 - 300	0.23					0.74	0.12
300 - 400	0.10					0.31	0.07
>400	0.02					0.06	0.34
Distance from roads		0.08	0.24	0.15	0.67		
0 - 50	0.36					1.15	0.27
50 - 100	0.30					0.97	0.19
100 - 150	0.23					0.74	0.17
150 - 200	0.10					0.31	0.16
>200	0.02					0.06	0.13
Landcover		0.01	0.24	0.23	2.98		
Urban	0.17					1.48	0.17
Rocks	0.10					0.90	0.09
Arable	0.01					0.07	0.01
Permanent cultivation	0.10					0.92	0.13
Forest	0.11					0.95	0.11
Grassland	0.24					2.11	0.14

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Shrubland	0.04					0.37	0.04
Sparse vegetation	0.12					1.08	0.21
Water body	0.12					1.05	0.09
TWI		0.17	0.25	0.08	1.00		
-2.12 - 1.52	0.19					1.01	0.20
1.52 - 3.35	0.20					1.04	0.20
3.35 - 5.70	0.18					0.92	0.18
5.70 - 9.62	0.17					0.90	0.18
9.62 - 20.06	0.25					1.30	0.24
TPI		0.00	0.31	0.31	1.35		
-1143.68 - -202.34	0.00					0.00	0.00
-202.34 - -17.33	0.18					0.74	0.21
-17.33 - -1.01	0.26					1.06	0.27
-1.01 - 20.75	0.24					0.98	0.26
20.75 - 243.84	0.31					1.24	0.27
TRI		0.00	0.34	0.34	1.47		
0 - 4.22	0.22					0.73	0.23
4.22 - 21.1	0.34					1.11	0.35
21.12 - 46.47	0.25					0.82	0.22
46.47 - 257.70	0.20					0.65	0.20
257.70 - 1077.30	0.00					0.00	0.00
Rainfall intensity		0.00	0.81	0.81	3.54		
84 - 110.83	0.81					11.29	0.32
110.83 - 127.38	0.08					1.15	0.27

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021

127.38 - 140.80	0.05					0.70	0.15
140.80 - 157.35	0.06					0.81	0.19
157.35 - 198.05	0.00					0.00	0.06
NDVI		0.14	0.25	0.11	0.48		
-0.66 - 0.15	0.14					0.70	0.13
0.15 - 0.34	0.22					1.13	0.21
0.34 - 0.52	0.25					1.26	0.25
0.52 - 0.66	0.21					1.07	0.21
0.66 - 0.99	0.18					0.89	0.20
Aspect		0.05	0.15	0.09	0.41		
Flat (-1)	0.11					1.02	0.10
North (0-22.5)	0.08					0.75	0.07
Northeast (22.5-67.5)	0.09					0.84	0.09
East (67.5-112.5)	0.11					1.08	0.11
Southeast (112.5-157.5)	0.14					1.31	0.14
South (157.5-202.5)	0.15					1.40	0.14
Southwest (202.5-247.5)	0.14					1.33	0.14
West (247.5-292.5)	0.08					0.76	0.09
Northwest (292.5-337.5)	0.05					0.50	0.07
North (337.5-360)	0.06					0.58	0.06

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Natural Hazards and Earth System Sciences
Discussions

Lithology		0.04	0.26	0.22	2.84		
Volcanites	0.26					3.45	0.16
Pre-Permian metamorphic sequence	0.11					1.50	0.11
Morainic	0.06					0.85	0.15
Gravels	0.04					0.52	0.04
Mix of alluvial deposits	0.05					0.70	0.03
Conglomerates	0.21					2.84	0.21
Limestone and dolomitic limestone	0.13					1.76	0.16
Calcareous shales	0.08					1.04	0.08
Shales and gypsums	0.06					0.76	0.07
Alternation of marls and sandstones	0.07					0.91	0.06
Water body	0.22					2.97	0.00

