Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

INGEGNERIA CIVILE, CHIMICA, AMBIENTALE E DEI MATERIALI

Ciclo 34

**Settore Concorsuale:** 04/A3 - GEOLOGIA APPLICATA, GEOGRAFIA FISICA E GEOMORFOLOGIA

**Settore Scientifico Disciplinare:** GEO/05 - GEOLOGIA APPLICATA

# LANDSLIDE SUSCEPTIBILITY IN THE BELT AND ROAD INITIATIVE

**Presentata da:** Giacomo Titti

**Coordinatore Dottorato**

Alessandro Tugnoli

**Supervisore**

Lisa Borgatti

**Co-supervisore**

Alessandro Pasuto

Luigi Lombardo

Peng Cui

**Esame finale anno 2022**

# Table of Contents

# List of Tables

# List of Figures

# Abstract

The Belt and Road Initiative (BRI) is a project launched by the Chinese Government whose main goal is to connect more than 65 countries in Asia, Europe, Africa and Oceania developing infrastructures and facilities. To support the prevention or mitigation of landslide hazards, which may affect the mainland infrastructures of BRI, a landslide susceptibility analysis in the countries involved has been carried out during the presented PhD research activity.

Due to the large study area, the analysis has been carried out using a multi-scale approach which consists of mapping susceptibility firstly at continental scale, in order to have an overview of the large study area, and then at national scale, where a detailed susceptibility map is required.

The study area selected for the continental assessment is the south-Asia, where a pixel-based landslide susceptibility map has been carried out using the Weight of Evidence method and validated by Receiving Operating Characteristic (ROC) curves. The results highlighted several areas which require a second landslide susceptibility analysis at national scale, such as: the 83% of Tajikistan, the 92% of Nepal, the 98% of Bhutan, the 84% of Myanmar and the 94% of Laos which are moderately to very highly susceptible. In particular, we selected the regions of west Tajikistan and north-east India to be investigated at national scale.

Data scarcity is a common condition for many countries involved into the Initiative. Therefore in addition to the landslide susceptibility assessment of west Tajikistan, which has been conducted using a Generalized Additive Model and validated by ROC curves, we have examined, in the same study area, the effect of incomplete landslide dataset on the prediction capacity of statistical models. Differently from what we expected, the variation in landslide presence significantly influences in a negative way the model prediction capacity only in the worst scenarios reproduced.

The entire PhD research activity has been conducted using only open data and open-source software. In this context, to support the analysis of the last years an open-source plugin for QGIS has been implemented. The SZ-tool allows the user to make susceptibility assessments from the data preprocessing, susceptibility mapping with statistic-based models, to the final classification. The SZ-tool has been tested in the study area of north-east India which demonstrated the possibility to compute a complete landslide susceptibility assessment in few steps.

All the output data of the analysis conducted during the presented PhD research are freely available and downloadable.

This text describes the research activity of the last three years which is summed up in three main chapters. Each chapter reports the text of the articles published in international scientific journal during the PhD, titled: 'Landslide susceptibility in the Belt and Road

Countries: continental step of a multi-scale approach', 'When enough is really enough? On the minimum number of landslides to build reliable susceptibility models' and 'Mapping susceptibility with open-source tools: a new plugin for QGIS'.

# 1. Introduction

The Belt and Road Initiative (BRI) is an almost world-wide project launched in 2013 by the President of the Popular Republic of China. The project aims at the development of facilities with commercial purposes such as infrastructures for: products, energy, oil & gas transportation and international programs to support economical relations between the involved countries which are more than 65 from 4 continents: Europe, Asia, Oceania and Africa (Lei et al., 2018).

The Silk Road Economic Belt together with the 21$^{st}$ Century Maritime Silk Road are the main components of the Belt and Road Initiative. They represent the principal routes of economic and cultural exchanges between Europe and Asia, the former on the mainland, the latter by the sea. The Silk Road Economic Belt is divided in 6 sections: China-Mongolia-Russia Corridor, New Eurasia Continental Bridge, China-Central Asia-West Asia Corridor, China-Indochina Peninsula Corridor, China-Pakistan Corridor, Bangladesh-China-India-Myanmar (Cui et al., 2017).

To support the development of new infrastructures or the renovation of already existing ones and prevent or mitigate the effects of natural hazards to people and buildings in the Belt and Road countries, a new sub-project of the Silk Road Economic Belt has been initiated: the Silk Road Disaster Risk Reduction (SiDRR) (Lei et al., 2018).

In this framework, the activity here presented is focused on the assessment, prevention and mitigation of landslide risk. The main goal of the PhD research was to map the landslide susceptibility of the Silk Road Economic Belt area. Susceptibility is probably the most common way to prevent natural hazard effects during the planning phase of land development. Assess the landslide susceptibility of a specific area means to estimate how much the study area is prone to landslide according to predisposing factors.

Due to the large extension of the study area, we preferred to investigate only a portion of the Silk Road Economic Belt area focused in south-Asia in order to prevent technical issues related to the management of a huge amount of heterogeneous data coming from about 65 countries. In total 14 countries have been selected (Fig. 2.1) where landslides produced 8377 deaths and more than 200 Billion $ of damages in the last 20 years according to the EM-DAT database (Guha-Sapir et al., 2016): China, Pakistan, India, Tajikistan, Bangladesh, Nepal, Afghanistan, Bhutan, Myanmar, Cambodia, Kyrgyzstan, Laos, Thailand and Viet Nam.

The selected area has been investigated using a multi-scale approach. The idea is to deal with the continental extension of the study area through a two steps analysis: the first at continental scale, the second at national scale. The continental analysis would investigate the 14 countries using coarse data, light enough to be computed in a reasonable time with common machines. The output of such analysis carried out a landslide susceptibility overview of the investigated region from where the sub-regions, which require more detailed analysis at national scale, may be highlighted.

In total three landslide susceptibility maps in south-Asia have been produced from the multi-scale analysis. The first is a continental landslide susceptibility map which has been used to highlight the most susceptible regions crossed by the Silk Road Economic Belt path (see Chapter 2). In particular the west-Tajikistan and north-east of India, have been selected to be analysed in detail at national scale. Their results have been described in Chapter 3 and Chapter 4 of this document respectively. Figure 1.1 shows the work-flow of the research activity conducted during the PhD resulting in three scientific papers.

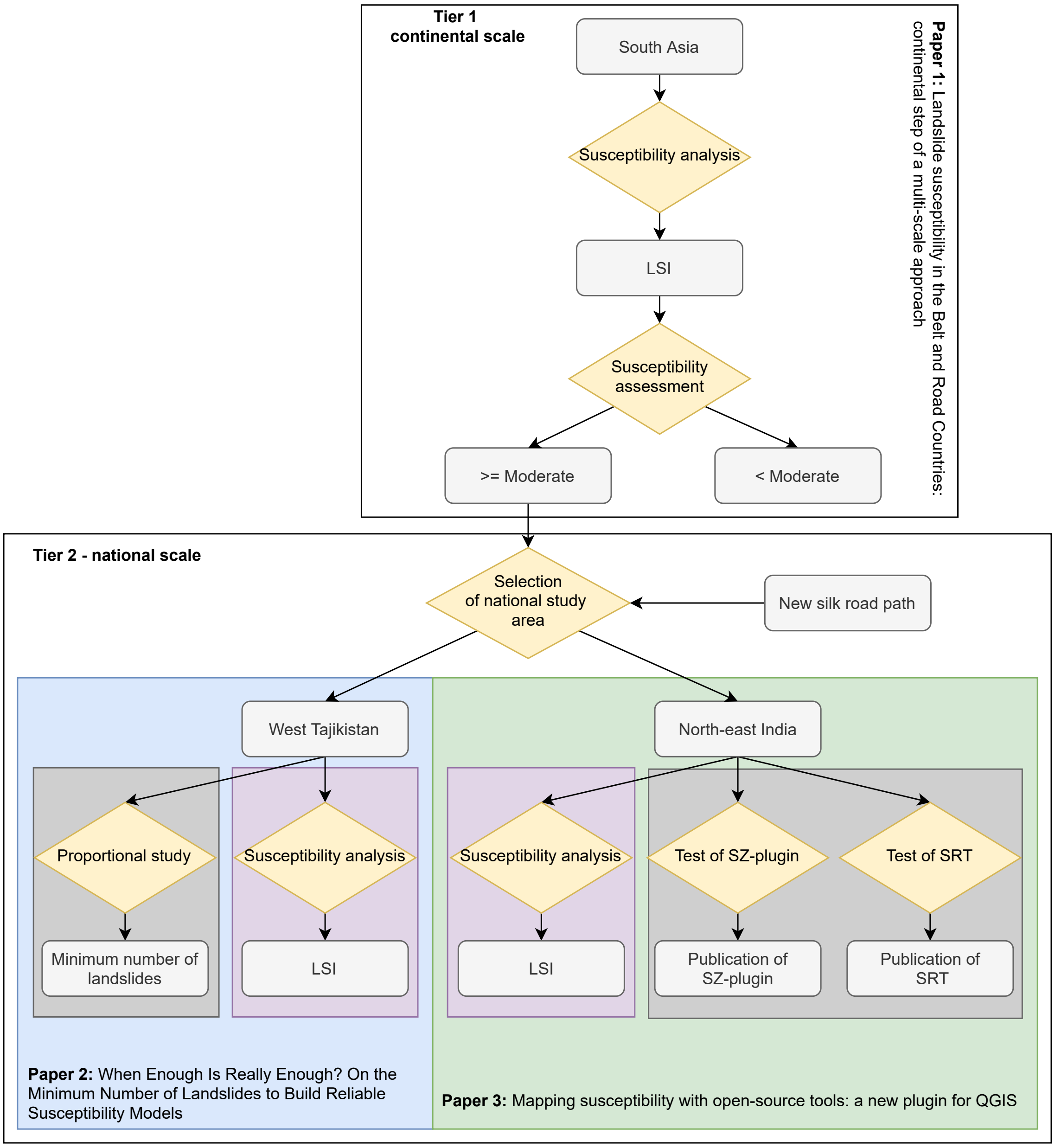

Tier 1
continental scale
South Asia
Susceptibility analysis
LSI
Susceptibility assessment
>= Moderate
< Moderate
Paper 1: Landslide susceptibility in the Belt and Road Countries: continental step of a multi-scale approach
Tier 2 - national scale
Selection of national study area
New silk road path
West Tajikistan
North-east India
Proportional study
Susceptibility analysis
Susceptibility analysis
Test of SZ-plugin
Test of SRT
Minimum number of landslides
LSI
LSI
Publication of SZ-plugin
Publication of SRT
Paper 2: When Enough Is Really Enough? On the Minimum Number of Landslides to Build Reliable Susceptibility Models
Paper 3: Mapping susceptibility with open-source tools: a new plugin for QGIS

Figure 1.1: Work-flow of the overall activity.

Taking advantage of these landslide susceptibility evaluations, additional analysis have been conducted in the same regions in order to improve and make the results so far reached reproducible. In particular, during the assessment in Tajikistan we asked ourselves: When enough is really enough? How many landslides are necessary to produce a reliable landslide susceptibility map with statistical models? To what extent should I collect landslides to have a 'complete' landslide inventory? In the Tajikistan study area we compared the prediction capacity of several landslide susceptibility maps carried out changing gradually the number of considered landslides. In this manner we reproduced realistic scenarios of scarcity data environments which is a common condition in the countries of the Silk Road Economic Belt (see Chapter 3).

During the last years, in order to conduct the analysis described in this text, several codes, tools, functions and graphs have been produced. All of them have been collected in two tools: the SZ-plugin for QGIS and the SRT tool for Google Earth Engine. The former allows the user to carry out a spatial susceptibility assessment using statistical/data-driven models, whereas the latter allows the user to collect and reduce spatially data from global datasets. Both have been tested in north-east India to produce a national scale landslide susceptibility map (see Chapter 4).

The basic idea behind this project is that the extension of the Silk Road Economic Belt area doesn't allow to investigate the landslide susceptibility of the countries involved into the initiative all at once. Therefore, in this research, the analysis carried out may be considered as a pilot model which could be replicated in the areas not analysed yet. In doing so, different approaches have been tested to evaluate the landslide susceptibility. Relevant differences between the three landslide susceptibility studies are represented by the mapping unit and the data-driven method used (Tab. 1.1). A mapping unit is the minimum geographic object which composes a study area and to which a unique value of susceptibility is assigned. Some example of mapping units are: grid cells, terrain units, Unique Condition Units, slope units and others. In this work three different mapping units have been adopted: grid cells for south Asia, Unique Condition Units for west Tajikistan and slope units for north-east India.

| | **Tier 1** | **Tier 2** | |
|---|---|---|---|
| | | **A** | **B** |
| **Study area** | south Asia | west Tajikistan | north-east India |
| **Mapping units** | Pixel | Unique Condition Units | Slope units |
| **Method** | Weight of Evidence | Generalized Additive Model | Logistic regression/Weight of Evidence |
| **Validation** | Simple cross-validation | 10-fold cross-validation | 10-fold cross-validation |
| **# of mapping units** | 2779933 | 16020 | 124553 |
| **Unstable units** | 759 | 677 | 3078 |

Table 1.1: The different methodologies used in the case studies

As regard the methods adopted, based on the literature, they can be distinguished in qualitative or knowledge-driven and quantitative or data-driven methods. The former are based on the experience of the operator which make the analysis. The latter replicates the correlation between the predisposing factors and the dependent variable which may be physically-based or statistics-based. In particular the statistics-based models determine the numerical relation between the spatial distribution of a binary classifier (landslide presence/absence) and the geo-environmental factors which predispose the landscape to failure. Under the assumption that "the past is the key to the future" (Carrara et al., 1995) the model evaluates the spatial probability of landslide occurrence and assigns a value from the Landslide Susceptibility Index (LSI) per mapping unit.

In respect to the principle of reproducebility, the models applied in the three study areas are quantitative and thus more objective than the qualitative models: Weight of Evidence, Generalized Additive Model and Logistic Regression.

Nowadays, the open science is a very current topic in the scientific community. The possibility to share information, knowledge and data has a priority role in the scientific and technological progress. This awareness led to the development of a number of IT media to share materials. Following this direction, this research activity promotes the open science using only open data and open-source software, providing source codes of any analysis conducted, source information about data and how to download them. Moreover, we published open access articles which summarize three years of researches. Most of the materials are available in the GitHub repositories: https://github.com/giactitti and https://github.com/CNR-IRPI-Padova. Finally, to support the user we published a tutorial video describing how do susceptibility with the tools implemented (https://www.youtube.com/watch?v=XpsiCkVF11s).

The three-years research activity has been finalized in three scientific articles reported in the next three chapters of the current text, which are titled "Landslide susceptibility in the Belt and Road Countries: continental step of a multi-scale approach" published in *Environmental Earth Sciences* (Chapter 2), "When Enough Is Really Enough? On the Minimum Number of Landslides to Build Reliable Susceptibility Models" published in *Geosciences* (Chapter 3) and "Mapping susceptibility with open-source tools: a new plugin for QGIS" published in *Frontiers in Environmental Sciences* (Chapter 4). Each of the three articles reports the state of the art related to their contents. The former article reports information regarding the landslide susceptibility at continental scale available in the literature, the second reports details about the available approaches to investigate the effect of lack of landslide data, and the latter article lists tools comparable with the category of the SZ-plugin. As regard the scientific articles, reports, repositories and other information sources cited into the text, their references have been listed in the end of this text in one single section: Chapter 5.

# 2. Landslide susceptibility in the Belt and Road Countries: continental step of a multi-scale approach

This chapter has been published in *Environmental Earth Sciences* journal in 2021 and authored by: Giacomo Titti[1,3], Lisa Borgatti[1,3], Qiang Zou[2], Peng Cui[2], Alessandro Pasuto[3] (Titti et al., 2021a).

## 2.1 Introduction

The Belt and Road Initiative (BRI) is a collaboration project launched by the Chinese Government to connect more than 65 countries all over the word by developing infrastructures, facilities and support among the involved Countries and to encourage innovation in less developed Countries (Cui et al., 2017; Liu and Dunford, 2016).

The Silk Road Disaster Risk Reduction (SiDRR) project (Lei et al., 2018) is one of the prioritized sub-projects of the BRI. The purpose of the SiDRR is to carry out a long-term research project dealing with natural hazard assessment and risk mitigation in the Belt and Road Countries. A group of experts, with the role of scientific coordination of the activities carried out by the involved Countries, as well as the dissemination of the results, has been created. The expected outcomes of the research activities of the group are the assessment of geo-hydrological hazards in the Belt and Road Countries and the definition of risk mitigation measures.

Risk, hazard and susceptibility zoning are three complementary approaches to support land planning. They implicate a decreasing complexity in method and in data types, respectively. Considering the small scale of the SiDRR analysis and the available data, a landslide susceptibility zoning has been proposed to build a map which should give a general overview of the landslide-prone areas in the Belt and Road Countries. The goal is to individuate the most susceptible areas where to focus further and more detailed assessments. "In mathematical form, landslide susceptibility, is the probability of spatial occurrence of known slope failures, given a set of geoenvironmental conditions" (Guzzetti et al., 2006). Otherwise, landslide susceptibility can be defined as the spatial component of the hazard (Reichenbach et al., 2018) which, in turn, is the combination of the frequency

[1] Department of Civil, Chemical, Environmental and Materials Engineering, Alma Mater Studiorum University of Bologna, Viale Risorgimento, 2, 40136 Bologna, Italy

[2] Institute of Mountain Hazards and Environment, Chinese Academy of Science, South Renmin Road, Section 4, 9, Chengdu 610041, China

[3] Research Institute for Geo-Hydrological Protection, Italian National Research Council, C.so Stati Uniti, 4, 35127 Padova, Italy

of landslide occurrence and the susceptibility map (Fell et al., 2008). Thus, landslide susceptibility can be considered a fundamental part of the process to reach landslide hazard and risk assessment. At the same time, it can be used in land use planning for large areas or in analyses characterized by scarcity of data (Corominas et al., 2014).

This work will support the development of landslide hazard prevention and mitigation measures proposing a multi-scale approach for landslide susceptibility zoning and discussing its first application in a test area.

The used scale classification has been derived by Glade and Crozier (2012) and introduced by Soeters and van Westen (1996). One class has been added to the original ranges: large scales (> 1:10,000), medium scales (1:15,000-1:100,000), regional scales (1:125,000-1:500,000), national scales (1:750,000-1:2,000,000) and continental scales (< 1:5,000,000).

An approach to landslide susceptibility assessment based on the analysis of the state-of-the-art is presented, where the terms "large area" and "small scale" are used as synonyms and they refer to continental scale. To show the feasibility and the robustness of the suggested approach, the case study of south-Asia has been analyzed. The results have been evaluated and validated by means of ROC analysis.

## 2.2 State of the art

Landsliding is a complex process driven by several possible predisposing and triggering factors. Numerous causative factors (geo-environmental factors) may predispose slope to failure, such as: geology, topography, tectonics, land cover and use, hydrology, and others (Eckelmann et al., 2006). On the other hand, the processes which trigger a landslide can be different and include: intense or prolonged rainfall, earthquakes, rapid snow melting, volcanic activity, human actions and others (Guzzetti et al., 2012).

The goal of landslide susceptibility zoning is to analyze the probability of landslide occurrence under the influence of a combination of factors, not including landslide frequency. Therefore, the temporal factor is not taken into consideration (Chacón et al., 2006).

Many different methods have been proposed for landslide susceptibility zoning in the scientific literature. The choice of a susceptibility mapping method significantly influences the prediction capacity of the analysis. According to Corominas et al. (2014), the methods can be categorized as qualitative (knowledge-driven methods) and quantitative (data-driven methods). The former take advantage of the theoretical and empirical knowledge of the researchers to make scientific analysis and judgment (Axing et al., 2010; Ayalew et al., 2004; Barredo et al., 2000; Günther et al., 2014; Saaty, 1990). The latter recreates the relation between landslides and their controlling factors using a mathematical model: physically (Chung and Fabbri, 2003; Goetz et al., 2011; Gorsevski et al., 2006) or statistics-based (Agterberg et al., 1989; Bonham-Carter et al., 1988; Bui et al., 2016; Carrara et al., 2008; Catani et al., 2013; Chen et al., 2016, 2017; Constantin et al., 2011; Eeckhaut et al., 2012; Gorsevski et al., 2000; Pham et al., 2016; Yao et al., 2008).

Reichenbach et al. (2018) pointed out that Logistic Regression is one of the most diffuse statistics-based models for landslide susceptibility on both large and small scales. Concerning landslide susceptibility assessment at continental, or similar scale, several different methodologies have been used so far. Most of them are knowledge-driven methods and statistic-based methods. In addition, physically based methods are excluded in small scale analyses, because they require a detailed knowledge of the landslide dynamics, which is not feasible for large areas.

The scarcity of data, which is very common in small-scale analyses, may be a driven factor in model selection. In particular, the lack of landslide inventories may affect the robustness and quality of the results. However, as stated by Hong et al. (2007) this is not always true: "more information does not necessarily lead to better results, depending on the quality of the data".

Due to the lack of a global landslide data set at that time, Hong et al. (2007) have produced a global landslide susceptibility map without landslide inventories. They have weighted the causative factors on the base of reference studies and information available combined through a linear combination method. Following a similar idea, Eeckhaut et al. (2012) proposed a statistical model application with limited landslide inventory data. They evaluated the landslide susceptibility over Europe with the Logistic Regression model.

The recent development of global landslide inventories has supported new analyses at global scale (Kirschbaum and Stanley, 2018; Stanley and Kirschbaum, 2017, e.g.) which have produced a global landslide susceptibility map for rainfall-triggered landslides based on the fuzzy overlay method. This approach combines landslide inventories with expert opinions to develop a heuristic model.

At the continental scale, Günther et al. (2014) and then Wilde et al. (2018) presented the landslide susceptibility maps of Europe, named ELSUSv1 and ELSUSv2, respectively. Despite that, they had numerous national landslide inventories heterogeneously distributed, they have proposed a qualitative model (Spatial Multi Criteria Evaluation) using Analytical Hierarchy Process.

A different approach at continental scale has been used by Broeckx et al. (2018) who analyzed the landslide susceptibility all over Africa based on a well-distributed inventory applying Logistic Regression model.

As a concern, the study area here considered (south-Asia) numerous landslide susceptibility maps at national or smaller scale have been produced in recent years. Some of those, based on the Neural Network method with about 1300 landslides all over China are reported in Liu et al. (2013). Recently, Saponaro et al. (2015) have covered Uzbekistan, Tajikistan and Kyrgistan, using the Weight of Evidence method.

In the context of the European Union's Thematic Strategy for Soil Protection (EC, 2006), the Soil Information Working Group (SIWG) of the European Soil Bureau Network (ESBN) has promoted a project for the identification of landslide hazard priority areas. The European Landslide Expert Group (Günther et al., 2013a) has then put forward a multi-Tier susceptibility analysis in the Guidelines for Mapping Areas at Risk of Landslides in Europe (Hervás, 2007). Taking inspiration from the latter, and considering, the extension of the study area and the geomorphological, geological, cultural, scientific heterogeneity of the context, the Weight of Evidence (WoE) method has been selected. Therefore, the application of the selected method and the Tier 1 approach at the south-Asia region is presented here.

## 2.3 Case study

The Belt and Road Initiative involves 3 continents (63% of the world), and more than 65 countries (Cui et al., 2017). The study area selected includes a relevant part of the Belt and Road Countries (Fig. 2.1), namely: China, Pakistan, India, Tajikistan, Bangladesh, Nepal, Afghanistan, Bhutan, Myanmar, Cambodia, Kyrgyzstan, Laos, Thailand, Viet Nam. The study area comprises a high density of landslides mainly triggered by heavy rainfall (more than 1600 mm/y) (Fig. 2.7) as well as some of the most disastrous earthquakes that have recently occurred in the world (i.e., Nepal 2015 and Wenchuan 2008). The diversity of climate, topography, geological features and land cover result in an extremely complex environment on which to assess landslide susceptibility.

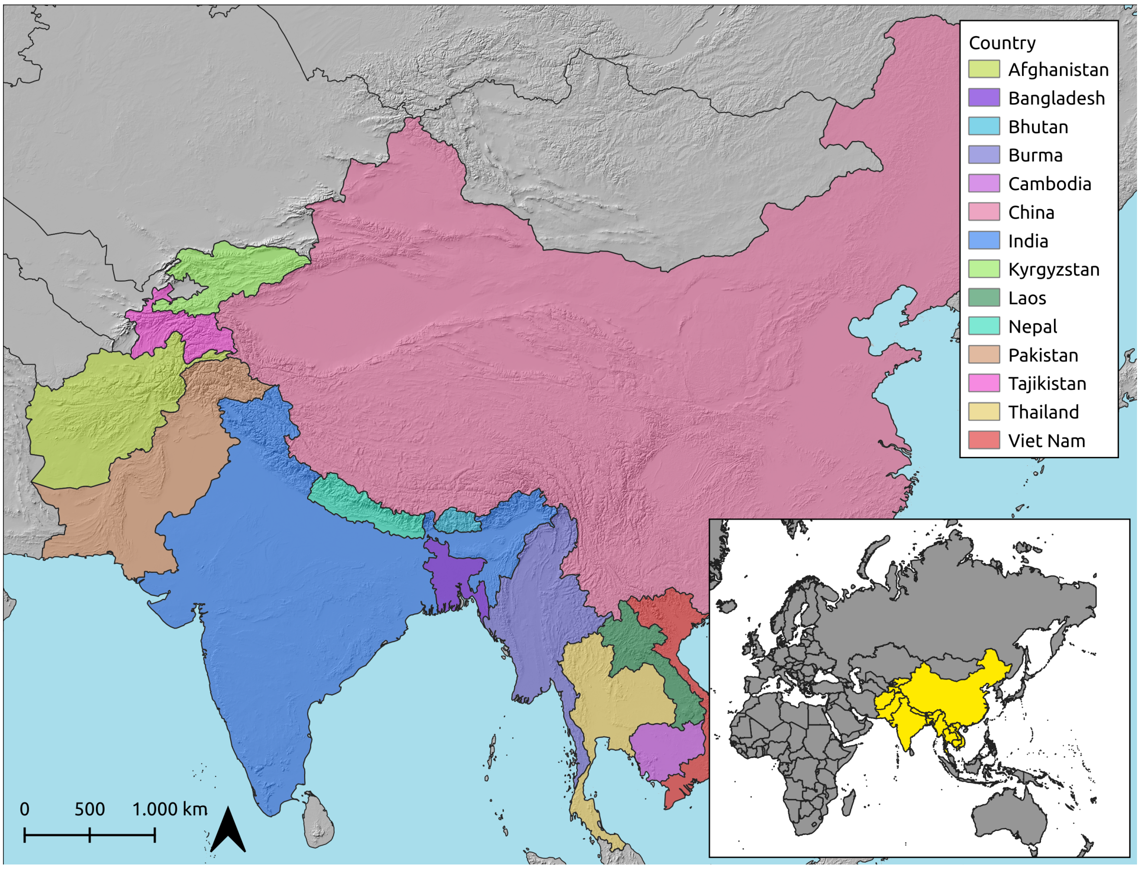


Figure 2.1: Study area of the south-Asia.

## 2.4 Materials and methods

The Tiers-based workflow produced consequential susceptibility zoning of the same study area by growing scales (Fig. 2.2). The smallest scale provides an overview of the object of study and it delineates the priorities, i.e., the most susceptible regions. Therefore, Tier 1 assessment exploits low-resolution data and incomplete spatial information. With the increase of the scale of analysis, the Tier 2 approach is intended to detail the landslide susceptibility analysis conducted by the Tier 1 (Günther et al., 2013b). The scale of Tiers cannot be defined a priori, since it depends on the available data resolution and the spatial extent of the study area. This means that the results should reflect, at least, the minimum resolution of the data input.