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021

Natural Hazards and Earth System Sciences
Discussions
Open Access

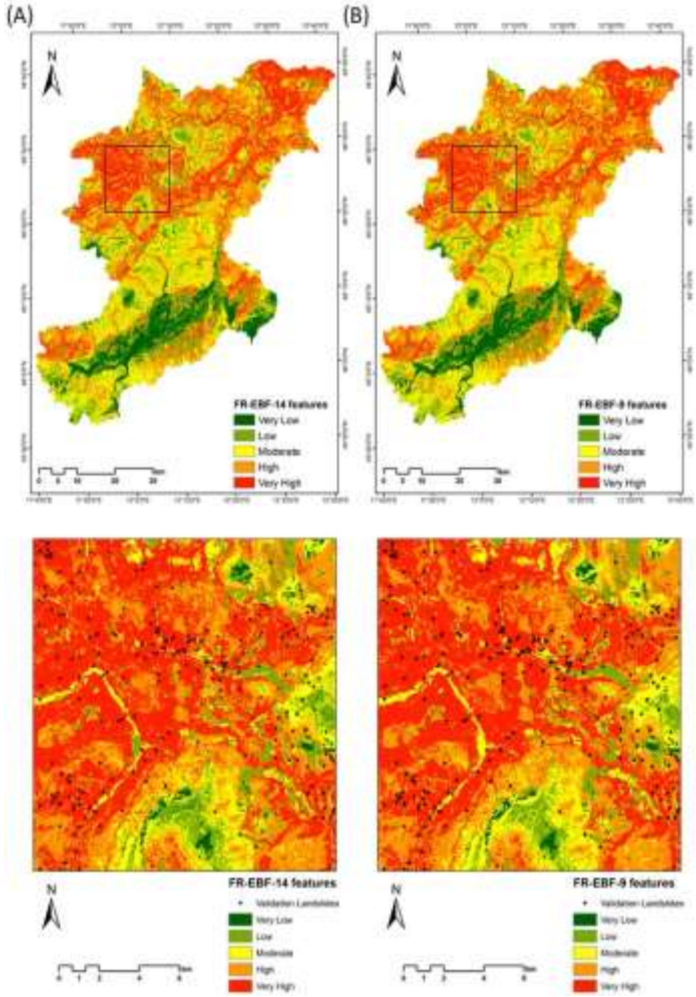


Figure 7: Landslide susceptibility maps derived using the ensemble of FR-EBF approaches for (A) 14 landslide features and (B) 9 landslide features (Black square represents the enlarged area).

Commentato [r154]: IN the map you show validation landslides. These are failures not used to prepare the model?

4.2 Machine learning models

The LSM was generated based on the conditioning factor data, where the model learnt the information from the feature maps, which helped identify areas of susceptibility. The final results of the ML models in generating the LSM are given in Table 3. We observe that the AUC scores of RF are not much apart from the XG-Boost model, indicating very good

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

prediction capability of both the models. Based on the information in Table 2, the number of pixels in the moderate susceptibility class is more in the XG-Boost model than the RF model. Visually the results show more susceptible areas near the landslide features (figures 8 and 9). The model performance in terms of the accuracy of AUC is relatively similar to the results after eliminating the lower degree of feature importance for both RF and XG-Boost. As discussed previously in section 3.3, the feature importance for the ML models is carried out using the impurity feature importance algorithm that enables to assess the relative relevance of the conditioning factors in the optimal prediction of the landslides in terms of accuracy. As seen in figure 10, the factors of Landcover, Profile Curvature, Plan Curvature, TWI and TPI have the lowest values for the RF model. After trial-and-error, a value of 0.03 was chosen as the threshold, and any factors above that were considered the "important" factors for landslide susceptibility. Hence, in figure 8, we see that the five factors mentioned above are removed and giving us 0.906 AUC as accuracy, which is better in AUC accuracy without removing the five factors (0.902 Table 3).

Similarly, the same was repeated for the XG-Boost ML model and referring to Table 3, and despite removing the lower valued conditioning factors of Profile Curvature, TPI, and Plan Curvature, the AUC accuracy score was similar (Table 3). We observe that Slope and Distance to Roads had a much bigger impact on the RF mode than the XG-Boost model. On the other hand, Lithology played a bigger role in estimating landslide occurrences in the XG-Boost model. These observations indicate interesting results which will be discussed further in the discussion section.

Table 3: Overall table with AUC results for landslide susceptibility of Belluno.

No.	Model	AUC
1	FR-EBF 14 features	0.836
2	FR-EBF 9 features	0.834
3	RF 14 features	0.902
4	RF 9 features	0.906
5	XG-Boost 14 features	0.910
6	XG-Boost 10 features	0.907

Commentato [r155]: I think this is the model skill not the prediction capability

Commentato [r156]: Not clear how it's possible to use table 2 to see the number of pixel

Commentato [r157]: This should be explained better because it's not clear

Commentato [r158]: Why?

Commentato [r159]: Why 10 and not 9?