The landslide prediction model represents the core of the entire analysis. The model proposed here is: (i) temporally/geographically reproducible; (ii) simple and thus clear to

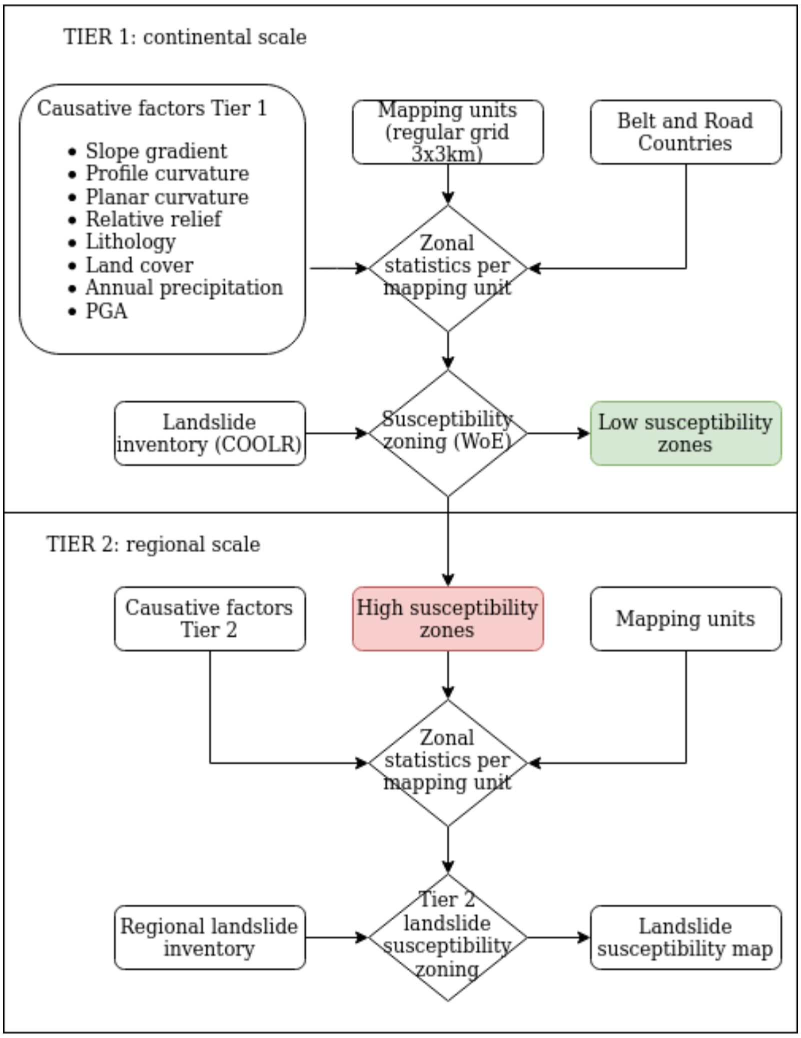

TIER 1: continental scale
Causative factors Tier 1
Slope gradient
Profile curvature
Planar curvature
Relative relief
Lithology
Land cover
Annual precipitation
PGA
Mapping units (regular grid 3x3km)
Belt and Road Countries
Zonal statistics per mapping unit
Landslide inventory (COOLR)
Susceptibility zoning (WoE)
Low susceptibility zones
TIER 2: regional scale
Causative factors Tier 2
High susceptibility zones
Mapping units
Zonal statistics per mapping unit
Regional landslide inventory
Tier 2 landslide susceptibility zoning
Landslide susceptibility map

Figure 2.2: Workflow of the Tiers-based approach.

the people involved; (iii) as realistic as possible. A qualitative assessment for Tier 1 level has been proposed by Günther et al. (2007). It was based on the expertise of the researchers responsible for the analysis, thus the reproducibility depends on the investigator. To make the results reproducible in time and in all areas, the assessment technique should be quantitative and as objective as possible. The use of physical predictors and of the validation procedure define the physical relevance of the model in accordance with the geological and geomorphological features of the study area. Therefore, to create a reproducible, simple, realistic landslide susceptibility map not affected by subjectivity, misunderstanding and abstraction, a limited number of causative factors related to all types of landslides and a quantitative susceptibility modeling technique, suitable to the specific Tier, have been assumed.

The landslide susceptibility concept is based on the simple principle that landsliding will occur more frequently in the most susceptible areas characterized by similar geo-environmental factors which predispose towards slope failures. A landslide inventory and selected causative factors are the preliminary requirements for susceptibility analysis (van Westen et al., 2008), especially if a statistic-based correlation analysis is applied (Bui et al., 2016) in accordance with the scale of mapping, the required usage, and the quality of the data available (Fell et al., 2008).

A landslide inventory is an essential part of the input dataset in landslide susceptibility mapping. It generally records the location, the date of occurrence and type of mass movements (Margottini et al., 2013). In this analysis, three different alternatives have been taken into consideration: (i) aggregation and homogenization of all local landslide datasets available in the Countries involved in the project; (ii) development of a new landslide dataset; and (iii) collection of global landslide datasets. The latter solution has been selected due to the complexity of the aggregation processes and the lack of information. Indeed, the regional slope failure database is sometimes not complete or, more often, it is absent completely. The few data available locally are often limited to recent years and, therefore, not representative of the instability condition as it is.

The NASA-COOLR dataset (Juang et al., 2019) has been selected for the purpose of this work. It is an open database for landslide events launched in 2018 which collects different inventories from different sources: Landslide Reporter Catalog (LRC) (Juang et al., 2019), NASA Global Landslide Catalog (GLC) (Kirschbaum et al., 2010, 2015) and collated inventories from external local sources. The LRC includes landslide reports by citizen scientists through the Landslide Reporter and checked by NASA. The GLC is a global inventory of rainfall-triggered landslides compiled by NASA since 2007 and is based on online media reports, disaster databases, scientific reports and more (Kirschbaum et al., 2015). The rest of the landslides are added into the COOLR by the LRC and other sources such as the SERVIR-Mekong team for Myanmar landslides (SMMML). The team collected landslides based on Google Earth imagery (Juang et al., 2019).

The landslide inventory of COOLR used for the analysis was downloaded in June 2020. Each point in the inventory has been assigned a radius of confidence between 1 and 75 km. In accordance with the spatial resolution of the analysis the landslide locations with 1 km of accuracy have been selected (1549 landslides). The landslide attributes are shown in Fig. 2.3 and Table 2.1. The heterogeneity of data sources implicates the variability of information. As a consequence, some information such as the exact location of the landslides have been collected but others remain unknown (e.g., landslide category and trigger).

Considering the limited information available on each event, a cross-check has been

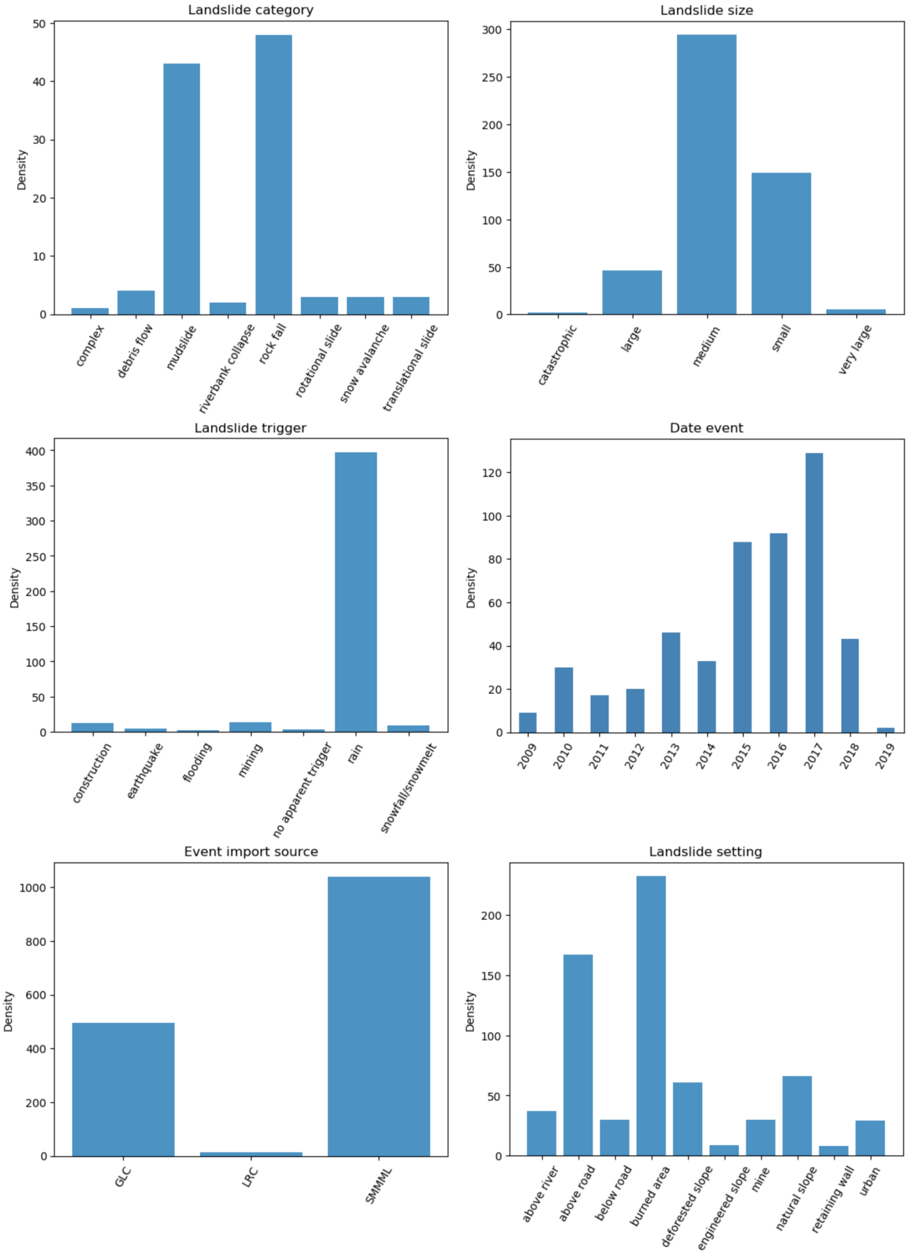


Figure 2.3: Landslide frequencies classified by attributes of 1 km location accuracy dataset (1549). The attributes include 'unknown', 'others' and empty records which are not reported into the graphs. They are 1142 of 'landslide category', 1053 of 'landslide size', 1109 of 'landslide trigger', 1040 of 'date of the event' and 880 of 'landslide setting'.

| Country | Number of COOLR landslides |
|---|---|
| China | 44 |
| Afghanistan | 1 |
| Burma | 1038 |
| Cambodia | 1 |
| Bangladesh | 14 |
| Viet Nam | 2 |
| Thailand | 4 |
| Tajikistan | 7 |
| Nepal | 51 |
| Pakistan | 27 |
| Kyrgyzstan | 6 |
| Laos | 2 |
| Bhutan | 1 |
| India | 351 |

Table 2.1: Number of COOLR landslides per country with location accuracy 'exact' or equal to 1 km

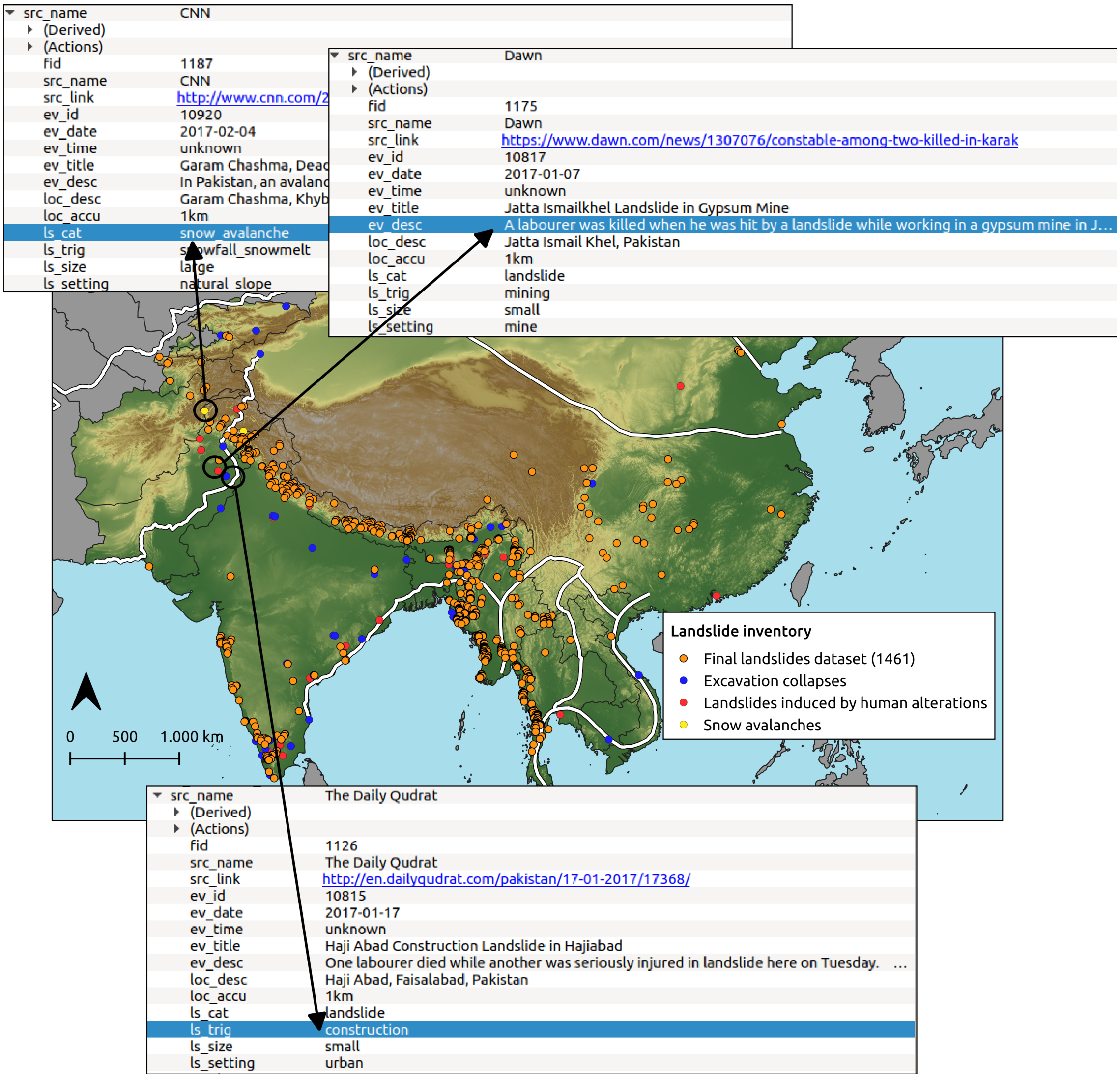
conducted to evaluate the reliability of the inventory for the purposes of this analysis. DEM and satellite images, along with a number of pictures, available for 118 landslides have been analyzed. The 250 m DEM has been downloaded from CIAT website (Reuter et al., 2007) which is the result of a resampling process from the 30 m SRTM data.

The sample is not properly representative of the inventory, but, it allows some possible incompatibilities with the goal of the analysis to be highlighted. As a result, some features have been removed from the inventory. For example: 3 events are classified as snow avalanches, whereas 47 landslides appear to be related with human alterations of the natural landscape (mining, engineered slopes and retaining walls) which are strictly site-specific. In particular, for 6 of the latter a photo link is available. They show that slope instability events occurred during mining activities and construction works which are different from slope cutting, these are probably triggered by anthropic activities. It could be argued that all the 47 landslides have been probably triggered by antrophic activities as well and caused by the same conditions. Therefore, they have been deleted from the inventory. Then, given a radius of confidence of approximately 1 km around each feature (9x9 cells of 250 m-side pixel), some features have been removed, since the slope of the terrain is lower than 3° , thus, they may be considered excavation collapses (38 events). At the end of this cross-check activity, 1461 landslides have been selected for the susceptibility zoning (Fig. 2.4).

The analysis focuses on the development of future scenarios based on the prediction of landslides spatial distribution. It reproduces spatially the combination of factors responsible for previous events.

The selection of the causative factors for the multi-scale analysis depends on the Tier. In the context of the Tier 1 analysis, the considered causative factors are: slope degree, plan curvature, profile curvature, relative relief, lithology, land cover, Peak Ground Acceleration (PGA) and annual rainfall (Table 2.2).

As stated by Fell et al. (2008), "areas with similar topography, geology and geomorphology as the areas which have experienced landsliding in the past are also likely to experience


Landslide inventory
● Final landslides dataset (1461)
● Excavation collapses
● Landslides induced by human alterations
● Snow avalanches
0 500 1.000 km

| src_name | CNN |
|---|---|
| (Derived) | |
| (Actions) | |
| fid | 1187 |
| src_name | CNN |
| src_link | http://www.cnn.com/2 |
| ev_id | 10920 |
| ev_date | 2017-02-04 |
| ev_time | unknown |
| ev_title | Garam Chashma, Dead |
| ev_desc | In Pakistan, an avalanc |
| loc_desc | Garam Chashma, Khyb |
| loc_accu | 1km |
| ls_cat | snow_avalanche |
| ls_trig | snowfall_snowmelt |
| ls_size | large |
| ls_setting | natural_slope |

| src_name | Dawn |
|---|---|
| (Derived) | |
| (Actions) | |
| fid | 1175 |
| src_name | Dawn |
| src_link | https://www.dawn.com/news/1307076/constable-among-two-killed-in-karak |
| ev_id | 10817 |
| ev_date | 2017-01-07 |
| ev_time | unknown |
| ev_title | Jatta Ismailkhel Landslide in Gypsum Mine |
| ev_desc | A labourer was killed when he was hit by a landslide while working in a gypsum mine in J… |
| loc_desc | Jatta Ismail Khel, Pakistan |
| loc_accu | 1km |
| ls_cat | landslide |
| ls_trig | mining |
| ls_size | small |
| ls_setting | mine |

| src_name | The Daily Qudrat |
|---|---|
| (Derived) | |
| (Actions) | |
| fid | 1126 |
| src_name | The Daily Qudrat |
| src_link | http://en.dailyqudrat.com/pakistan/17-01-2017/17368/ |
| ev_id | 10815 |
| ev_date | 2017-01-17 |
| ev_time | unknown |
| ev_title | Haji Abad Construction Landslide in Hajiabad |
| ev_desc | One labourer died while another was seriously injured in landslide here on Tuesday. … |
| loc_desc | Haji Abad, Faisalabad, Pakistan |
| loc_accu | 1km |
| ls_cat | landslide |
| ls_trig | construction |
| ls_size | small |
| ls_setting | urban |

Figure 2.4: Landslide dataset used for the analysis and the features removed after the cross-check.

| | Class factor | Factor |
|---|---|---|
| Tier 1 | Morphological | Slope angle |
| | | Plan curvature |
| | | Profile curvature |
| | | Relative relief |
| | Geological | Lithology |
| | Environmental | Land cover |
| | | PGA |
| | | Annual rainfall |

Table 2.2: Causative factors.

landsliding in the future". To identify the causes of past landslides and predict future scenarios, the statistical approach proposed in this work requires the classification of the causative factors. The classification significantly affects the prediction skill of the analysis. Therefore, classifications previously proposed in papers and technical reports have been assigned to the causative factors considered here. The diagram in Fig. 2.5 shows how the causative factors have been pre-processed.

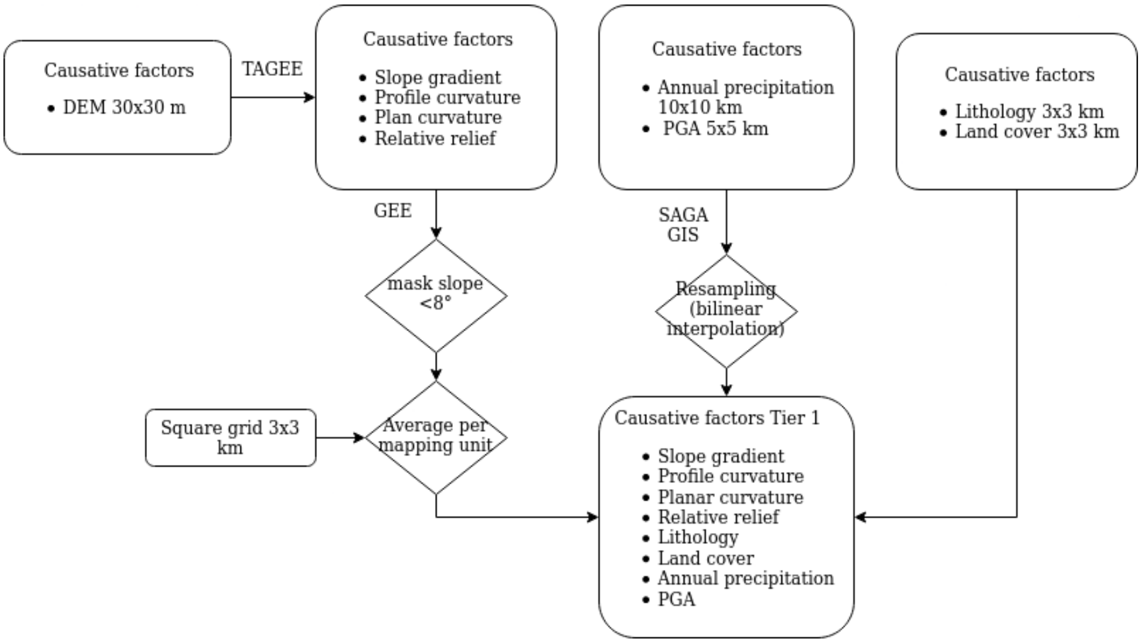


Figure 2.5: Pre-processing of the causative factors.

The morphological factors for Tier 1 assessment have been classified according to the slope angle, plan curvature (curvature tangent to the contour line), profile curvature (curvature tangent to the slope line) and relative relief (the maximum range of elevation in a neighborhood of 1 km of radius).

The morphological factors have been derived from the 30 m Shuttle Radar Topography Mission (SRTM) DEM (Farr et al., 2007; Florinsky et al., 2019). The morphological factors have been calculated in Google Earth Engine (GEE) (Gorelick et al., 2017) using Terrain Analysis in Google Earth Engine (TAGEE) a GEE package for terrain analysis (Safanelli et al., 2020). GEE allowed us to calculate slope angle, plan and profile curvature and relative relief with a pixel size of 30 m then resampled into a square grid of 3x3 km by average calculation. To reduce the computational cost and balance the amount of stable and unstable cells of the dataset, the 30 m cells with slope degree lower than 8° have been masked for all the causative factors. This value has been selected considering the average slope of the debris flow fans which represent the steepest terrain in which landslides are not expected but accumulation landforms. Therefore, the valley bottoms along with all the depositional forms that couldn't be affected by instability processes in the mountain areas have been excluded from the analysis. The relief has been calculated as the range between the maximum and the minimum elevation in a buffer radius of 1 km around each 30 m pixel.

The slope angle (terrain gradient) classification reflects the classes used for the ELSUSv1 (Günther et al., 2014): 0° , 1-3°, 4-6°, 7-10°,11-15°, 16-20°, 21-30°, > 30° . The curvatures have been classified in quartiles. Therefore, profile curvature has been classified in:< -2.6 $10^{-4}$, -2.6 $10^{-4}$ to -1.4 $10^{-4}$, -1.4 $10^{-4}$ to -5.4 $10^{-5}$, -5.4 $10^{-5}$ to 4.2 $10^{-5}$, > 4.2 $10^{-5}$. Plan curvature has been classified in: < 1.3 $10^{-7}$, 1.3 $10^{-7}$, -1.2 $10^{-4}$, 1.2 $10^{-4}$, -2.2 $10^{-4}$, 2.2 $10^{-4}$, -3.8 $10^{-4}$, > 3.8 $10^{-4}$. Relative relief has been divided in deciles: 0-9 m, 9-18 m, 18-34 m, 34-65 m, 65-120 m, 120-194 m, 194-288 m, 288-424 m, 424-625 m, > 625 m (Fig. 2.6).

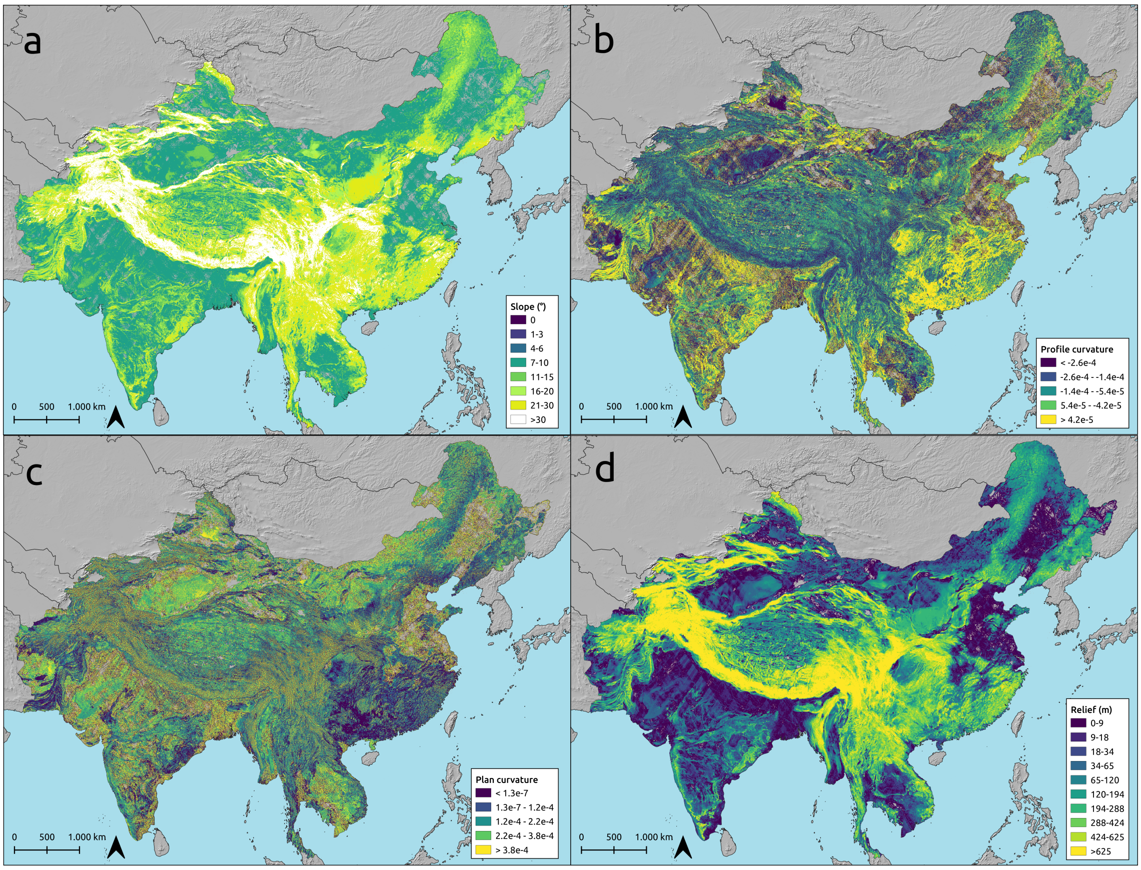


Figure 2.6: Causative factors selected for the Tier 1 landslide susceptibility: **a** slope, **b** profile curvature **c** plan curvature and **d** relief.

The geological factor has been proposed by Hartmann and Moosdorf (2012). They have mapped the lithology of the globe into 16 classes, all of them present in south-Asia.

They have been grouped into 7 classes: (1) ice, glaciers and water bodies; (2) siliciclastic sedimentary rocks, mixed sedimentary rocks, carbonate sedimentary rocks and pyroclastics; (3) mixed sedimentary rocks; (4) evaporites; (5) acid volcanic rocks, intermediate volcanic rocks, basic volcanic rocks; (6) acid plutonic rocks, intermediate plutonic rocks, basic plutonic rocks; and (7) metamorphic rocks (Fig. 2.7e).

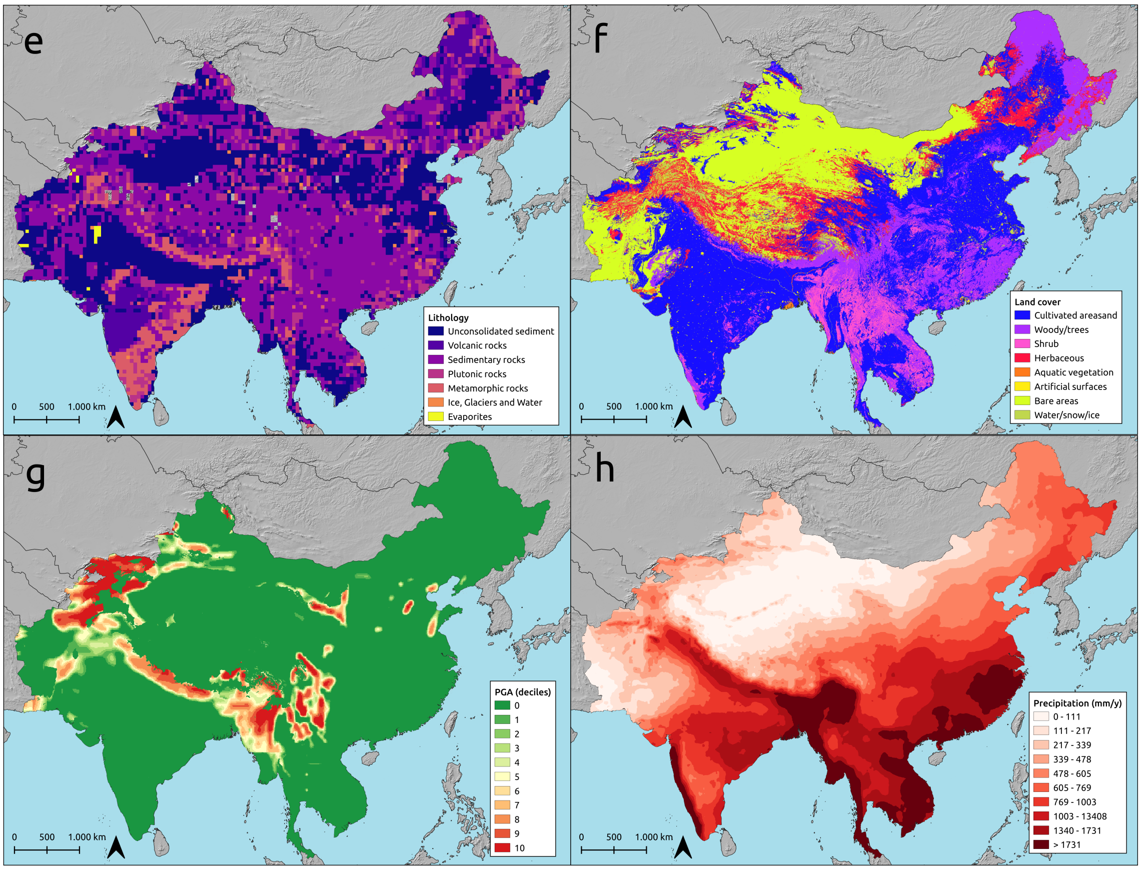


Figure 2.7: Causative factors selected for the Tier 1 landslide susceptibility: **e** lithology, **f** land cover, **g** PGA and **h**. precipitation.