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021

Natural Hazards and Earth System Sciences
Discussions

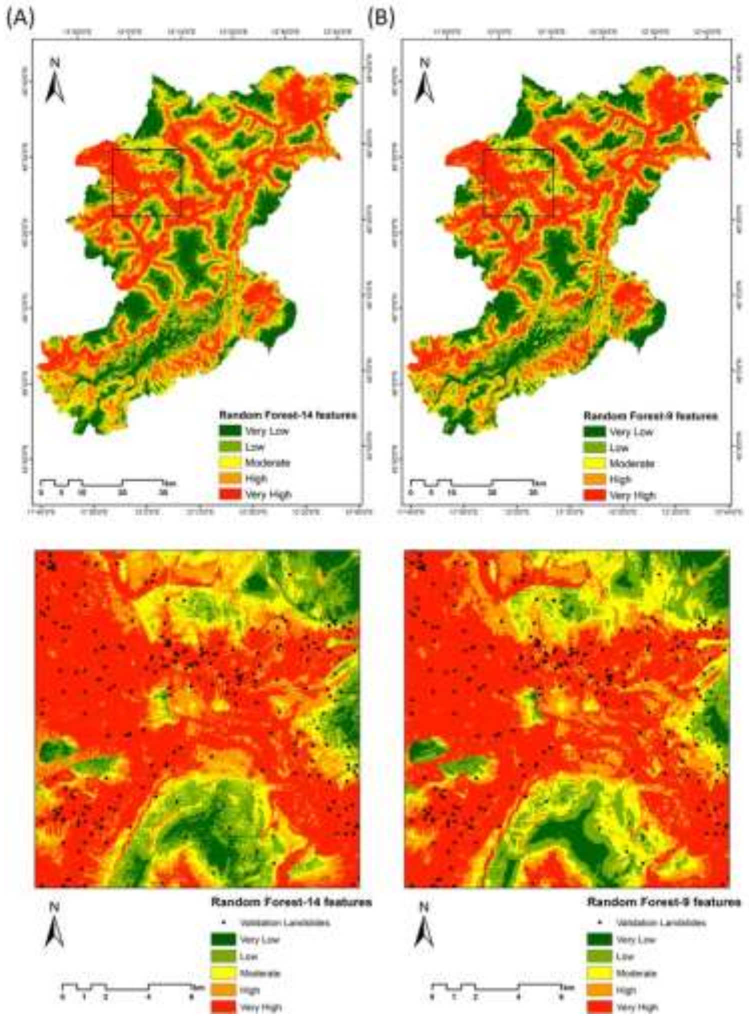


Figure 8: LSMs derived using the Random Forest approach for (A) 14 landslide features and (B) 9 landslide features (Black square represents the enlarged area).

Commentato [r160]: IN the map you show validation landslides. These are failures not used to prepare the model?

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Natural Hazards and Earth System Sciences Discussions
Open Access
EGU

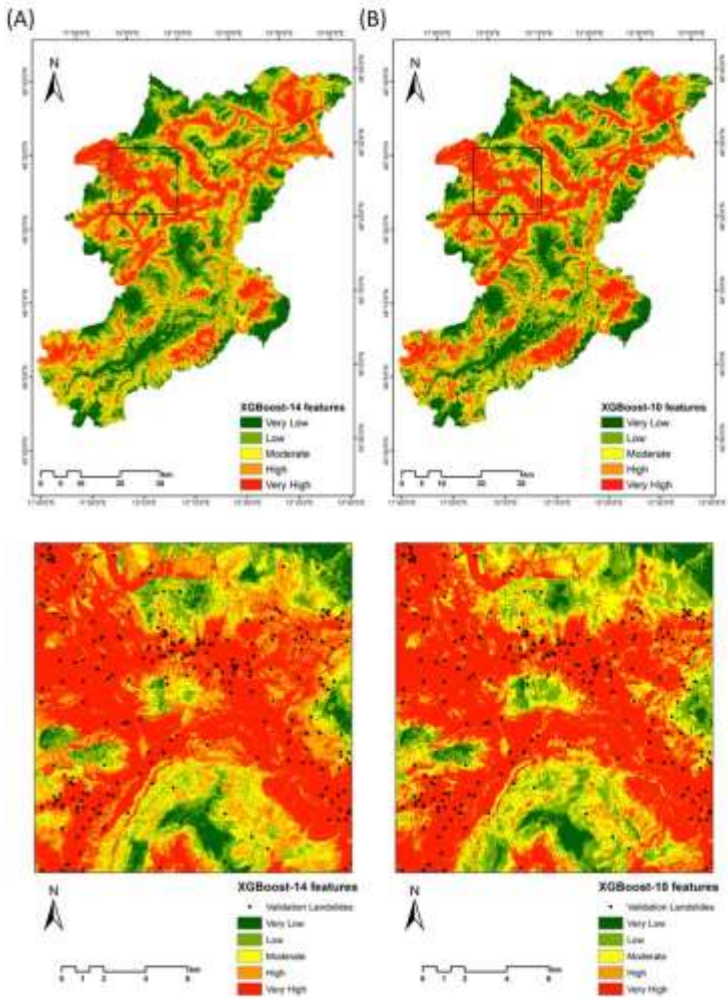


Figure 9: LSMs derived using the XG-Boost approach for (A) 14 landslide features and (B) 9 landslide features (Black square represents the enlarged area).

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Natural Hazards and Earth System Sciences Discussions Open Access EGU

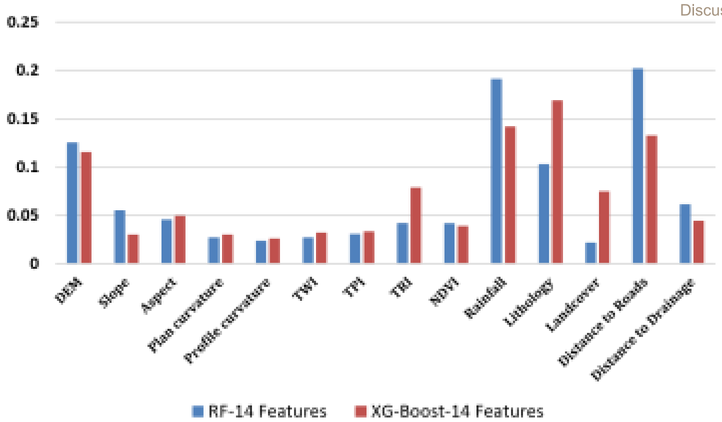


Figure 10: Feature importance of the RF and XG-Boost models.

Commentato [r161]: Please add the legend of the y-axes

5. Validation

Commentato [r162]: Is this the model skill or the validation skill?

Validation is crucial in producing quality LSMs for natural hazards where the information presented in the map is beneficial for planners (Goetz et al., 2015) A number of validation approaches may be used to assess the quality of the LSMs. We compare the landslide inventory data to the resultant maps derived using the ensemble of FR-EBF, machine learning RF and XG-Boost models. The efficiency of any model for LSM is calculated by comparing the inventory data to the produced maps. This reflects if the models in use can accurately forecast which areas are susceptible to landslides (Pourghasemi et al., 2018). The findings from the total landslide input events were validated using 30% of the landslide occurrences. Validation for this study was done using the Receiver Operating Characteristics (ROC) and the Relative Landslide Density (R-Index) approaches.

Commentato [r163]: Did you define a training set and a validation set? How did you define them?

5.1 Receiver Operating Characteristics (ROC)

The receiver operating characteristics (ROC) approach was used for this study to corroborate the six resultant LSMs from statistical and machine learning using the validation data. The ROC approach demonstrates the assessment between the true positive rate (TPR) and the false positive rate (FPR) in the resulting LSMs (Ghorbanzadeh et al., 2018; Linden, 2006). TPRs are pixels in the landslide validation data that are correctly categorised as high susceptibility, whereas FPRs are pixels that are erroneously labeled. TPRs versus FPRs are shown to create ROC curves. The AUC refers to the degree to which the generated LSMs are accurate. The AUC indicates whether more correctly labelled pixels were

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

present than incorrectly labelled pixels. Greater AUC values indicate that the susceptibility map is more accurate and the viceversa. If the AUC values are near to unity or one, the susceptibility map is meaningful. A map with a value of 0.5 is considered insignificant since it was created by chance. (Baird, 2013).

Commentato [r164]: Need to be rephrased

Figure 11 shows the accuracy values obtained using the ROC technique for the statistical approaches of FR-EBF and machine learning approaches of RF and XG-Boost. XG-Boost shows the highest accurate results with an AUC value of 0.91 and RF with 0.906, and FR-EBF with 0.836 (refer to Table 3). These results are quite good as it is closer to unity or one. The ensemble of FR-EBF shows lower AUC values than the machine learning-based XG-Boost and Random Forest. Machine learning results may vary as the models used landslides and non-landslides features as training data, whereas results of FR-EBF are derived only from the landslide data. The results could vary based on the geographical location and the selection of landslide conditioning factors as well.

Commentato [r165]: Not clear

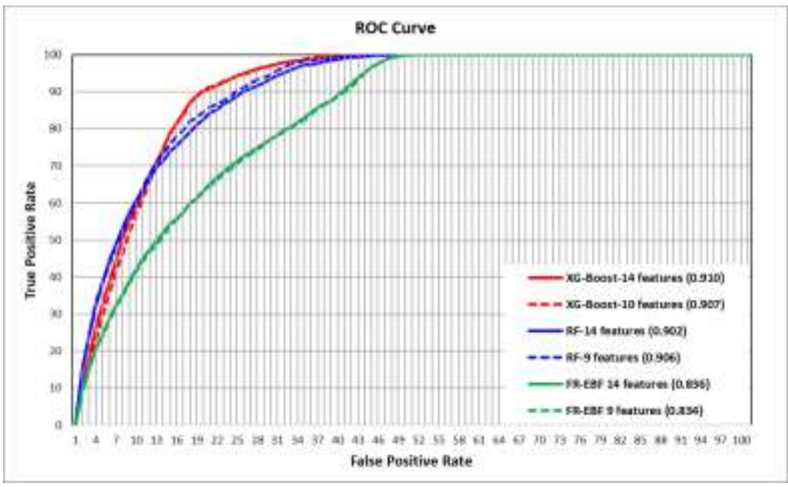


Figure 11. The ROC represents the success rate curves for the statistical based and machine learning models for LSM in Belluno province, Italy.

Commentato [r166]: ROC curve and success rate curve are two different thinks.

5.2 Relative Landslide Density (R-Index)

The relative landslide density index was also used to assess the accuracy of the LSMs ~~that resulted~~ (R-index). Equation (4) is used to get the R-index:

$$R = (ni/Ni)/\Sigma(ni/Ni) \times 100 \quad \text{(Eq.4)}$$

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

where Ni is the percentage of landslides in each susceptibility class and ni is the percentage of land susceptible to landslides in each susceptibility class. Table 4 shows the quantile classification approach to classify the six landslide susceptibility maps into five susceptible ~~groups~~classes. In comparison to the RF and FR-EBF models, the XG-Boost model with 14 and 10 features has a higher R-index for very high susceptibility classes. The R-index findings show that FR EBF has a better R-index value for high susceptibility class than XG-Boost, which has the lowest R-index for high susceptibility class. FR-EBF has a higher r-index value for the high susceptibility class than the other three approaches. In addition, the R-index of FR-EBF is higher for the very low susceptible class. Table 4 shows the R-index values for susceptibility class in FR-EBF, RF, and XG-Boost, as well as plots of the same in figure 12.