In regards to the land cover classification, the ESA GlobCover 2009 Project has classified the land cover information into 22 classes (Bontemps et al., 2011) which have been grouped into 8 categories (Table 2.3) according to the United Nations (FAO) Land Cover Classification System (LCCS) (Di Gregorio, 2016). Therefore, the 8 LCCS classes have been suggested for Tier 1 land cover factor subdivision (Table 2.3) (Fig. 2.7f).

Since most of the landslides in Asia are mainly triggered by rainfall and earthquakes, two factors have been included: annual rainfall and PGA.

The annual rainfall factor has been calculated from the annual sum of the daily precipitation measured by the Multi-satellitE Retrievals for Global Precipitation Measure (IMERG) (Huffman et al., 2019a) with a cell size of 10x10 km. The final result is the average of the annual precipitation over 11 years (2009-2019) resampled to a 3 km square grid using the bilinear resampling method of SAGA GIS. The map has been classified into deciles: 0-111 mm/y, 111-217 mm/y, 217-339 mm/y, 339-478 mm/y, 478-605 mm/y, 605-769 mm/y, 769-1003 mm/y, 1003-1340 mm/y, 1340-1731 mm/y, >1731 mm/y (Fig. 2.7h).

| GlobCover legend | LCCS |
|---|---|
| Post-flooding or irrigated croplands<br>Rainfed croplands<br>Mosaic Cropland (50-70%) / Vegetation (grassland, shrubland, forest) (20-50%)<br>Mosaic Vegetation (grassland, shrubland, forest) (50-70%) / Cropland (20-50%) | A11 Managed Lands Cultivated Terrestrial Areas |
| Closed to open (>15%) broadleaved evergreen and/or semi-deciduous forest (>5 m)<br>Closed (>40%) broadleaved deciduous forest (>5 m)<br>Open (15-40%) broadleaved deciduous forest (>5 m)<br>Closed (>40%) needleleaved evergreen forest (>5 m)<br>Open (15-40%) needleleaved deciduous or evergreen forest (>5 m)<br>Closed to open (>15%) mixed broadleaved and needleleaved forest (>5 m)<br>Mosaic Forest/Shrubland (50-70%) / Grassland (20-50%)<br>Mosaic Grassland (50-70%) / Forest/Shrubland (20-50%) | A12 Natural and Semi-Natural Terrestrial Vegetation:Woody—Trees1 |
| Closed to open (>15%) shrubland (< 5 m) | A12 Natural and Semi-Natural Terrestrial Vegetation:Shrub |
| Closed to open (>15%) grassland<br>Sparse (>15%) vegetation (woody vegetation, shrubs, grassland) | A12 Natural and Semi-Natural Terrestrial Vegetation:Herbaceous |
| Closed (>40%) broadleaved forest regularly flooded—Fresh water<br>Closed (>40%) broadleaved semi-deciduous and/or evergreen forest regularly flooded—Saline water<br>Closed to open (>15%) vegetation (grassland, shrubland, woody vegetation) on regularly flooded or waterlogged soil—Fresh, brackish or saline water | A24 Natural and Seminatural Aquatic Vegetation |
| Artificial surfaces and associated areas (urban areas > 50%) | B15 Artificial surfaces |
| Bare areas | B16 Bare areas |
| Water bodies<br>Permanent snow and ice | B28 Inland waterbodies, snow and ice |

Table 2.3: Land cover classification conversion from GlobCover classification to LCCS (Bontemps et al., 2011; Di Gregorio, 2016).

The PGA map has been developed from a collaboration among the Columbia University Center for Hazards and Risk Research (CHRR) and Columbia University Center for International Earth Science Information Network (CIESIN) using Global Seismic Hazard Program (GSHAP) data. It includes areas with a probability to exceed at least 10% the PGA in a time span of 50 years (>2 m/s$^2$). The PGA have been classified into deciles from the 1th to the 10th (Dilley et al., 2005; CHRR and CIESIN, 2005). The zero class has been added (Fig. 2.7g).

The details about the data type, source and quality are reported in Table 2.4. All the data used for the analysis are freely available from different global databases (Table 2.4) (Figs. 2.6, 2.7).

The WoE technique has been proposed for the mathematical evaluation of the Landslide Susceptibility Index (LSI). The WoE model introduced by Agterberg et al. (1989) and then by Bonham-Carter et al. (1988) is a bivariate statistical analysis, which compares dependent (landslide inventory) and independent variables (causative factors), it is used to evaluate landslide susceptibility (Pasuto and Tagliavini, 2007). It assigns two weights

| Data type | Data set | Extent | Data extension | Pixels size | EPSG[4] | Data source |
|---|---|---|---|---|---|---|
| Morphological | DEM | global | GEOTIFF | 30x30 m | 4326 | SRTM[5] |
| Geological | Vector | global | WFS | - | 4326 | GLIM[6] (Hartmann and Moosdorf, 2012) |
| Environmental | Raster | global | GEOTIFF | 300x300 m | 432 | ©ESA 2010 and UCLouvain[7] |
| PGA | Raster | global | GEOTIFF | 90x90 m | 4326 | CHRR-CIESIN[8] (CHRR and CIESIN, 2005) |
| Precipitation | Raster | global | txt | 10x10 km | 4326 | IMRG[9] (Huffman et al., 2019a) |
| Landslide inventory | Vector | global | shp | - | 4326 | NASA-COOLR[10] (Kirschbaum et al., 2010) |

Table 2.4: Data freely available from published databases for Tier 1 application.

($W^+$, $W^-$) to the classes of each causative factor. The weights $W^+$ and $W^-$ mean that the presence of the factor is favorable to slope instability and the presence of the factor is favorable to slope stability, respectively. The general formulations of Agterberg et al. (1989) are the following:

$$W^+ = \ln \frac{P(B|D)}{P(B|D_1)} \tag{2.1}$$

$$W^- = \ln \frac{P(B_1|D)}{P(B_1|D_1)} \tag{2.2}$$

$$W_f = W^+ - W^- \tag{2.3}$$

where $P$ is the probability, $B$ is the presence of a potential landslide causative factor, $B_1$ is the absence of a potential landslide causative factor, $D$ is the presence of a landslide and $D_1$ represents the absence of a landslide. $W_f$ is called weight contrast: the magnitude of the contrast reflects the overall spatial relation between causative factors and landslides (Dahal et al., 2008).

The landslide susceptibility is mapped by the sum of $ith$ weights contrast of the classified maps for the $n$ causative factors:

$$SI = \sum_{i=1}^{n} W_{fi} \tag{2.4}$$

The result is the Landslide Susceptibility Index (LSI). Once standardized, it represents a measure of the landslide likelihood of occurrence or a measure of the potential spatial distribution of future landslides.

To evaluate the ability of the susceptibility model to predict the spatial distribution of the landslides and to evaluate the robustness of the model fitting capacity, the Area Under the Curve (AUC), calculated from the Receiving Operating Characteristic (ROC) curve (Chung and Fabbri, 2003; Fawcett, 2006) has been proposed. The ROC curve explores the relation between the True Positive Rate and the False Positive Rate by consecutive cutoffs of the LSI. Formally, each map-unit of the susceptibility map is labeled with True Positive ($tp$), False Positive ($fp$), True Negative ($tn$) and False Negative ($fn$) tags. In the susceptibility map, a map-unit is True or False according to the presence or the absence of landslides, respectively. Moreover, the unit is considered Positive or Negative if the relative susceptibility value is higher (stable unit) or lower (unstable unit) than the cutoff. The ROC curves are graphed coupling $tp_{rate}$ (y-axis) and $fp_{rate}$ (x-axis):

$$tp_{rate} = \frac{tp}{tp - fn} \tag{2.5}$$

$$fp_{rate} = \frac{fp}{fp - tn} \tag{2.6}$$

The area underlying the ROC curve (AUC) can be used as a metric to assess the overall quality of a model: the larger the area, the better the performance of the model over the whole range of possible cutoffs. Therefore, if the AUC is equal to 1 it means that the results are perfect, whereas if it is equal to 0.5 the scenario predicted is unlikely.

Considering the accuracy of the landslide inventory, the analysis has been conducted with the pixel size of 3 km x 3 km. Thus, the causative factors have been resampled to the same size before the statistical analysis. The categorical factors have been resampled taking into account the predominant class, while for the continuous factors, the average of the included values have been calculated. The landslide inventory has been divided randomly in two datasets: 70% and 30%, to train and validate the model. The software used for the analysis are QGIS, SAGA-GIS. The simulation based on the WoE and the validation have been processed with the SZ-plugin (Titti and Sarretta, 2020) developed for QGIS. The ROC analysis has been carried out using the Scikit-learn module(Pedregosa et al., 2011).

## 2.5 Results and discussion

The landslide susceptibility map, resulting from the analysis, is shown in Fig. 2.8a and the class weights are reported in Table 2.5.

The WoE is a bivariate approach which evaluates the single predisposing factor in relation with the dependent variable, Table 2.5 reports the $W^{+}$, $W^{-}$, W$_f$ values and the percentage of landslide cells and area of each class factor. The weight contrasts of slope factor show an almost constant increase in instability from 0° to > 30° , with a peak around 21° - 30° and a negative value between slope angle of 7° and 15° . Noticeably, the mask until 7° of slope adopted for the factors has excluded the first three classes of the slope factor, since their cell value is equal to the average of 30 m slope. Even though the pixel size of the analysis cannot perfectly describe the land surface morphology, the trend of the W$_f$ is realistic. Moreover, the highest number of landslides is present in the class 21° - 30° , which includes the 52% of the landslide cells, revealing the highest W$_f$ of the slope factor. The slope factor also includes the most stable class of all factors which is the class 7° - 10° with W$_f$ equal to -6.

Plan and profile curvatures represent the convexity and concavity of the surface tangent to the contour line and to the slope line, respectively. The former is related to the lateral flow convergence or divergence, while the second to the acceleration and deceleration of a flow along the gravity direction. Based on the W$_f$ values of these factors, there is not a relevant difference between them. The landslides percentage per class and the area percentage are very balanced. (Table 2.5).

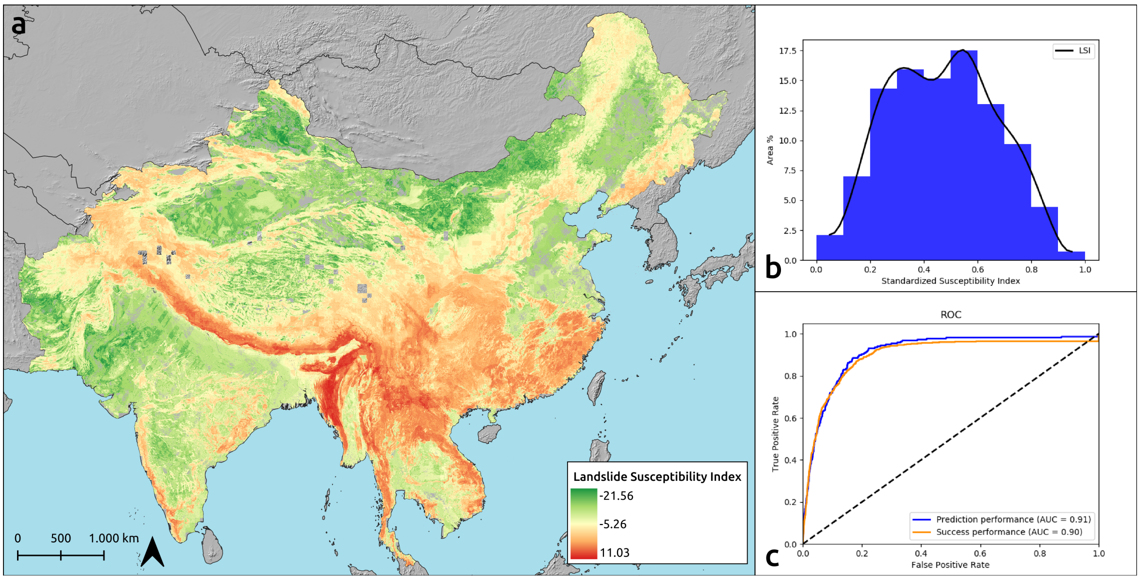


Figure 2.8: **a** 3x3 km landslide susceptibility map of south-Asia. **b** Spatial distribution of the LSI. **c** Prediction rate curve and Success rate curve of the landslide susceptibility map.

The relative relief presents an increasing trend from 18 m to > 625 m. Since the classes represent deciles, the area of each class is almost 10% of the total. Therefore, the trend of the $W_f$ is dependent on the landslide included. The most unstable classes are the 424-625 m and > 625 m which include the 58% of the total landslide cells (Table 2.5).

As regards the lithology factor, the results confirm that $W_f$ parameters must be analyzed in relation to the other classes and with reference to the specific study area. Indeed, depending on the geological context, some unconsolidated lithologies might have lower strengths compared to metamorphic ones. Here, "unconsolidated sediment" is the most stable class, while "metamorphic rocks" is the least stable. In particular, the stability of the former comes from the high extension of the area (31% of the total area), although it includes the 10% of the landslide cells, while the instability of the latter is due to the balance between the number of landslides included (15% of the total landslide cells) and the area covered by the class (7% of the total). The second highest $W_f$ is the "sedimentary rocks" which covers about 42% of the total area and the 60% of the landslide cells (Table 2.5).

Regarding the land cover, it contains one of the most stable classes and the most unstable class of all considered factors. Indeed, the land covered by "shrub" reflects a $W_f$ value equal to 2.35, while "bare areas" display the lowest value, equal to -3.77 (Table 2.5).

The $W_f$ values of the PGA and precipitation classes reveal that the landslides included in the inventory are mainly triggered by precipitation (Fig. 2.3). The PGA $W_f$ are variable between -1.57 and 2.01 without a precise trend. The precipitation classes have an increasing trend similar to relief and slope. The higher the annual precipitation, the higher the instability up to a $W_f$ value of 2.55.

The standardized LSI (0-1) has been divided into 5 classes (Fig. 2.9) to fit the success curve as best as possible: 0-0.54 "very low", 0.54-0.64 "low", 0.64-0.74 "moderate", 0.74-0.80 "high", 0.80-1 "very high". Statistically, in the 5-classes susceptibility map (Fig. 2.9a), the highly susceptible terrain covers 5% (Fig. 2.9b) of the total mapped area, more than the 93% of that is steeper than 15° and the 92% has a relief higher than 194 m. On the contrary, the 68% of the "very low" susceptibility areas, which cover the 68% (Fig. 2.9b) of the study area, respectively, has a relief lower than 120 m (87% lower than 288 m).

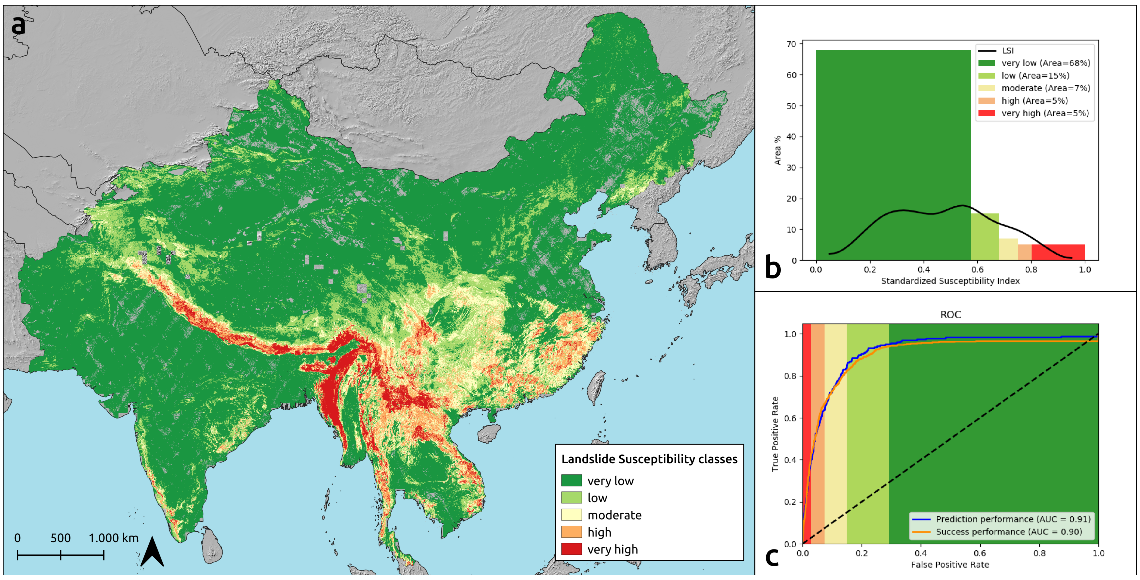


Figure 2.9: **a** Classified 3x3 km landslide susceptibility map of south-Asia. **b** Spatial distribution of the LSI and the area covered by the relative classes. **c** Prediction rate curve and Success rate curve of the landslide susceptibility map.

The weights $W^+$ and $W^-$ are calculated from the relation between the number of landslides included or excluded into the class factor and the size area of the class factor. The result of the difference between $W^+$ and $W^-$ plays a significant role to determine if the specific class is favorable to the instability of the slope. In particular, the areas represented by negative $W_f$ values can be considered stable and the areas represented by positive $W_f$ values, unstable with respect to a specific predisposing factor. Therefore, the LSI resulting from the sum of the $W_f$ may range between negative and positive values which allow for evaluating the stability or instability of the area.

To select the most susceptible area to analyze in detail in the Tier 2 assessment, a susceptibility class has been assigned to each administration unit of the study area. Different levels of administration units are available in GADM website.Footnote 1 Since different levels are available for different countries, a specific level has been assigned to each country to homogenize the dimension of the administration units all over the study area. Taking inspiration from Arup (2020) the relative landslide susceptibility of the single administrative unit has been evaluated as the 80th percentile of the LSI pixel-based map (Fig. 2.8) and then classified from very low to very high to optimize the ROC curve weighted over the extension area of the administrative units area. The result is shown in Fig. 2.10.

The prediction performance and the success of the Tier 1 analysis have been evaluated by the ROC curves (Fawcett, 2006), which are reported in Fig. 2.9c. The curves have been plotted using the validation dataset and the training data set, respectively. The resulting AUC is equal to 0.91 for the prediction curve and equal to 0.90 for the success curve. Overall, the model applied to the selected study area has demonstrated good reliability to evaluate potential instability areas.

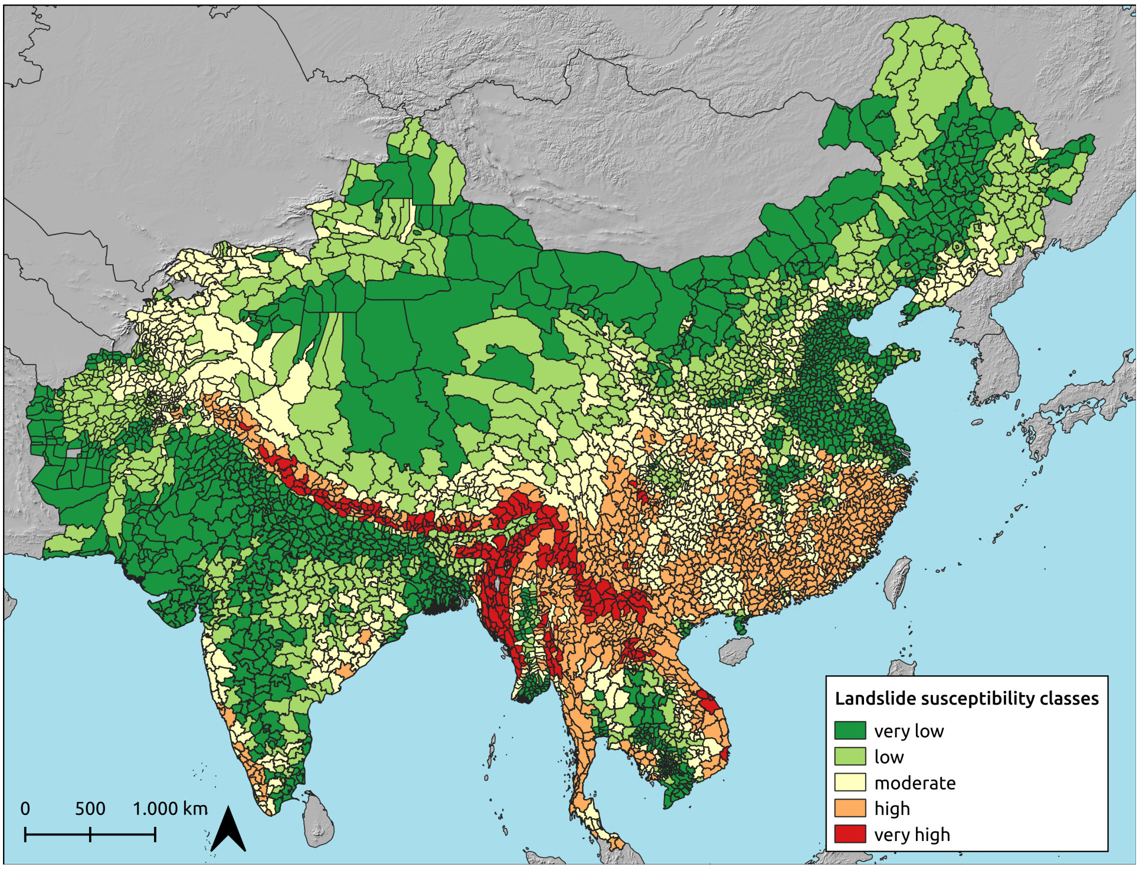


Figure 2.10: Landslide susceptibility map based on administrative units.

An additional way to evaluate the prediction capacity of the model is presented in Fig. 2.11. It compares the ROC curve of the single causative factor with the curve of the landslide susceptibility map. The significant differences among the ROC curve of the susceptibility map (AUC=0.91) and the curve of the slope (AUC=0.77), relief (AUC=0.77), precipitation (AUC=0.90) along with the lower values of the AUC of the other causative factors, confirm the goodness of the prediction performance based on the combination of multiple factors (slope, curvatures, relief, land cover, lithology, PGA and precipitation) in comparison with the use of single factor alone (Günther et al., 2013b; Remondo et al., 2003). The factors that appear to play a major role in predisposing slope instability phenomena with respect to the landslide inventory used are slope, relief and precipitation.

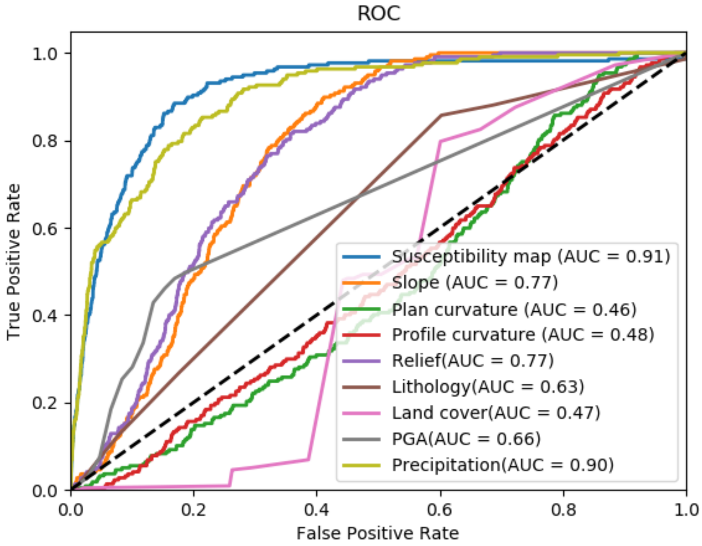


Figure 2.11: ROC curves and relative prediction performances (AUC) of the causative factors and of the susceptible map.

## 2.6 Conclusions

The work is aimed to map the landslide susceptibility in the Belt and Road Countries. In this framework, the landslide susceptibility zoning through the multi-Tier approach has been carried out.

The landslide susceptibility map of south-Asia has been modeled using a quantitative, statistical method. Eight independent variables, i.e., Slope, Plan curvature, Profile curvature, Relative relief, Lithology, Land cover, PGA, Precipitation, have been classified and then weighted by the WoE. The analysis has been based on the NASA-COOLR landslides inventory. It is a global landslides catalog that collects data from online media reports, disaster databases, scientific reports, citizen reports, and others. All the data and software used in this work are open and open-source.