Commentato [r167]: How do you compute the percentage of land using points?

Commentato [r168]: Is this index reliable using points to compute ni?

Table 4: R-indices for the FR-EBF, RF, and XG-Boost models' landslide susceptibility mappings (LSMs).

Validation methods	Susceptibility class	Number of pixels	Area (km²)	Area (%) (ni)	Number of landslides	Landslide (%) (Ni)	R- index
FR-EBF-14 Features	Very Low	21875	334248750	9.28	48	2.71	6
	Low	90000	570760000	15.85	171	9.66	13
	Moderate	165000	896709375	24.90	308	17.40	15
	High	263750	1026578125	28.50	460	25.99	20
	Very High	444375	773585000	21.48	783	44.24	45
FR-EBF-9 Features	Very Low	19375	323332500	8.98	38	2.15	5
	Low	91875	541371875	15.03	179	10.11	15
	Moderate	153125	894758125	24.84	289	16.33	15
	High	276875	1041846875	28.93	480	27.12	21

Commentato [r170]: There is a problem with the dimension (km2???)

Commentato [r169]: You have mentioned the quantile classification to define the susceptibility classes (the classification separates the values into groups with an equal number of values). Why the number of pixel in each class is so different?

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Natural Hazards and Earth System Sciences Discussions Open Access EGU

	Very High	443750	800571875	22.23	784	44.29	44
RF-14 Features	Very Low	6875	682346250	18.94	11	0.62	1
	Low	34375	658375000	18.28	55	3.11	4
	Moderate	75625	619031875	17.19	122	6.89	9
	High	159375	749470625	20.81	264	14.92	17
	Very high	712500	892657500	24.78	1318	74.46	69
RF-9 Features	Very Low	7500	735246875	20.41	12	0.68	1
	Low	30000	632679375	17.57	48	2.71	4
	Moderate	75000	581844375	16.15	120	6.78	10
	High	147500	692276250	19.22	245	13.84	17
	Very High	729375	959834375	26.65	1345	75.99	68
XG-Boost-14 Features	Very Low	11250	1076978750	29.90	18	1.02	1
	Low	6875	330045625	9.16	11	0.62	3
	Moderate	11875	278243750	7.72	19	1.07	5
	High	11250	352568125	9.79	18	1.02	4
	Very High	947500	1564045000	43.42	1704	96.27	87
	Very Low	12500	1094226250	30.38	20	1.13	1
	Low	7500	297782500	8.27	12	0.68	3
XG-Boost-10 Features	Moderate	8125	242914375	6.74	13	0.73	4
	High	15625	314181875	8.72	25	1.41	7
	Very High	945000	1652776250	45.89	1700	96.05	84

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Natural Hazards and Earth System Sciences Discussions
Open Access
EGU

6. Discussion

Landslides are very dynamic in nature, meaning that their behaviour, movement, and spatial distribution changes over space and time. Therefore, it is vital to analyse the significance of the conditioning factors that lead to landslide occurrence. The relevance of the conditioning features for LSM is essential to realize which of the features had impact on the prediction of landslide occurrences. As not all features can be available globally, or even locally due to various restriction or data unavailability, it is essential to choose the important features which could be available for most use cases. For example, topographical features derived from digital elevation models such as Elevation, Slope, aspect, Plan curvature, Profile curvature, TWI, TPI, TRI. Other features, such as distance to roads and drainage networks, that might have direct or indirect influence on the occurrence of landslides, can also be easily accessed through numerous open-source platforms. In this study we used fourteen features for landslide susceptibility assessment and ~~caried~~carried out the feature importance test using traditional statistical ensemble model of FR-EBF and machine learning models RF and XG-Boost. The feature selection approach from statistical model is dependent upon the landslide data and its relation to each feature and their classes. On the other hand, feature selection and determining their importance using machine learning models depends upon the landslide and non-landslide samples used to train the models. We used the in-built impurity feature importance algorithm to assess the importance of the features during the model training phases. Based on literature review for this sort of study, there is no standard threshold values available for discarding or selection of features for LSM. In this study, we used a trial-and-error approach to determine a threshold of 0.30 for selection of features used for landslide susceptibility for all the three models.