The result is a 5 classes landslide susceptibility map. The ability of the susceptibility model to predict the spatial distribution of the landslides and the goodness of the model fitting have been evaluated by the comparison of the ROC curves calculated from the

| Causative factors | Classes | $W^+$ | $W^-$ | $W_f$ | Landslides (%) | Area (%) |
|---|---|---|---|---|---|---|
| Slope (°) | 0 | - | - | - | - | - |
| | 1-3 | - | - | - | - | - |
| | 4-6 | - | - | - | - | - |
| | 7-10 | -5.55 | 0.46 | -6.01 | 0.14 | 36.78 |
| | 11-15 | -1.10 | 0.14 | -1.24 | 6.27 | 18.74 |
| | 16-20 | 0.08 | -0.01 | 0.09 | 13.96 | 12.90 |
| | 21-30 | 0.94 | -0.51 | 1.44 | 52.14 | 20.43 |
| | >30 | 0.90 | -0.20 | 1.11 | 27.49 | 11.15 |
| Plan curvature | <1.3 $10^{-7}$ | -0.28 | 0.06 | -0.33 | 15.10 | 19.93 |
| | 1.3 $10^{-7}$ to 1.2 $10^{-4}$ | 0.30 | -0.09 | 0.39 | 27.07 | 20.09 |
| | 1.2 $10^{-4}$ to 2.2 $10^{-4}$ | 0.16 | -0.04 | 0.20 | 23.50 | 20.05 |
| | 2.2 $10^{-4}$ to 3.8 $10^{-4}$ | 0.10 | -0.03 | 0.13 | 22.22 | 20.02 |
| | >3.8 $10^{-4}$ | to 0.50 | 0.09 | -0.59 | 12.11 | 19.92 |
| Profile curvature | <-2.6 $10^{-4}$ | 0.16 | -0.04 | 0.20 | 23.50 | 20.01 |
| | -2.6 $10^{-4}$ to -1.4 $10^{-4}$ | 0.18 | -0.05 | 0.24 | 24.07 | 20.02 |
| | -1.4 $10^{-4}$ to -5.4 $10^{-5}$ | -0.11 | 0.03 | -0.14 | 17.95 | 20.04 |
| | -5.4 $10^{-5}$ to 4.2 $10^{-5}$ | -0.09 | 0.02 | 0.01 | 18.38 | 20.04 |
| | >4.2 $10^{-5}$ | -0.21 | 0.05 | -0.25 | 16.10 | 19.89 |
| Relief (m) | 0-9 | 0.00 | 0.11 | -0.11 | 0.00 | 10.50 |
| | 9-18 | 0.00 | 0.10 | -0.10 | 0.00 | 9.46 |
| | 18-34 | -4.22 | 0.10 | -4.33 | 0.14 | 9.74 |
| | 34-65 | -1.95 | 0.09 | -2.04 | 1.42 | 10.05 |
| | 65-120 | -0.86 | 0.06 | -0.92 | 4.27 | 10.10 |
| | 120-194 | -0.29 | 0.03 | -0.32 | 7.41 | 9.92 |
| | 194-288 | 0.08 | -0.01 | 0.09 | 10.83 | 10.00 |
| | 288-424 | 0.50 | -0.07 | 0.57 | 16.52 | 10.05 |
| | 424-625 | 1.17 | -0.29 | 1.46 | 32.48 | 10.07 |
| | >625 | 0.98 | -0.21 | 1.18 | 26.92 | 10.12 |
| Lithology | Unconsolidated sediment | -1.10 | 0.27 | -1.37 | 10.40 | 31.40 |
| | Volcanic rocks | -0.89 | 0.05 | -0.94 | 3.42 | 8.31 |
| | Sedimentary rocks | 0.35 | -0.37 | 0.72 | 60.54 | 42.64 |
| | Plutonic rocks | 0.02 | -0.00 | 0.02 | 10.11 | 9.88 |
| | Metamorphic rocks | 0.78 | -0.09 | 0.87 | 15.53 | 7.17 |
| | Ice, Glaciers and Water | 0.00 | 0.00 | 0.00 | 0.00 | 0.40 |
| | Evaporites | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 |
| Land cover | Cultivated areas | -0.56 | 0.25 | -0.81 | 22.79 | 39.81 |
| | Woody/trees | 0.69 | -0.21 | 0.90 | 31.77 | 15.87 |
| | Shrub | 1.92 | -0.42 | 2.35 | 38.32 | 5.62 |
| | Herbaceous | -1.69 | 0.11 | -1.79 | 2.28 | 12.30 |
| | Aquatic vegetation | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 |
| | Artificial surfaces | 2.08 | -0.03 | 2.11 | 3.42 | 0.43 |
| | Bare areas | -3.51 | 0.26 | -3.77 | 0.71 | 23.77 |
| | Water/snow/ice | -1.10 | 0.01 | -1.11 | 0.71 | 2.13 |

| Causative factors | Classes | $W^+$ | $W^-$ | $W_f$ | Landslides (%) | Area (%) |
|---|---|---|---|---|---|---|
| PGA (deciles) | 0th | -0.54 | 1.03 | -1.57 | 47.29 | 81.21 |
| | 1th | -0.60 | 0.01 | -0.61 | 1.00 | 1.83 |
| | 2th | -0.24 | 0.0035 | -0.25 | 1.28 | 1.63 |
| | 3th | 0.53 | -0.01 | 0.54 | 3.13 | 1.84 |
| | 4th | 1.66 | -0.07 | 1.74 | 8.69 | 1.65 |
| | 5th | 1.18 | -0.04 | 1.23 | 5.70 | 1.74 |
| | 6th | 1.00 | -0.023 | 1.03 | 4.56 | 1.68 |
| | 7th | 1.47 | -0.06 | 1.54 | 7.98 | 1.83 |
| | 8th | 1.90 | -0.11 | 2.01 | 12.25 | 1.84 |
| | 9th | 1.01 | -0.04 | 1.05 | 5.56 | 2.02 |
| | 10th | -0.06 | 0.001 | -0.06 | 2.56 | 2.71 |
| Precipitation (mm/y) | 0-111 | 0 | 0.10 | -0.10 | 0.00 | 9.96 |
| | 111-217 | -3.55 | 0.10 | -3.65 | 0.28 | 9.95 |
| | 217-339 | -3.55 | 0.10 | -3.65 | 0.28 | 9.94 |
| | 339-478 | -1.84 | 0.09 | -1.93 | 1.57 | 9.85 |
| | 478-605 | -2.06 | 0.09 | -2.15 | 1.28 | 10.02 |
| | 605-769 | -2.86 | 0.1 | -2.96 | 0.57 | 9.93 |
| | 769-1003 | -0.81 | 0.06 | -0.87 | 4.42 | 9.94 |
| | 1003-1340 | 0.22 | -0.03 | 0.24 | 12.68 | 10.21 |
| | 1340-1731 | 0.67 | -0.11 | 0.79 | 19.80 | 10.08 |
| | >1731 | 1.77 | -0.79 | 2.55 | 59.12 | 10.11 |

Table 2.5: Causative factors.

validating and training datasets. The prediction and success performance are 0.91 and 0.90, respectively. Among the causal factors slope, relief and precipitation play a major role. The administrative units, displaying moderate to very high susceptibility class, have been selected for further analysis to be carried out at a national scale (Tier 2).

# 3. When Enough Is Really Enough? On the Minimum Number of Landslides to Build Reliable Susceptibility Models

This chapter has been published in *Geosciences* journal in 2021 and authored by: Giacomo Titti[1,3], Cees van Westen[2] Lisa Borgatti[1,3], Alessandro Pasuto[3] , Luigi Lombardo[2] (Titti et al., 2021c).

## 3.1 Introduction

Landslide susceptibility assessment is the most common approach to assess how prone a given landscape is to generate landslides. At the scale of a single slope this is commonly achieved by solving geotechnical relations that express the equilibrium of a potential unstable mass (Van Den Bout et al., 2018, 2021). For broader geographic contexts, i.e., from catchments to regional and even continental scales, landslide susceptibility is commonly generated via data-driven or expert-driven methods. Expert-driven models are based on the standardization and weighting of causative factor maps, even without the availability of a sufficiently complete landslide inventory. Data-driven methods use binary classifiers, of a statistical or machine learning origin. Irrespective of the specific algorithm at hand, certain conditions need to be met for a data-driven-based susceptibility assessment. The first condition is to have access to a series of spatially-explicit instances describing landslides that have already occurred. Under the assumption that “the past is the key to the future” (Carrara et al., 1995), the model learns how to discriminate landslide presences from the absences on the basis of a set of predisposing factors. Any classifier, essentially, measures the correlations existing among the proportion of presence/absence data and the range or classes described by each predisposing factor under consideration. Due to this, variations in the distribution and proportion of presence/absence data have an impact on the model results. Some articles have already investigated this effect. The first attempt was made by Guzzetti et al. (2006), who assessed the variation in model performance as the input data changed by randomly sub-sampling the whole number of mapping units available in their study area. A mapping unit corresponds to the basic geographic objects in which a study area is partitioned and to which probability values of landslide occurrence are assigned to Guzzetti et al. (2006) concluded that the random sampling did not have any major effect

[1] Department of Civil, Chemical, Environmental and Materials Engineering, Alma Mater Studiorum University of Bologna, Viale Risorgimento, 2, 40136 Bologna, Italy

[2] Department of Earth Systems Analysis, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, P.O. Box 6, 7514 AE Enschede, The Netherlands

[3] Research Institute for Geo-Hydrological Protection, Italian National Research Council, C.so Stati Uniti, 4, 35127 Padova, Italy

on the proportion of stable and unstable units, which remained relatively constant for the replicates they produced. A similar conclusion was later reached by Heckmann et al. (2014) who tested a similar assumption but adding a step where the size of the dataset changed while contextually checking auto-correlation issues as many susceptibility replicas were built. Frattini et al. (2010) initially constrained the input data, combining an equal amount of mapping units with absence and presence of landslides. They highlighted variations in the model performance as the proportion of absences became increasingly unbalanced. This topic was further investigated in several contributions such as Van Den Eeckhaut et al. (2006); Petschko et al. (2014); Conoscenti et al. (2016); Lombardo and Mai (2018).

It must be highlighted that all the scientific efforts mentioned above were carried out by keeping the presence information essentially constant and varying the number of the absence instances putted into the model. This is clearly a reasonable assumption because it is important to use as much landslide data as possible in building the best susceptibility model. Fewer articles have tested the implications of varying the presence conditions. This specific topic involves a slightly different scientific domain, where the research question is related to the effects of incomplete inventories onto the resulting susceptibility map (Steger et al., 2017), which means also investigating the minimum requirement with respect to landslide data in order to be able to use data-driven models at all.

In this case, the thread linking the scientific research gravitates around landslide positional errors (Steger et al., 2016) or the missing information of landslide presences at certain locations (Lin et al., 2021). The common denominator among these is to assess the effect of incomplete landslide inventories and test approaches aimed at removing the bias induced by the incompleteness, from an optimal reference model (Steger et al., 2021).

From this particular setting we took inspiration for the present work. However, instead of seeking ways to adjust a model built on the basis of a less numerous landslide inventory, here we pose a different key question: when would a model built on a small number of landslides be sufficiently capable to describe the landslide susceptibility theoretically obtained by training a model on the basis of a much larger dataset, without any specific bias adjustment? The reason for such a question is the fact that in data scarcity situations (Jacobs et al., 2016), it is very common to have limited information on the distribution of landslides (Jacobs et al., 2017), and globally, it is very common to hardly have access to any landslide data at all (Jacobs et al., 2018). Therefore, mapping landslides is a mandatory step without which no data-driven susceptibility model could be built in these areas (Dewitte et al., 2021; Depicker et al., 2021).

More specifically, in the context of the research project 'Silk Road Disaster Risk Reduction', aimed at assessing unstable slopes along the new silk road (Belt and Road Initiative) (Lei et al., 2018), the already existing landslide inventories in this famous pathway are very limited. Based on the landslide susceptibility evaluation carried out at a continental scale in central Asia, Titti et al. (2021a) show the necessity to build more detailed susceptibility models at a national scale even where national scale landslide inventories are not available or very incomplete at best. Considering the vast area and the complexity of the mountainous landscape in the countries of central Asia, to map a complete landslide inventory would take a very long time.

In this context we tried to answer our key question, evaluating the landslide susceptibility in Tajikistan applying several constraints. At the beginning of this work, a limited number of landslide inventories were already available in the study area. To safely consider a landslide inventory that is enough populated to support our analysis, we decided to integrate the existent catalogs with an intense landslide mapping activity from satellite images that required long additional time, after which, we assumed the inventory to be reliable (see Figure 3.1). Then, the analysis was carried out, reducing the number of presence cases in our dataset and actually converted them to absences, mimicking a situation where we did not know that landslides were there. After that, we compared the effect of an increasingly smaller number of landslide presences to a landslide susceptibility model built by using a binomial Generalized Additive Model (Goetz et al., 2011, 2015; Lombardo and Tanyas, 2020), implemented in INLA (Lombardo et al., 2018, 2019, 2020a). INLA is an R (R Core Team, 2020) package developed by Lindgren and Rue (2015) that allows users to implement a number of statistical models in a Bayesian framework, providing analogous and much faster results than traditional MCMC (Bakka et al., 2018; Krainski et al., 2018).

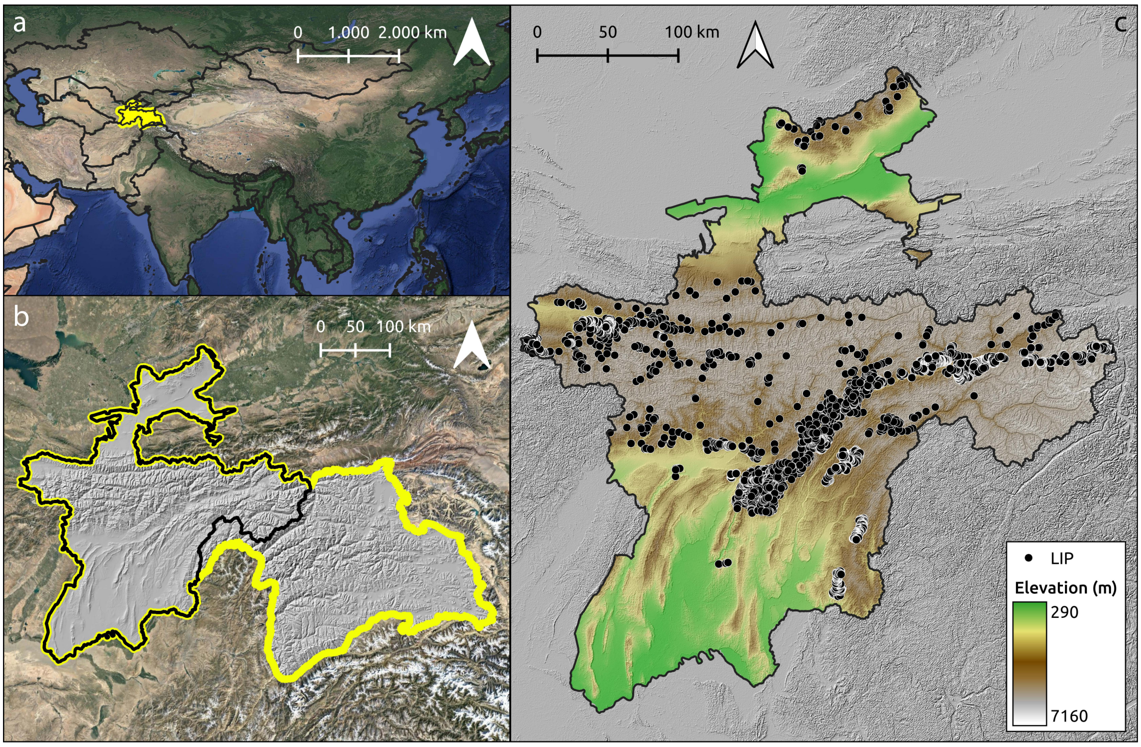


Figure 3.1: Panel (**a**) locates Tajikistan in yellow. Panel (**b**) shows the study area corresponding to the western sector of Tajikistan. Panel (**c**) reports the Landslide Identification Points (LIP) for the mapped landslides.

## 3.2 Materials and methods

### 3.2.1 Tajikistan and its reference landslide inventory

Tajikistan is characterized by an extremely complex rough terrain, with approximately 44% of its territory potentially hosting permafrost, which is rapidly thawing (Mergili et al., 2012). Thick soils with a relatively high erodibility index (Bühlmann et al., 2010) characterize the area. In conjunction with these peculiarities, high intensity precipitation affects the whole territory and strong earthquakes have been reported throughout the history of the country. This combination makes Tajikistan particularly prone to landslides.

Shallow and fast landslides are mainly visible in Tajikistan, such as debris slides and debris flows.

Numerous examples of earthquake-induced landslides (Havenith et al., 2003; Evans et al., 2009; Havenith et al., 2015), rainfall-induced ones (Wang et al., 2021; Schneider et al., 2010; Wolfgramm et al., 2014), and a combination of both effects exist (Torgoev et al., 2013). This has even produced the highest landslide dam in the world (Hanisch, 2018) after a coseismic failure.

Due to the large area of the country, we focused on the western part, excluding the eastern region of Gorno-Badakhshan, which is characterized by the presence of numerous glaciers and by the very low density of population and infrastructure.

Geologically, the western part of Tajikistan is extremely complex, with 12% of its lithologies being of igneous origins, 11% metamorphic, 72% sedimentary, and 3% covered by glaciers (see Figure 3.2a). In terms of land use, the anthropic control on the natural environment is very limited, with relatively small pockets of croplands and large areas with bare land or sparse vegetation (see Figure 3.2b).

As for the main landslide triggers, the strongest seismic activity is registered in the central eastern regions, partly overlapping with the areas with the highest amount of rainfall (see Figure 3.2c,d).

Landslides were mapped as individual points, without any attributes on the classification with respect to type and age, focusing on the initiation areas. Some information has been collected from Havenith et al. (2015); Nazirova and Saidov (2015), and others have been mapped using image interpretation from Google Earth Images. In addition to this, we used the earthquake-induced landslide inventories for some of the historical earthquakes (e.g., the Khait earthquake 1949) (Yablokov, 2001; Evans et al., 2009). Each point representative of a landslide was assigned to the highest point along the landslide crown. We chose this representation instead of the common alternative of a centroid (see (Hussin et al., 2016)) because of the shallow nature of the recognized landslides. This is in fact the most conservative choice in case of shallow flow-like landslides although a laboratory test have also highlighted that the same processes can also exhibit a marked retrogressive behavior (e.g., (Acharya et al., 2009)). Nevertheless, due to the relatively coarse size of the mapping unit we chose (see Section 3.2.3), the models we tested would not be sensitive to either point representation criteria. In other words, both the centroid or the highest point along the landslide crown would fall within the same mapping unit, thus labeled as unstable for the following analyses.

### 3.2.2 Covariates

Nine morphometrical, seismic, and meteorological covariates were selected and pre-processed (see Table 3.1), seven of which are continuous and two categorical.

Relative relief, slope degree, plan curvature (curvature tangent to the contour line), and profile curvature (curvature tangent to the slope line) have been derived from the 30-m Shuttle Radar Topography Mission (SRTM) DEM (Farr et al., 2007). The former has been calculated as the range between the maximum and minimum elevation in a neighborhood of 1 km of radius, the others using the formula by Zevenbergen and Thorne (1987).

The annual rainfall map is the result of the cumulative daily precipitation of one year averaged over 10 years of data (2010–2019). The precipitation data was obtained from

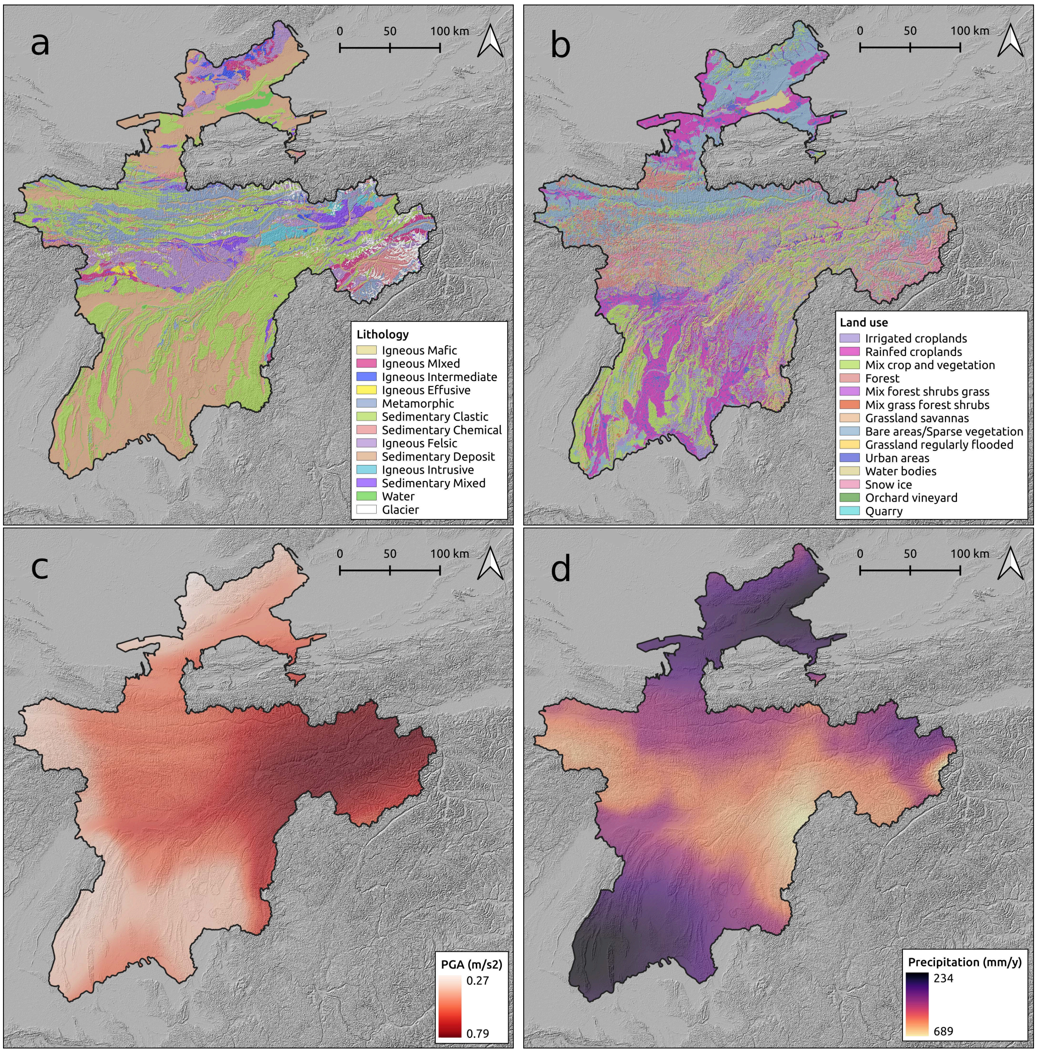


Figure 3.2: Four of the nine covariates used in this study: (**a**) Land use, (**b**) lithology, (**c**) PGA maps of the 475-year Return Period, and (**d**) average annual precipitation (over a 10-year period).

the Multi-Satellite Retrievals for Global Precipitation Measure (IMERG) (Huffman et al., 2019b) with a cell size of 10 × 10 km.

The PGA has been mapped with the 475-year Return Period by Ischuk et al. (2017).

Lithology provided by the Tajikistan General Office of Geology (1974), which has been classified into 13 classes (see Figure 3.2a) and Land use in 14 classes (see Figure 3.2b).

| Covariate | Original Type | Acronym | Unit |
|---|---|---|---|
| Slope degree | Continuous | Slope | degree (°) |
| Relative relief | Continuous | Rlf | m |
| Plan curvature | Continuous | PlC | $m^{-1}$ |
| Profile curvature | Continuous | PrC | $m^{-1}$ |
| Peak Ground Acceleration | Continuous | PGA | $m/s^2$ |
| Annual precipitation | Continuous | Rn | mm/y |
| Land use | Categorical | LU | unitless |
| Lithology | Categorical | Litho | unitless |
| Area with Slope > 10° per map unit | Continuous | Area | $m^2$ |

Table 3.1: Covariates used in the analysis, types, acronyms, and units.

Grid-cells with a slope steepness less than 10 degrees were masked out before computing any of the covariates mentioned above, as these were considered a priori as not being susceptible to landslide initiation. Before any susceptibility analyses was run, a mean zero, unit variance rescaling procedure was applied to all numerical variables.

### 3.2.3 Mapping units

Different methods may be used to obtain the terrain objects in which a given study area is partitioned. In the landslide susceptibility literature, these objects commonly consist of grid cells, terrain units, Unique Condition Units (UCUs), slope units, geo-hydrological units, topographic units, and administrative units (Reichenbach et al., 2018).

To evaluate the landslide susceptibility in Tajikistan, the study area was subdivided into mapping units that take into account the landslides predisposing factors according to the concept of the Unique Condition Unit (UCU) (Ermini et al., 2005; Chiessi et al., 2016). The reason behind such a choice was meant to partition the study area into spatial object that are as independent as possible from the adjacent once while respecting the natural hydrological, geomorphological, and geological structure of the landscape under consideration. This is particularly relevant in case of statistical models, for they assume independence among each instance of the target variable.

Specifically, we intersected a catchment partition (obtained via 'r.watersheds' in QGIS) with the lithology map (scale 1:2,800,000). This operation produced 16020 UCUs with a maximum size of 738 km $^2$. These UCUs were assigned with a binary status 1/0 according to the presence/absence of landslides. After this operation, 677 UCUs contained at least one landslide and 15,343 were considered stable (Figure 3.3).

As for the covariate information associated to each of the UCUs, we opted to summarize the distribution of each continuous parameter via its mean and standard deviation (with the exception of the area). Conversely, the categorical covariates have been expressed in terms of the predominant class per UCU. Overall, this operation led to the generation of 15 covariates. Their specific use within the model will be explained below.

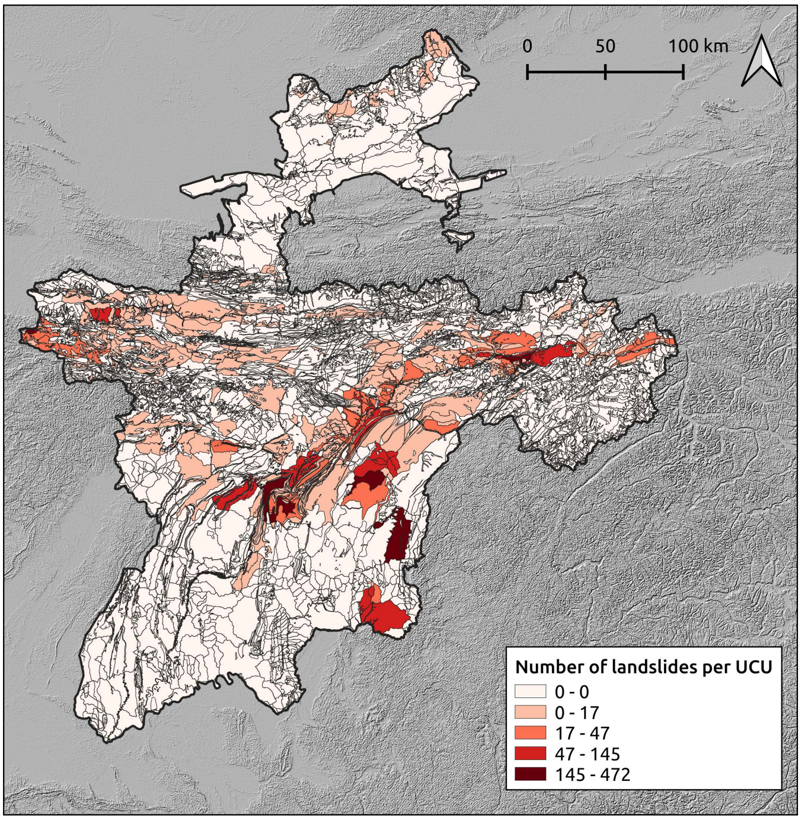


Figure 3.3: Landslide distribution per UCU.

### 3.2.4 Modeling strategy

**Generalized Additive Model**

To estimate the landslide susceptibility of Tajikistan, we chose a Bayesian version of a binomial Generalized Additive Model (GAM) (Goetz et al., 2011; Lombardo et al., 2020b). This is being implemented in INLA (Rue et al., 2017; Lombardo and Tanyas, 2021).

The binomial GAM we implemented can be denoted as follows:

$$\eta(P) = \beta_0 + \sum_{j=1}^{J} \beta_j x_j + f(Litho) + f(LU) + f(Rn_\mu) + f(Slope_\mu), \qquad (3.1)$$

where:

- $\eta$ is the logit link;
- $P$ is the probability of landslide occurrence;
- $\beta_0$ is the global intercept;

- $\beta_j$ are the $j$th regression coefficients estimated for the $x$th covariates which we modeled as fixed effects (or linear properties);
- $f(Litho)$ and $f(LU)$ are two random effects (non linear properties), which we modeled as independent and identically distributed (iid) covariates. This implies that the regression coefficient associated with each class is estimated independently from the other classes;
- $f(Rn_\mu)$ and $f(Slope_\mu)$ are two random effects (non linear properties) that we modeled as random walks of the first order ($rw1$) covariates. This implies that the regression coefficient associated with each class is estimated with an adjacent class dependence. In other words, the coefficient of a single class depends on the coefficient estimated for the class before and after. The use of a random walk allows one to retain the ordinal structure of a covariate that was originally continuous in nature, which we reclassified to obtain a non linear function of the same.

The choice of which variable have to be used linearly or non linearly was tested in a pre-processing step where a bivariate binomial GAM was fitted by using one covariate at a time, with a $rw1$ structure. Then, the covariates that exhibited a nonlinear relation with respect to landslide occurrences have been used accordingly, leaving the others as linear properties.

### Performance assessment

To measure the modeling performance throughout the analyses implemented in this work, we use the Receiver Operating Characteristic (ROC) curves and their Area Under the Curve (AUC), following the performance classification scheme proposed by Hosmer and Lemeshow (2000).

Specifically, we tested the performance of the reference susceptibility model by using a standard 10-fold cross-validation. This means that 10 times the 90% of the UCUs is used to calibrate the model and the remaining 10% of the whole data are used to test the prediction skill. Notably, each 10% subset was randomly selected with the constriction to not appear in the subsequent sample, ensuring that each testing dataset was mutually exclusive with the others and each sample maintain the exact proportion between presence/absence UCUs of the original dataset.

For models built by using a grid-cell partition, it is more advisable that the cross-validation subsets are also spatially constrained. When this condition is met, we refer to this procedure as spatial cross-validation (Brenning, 2005, see ), and it is of fundamental importance to remove any residual dependence in space from the performance assessment. In our case, being the mapping unit a UCU, which is assumed to be independent from the neighboring ones by definition, we did not use an explicit spatial cross-validation although we stress here that this is of fundamental importance whenever the mapping units finely discretize the study area at hand.

Taking one step back, the idea behind a ROC curve is that the spectrum of susceptibility values can be cut into two parts, turning the continuous probability values into two binary label, one consisting of stable slope conditions and the other of unstable ones. These instances can then be matched with the observed data on mapping units that actually contain at least a landslide or not, allowing one to compute confusion matrices reporting

True and False Positives as well as True and False Negatives (Rahmati et al., 2019). A ROC curve is then defined by slicing the susceptibility spectrum via multiple cutoffs, storing the True and False Positives Rates each time.

In this work, not only were the ROC curves used to assess the model performance but we also took inspiration to produce a robust classification of the susceptibility map. The literature is rich of articles where different approaches to classify the susceptibility are used. These span from using a quantile description to more robust statistical criteria with no justification at all (Lombardo et al., 2020a, details in ). In this study we decided to classify the maps in four susceptibility classes: Very low, low, high, and very high. To select the best four probability cutoffs, we tested a large number of combinations through an optimization algorithm (Genetic Algorithm-based) and extracted the four thresholds that yielded the highest AUC value (Titti et al., 2021b). This was done both by using a cutoff that best replicates the number of unstable UCUs per probability bin (classical ROC), and by using the cutoff that best replicates the area (with slope >10°) of unstable UCUs per probability bin (ROC weighted over the UCUs area).