Feature importance algorithms used in this study are different, however there is similarity in the importance of the features in both statistical and machine learning algorithms (See figure 6 and 10). As we look at the figures 7, 8, and 9 in the enlarged region, we observe that there are not many differences despite removing the least important features. The reason for such observation can be linked to the lower impact of least important factors on overall LSM results. Furthermore, there are several factors that determine the importance of features for carrying out LSM such as (1) completeness and quality of the landslide inventory dataset used for analysis, (2) mapping scale of the features maps like landcover, lithology, or other geological features. If the spatial locations of landslides in an inventory does not represent the ground truth phenomenon, then there can be negative impact of landslide input data for feature selection. Most importantly, the type of landslide inventory data also impacts the landslide feature selection algorithms, such as landslides mapped as points and polygons. Sampling methodology of- landslide selection is important, there are various ways to use landslides in carrying out susceptibility assessment, many studies have used 70-30 ratio and others have

Commentato [r171]: The chapter should be reorganized and revised by an English speaking person

Commentato [r172]: Explain better the importance of feature selection in Landslide Susceptibility if it doesn't affect the results.

Commentato [r173]: IN the article you do not provide this type of information

Commentato [r174]: How can you confirm this statement?

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Natural Hazards and Earth System Sciences Discussions
Open Access
EGU

used random sampling or K-fold sampling methods (Chen et al., 2018; Merghadi et al., 2018). One of the most important observations from this study was the reclusion of the "least important features" in the context of LSM. The fact that despite removal of certain factors, we still get very good results or comparable results post feature removal. This observation annotates the use of very important features for LSM which can be obtained for most of the use cases. The use of landslide samples along with non-landslide samples can affect the landslide feature importance as can be seen in results in this study. In the case of the statistical model, one of the reasons for the lower AUC performance can be accredited to the absence of the non-landslide samples. Therefore, the model's ability to discriminate between the non-landslide and landslide pixels is affected hence, predicting landslide occurrences over potentially non-landslide locations. Thus, this exhibits the homogeneous distribution of predicted landslide pixels (see figure 7). We used landslides and non-landslide samples for training the ML models which shows varying results from that of the statistical ensemble model (See figure 8 and 9). There is more homogeneous distribution of landslide susceptibility classes in statistical model results, but it is evident from the machine learning results that the non-landslide samples have a greater impact on final landslide susceptibility results.

Commentato [r175]: This means that feature selection is not relevant?

Commentato [r176]: What do you mean?

7. Conclusions

Commentato [r177]: The chapter should be reorganized and revised by an English speaking person

In context of the current state-of-the-art approaches for LSM, the contemporary literature lays emphasis on the advent of different models for improving accuracy of landslide occurrences against the test data. However, this study investigated how the conditioning factors affect the overall prediction of landslides in the context of northeast Italy, Belluno province. An important aspect of this study was to identify if at all, removing the “least important” conditioning factors in the modelling process affects the performance in predicting new unknown landslides.

As understood, ML models require conditioning factors as input for LSM, however, investing on the importance of the features (conditioning factors) could possibly direct a better understanding of landslide occurrences with respect to the available factor/feature maps for LSM. This study indicates that various models behave differently with different features, whereby the same features that are important in one instance of a particular model, can be the least important (even null-void) in other models. Therefore, this study gave new insights towards the application and use of already available maps, without spending/exhausting resources for generating other maps/features that would otherwise not be available, thus suggesting a streamlined acquisition of data and modelling of landslide occurrences for future events.

In this study we also concluded that the landslides and non-landslides samples impacts the feature importance, especially in the ML models as these models use inputs in the form of landslides and non-landslides samples. Therefore, it was

Natural Hazards and Earth System Sciences Discussions

found to be crucial in asserting a balance between the two data samples to avoid overfitting or underfitting. This study illustrates that feature selection is very important step of carrying out LSMs. We found that there are differences in the final LSMs derived from the statistical and ML models, which are attributed to the above-mentioned sample selection techniques.

This research introduces the importance of post-training feature importance algorithms for LSM. This approach can also be used to assess the susceptibility of other natural disasters. The results can eventually comment whether certain conditioning factors can be discarded while modelling landslide occurrences. In many parts of the globe, the availability of data is scarce and therefore, with the ability to model landslides without relying on the conventional factors, we can still predict landslides spatially over a given region. Although there are certain drawbacks like (1) the same factor maps will not be available everywhere, (2) factors that are least important in one region might not repeat the same behaviour in other regions of the world, and (3) model capability changes with respect to different regions, the resulting susceptibility maps can still give quality information for local emergency relief measures, planning of disaster risk reduction, mitigation, and to evaluate potentially affected areas.

Funding: This research was funded by the Veneto Region, VAIA-LANDslides project, Research Unit UNIPD-GEO, Principal Investigator Mario Floris.

Natural Hazards and Earth System Sciences
Open Access
EGU
Discussions

References:

Commentato [r178]: Check the list. Some references are not completed

Commentato [r179]: Non completed

Commentato [r180]: Non completed

Baglioni, A., Tosoni, D., De Marco, P., Arziliero, L., 2006. Analisi del dissesto da frana in Veneto.

Baird, C., 2013. Comparison of Risk Assessment Instruments in Juvenile Justice.

Boretto, G., Crema, S., Marchi, L., Monegato, G., Arziliero, L., Cavalli, M., 2021. Assessing the effect of the Vaia storm on sediment source areas and connectivity storm in the Liera catchment (Dolomites). Copernicus Meetings.