### Fitting different presence data proportions

To investigate the effect of partially-mapped landslides in a data scarce environment, we fitted several binomial GAMs at varying proportions of presence information in the model. Specifically, we used a reference model fitted by using the whole landslide presence/absence information and then, we compared to the reference, a series of models fitted by randomly changing the presence instances into absences. This operation numerically mimics a situation where an actual mapping unit containing a landslide is not properly scanned, thus leading one to miss the landslide occurrence at that location and assigning an absence condition to that UCU. Our random substitution of presences into absences has been gradually run, switching 5% of the total presences at a time, from 0% of the total presence to the 95%. This means that the second run has included 95% of the landslide inventory (5% turned into absences) and the last run with 5% of the inventory. In this manner, each run is linked to the previous by the same landslides sample minus its 5% (as absolute value, the 5% is always relative to the total presences with no switches). Each substitution has been tested with a 10-fold cross-validation. The entire operation has been repeated 10 times for each substitution, creating a routine for which the prediction was tested with respect to the remaining data, which was left unchanged (landslide presences were still presences). In total, 2000 runs has been completed.

As a result of this procedure, we monitored the variation in AUC as the landslide presence information decreased to just 5% of the data we used in the reference model. We also monitored the variation in the regression coefficients estimated for each covariate, being either linearly or nonlinearly used in our model.

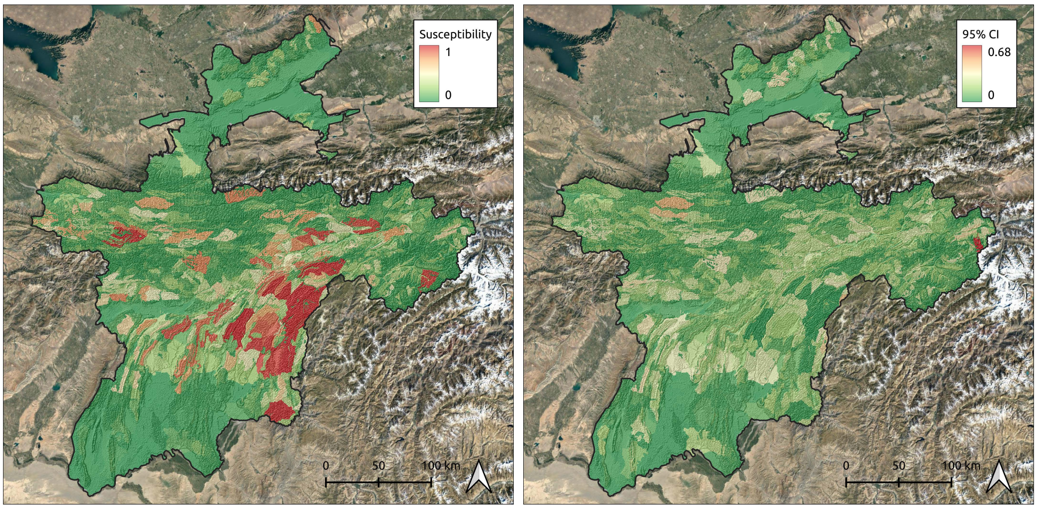
A graphical summary of the analytical protocol we proposed is shown in the flow-chart of Figure 3.12.

## 3.3 Results and discussion

### 3.3.1 Reference susceptibility model

The first model we tested represents the benchmark for the subsequent analyses. It is trained by fitting a model informed of the total number of unstable UCUs and by featuring all the stable UCUs at the same time. The associated area under the ROC curve (AUC) is 0.88, which is considered an outstanding performance class according to the criteria proposed by Hosmer and Lemeshow (2000). The corresponding susceptibility map (measured in terms of posterior mean and its 95% credible interval) is shown in Figure 3.4, where we use the continuous spectrum of probability values to plot it in the left panel.

We present each model component in Figure 3.5. Here the three panels in the first row report the nonlinear effects of the mean slope steepness (Figure 3.5a), mean relief (Figure 3.5b), and mean annual precipitation (Figure 3.5c). The first two appear to share a similar effect: Relatively low steepness and relief values being associated with high regression coefficients whereas the remaining distribution appears to be either not correlated or even negatively correlated to landslide occurrences. As for the rainfall, the pattern appears to be reasonable, with low annual average precipitations being associated with negative regression coefficients, which rapidly increase as the rainfall amount increases.



(a) (b)

Figure 3.4: (**a**) Landslide susceptibility map of Tajikistan. (**b**) 95% of Confidence Interval (CI) of the landslide susceptibility map of Tajikistan.

In Figure 3.5d, the linear contributions of the UCU area and PGA appear to be significant, increasing the probability of landslide presences per mapping unit. It should be noted, however, that there is only one major earthquake-induced landslide inventory available (for the Khait earthquake in 1949) and that therefore the mapped landslides do not reflect earthquake-triggered events in their majority. Conversely, the plan and profile curvatures were estimated to be not significant and with a posterior mean value around zero. This is related to the selection process of the UCU, which contains often fairly large units with large topographical variations.

The categorical variables, land use (Figure 3.5e) and lithology (Figure 3.5f), present a

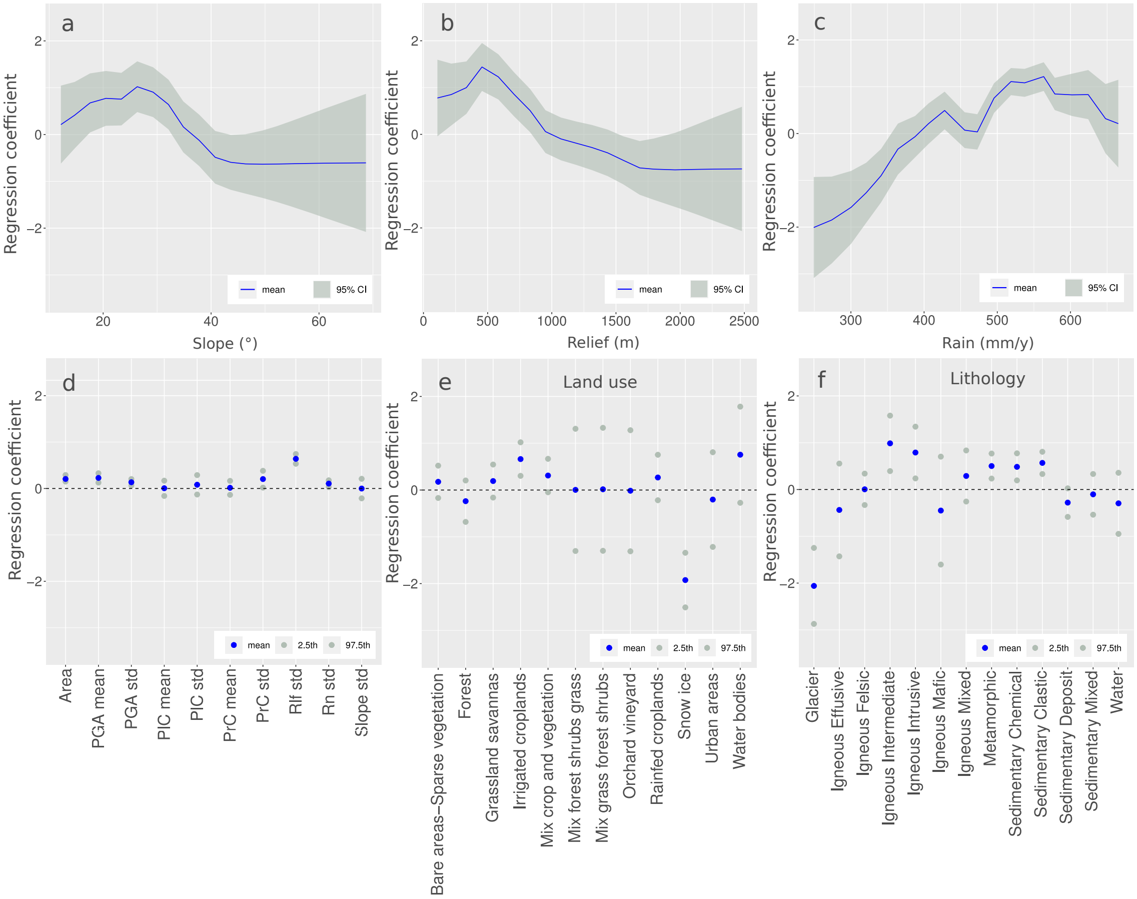


Figure 3.5: Performance of the single covariate used to model the landslide susceptibility map: a) slope, b) relative relief, c) annual precipitation, d) covariates modeled as fixed effects, e) land use, f) lithology. See Table 3.1 for the explanation of the abbreviations of the covariates.

large variation in significance per each category. In detail, land use reveals a significant contribution to increase the probability of landslide presence per UCU with several land use types that have a positive contribution to the presence of landslides. The class with bare or sparse vegetation seems to have less contribution to landslide occurrence that we originally expected. The units with waterbodies have a higher contribution than expected, due to the fact that the slopes around existing hydropower dams have a relatively high landslide density, and the available inventories were more detailed for the areas surrounding hydropower projects. The areas covered by snow and ice decrease the probability of landslide presence per mapping unit. This can also be explained by the difficulty of mapping landslides in these high mountain terrains, where many of the available satellite images have a snow cover, thus leading to a possible (or not) under representation of landslides mapped in these areas.

As regard the role of lithological classes, a large number was estimated to be significant and most were recognized to positively contribute to the susceptibility estimates. These correspond to: Igneous intermediate, igneous intrusive, metamorphic, sedimentary chemical, and sedimentary clastic rocks. Among these, the highest mean regression coefficient is estimated for an igneous intermediate substratum. This could imply that the detrital mantle and soil draping over this bedrock type can be rich in clay as this mineral is the natural product of the long-term weathering effect on igneous rocks (Cawsey and Mellon, 1983). In turn, clays could be responsible for shallow landslides (Ubaidulloev et al., 2021) as they are subject to shrink and swell processes at varying soil moisture conditions. Similar to the land use class, the posterior mean value of glaciers in the lithology map is highly negative.

### 3.3.2 First set of cross-validations

From the 10-fold cross-validation, we extracted the posterior mean to compute the AUC for each replicate. The average AUC out of the 10 calculations is 0.87, and a variability measured with a two-times standard deviation is equal to 0.02. The continuous spectrum of probability which composes the susceptibility map resulting from the first set of traditional cross-validation runs has been classified using two different criteria. The maps are shown in Figure 3.6a,c. We classified the probabilities using the SZ-plugin developed by Titti et al. (2021b). The upper map (Figure 3.6a) is classified by selecting the edges of each class equal to the cutoffs that maximize the AUC of the segmented ROC curve, which means to maximize the number of unstable UCUs per probability bin. Therefore, being the UCUs irregular spatial objects, the map appears dominated by very high susceptibility conditions. Conversely, the bottom map is generated by cutoffs that maximize the area (with slope $>10°$) of unstable UCUs per probability bin (the relative ROC curve is weighted over the area extension). Here the maps appear much more realistic and similar to the benchmark presented in Figure 3.4 with the main difference being expressed in the extension of the red area, which covers 43,058 km$^2$ (43% of total area, see Figure 3.6b) in the first classification and 10012 km$^2$ (13% of total area, see Figure 3.6d) in the second classification.

Figure 3.7a provides two interesting perspectives on the analyses we have run. It demonstrates that the susceptibility pattern (taken from Figure 3.6c) closely reproduce the spatial distribution of landslides at the scale of the UCUs. Figure 3.7b complements this information by highlighting that the highest susceptible zones are not necessarily those with the highest relief, but with the most susceptible lithologies, in combination with the highest levels of rainfall and earthquake acceleration. The relatively lower susceptibility

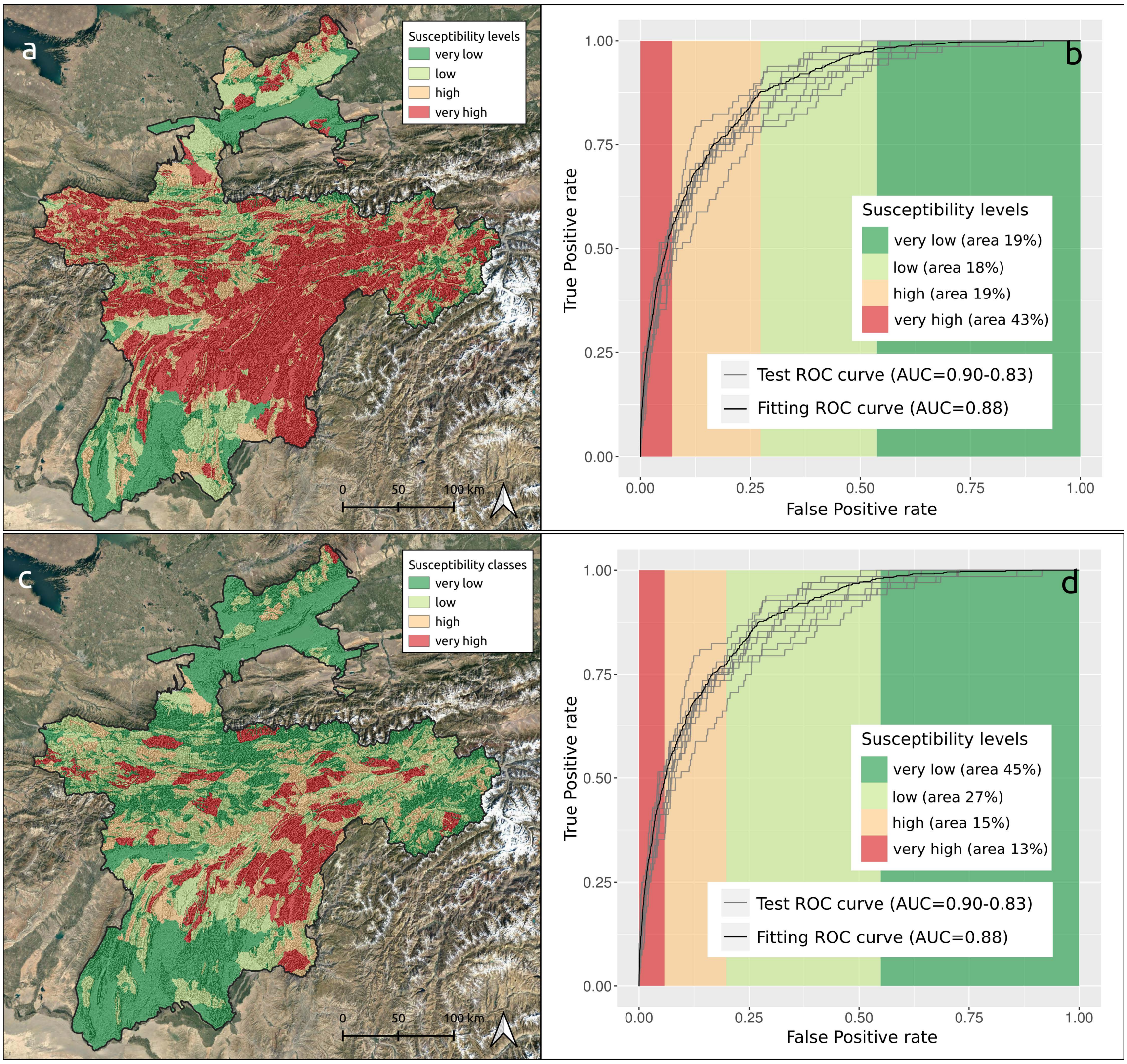


Figure 3.6: (**a**) Tajikistan landslide susceptibility map classified by maximizing the AUC and the relative ROC curves with respect to the unstable UCUs. (**c**) Tajikistan landslide susceptibility map classified by maximizing the AUC of the area-weighted ROC curve and the relative ROC curves. Panels (**b**,**d**) show the corresponding classification schemes and the relative percentage area per class.

of the high elevation, snow- and ice-covered regions is due to the limited availability of inventories for these areas, and may not reflect the actual situation, or the possible future situation, where climate change will contribute to decreasing glacial coverage.

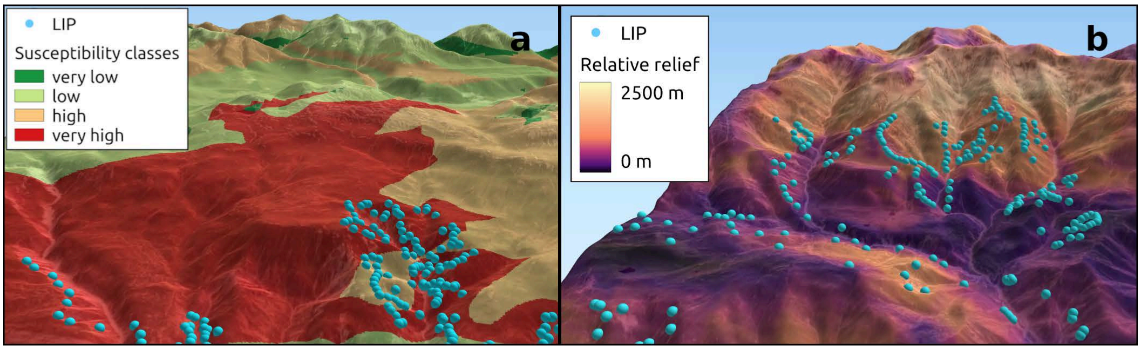


Figure 3.7: (**a**) 3D representations of the landslides distribution on the classified, area-based (Figure 3.6c) susceptibility map. (**b**) 3D representation of the landslides distribution and of the relative relief.

### 3.3.3 Sensitivity analyses at varying landslide presence

In this section, we evaluate the results from the varying proportions of landslide presences per UCU, and compare them with the reference susceptibility model presented in Section 3.3.1. Figure 3.8 reports the AUC values obtained for each cross-validation routine from an initial landslide presence sample of 50 UCUs (5% of the reference inventory) up to 677 UCUs (100% of the reference inventory). Surprisingly, the corresponding difference in performance is not as evident as we initially assumed. In fact, just 5% of the reference inventory is enough to reach a median AUC of approximately 0.79. Conversely, 100% of the reference inventory produced a median AUC of 0.87. The latter corresponds to an outstanding performance according to the classification scheme proposed by Hosmer and Lemeshow (2000), being clearly preferable to a 0.79 (all the ROC curves are visible in Figure 3.19). However, in reality the difference in the time and efforts required to map less than one hundred or several hundreds of mapping units with landslides is very large. Thus, specifically for Tajikistan, such a result inevitably raises the question why this happens and whether it is worth to invest resources in mapping landslides when fewer mass movements already produce suitable predictive performances.

Comparing the classified susceptibility maps assessed as the median of the 10 runs with 100% (Figure 3.9a) and 5% (Figure 3.9b) of total landslide inventory, it is evident how much the latter is less reliable than the former. In particular, this is clearly visible in the southern and northern part of Tajikistan where, on the contrary to the map in Figure 3.6c and to the map in Figure 3.9a, in the susceptibility map modeled with the 5% of total landslides available (Figure 3.9b), the classes ‘high’ and ‘very high’ cover a larger area.

To investigate this result further, we carefully checked the area and recognized, as also shown in Figure 3.5, that landslides generally occur in very similar morphometric situations in Tajikistan. At a medium slope and medium relative relief values they are abundant, whereas they tend not to be represented where snow and ice cover the landscape (mainly at high relative relief values). This is interesting because we cannot know whether this effect is real. In fact, snow cover tend to hide landslides or make it impossible to separate a landslide from other active processes, such as snow avalanches, that dominate a high portion of relief. Therefore, with the same probability, landslides may actually be considered absent and present.

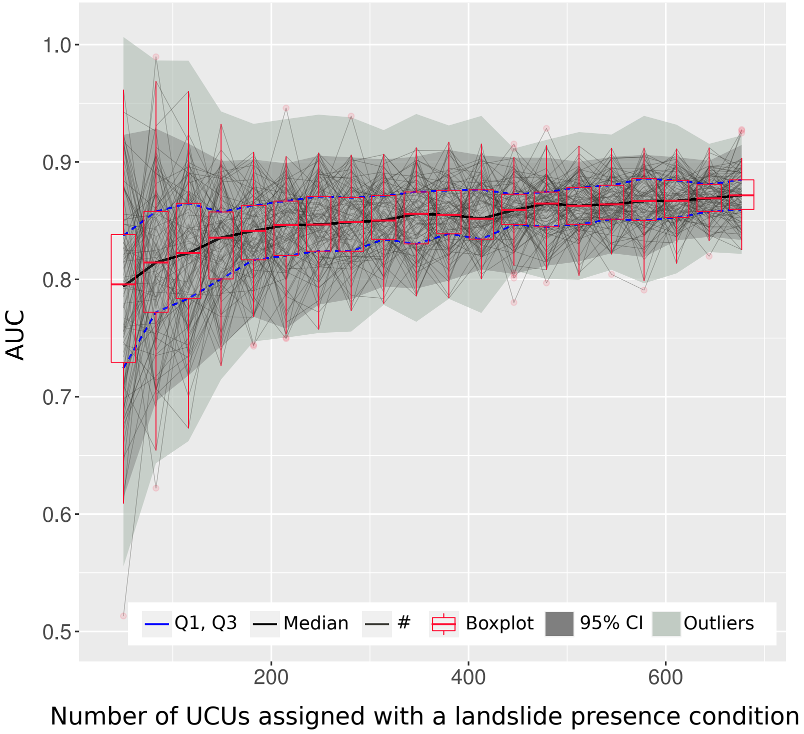


Figure 3.8: Distribution of the Area Under the ROC Curve (AUC) values of the 2000 model runs with different proportions of number presence condition.

Landslides tend to cluster in the wettest portion of the Tajikistan territory. As for the signal captured by categorical properties, many lithologies are quite susceptible to landslides, and contribute more to susceptibility than land use types. The lithological units were grouped into rather broad classes, and therefore the combination of covariates that results in high susceptibility can be characterized by using a relatively small number of UCUs. However, the high AUC values could also be related to the large dimension of the UCUs. Lima et al. (2021) states that a “too detailed representation of the terrain may be detrimental in the presence of inaccurate landslide data”. In our case, since the UCUs are rather large terrain units, this could attenuate the effect of the incompleteness we induced by decreasing the number of presences.

We justify the relatively similar predictive performance through the considerations provided above. However, to support them numerically we checked their validity by examining the variations in the estimated regression coefficients as the proportion of presences changed. For conciseness, a brief overview of this procedure is provided in the main text through Figure 3.10. There we exemplify for the reader the variation of the regression coefficients related to six covariates (one fixed effect, and five classes extracted from five respective random effects), plotted as the information on landslide presence data increases. In Appendix 3.5, we report the variation of the regression coefficients for all the model components estimated to be significant. Specifically, Figure 3.13 shows the variation for the four fixed effects. The PGA and the two curvatures essentially exhibit the same regression coefficient irrespective of the number of presence conditions in the dataset and among each of the bootstrap replicates. Conversely, the UCU area over 10° of the slope

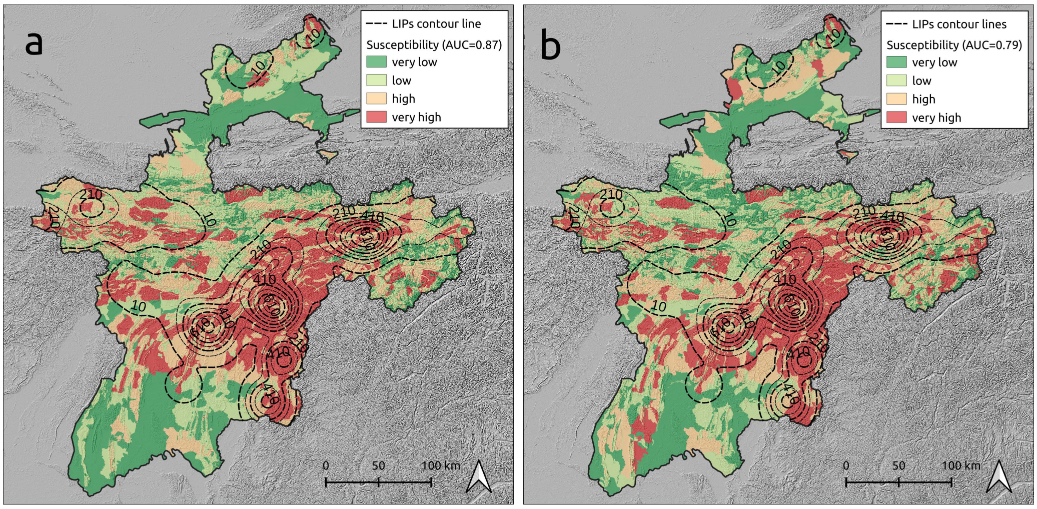


Figure 3.9: (**a**) 10-runs median susceptibility map with 100% of total landslide inventory (AUC = 0.87) classified in quartiles and the contour lines of the landslides density distribution; (**b**) 10-runs median susceptibility map with 5% of total landslide inventory (AUC = 0.79) classified in quartiles and the contour lines of the landslides density distribution.

becomes more positive as the proportion of presences increases.

To provide an overview of what happens with a covariate used in a non linear fashion, we report the effects estimated for each of the mean slope steepness classes, mean relative relief classes, and mean annual rainfall classes. The behavior of the slope steepness is shown in Figure 3.14 where the variations as the number of presence instances in the dataset appear to be stronger than for the linear covariates mentioned before. Specifically, as the number of landslide presences increase, the mean steepness classes 20°–22°, 23°–25°, and 26°–28° are estimated with an increasingly positive contribution to the susceptibility. This trend appears to be interrupted from the class 29°–31° and to invert itself for the classes from 38° to 69° where the regression coefficients become increasingly negative as the proportion of landslide presences increases.

Mean relative relief has been subdivided into 20 classes (Figure 3.15). The first 7 classes show positive regression coefficients, whereas the others became increasingly negative as the relative relief increases. As the proportion of landslide presences change in each relief class, the regression coefficients remain constant.

On the contrary of the sinusoidal behavior of the slope covariate, precipitation presents a discontinuous trend between the minimum and maximum annual rainfall recorded in 10 years in Tajikistan (Figure 3.16). In particular, increasing the number of landslides in sample, it contributes negatively to susceptibility from 243 mm/y to about 375 mm/y where the trend inverts itself, the regression coefficients became increasingly positive as the proportion of landslide presences increases from 486 mm/y. The classes between 375 mm/y and 486 mm/y do not significantly contribute to estimate the susceptibility.

The situation for the categorical properties is also quite interesting (see Figure 3.17 and 3.18). Metamorphic lithotypes (Figure 3.17) exhibited the largest variation as the proportion of landslide presences increases. Specifically, they increasingly contribute in a negative manner to the susceptibility model. Other examples exist where the situation is the opposite albeit with a less striking pattern, these being mixed igneous rocks and chemical sedimentary

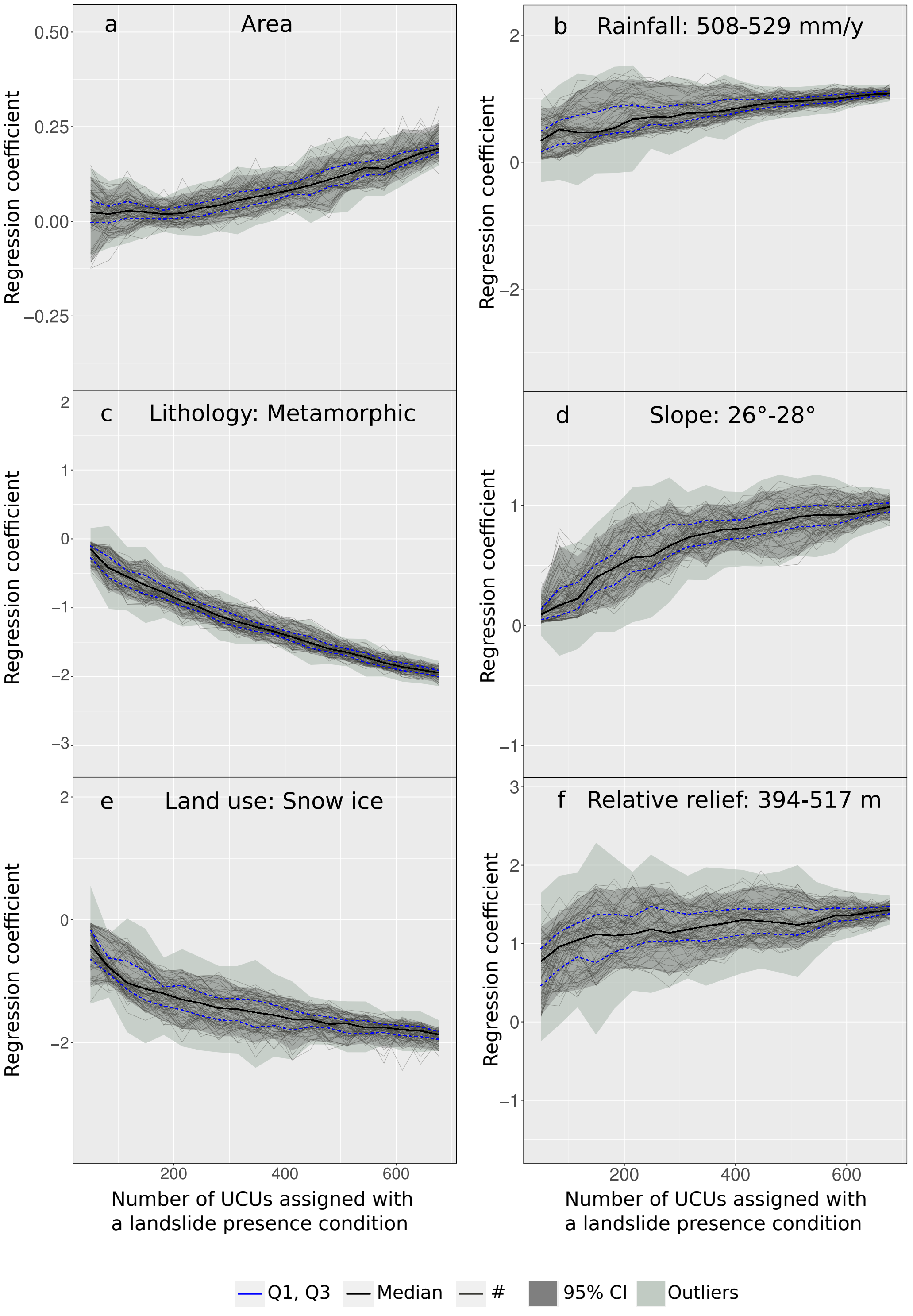


Figure 3.10: Example of covariate effects estimated at a varying proportion of landslide presence data for: (**a**) Area, (**b**) annual rainfall, (**c**) lithology, (**d**) slope, (**e**) land use, and (**f**) relative relief.

materials.

The land use classes (Figure 3.18) show a much more stable response to the landslide presence content in the dataset and across cross-validation replicates. The only exception consists of the snow/ice class, which contributes to decreasing the estimated probability of landslide occurrences per UCU.

Notably, across all these plots not a single example exists where we monitored an inversion of the regression coefficient sign as the proportion of landslides increased. This is a fundamental point which justifies the fact that performance-wise, no extreme changes can be observed at varying presence conditions in the training dataset.

To answer the main question of this paper, starting from a complete landslide inventory, we mapped the landslide susceptibility of west Tajikistan and repeated the same analytical protocol, gradually reducing the amount of landslides presence condition of each UCU, labeling them as stable. As a result, we induced systematic incompleteness and associated biases (Lima et al., 2021) in the the analyses. We did this in the hope of finding a numerical threshold above which the susceptibility map is not sensitive to the addition of new unstable mapping units. However, the AUC distribution plotted in Figure 3.8 was quite smooth. Although, the transition from median AUC values of 0.79 to 0.87 is an interesting trend to monitor, recognizing a single number of unstable UCUs that separates two different predictive behaviors could not be found.

To further deepen our observation, we opted to calculate the rate at which the median AUC shown in Figure 3.8 varies from small to large samples of unstable UCUs passed to the model. This is shown in Figure 3.11 where the rate of change between the subsequent AUC median has been measured with a function $f(x)$, explained as follows:

$$f(x) = \Delta g(x)'_k - \Delta g(x)'_{k-1} \tag{3.2}$$

where, $g(x)$ is the median function, $\Delta g(x)'$ is the gap between two consecutive first order derivatives of $g(x)$ in sliding windows, and $f(x)$ evaluates how fast the curve trend changes by the difference between the gaps $\Delta g(x)'$ of the sliding windows and the previous. Interestingly, the maximum of $f(x)$ corresponds to 220 UCUs assigned with a presence condition (32% of unstable UCUs and 1.5% of the total UCUs). After this, the model does not significantly improve and therefore, we consider this value as the minimum amount of presence condition required for this study area.

This framework raised the question of when shall the inventory be considered complete and what would have been the implications if we would have stopped mapping earlier than the deadline we set. All the analyses we then performed have been voted to address this issue. A series of sensitivity checks have been tested at decreasing numbers of unstable UCUs.

This observation is much more interesting than the one based on Figure 3.8 and it does answer the main question posed at the beginning. However, this result may be site-specific and not generalizable to other study sites or to other kinds of mapping units.

## 3.4 Conclusions

This research was part of a large international project for which a landslide susceptibility assessment was required for landscapes intersected by the new silk road. The first limitation

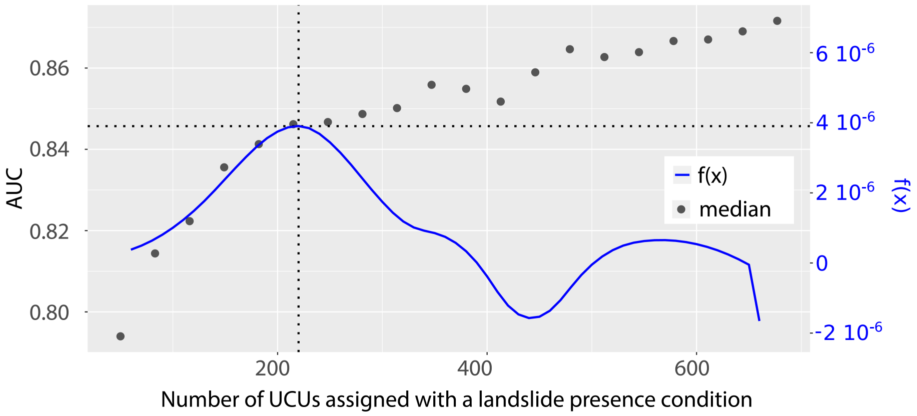


Figure 3.11: Graph in dots of the median AUC also visible in Figure 3.8 and the relative $f$ function (Equation 3.2) which measure how much the trend of the median graph varies.

we found was to access landslide inventories in countries with scarce data availability, among which is Tajikistan. As a result, we mapped landslides under the assumption that we could not map them all and that at a certain point, we should have stopped and considered our inventory suitable for susceptibility modeling.

This framework raised the question when the landslide inventory can be considered complete enough to serve as a basis for susceptibility modeling and what would have been the implications if we would have stopped mapping earlier than the deadline we set. All the analyses we performed were aimed to address this issue, including sensitivity checks tested at decreasing numbers of unstable UCUs. The susceptibility patterns and associated performances appeared quite robust, with relatively small differences when using a small or a large sample of unstable UCUs. This may be due to the relatively coarse spatial partition we adopted, and the use of the UCUs, which were defined as the intersection between catchments and geological units.

The susceptibility analysis measures the propensity of the territory to generate landslides, from the presence/absence information associated to the mapping units and from the selected covariates. The small variations we found may imply that UCUs are already sufficient to capture the landslide distribution even with a relatively small sample of landslides, and that UCUs that are unstable share very specific characteristics. In fact, the difference in AUC between a median model yielding 0.79 and a median model yielding a 0.87 approximately corresponds to just 80 km$^2$, for the very high susceptible class in quartile classification. Therefore, from a pure performance perspective, it seems that it could have been convenient to stop mapping long before the deadline we set for ourselves. However, merely looking at the absolute performance would be a mistake. In particular, the spatial distribution of the probability with an AUC of 0.79 shows high and very high levels of susceptibility in areas where we expected stable conditions.

Therefore, we further investigated the best compromise in between the two extremes. This is shown in Figure 3.11 where we measured the rate of variations in AUC as the unstable sample increased. There, the largest variation coincides with 220 unstable UCUs, after which the improvement rate of the susceptibility significantly drops.

One of the limitations in the generalization of the results we obtained may be due to the mapping unit we chose. In fact, coarser mapping units are less sensitive to variations in the number and distribution of landslides. Conversely, smaller mapping units would likely capture the variations induced by our modeling strategy at varying landslide proportions. This in turn may have led to variations in model performance and covariate effects. We are currently testing the same analytical protocol on the basis of a slope unit-based partition to monitor whether the choice of finer mapping units may yield to slightly different conclusions.

All these considerations could still be very local and not relevant to any landslide susceptibility studies outside Tajikistan. However, we believe the protocol developed is a very interesting tool to investigate a problem that is heavily underestimated in the literature. We envision future development to measure whether the spatial incompleteness we tested here can be extended to its spatio-temporal counterpart.

To promote reproducible and repeatable analyses and to stimulate further research on this topic, we have shared the R codes at this repository GAM-Tajikistan.

## 3.5 Appendix

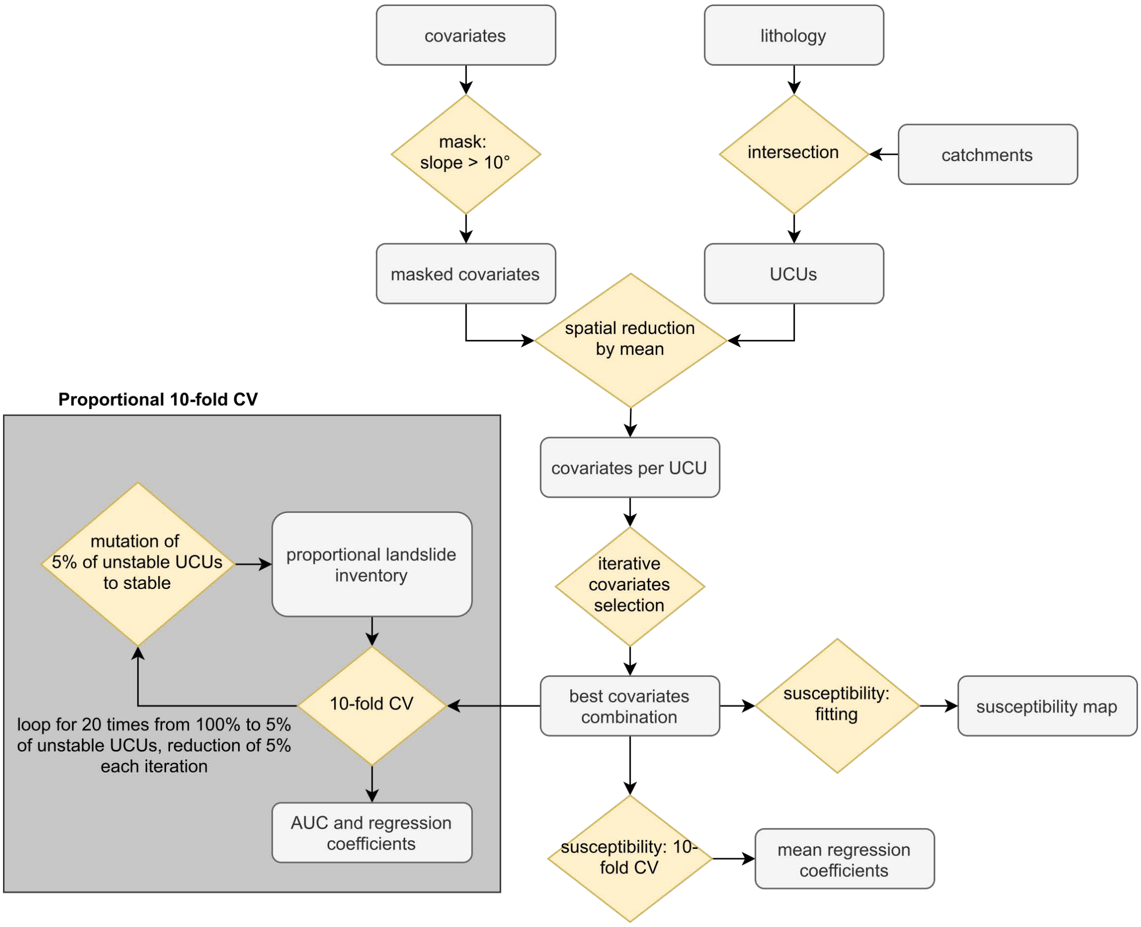


Figure 3.12: Flow-chart summarizing the analytical protocol we implemented.

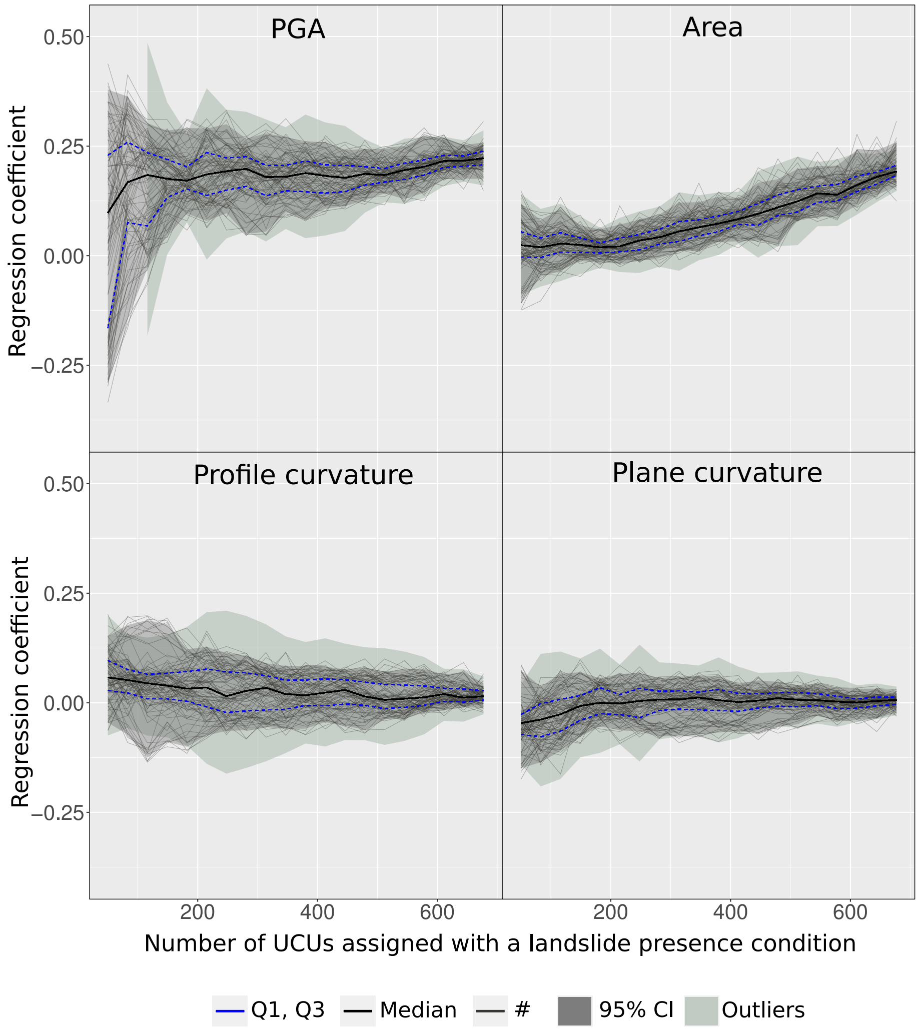


Figure 3.13: PGA, area, profile curvature, and plan curvature effects graphical summary at a varying proportion of landslide presence data.

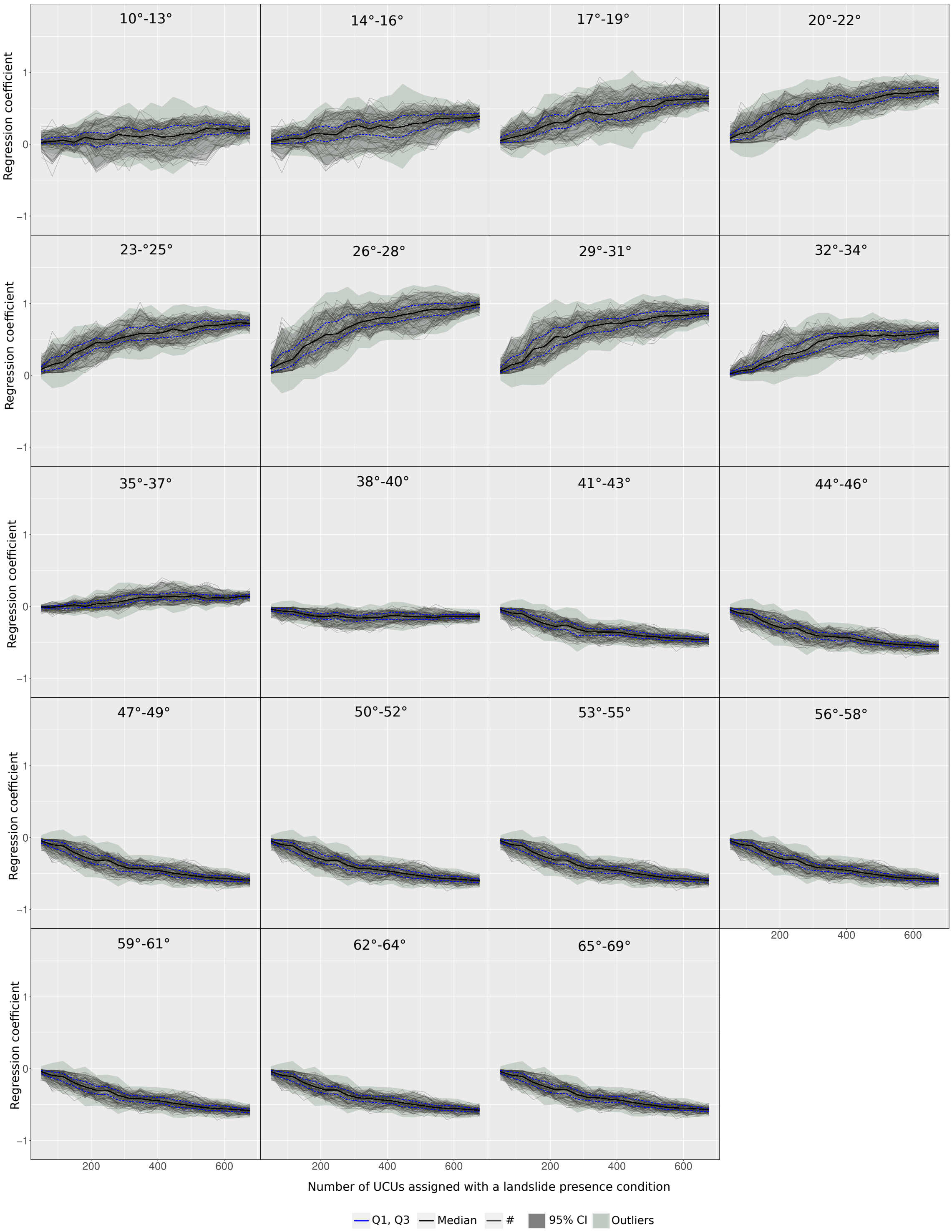


Figure 3.14: Mean slope effect graphical summary at a varying proportion of landslide presence data.

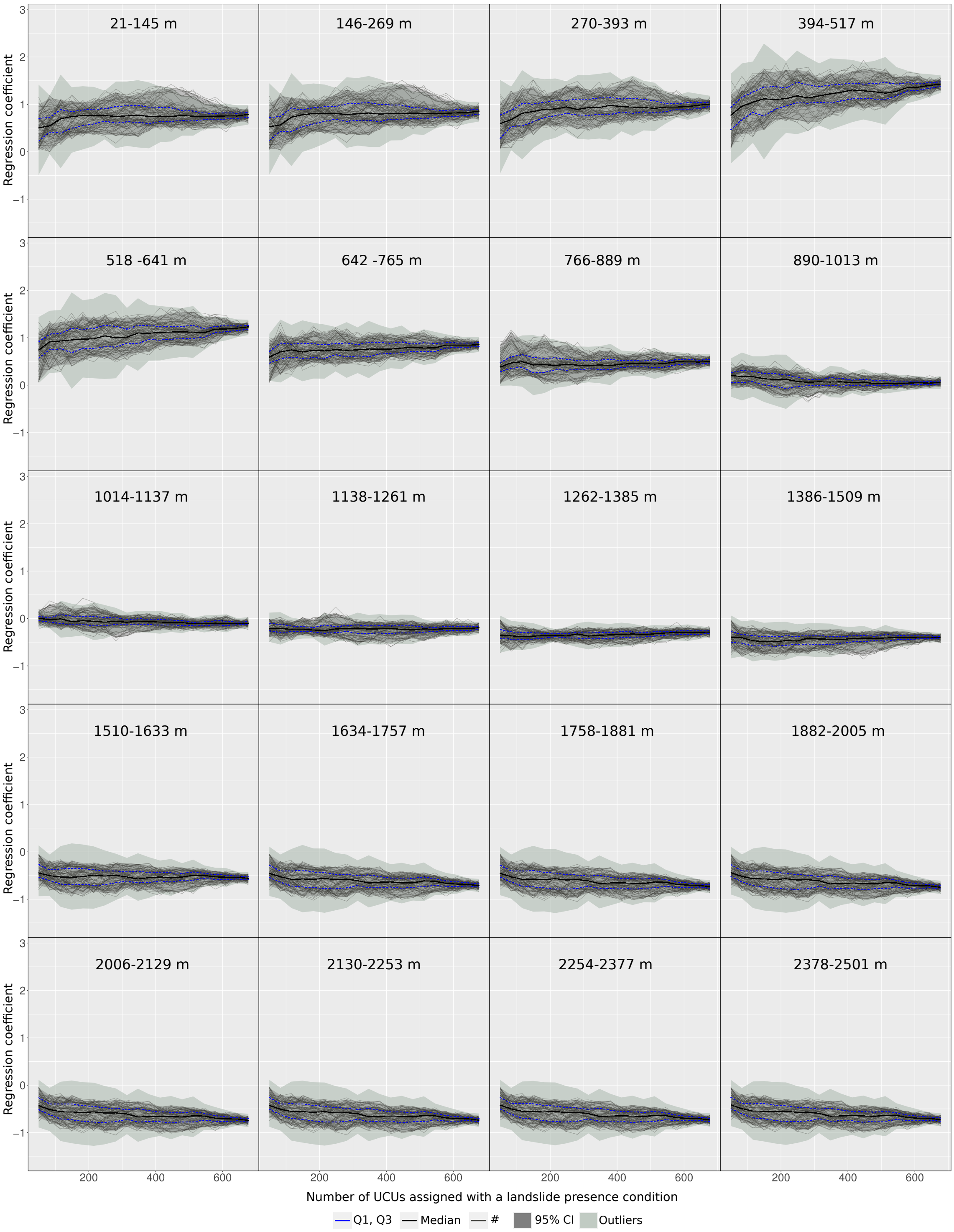


Figure 3.15: Mean relative relief effect graphical summary at a varying proportion of landslide presence data.

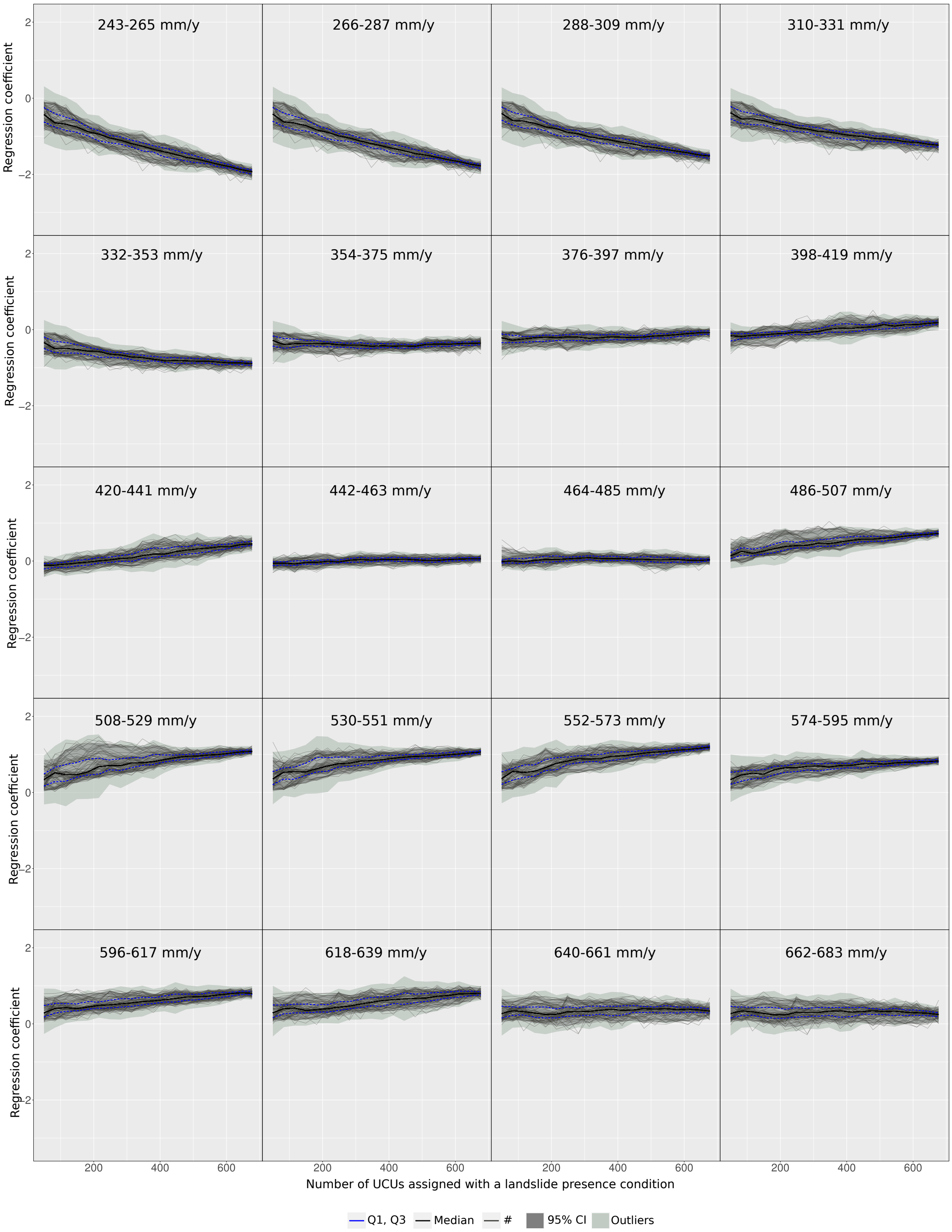


Figure 3.16: Mean annual precipitation effect graphical summary at a varying proportion of landslide presence data.