Brabb, E.E., Pampeyan, E.H., Bonilla, M.G., 1972. Landslide susceptibility in San Mateo County, California, Miscellaneous Field Studies Map, Reston, VA.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-32.

Can, R., Kocaman, S., Gokceoglu, C., 2021. A Comprehensive Assessment of XGBoost Algorithm for Landslide Susceptibility Mapping in the Upper Basin of Ataturk Dam, Turkey. Applied Sciences 11, 4993.

Castellanos Abella, E.A., Van Westen, C.J., 2008. Qualitative landslide susceptibility assessment by multicriteria analysis: A case study from San Antonio del Sur, Guantánamo, Cuba. Geomorphology 94, 453-466.

Catani, F., Lagomarsino, D., Segoni, S., Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Natural Hazards and Earth System Sciences 13, 2815-2831.

Chacón, J., Irigaray, C., Fernández, T., El Hamdouni, R., 2006. Engineering geology maps: landslides and geographical information systems. Bulletin of Engineering Geology and the Environment 65, 341-411.

Chen, T., Trinder, J.C., Niu, R., 2017. Object-oriented landslide mapping using ZY-3 satellite imagery, random forest and mathematical morphology, for the Three-Gorges Reservoir, China. Remote sensing 9, 333.

Chen, W., Peng, J.B., Hong, H.Y., Shahabi, H., Pradhan, B., Liu, J.Z., Zhu, A.X., Pei, X.J., Duan, Z., 2018. Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. Science of the Total Environment 626, 1121-1135.

Chung, C.-J.F., Fabbri, A.G., 2003. Validation of Spatial Prediction Models for Landslide Hazard Mapping. Natural Hazards 30, 451-472.

Compagnoni, B., Galluzzo, F., Bonomo, R., Tacchia, D., 2005. Carta geologica d'Italia. Dipartimento difesa del suolo.

Corò, D., Galgaro, A., Fontana, A., Carton, A., 2015. A regional rockfall database: the Eastern Alps test site. Environmental Earth Sciences 74, 1731-1742.

Dahal, R.K., Hasegawa, S., Nonomura, A., Yamanaka, M., Masuda, T., Nishino, K., 2008. GIS-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping. Environmental Geology 54, 311-324.

Dai, F.C., Lee, C.F., Ngai, Y.Y., 2002. Landslide risk assessment and management: an overview. Engineering Geology 64, 65-87.

Desiato, F., Lena, F., Baffo, F., Suatoni, B., Toreti, A., di Ecologia Agraria, U.C., Romagna, A.E., 2005. Indicatori del clima in Italia. APAT, Roma.

Doglioni, C., 1990. Thrust tectonics examples from the Venetian Alps.

Dunning, S., Massey, C., Rosser, N., 2009. Structural and geomorphological features of landslides in the Bhutan Himalaya derived from terrestrial laser scanning. Geomorphology 103, 17-29.

Dury, G., 1972. Hillslope form and Process. M.A. Carson and M.J. Kirkby, 1972. Cambridge University Press, London, vii + 475 pp., ~~£ 6.60~~.

Ercanoglu, M., Gokceoglu, C., 2002. Assessment of landslide susceptibility for a landslideprone area (north of Yenice, NW Turkey) by fuzzy approach. Environmental Geology 41, 720730.

Floris, M., Iafelice, M., Squarzoni, C., Zorzi, L., De Agostini, A., Genevois, R., 2011. Using online databases for landslide susceptibility assessment: an example from the Veneto Region (northeastern Italy). Nat. Hazards Earth Syst. Sci. 11, 1915-1925.

Gariano, S.L., Verini Supplizi, G., Ardizzone, F., Salvati, P., Bianchi, C., Morbidelli, R., Saltalippi, C., 2021. Long-term analysis of rainfall-induced landslides in Umbria, central Italy. Natural Hazards 106, 2207-2225.

Ghorbanzadeh, O., Rostamzadeh, H., Blaschke, T., Gholaminia, K., Aryal, J., 2018. A new GIS-based data mining technique using an adaptive neuro-fuzzy inference system (ANFIS) and k-fold cross-validation approach for land subsidence susceptibility mapping. Natural Hazards 94, 497-517.

Glade, T., Anderson, M.G., Crozier, M.J., 2006. Landslide hazard and risk. John Wiley & Sons. Goetz, J.N., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Computers & Geosciences 81, 1-11.

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., Galli, M., 2006. Estimating the quality of landslide susceptibility models. Geomorphology 81, 166-184.

Huang, F., Chen, J., Du, Z., Yao, C., Huang, J., Jiang, Q., Chang, Z., Li, S., 2020. Landslide Susceptibility Prediction Considering Regional Soil Erosion Based on Machine-Learning Models. ISPRS International Journal of Geo-Information 9, 377.

Iadanza, C., Trigila, A., Starace, P., Dragoni, A., Biondo, T., Roccisano, M., 2021. IdroGEO: A Collaborative Web Mapping Application Based on REST API Services and Open Data on Landslides and Floods in Italy. ISPRS International Journal of Geo-Information 10, 89.