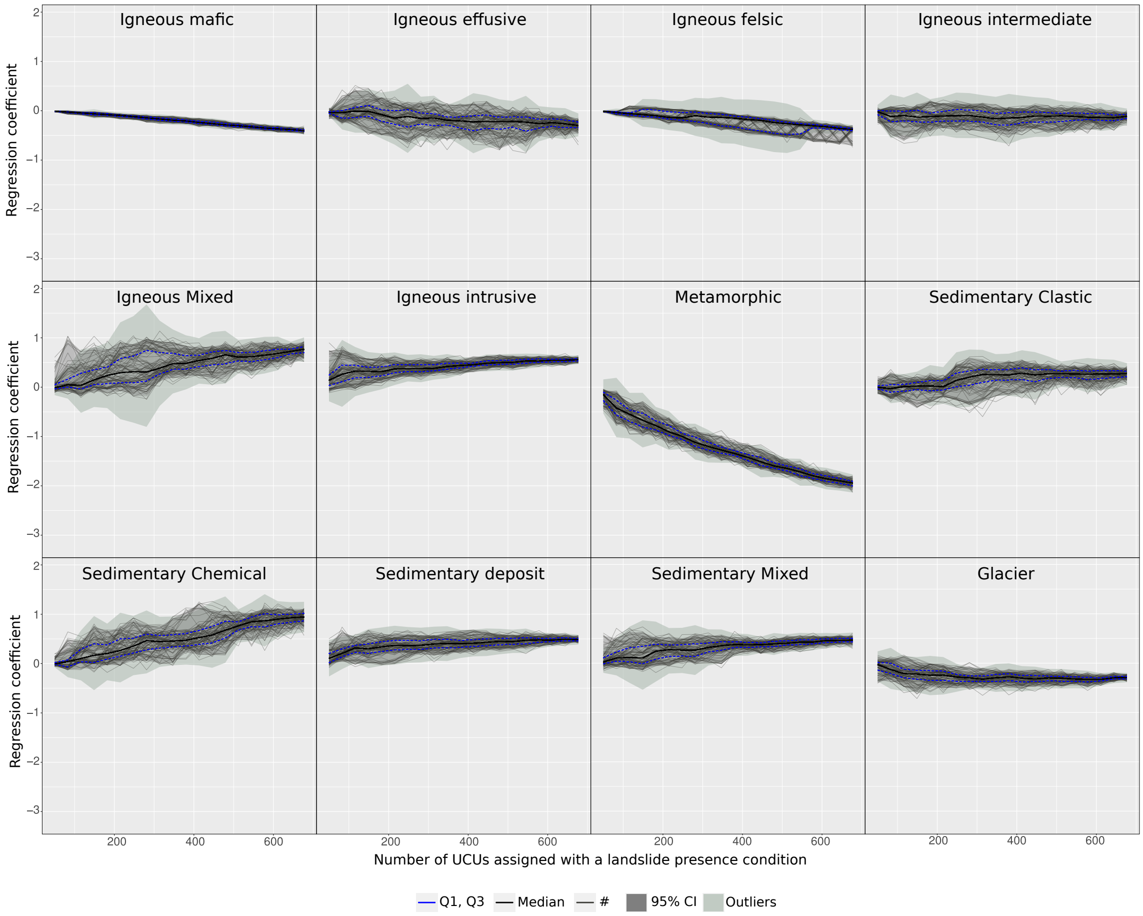


Figure 3.17: Lithology effect graphical summary at a varying proportion of landslide presence data

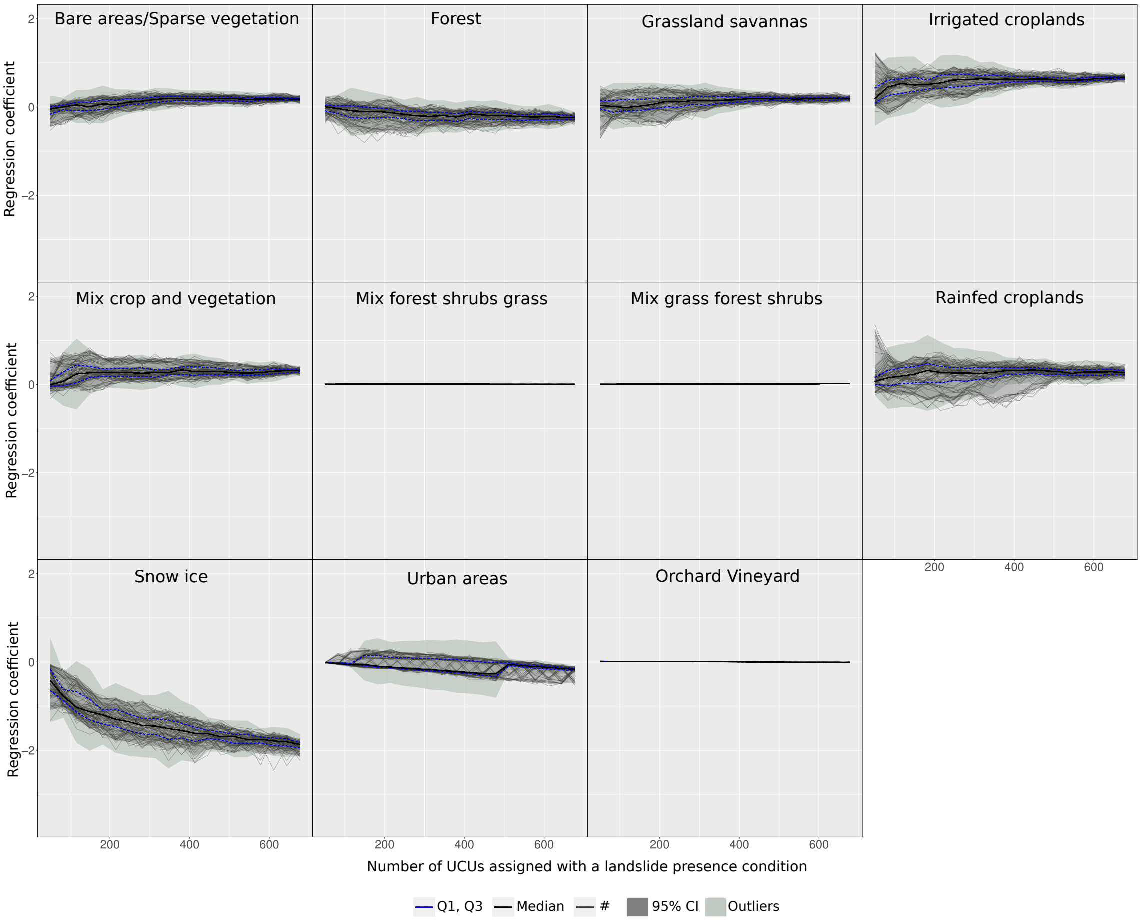


Figure 3.18: Land use effect graphical summary at a varying proportion of landslide presence data.

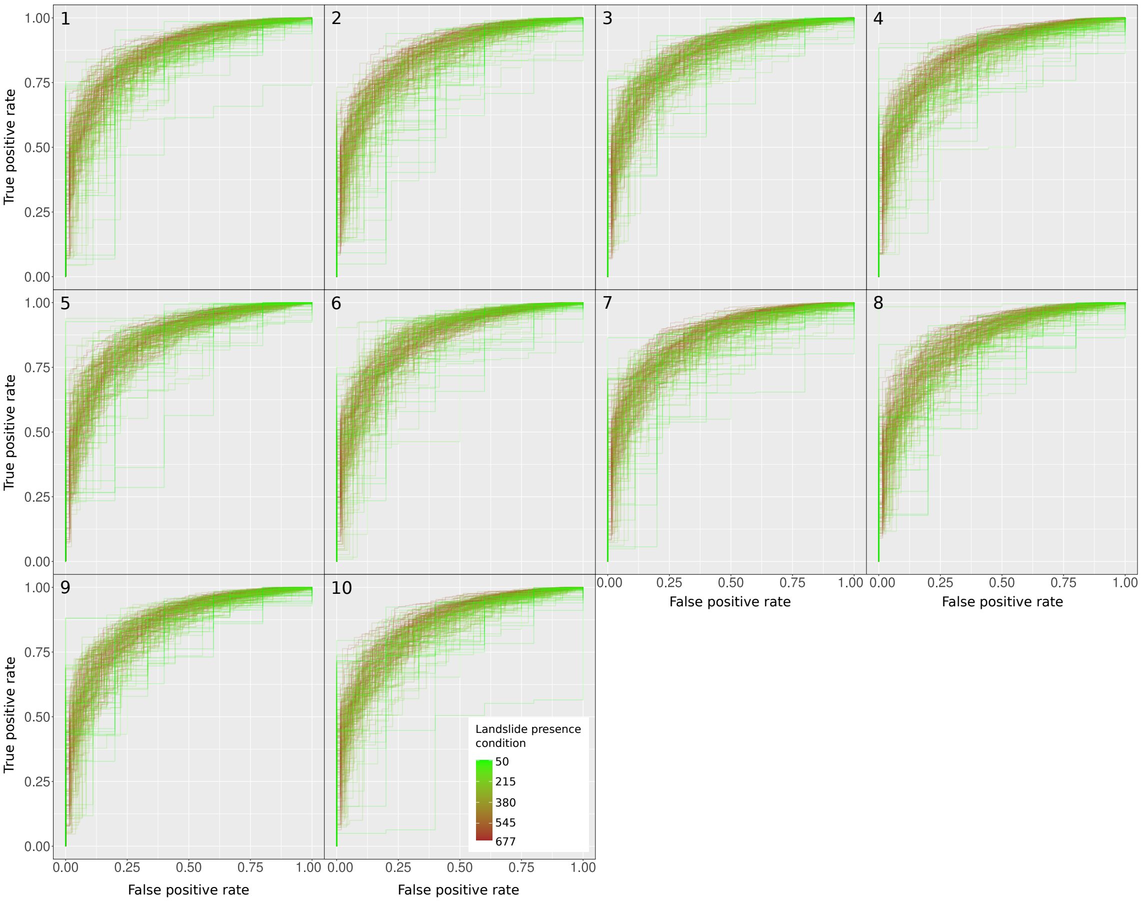


Figure 3.19: ROC curves from the 1st to 10th run at a varying proportion of landslide presence data. Each figure shows 10 ROC curves of cross-validation per 20 different landslide presence proportions (200 curves in each figure).

# 4. Mapping susceptibility with open-source tools: a new plugin for QGIS

This chapter is under review in *Frontiers in Earth Sciences* journal in 2022 and authored by: Giacomo Titti[1,2], Alessandro Sarretta[2], Luigi Lombardo[3], Stefano Crema[2], Alessandro Pasuto[2], Lisa Borgatti[1,2]. The paper is waiting for the final decision of the editor after a first minor review.

## 4.1 Introduction

The measure of how much a specific area is prone to natural hazards is called susceptibility. It does not evaluate when or how often the given hazard may occur (Guzzetti et al., 2006) but it provides the expected locations where such processes may take place in the future. Mathematically, the susceptibility is the estimation of the likelihood of spatial occurrence of natural hazard evaluated on the basis of terrain and environmental conditions (Brabb, 1985). In most cases, this likelihood can be obtained via rigorous probabilistic models, although other tools are also able to convey a similar information without relying on complex multivariate statistics (e.g., Lombardo et al., 2020a; Ciurleo et al., 2017). All these methods fall under the definition of data-driven models and they empirically classify a landscape, labeling it as prone or not prone to slope failures. The way a classifier specifically works is to weigh the contribution of each predisposing factor to the occurrence of natural hazards, taking into account the presence/absence proportion of past records, given other predisposing factors in the model. The basic idea behind data-driven models is that “the past is the key to the future” (Carrara et al., 1995). Thus, an area that has been affected by natural hazards in the past under certain circumstances, may undergo similar environmental stresses and suffer from analogous hazards in the future. Therefore, the statistical analysis of susceptibility is based on a spatial dataset of past events, which acts as the dependent variable of any given model, together with a set of geo-environmental factors acting as explanatory variables.

This paper presents the Susceptibility Zoning (SZ) plugin, a new tool for susceptibility analysis integrated within one of the most common open-source GIS platforms, QGIS (QGIS.org, 2021). Specifically, the SZ-plugin is a collection of functions implemented as a QGIS plugin, supporting a number of pre-processing requirements, as well as the

[1] Department of Civil, Chemical, Environmental and Materials Engineering, Alma Mater Studiorum University of Bologna, Viale Risorgimento, 2, 40136 Bologna, Italy

[2] Research Institute for Geo-Hydrological Protection, Italian National Research Council, C.so Stati Uniti, 4, 35127 Padova, Italy

[3] Department of Earth Systems Analysis, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, P.O. Box 6, 7514 AE Enschede, The Netherlands

susceptibility mapping and validation itself. Moreover, the plugin is equipped with a series of plotting routines aimed at exploring and interpreting each model components as well as estimating the predictive ability of the model when dealing with unknown data.

A number of tools for susceptibility zoning are already available in the literature. These among others LSAT (ArcGIS toolbox) (Torizin, 2012; Polat, 2021), BSA tool (ArcMap tool) (Jebur et al., 2015), LAND-SE (R script) (Rossi and Reichenbach, 2016), frmod (Python script) (Dávid, 2021), GeoFIS (standalone) (Osna et al., 2014). The SZ-plugin, to our knowledge, is the first tool that enables susceptibility routines within QGIS.

This paper describes in detail the SZ-plugin graphical user interface together with all its functions and provides a sample application to landslide susceptibility in north-east India. A previous version of this plugin (v0.1) was already published in 2020 by Titti and Sarretta (2020) but here we have extended the available options encompassing other modeling approaches within the same plugin and we have equipped the SZ-plugin with a suite of plotting and performance evaluation tools. The current version (v1.0) is available in the following GitHub repository CNR-IRPI-Padova/SZ.

## 4.2 Plugin description

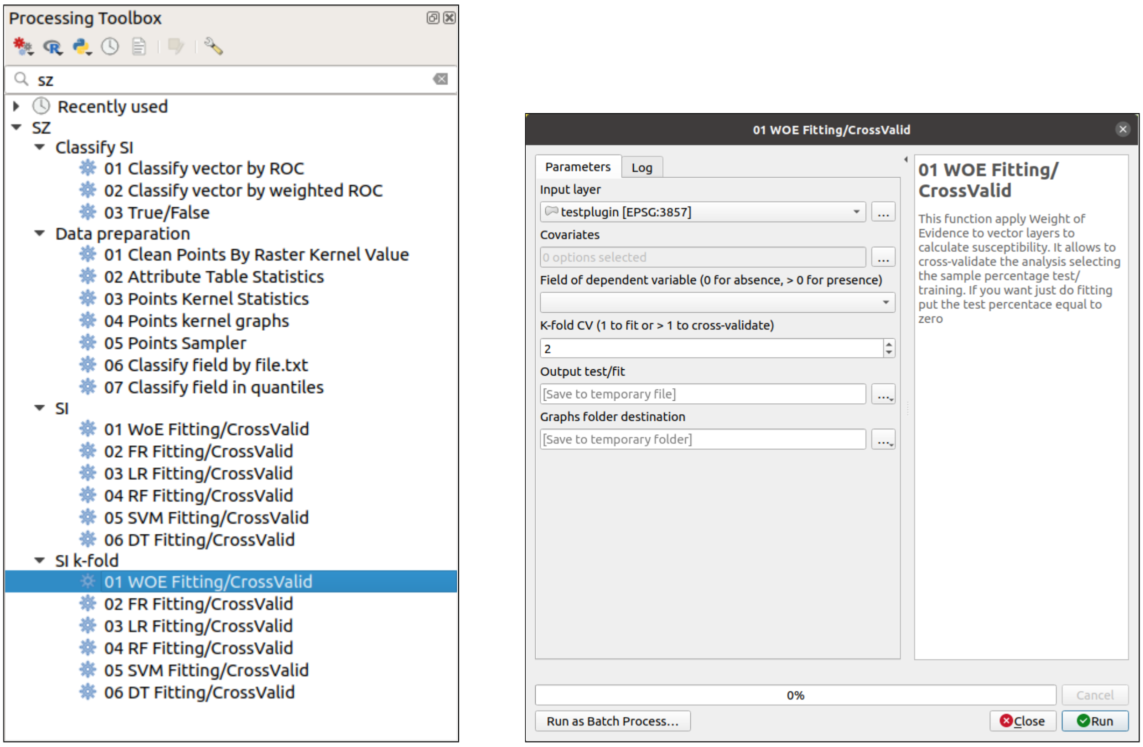
The SZ-plugin has been developed specifically for landslide susceptibility zoning, however, it can be used to map any kind of susceptibility. The code has been written in Python language and developed as a QGIS plugin. QGIS is a software for Geographic Information System (GIS) that is completely open-source and supported by a large community of users and developers. A positive consequence of this open approach is that anyone can develop their own plugin to address specific needs. Hundreds of plugins are freely available from official and non official repositories, but none has focused on susceptibility modeling.

In order to better integrate the plugin with the Graphical User Interface (GUI) of QGIS and simplify its usability, the SZ-plugin can be accessed from the QGIS processing toolbox, the main element of the processing GUI. In detail, the SZ-plugin is a collector of QGIS processing scripts. Some functions can pre-process data according to the asset required by the core model functions, which can estimate and validate the susceptibility using a suite of possible models. These include: Weight of Evidence (WoE, Hussin et al., 2016), Frequency Ratio (FR, Arabameri et al., 2019), Logistic Regression (LR, Lombardo et al., 2020b), Random Forest (RF, Catani et al., 2013), Support Vector Machine (SVM, Lin et al., 2017) and Decision Trees (DT, Yeon et al., 2010). The evaluation of the results and the classification of the final map proposed are based on the Receiving Operating Characteristic (ROC) curves (see Sec. 4.3.2).

Working mainly with vector layers, the SZ-plugin allows to use any shapes or form of mapping unit. In landslide susceptibility, the most common ones consist in grid cells (Reichenbach et al., 2018), terrain units (Van Westen et al., 1997), unique condition units (UCU, Ermini et al., 2005), slope units (Alvioli et al., 2016), geo-hydrological units (Zêzere et al., 2017), topographic units (Eeckhaut et al., 2009) and administrative units (Lombardo et al., 2019).

Figure 4.1 shows the GUI of the plugin inside QGIS, listing the implemented functions, which are separated into four groups: Data preparation, SI, SI k-fold, Classify SI.

’Data preparation’ includes pre-processing functions for vector data. ’SI’ and ’SI k-fold’ are the core groups, which allow users to choose among a number of possible statistical



Input layer: testplugin [EPSG:3857]

Covariates: 0 options selected

Field of dependent variable (0 for absence, > 0 for presence)

K-fold CV (1 to fit or > 1 to cross-validate): 2

Output test/fit: [Save to temporary file]

Graphs folder destination: [Save to temporary folder]

01 WOE Fitting/ CrossValid

This function apply Weight of Evidence to vector layers to calculate susceptibility. It allows to cross-validate the analysis selecting the sample percentage test/ training. If you want just do fitting put the test percentace equal to zero

Figure 4.1: SZ-plugin GUI.

models. 'SI' and 'SI k-fold' allow to fit or cross-validate (CV) the selected model. If a cross-validation is selected, the 'SI' functions use a binomial sampler splitting randomly the dataset into train and test samples. While, 'SI k-fold' functions use k-fold cross-validation (see Sec. 4.3.1) method where the user can choose the number of subsamples. 'Classify SI', instead, provides performance metrics to evaluate the goodness-of-fit or predictive skills in case of cross-validation. The former case returns a single performance value whereas the latter provides summary statistics for the number of cross-validations opted by the user.

It must be stresses that, to maximize the model performance, its assessment offers a new ROC-based classification method which selects the cutoffs required to build the ROC curves to maximizes the relative Area Under the Curve (AUC). Details are reported in Section 4.3.2.

### 4.2.1 Functions description

The sub-section of SZ-plugin called 'Data preparation' is useful to pre-process and set the data to be used by the 'SI' and 'SI k-fold' functions. These functions are:

- 01 Clean points by raster kernel value: it is a filter function that removes all the points of a layer that do not satisfy the minimum value selected in a fixed neighborhood. The values of the neighborhood are collected from an overlapping raster layer.

- 02 Attribute table statistics: this function detects, field by field, the unique values and lists the ID of the feature which reports the same value. Moreover, it produces histograms of the unique values frequency using the Plotly library.

- 03 Points kernel statistics: it calculates the effective, maximum, minimum, standard deviation, sum, average and maximum range value of the points neighborhood.

- 04 Points kernel graphs: the results of the previous function are plotted by this function as frequency graphs.
- 05 Points sampler: this function randomly samples the vector points according to the train/test scheme selected by the user.
- 06 Classify field by file.txt: to apply the WoE or FR method, the covariates should be cut in a number of classes. This function classifies the vector fields according to the bin limiting values that the operator may choose. These need to be reported in a text file.
- 07 Classify field in quantiles: if the operator does not wish to provide the bin limiting values, the vector fields can be classified according to any quantile representation (i.e., deciles, quartiles).

'SI' and 'SI k-fold' functions applies WoE, LR, DT, RF or FR to calculate susceptibility. They require a polygonal layer which includes one field per each covariate and one field with the dependent variable (number of i.e., landslides, tornado, floods) per mapping unit. The 'SI' functions allow to cross-validate the results selecting the sample percentage of training and test or allow to fit the model to the whole dataset. Whereas, 'SI k-fold' functions allow to cross-validate the results with a k-fold method (see Sec. 4.3.1) or also fit the model to the whole dataset.

The functions produce vectors of training/testing or fitting results and report in a text file the relative weights or regression coefficients (depending on the model the user has chosen). And, it produces a graph of the ROC curves with the associated AUCs. The close the AUC is to 1, the higher the capacity of the given model is to suitably classify the study area into stable or unstable conditions. The ROC analysis as well as all the probabilistic models available within the SZ-plugin (LR, DT, RF and SVM) are based on the library Scikit-learn (Pedregosa et al., 2011). As for the bivariate statistical models (WoE and FR), they have been implemented manually and added to the collection because largely used in the literature from many years (van Westen et al., 2000).

The last group of functions is the 'Classify SI' which includes: '01 Classify vector by ROC', '02 Classify vector by weighted ROC' and '03 True/False'. Using Genetic Algorithms (GA) the 01 and 02 'Classify SI' functions classify the Susceptibility Index (SI) into the indicated number of classes through the reconstruction of the segmented ROC curve in order to maximize the AUC (for more details see Sec. 4.3.2).

The only difference between the '01 Classify vector by ROC' and the '02 Classify vector by weighted ROC' is the use of weighted ROC curve. The weights can be established among the input vector fields.

'03 True/False' produces a map of mapping units labeled as: True Positive, True Negative, False Positive, False Negative according to a selected cutoff of the susceptibility index.

### 4.2.2 Software availability

The SZ-plugin has been implemented in Python 3. It requires many dependencies such as: Numpy, Scipy, GDAL, Scikit-learn (Pedregosa et al., 2011), Pandas (The pandas development team, 2019), Matplotlib (Hunter, 2007), Plotly (Plotly Technologies Inc., 2015). The version of SZ-plugin used in this work is the v1.0 (Titti et al., 2021b). The latest version of the SZ-plugin is always available on the GitHub repository CNR-IRPI-Padova/SZ.

The plugin has been tested with QGIS 3.16 on Ubuntu 20.04 and Windows 11. To support and increase the usability of the plugin a video tutorial has been published here https://www.youtube.com/watch?v=XpsiCkVF11s.

## 4.3 Background

To understand the functionalities of the plugin, the explanation on how the plugin handles the cross-validation, performance assessment and susceptibility index is reported hereafter. Moreover, the Section 4.4 shows the application of the tool to landslide susceptibility zoning using the WoE and LR.

### 4.3.1 Cross-validation

Validation routines involve the test of the performance of a data-driven model with respect to unknown data. Ideally, the unknown data should belong to a temporal replicate of the modeled process. However, geomorphological studies often lack of multitemporal inventories. Thus, testing the model performance most of the times requires the implementation of cross-validation routines. These are commonly performed by splitting the entire input dataset into two subsets, one used for training the given data-driven model and the other one to test. This structure revolves around considering a subsample of the whole dataset in the same way as one would consider future landslide occurrences, thus offering the chance to compare locations labeled to be stable/unstable with respect to a set of actual stable/unstable instances.

The literature reports few ways to extract the testing subset. The most common in the geomorphological literature is to extract a random sample from the full dataset. Most of the times this is done just once (Arabameri et al., 2020), in this case most of the variability of a study area is disregarded. In other cases, the random samples are extracted, without any constraint, a large number of times with the purpose of depicting the potential variability of a test site (Amato et al., 2019). In fewer cases, the variability of a given study area is accounted for by extracting samples that are constrained to be selected just once across replicates, leading to two-fold (Yeon et al., 2010), five-fold (Dang et al., 2019) or ten-fold (Lombardo and Tanyas, 2021) cross-validations. All the examples mentioned above adopt a cross-validation scheme where the extraction, constrained or unconstrained, is randomized in space. However, this operation is statistically appropriate only if one assumes that the presence/absence label assigned to a given mapping unit of choice is independent from the labels of the surrounding mapping units. In other words, these procedures assume that there is no spatial structure in the data other than the one captured by the selected explanatory variables. This is an acceptable assumption for medium (e.g., slope units) to large (e.g., catchments) mapping units but it is not valid for fine mapping units such as grid-cells. In such cases, the most appropriate way to implement cross-validation routines is to constrain the testing subsets in space, each one being representative of a specific sector of the study area. In turn, this operation ensures that any residual spatial structure in the data, not captured by the explanatory variables, would not affect the validation estimates. In other words, the testing is free or as free as possible from any spatial effect that may bias the performance toward results that are forcefully better than what they should be.

This operation is commonly referred to as spatial-cross validation (e.g., Petschko et al., 2014).

Out of the cross-validation schemes described above, the current version of the SZ-plugin offers two options. The first is a cross-validation where the train/test split is performed just once as per the majority of cases in the geoscientific literature ('SI' functions). This is achieved by using the *train_test_split* function in Scikit-learn. The second is a $k$-fold cross-validation where from the whole dataset, the test data is randomly extracted according to number ($k$) of mutually exclusive subsets ('SI k-fold' functions). The training data is represented by the complementary subsets. In other words, if $k$ is equal to 10, then ten non-overlapping test sub-samples (each one made of 10% of the total mapping units) are created to test the prediction skill of the model and the ten complementary 90% subsets are used to calibrate the model instead. Thus, the union of the ten 10% subsets returns the whole study area, in such case the resulting susceptibility map would have been generated by fully predicted instances. Any test sample is balanced in terms of presence/absence. The presence/absence proportion of the test sample is equal to the proportion of presence/absence of the complete dataset. The spatial-cross validation is not implemented in the current version of the SZ-plugin but it is part of the development plan for the subsequent versions.

### 4.3.2 Performance assessment

The model performance is evaluated via the Receiving Operating Characteristic (ROC) curves and their relative AUC (Fawcett, 2006; Chung and Fabbri, 2003).

Each mapping unit of a susceptibility map can be labeled as True Positive ($TP$), True Negative ($TN$), False Positive ($FP$) and False Negative ($FN$) unit (Rahmati et al., 2019) according to the presence/absence of the dependent variable (landslides in our application) and to positive/negative (stable/unstable) label assigned. After sorting the susceptibility index by descending order, the spectrum of susceptibility values can be split into two categories, prone or not to the occurrence of the hazard. The susceptibility cutoffs, which assign the binary label (stable/unstable), are assigned continuously until the lowest value of susceptibility. The ROC curve then plots the relation between $TP_{rate}$ and $FP_{rate}$:

$$TP_{rate} = \frac{TP}{TP + FN} FP_{rate} = \frac{FP}{FP + TN} \tag{4.1}$$

### 4.3.3 Susceptibility index classifier

In the literature, several methods have been used to segment the susceptibility index into discrete classes. The continuous spectrum of susceptibility values has been sliced into quantiles, other cases report their classification on the number of landslides per class or on the ratio between the landslide area and the surface area of each class in comparison to the entire study area (Lombardo et al., 2020a). The SZ-plugin proposes a new Genetic Algorithm-based classifier.

The Genetic Algorithm (GA) is an iterative meta-heuristic algorithm based on the numerical reproduction of the Charles Darwin natural selection theory (Chatterjee et al., 1996). The meta-heuristic algorithms are designed to explore the search space from several points of view and to get the solution as near as possible to the optimal (Said et al., 2014).

A GA near-optimal solution is reached through the iterative delineation of the best object selected from a group of admissible solutions which are evolved by operators such as: crossover, mutation, inversion and others (Chatterjee et al., 1996). The quality evaluation of the solutions is indicated by the fitness function that assigns a score or fitness determining the minimum requirement for the potential solutions (Mitchell, 1995). During the iterations, the best subsequent population is selected and the worst excluded to avoid future reproduction (Razali, 2015). The result may be an exception of the population (as unexpected genetic mutation) or the effect of continuous and slow improvements.

The idea behind the new GA tool featured in the SZ-plugin is to classify the SI into a number of classes, by optimizing their respective boundaries. These represent the cutoffs necessary to build the ROC curve and to maximize the relative AUC. The maximization of the AUC of the segmented ROC is the fitness function of the iterative meta-heuristic algorithm. Figure 4.2 shows the classifier workflow. The ROC curve maybe weighted or none.

## 4.4 Application to landslide susceptibility

The SZ-plugin was born from the necessity of specific functions not available in QGIS with the goal of producing a landslide susceptibility map of south-Asia using WoE method (Titti et al., 2021a) in the context of the Belt and Road Initiative (Lei et al., 2018). After that, several other applications have supported the SZ-plugin development. Among these, in Titti et al. (2021c) the classifier ('Si classify' functions) of the plugin was built to reclassify the landslide susceptibility in Tajikistan. Subsequent experiments have then stimulated the development of additional functions, leading to the current version of the SZ-plugin.

Here we present the current set of functionalities offered through the SZ-plugin in the context of landslide susceptibility, although we recall that these can be used even outside the geomorphological context. The selected study area corresponds to the north-eastern sector of India, including Assam, Manipur, Meghalaya, Mizoram, Nagaland and Tripura (Fig. 4.3). This area has been selected because it appears to be one of the areas most susceptible to landslides in south Asia, according to the analysis conducted by Titti et al. (2021a). In their work, the authors' goal was to highlight regions mostly prone to landslides across countries involved into the Belt and Road Initiative. The application here presented and the analysis conducted in Tajikistan (Titti et al., 2021c) follow an analogous criterion.

All the data used in the application here presented are open-data. The landslide inventory used for the analysis was provided by the Gological Survey of India and is available at the Bhukosh website ( https://bhukosh.gsi.gov.in , accessed 15 November 2021). It is an open database developed to evaluate the spatial distribution of natural hazards in India. In the selected study area, it includes 5759 Landslide Identification Points (LIP). The catalog includes landslides triggered by: rainfall, anthropogenic activity, road/slope cut, quarrying, toe erosion and ground motion.