Komac, M., 2006. A landslide susceptibility model using the Analytical Hierarchy Process method and multivariate statistics in perialpine Slovenia. Geomorphology 74, 17-28. Linden, A., 2006. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. Journal of evaluation in clinical practice 12, 132-139.

Liu, L.-L., Yang, C., Wang, X.-M., 2021. Landslide susceptibility assessment using feature selection-based machine learning models. Geomechanics and Engineering 25, 1-16.

Melville, B., Lucieer, A., Aryal, J., 2018. Object-based random forest classification of Landsat ETM+ and WorldView-2 satellite imagery for mapping lowland native grassland communities in Tasmania, Australia. International journal of applied earth observation and geoinformation 66, 46-55.

Merghadi, A., Abderrahmane, B., Tien Bui, D., 2018. Landslide Susceptibility Assessment at Mila Basin (Algeria): A Comparative Assessment of Prediction Capability of Advanced Machine Learning Methods. ISPRS International Journal of Geo-Information 7.

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Natural Hazards and Earth System Sciences Discussions

Meten, M., PrakashBhandary, N., Yatabe, R., 2015. Effect of Landslide Factor Combinations on the Prediction Accuracy of Landslide Susceptibility Maps in the Blue Nile Gorge of Central Ethiopia. Geoenvironmental Disasters 2, 9.

Micheletti, N., Foresti, L., Robert, S., Leuenberger, M., Pedrazzini, A., Jaboyedoff, M., Kanevski, M., 2014. Machine Learning Feature Selection Methods for Landslide Susceptibility Mapping. Mathematical Geosciences 46, 33-57.

Pham, B.T., Tien Bui, D., Pourghasemi, H.R., Indra, P., Dholakia, M.B., 2015. Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. Theoretical and Applied Climatology 128, 255-273.

Pham, B.T., Tien Bui, D., Prakash, I., 2018. Bagging based Support Vector Machines for spatial prediction of landslides. Environmental Earth Sciences 77, 146.

Pourghasemi, H.R., Gayen, A., Park, S., Lee, C.W., Lee, S., 2018. Assessment of LandslideProne Areas and Their Zonation Using Logistic Regression, LogitBoost, and NaiveBayes Machine-Learning Algorithms. Sustainability 10.

Pourghasemi, H.R., Pradhan, B., Gokceoglu, C., 2012. Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. Natural Hazards 63, 965-996.

Pradhan, B., 2010. Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches. Journal of the Indian Society of Remote Sensing 38, 301-320.

Raja, N.B., Çiçek, I., Türkoğlu, N., Aydin, O., Kawasaki, A., 2017. Landslide susceptibility mapping of the Sera River Basin using logistic regression model. Natural Hazards 85, 13231346.

Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M., Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. Earth-Science Reviews 180, 60-91.

Riley, S.J., DeGloria, S.D., Elliot, R., 1999. Index that quantifies topographic heterogeneity. intermountain Journal of sciences 5, 23-27.

Rossi, M., Guzzetti, F., Salvati, P., Donnini, M., Napolitano, E., Bianchi, C., 2019. A predictive model of societal landslide risk in Italy. Earth-Science Reviews 196, 102849.

Sahin, E.K., 2020. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences 2, 1308.

Sauro, F., Zampieri, D., Filipponi, M., 2013. Development of a deep karst system within a transpressional structure of the Dolomites in north-east Italy. Geomorphology 184, 51-63.

Schönborn, G., 1999. Balancing cross sections with kinematic constraints: The Dolomites (northern Italy). Tectonics 18, 527-545.

Senouci, R., Taibi, N.-E., Teodoro, A.C., Duarte, L., Mansour, H., Yahia Meddah, R., 2021. GIS-Based Expert Knowledge for Landslide Susceptibility Mapping (LSM): Case of Mostaganem Coast District, West of Algeria. Sustainability 13, 630.

Shahabi, H., Hashim, M., 2015. Landslide susceptibility mapping using GIS-based statistical models and Remote sensing data in tropical environment. Scientific reports 5, 9899.

https://doi.org/10.5194/nhess-2021-299 Preprint.
Discussion started: 18 November 2021 c Author(s) 2021.

Natural Hazards and Earth System Sciences Open Access EGU Discussions

Stanley, T.A., Kirschbaum, D.B., Benz, G., Emberson, R.A., Amatya, P.M., Medwedeff, W., Clark, M.K., 2021. Data-Driven Landslide Nowcasting at the Global Scale. Frontiers in Earth Science 9.

Trigila, A., Iadanza, C., 2018. Landslides and floods in Italy: hazard and risk indicators - Summary Report 2018.

Trigila, A., Iadanza, C., Spizzichino, D., 2010. Quality assessment of the Italian Landslide Inventory using GIS processing. Landslides 7, 455-470.

van Westen, C.J., Castellanos, E., Kuriakose, S.L., 2008. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. Engineering Geology 102, 112-131.

Youssef, A.M., Pourghasemi, H.R., 2021. Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. Geoscience Frontiers 12, 639-655.