Slope units have been used as the reference mapping unit. A slope unit is a terrain unit derivable from a DEM, under the constraint of internal aspect homogeneity in areas defined between ridges and streamlines (Alvioli et al., 2020). The shapes have been calculated using *r.slopeunits*, a tool developed in GRASS GIS by Alvioli et al. (2016). In this case ,we parameterized *r.slopeunits* with a flow accumulation threshold of 5,000,000 $m^2$, a minimum unit area of 500,000 $m^2$ and a circular variance of 0.3. The alluvial plains have been excluded from the analysis because considered not susceptible a-priori. As a result,

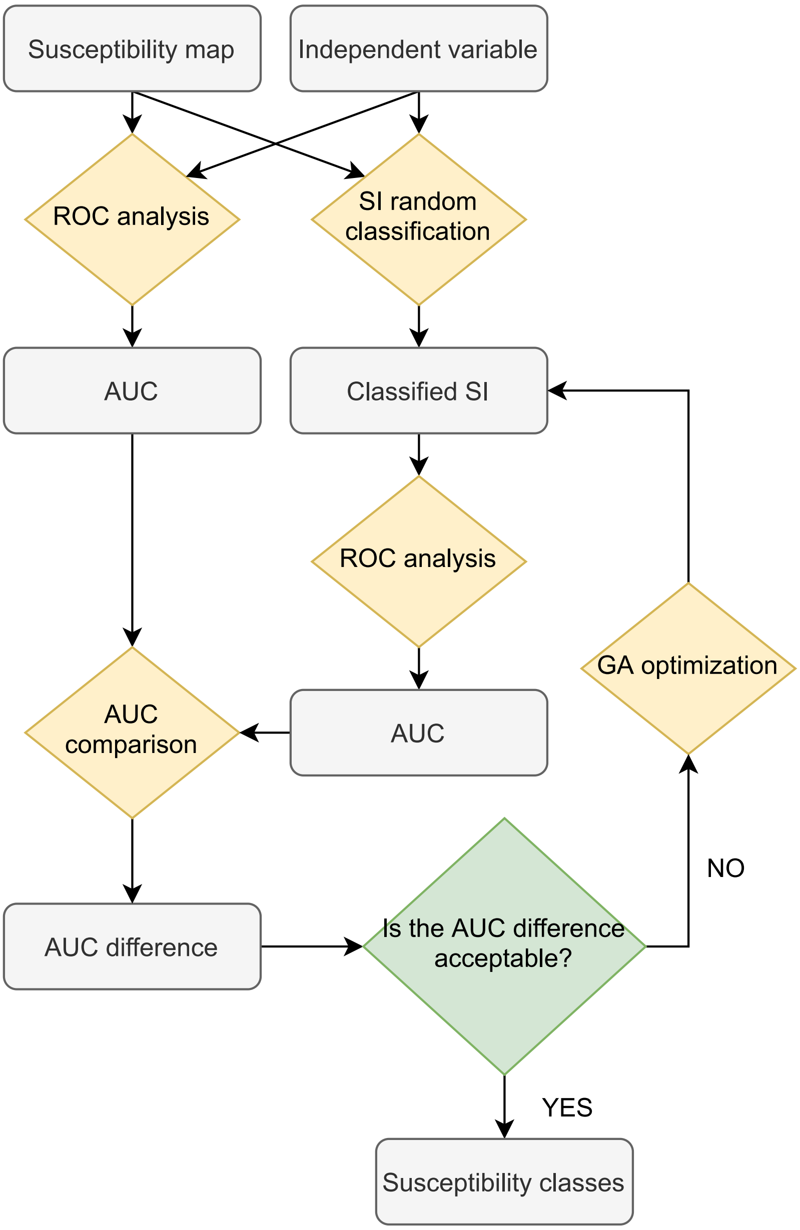


Figure 4.2: Workflow of the GA-based classifier functions.

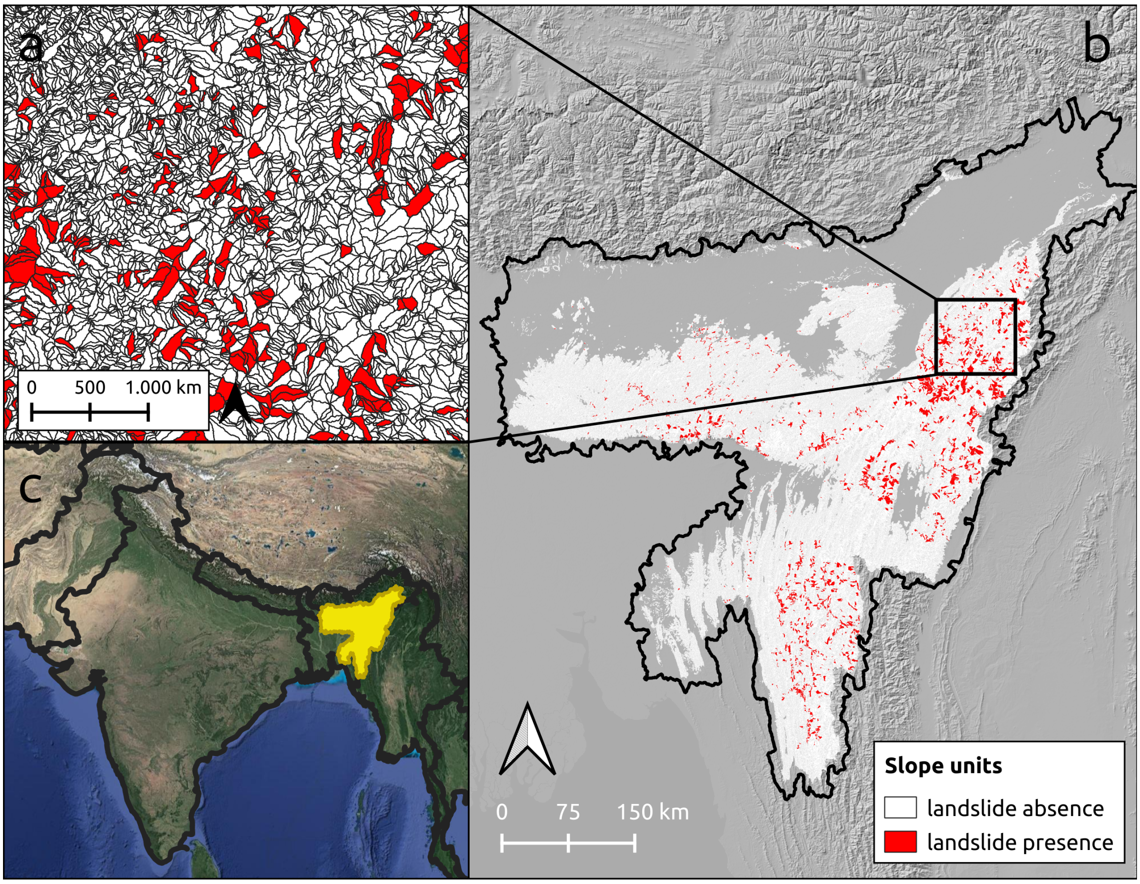


Figure 4.3: Study area and relative slope units of the SZ-plugin application to landslide susceptibility (Map data panel c ©Google). Panel b shows in white and red the slope units investigated after the exclusion of the alluvial plains. The red zones represent the unstable slope units, whereas the white zones the stable ones.

124,553 slope units were generated with a maximum surface extension of 27 $km^2$, a mean of 1 $km^2$ and a standard deviation of 1.1 $km^2$.

All the data used as predisposing factor are also open data. They consist of: lithology, land cover, slope, plan curvature (tangent to the contour line), profile curvature (tangent to the slope line), relative relief (maximum elevation range in a circular neighborhood of 1 km of radius), Peak Ground Acceleration (PGA), annual rainfall, Normalized Difference Vegetation Index (NDVI) and area of each mapping unit (see Tab. 4.1 for the respective data sources).

| | Data type | Layer | Pixels size | Data source |
|---|---|---|---|---|
| 1 | Morphology | Raster | 30x30 m | SRTM (Farr et al., 2007) |
| 2 | Geology | Vector | - | USGS |
| 3 | Land cover | Raster | 300x300 m | ESA 2010 and UCLouvain |
| 4 | PGA | Raster | 90x90 m | CHIRPS (Funk et al., 2015) |
| 5 | Precipitation | Raster | 5x5 km | IMRG (Huffman et al., 2019a) |
| 6 | NDVI | Raster | 30x30 m | Landsat 7 C1 Tier 1 |

Table 4.1: Predisposing factors and their acronyms. See more details at the end of the Section 4.4

The landslide susceptibility zoning of the study area is carried out using the methods Weight of Evidence (WoE) and Logistic Regression (LR). The Weights of Evidence technique is a bivariate statistical approach (Bonham-Carter et al., 1988; Bonham-Carter, 1989). It quantifies how prone an event occurrence is according to the proportion of presence/absence for each predisposing factor class. The WoE assigns two weights per class: W$^+$ and W$^-$. The weights represent, respectively, the positive and negative influence of the predisposing factors on a potential natural hazard. They are calculated by:

$$W^+ = \ln \frac{\frac{M_1}{M_1+M_2}}{\frac{M_3}{M_3+M_4}} \tag{4.2}$$

$$W^- = \ln \frac{\frac{M_2}{M_1+M_2}}{\frac{M_4}{M_3+M_4}} \tag{4.3}$$

$$W_f = W^+ - W^- \tag{4.4}$$

where, $M_1$ is the number of mapping units where the factor class and the event are both present; $M_2$ is the number of mapping units where the factor class is absent while the event is present; $M_3$ is the number of mapping units where the factor class is present while the event is absent; $M_4$ is the number of mapping units where the factor class and the event are both absent. The weight contrast ($W_f$) is the final weight assigned to each class factor. It evaluates the relation between the spatial distribution of the causes and the spatial distribution of the events (Dahal et al., 2008).

The local sum of the factor weight contrasts produces the Susceptibility Index (SI):

$$SI = \sum_{i=1}^{n} W_{fi} \tag{4.5}$$

The second model tested is the Logistic regression. It is an extension of the classical linear regression analysis. The latter commonly requires a continuous target variable ($y$), whose estimation is achieved as a linear combination of $n$ input covariates ($x$):

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{4.6}$$

where $\beta_0$ is the intercept and $\beta_i$ are the covariate coefficients.

However, the above scheme is not suitable to model a discrete target variable. In such cases, and specifically for a target variable that can take only two discrete values, a logistic regression represents the most common solution. Its structure takes a vector ($y$) reporting a series of zero and one. The zero conventionally conveys the absence of a given process in space, time or both, whereas one conveys the opposite case where the process is present. The model still linearly regresses, this time the odd ratio (see, Szumilas, 2010), with respect to $n$ input covariates $x$. The use of the logit link function, then transforms this quantity into a probability (Menard, 2002), commonly referred to the actual occurrence of the process of interest:

$$P(y = 1) = \frac{1}{1 + \exp(-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i))} \tag{4.7}$$

As a result, one can interpret an increase in the estimated regression coefficients as a linear increase or decrease of the probability $P(y = 1)$, determined by the sign of $\beta_i$.

The application and validation of the WoE and LR results are described in detail by the schema in Figure 4.5. The data pre-process, instead, is described in Figure 4.4. The flow-charts outline step by step how the SZ-plugin functions have been used to carry out the landslide susceptibility map of the north-east of India.

As described in Figure 4.4, the average per mapping unit of slope, plan curvature, profile curvature, relative relief, annual precipitation and NDVI have been calculated using a script implemented by us in Google Earth Engine (GEE) (Gorelick et al., 2017) called Spatial Reduction Tool (SRT). Using the TAGEE package for terrain analysis developed by Safanelli et al. (2020), the SRT allows to calculate, terrain variables that are DEM derived such as: slope, elevation, aspect, northness, eastness, mean curvature, gaussian curvature, minimal curvature, maximal curvature, shape index, horizontal curvature and vertical curvature. Moreover, the SRT allows to calculate the relative relief and collect data from various databases such as: precipitation, temperature and NDVI. Finally, the SRT spatially reduces the pixel based variables into their mean and standard deviation per selected mapping unit. Notably, any shape can be used as a reference mapping unit.

As regards other predictors, the majority of the continuous variable (PGA) and the majority of the categorical variables (lithology and land cover), per mapping units, have been calculated using QGIS basic functions and aggregated in one vector layer.

The next steps are described in Figure 4.5. To verify the quality of the landslide inventory and to avoid mistakes related to the landslides survey, the landslide catalogue has been filtered using the function '01 Clean points by raster kernel value' assuming a minimum slope degree of 8° in a neighborhood of 500 m of radius. One landslide only has been deleted. Then, the resulting inventory attributes has been investigated using the functions: '02 Attribute table statistics' i.e. to know the landslide triggers, then '03

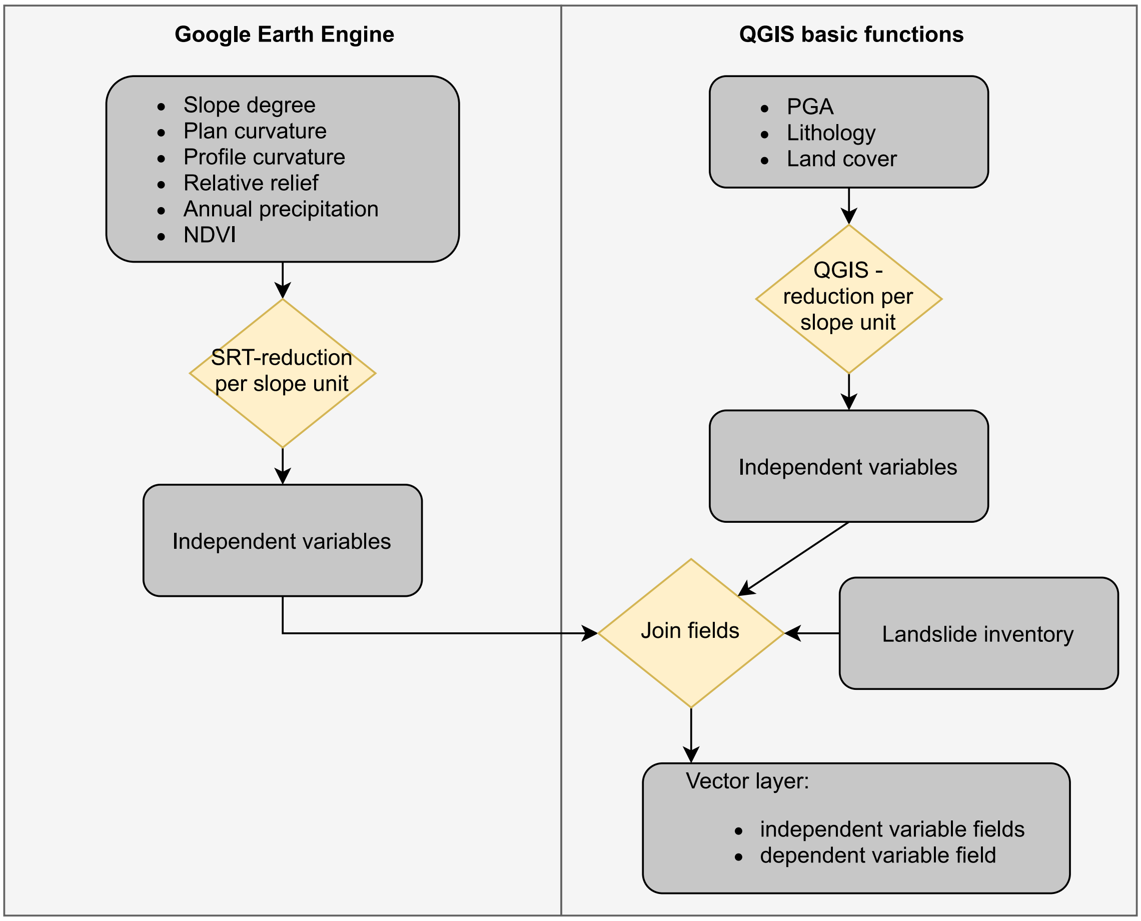


Figure 4.4: Flow-chart of the data collection and pre-processing using the SRT (GEE) and QGIS.

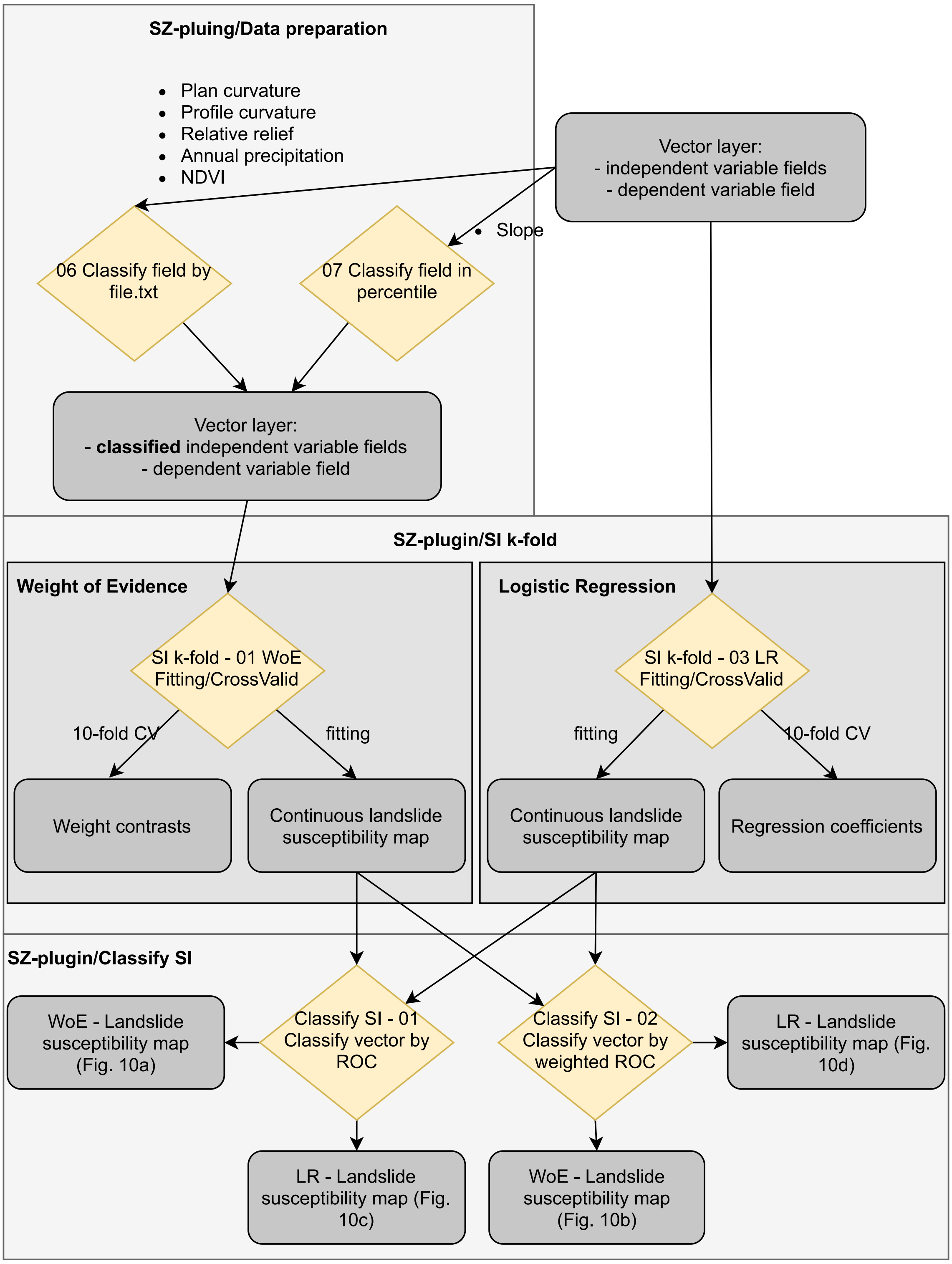

SZ-pluing/Data preparation
• Plan curvature
• Profile curvature
• Relative relief
• Annual precipitation
• NDVI
Vector layer:
- independent variable fields
- dependent variable field
• Slope
06 Classify field by file.txt
07 Classify field in percentile
Vector layer:
- classified independent variable fields
- dependent variable field
SZ-plugin/SI k-fold
Weight of Evidence
SI k-fold - 01 WoE Fitting/CrossValid
10-fold CV
fitting
Weight contrasts
Continuous landslide susceptibility map
Logistic Regression
SI k-fold - 03 LR Fitting/CrossValid
fitting
10-fold CV
Continuous landslide susceptibility map
Regression coefficients
SZ-plugin/Classify SI
WoE - Landslide susceptibility map (Fig. 10a)
Classify SI - 01 Classify vector by ROC
Classify SI - 02 Classify vector by weighted ROC
LR - Landslide susceptibility map (Fig. 10d)
LR - Landslide susceptibility map (Fig. 10c)
WoE - Landslide susceptibility map (Fig. 10b)

Figure 4.5: Step by step flow-chart of the functions used form the SZ-plugin.

Points kernel statistics' and '04 Points kernel graphs' to plot the frequency distribution of the effective, maximum, minimum, standard deviation, sum, average and maximum range value for a specific thematic map of the LIPs neighborhood. To provide an example of the functions '03 Points kernel statistics' and '04 Points kernel graphs', Figure 4.6a shows the cumulative effective PGA value across LIPs, whereas Figure 4.6b depicts the cumulative average PGA value in a LIP around of 5x5 km.

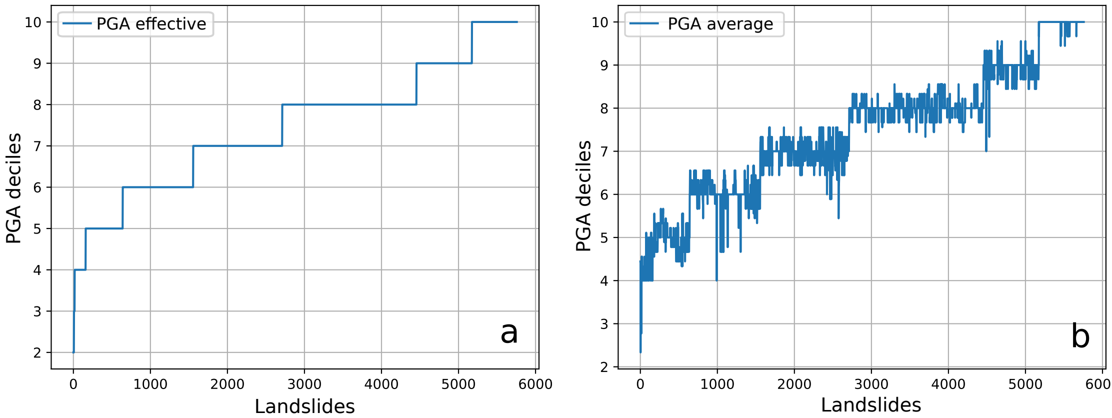


Figure 4.6: Results of the functions '03 Points kernel statistics' and '04 Points kernel graphs'. a) shows the cumulative effective LIP PGA value; b) cumulative average PGA value in a LIP around of 5x5 km.

The landslide susceptibility zoning of the study area is carried out using the methods WoE and LR. To apply the WoE, the variables need to be classified. This step is done by using the functions '06 Classify field by file.txt' and '07 Classify field in percentile'. The slope has been classified in eight classes: 0°, 1–3°, 4–6°, 7–10°, 11–15°, 16–20°, 21–30°, >30°, whereas the other continuous factors, in deciles.

After that, the susceptibility can be modeled with WoE. At the beginning the reference map has been built fitting the entire dataset with an AUC equal to 0.79, then the model has been 10-fold cross-validated. The relative ROC curves are visible in Figure 4.7a with an AUC mean equal to 0.8 and a variability measured with a standard deviation equal to 0.07. Both the runs were tested by using 'SI k-fold: 01 WoE Fitting/CrossValid', selecting different input parameters. The weight contrasts of the slope, relative relief, area and annual precipitation resulting from the cross-validation are reported in Figure 4.8.

To apply the LR method, the categorical covariates have been processed to generate one variable per category, measuring the percentage of surface covered by each category with respect to the extent of each slope units. Then, to avoid multi-collinearity issues (the sum of all percentage classes always returns 100%, thus being a linear combination by definition), the most representative class across all slope units has been removed. This operation should sufficiently perturb the dependence structure among the remaining classes, ensuring that a linear combination of their respective values would not yield the same result.

Finally, the susceptibility can be modeled, first by fitting the complete dataset (AUC=0.81) and then through a 10-fold cross-validation (AUC mean equal to 0.81 and standard deviation equal to 0.06). Both the analysis have been carried out via the 'SI k-fold: 03 LR Fitting/CrossValid'. The ROC curves of the cross-validation are shown in Figure 4.7b whereas the regression coefficients are reported in Figure 4.9.

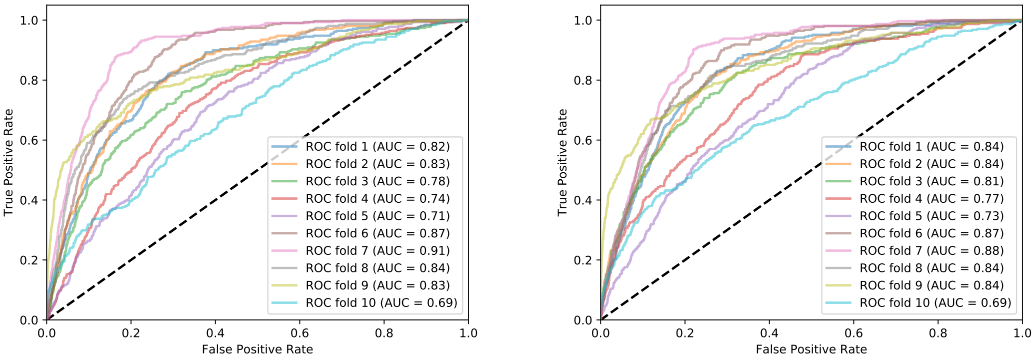


Figure 4.7: ROC curves of the 10-fold cross-validation. On the right the ROC of the WoE cross-validation, on the left the ROC of the LR cross-validation

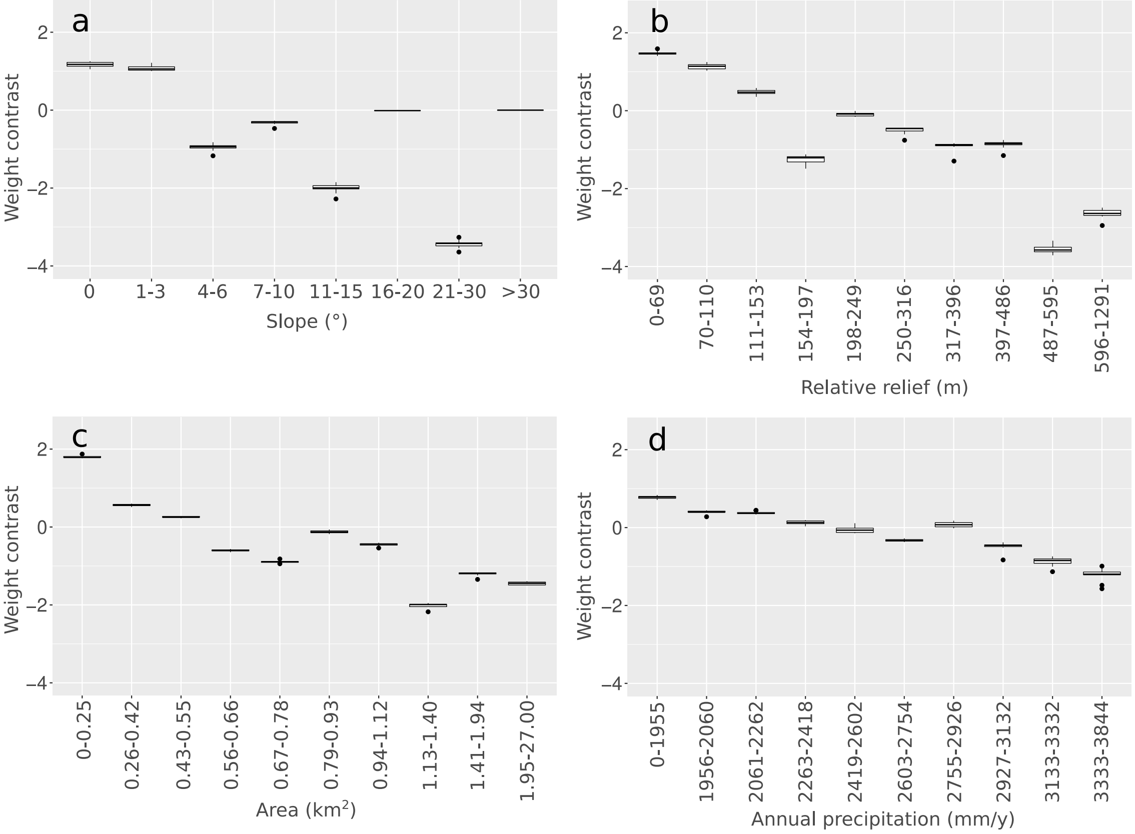


Figure 4.8: Boxplots from the WoE 10-fold cross-validation weight contrast of: a) slope, b) relative relief, c) area, d) annual precipitation.

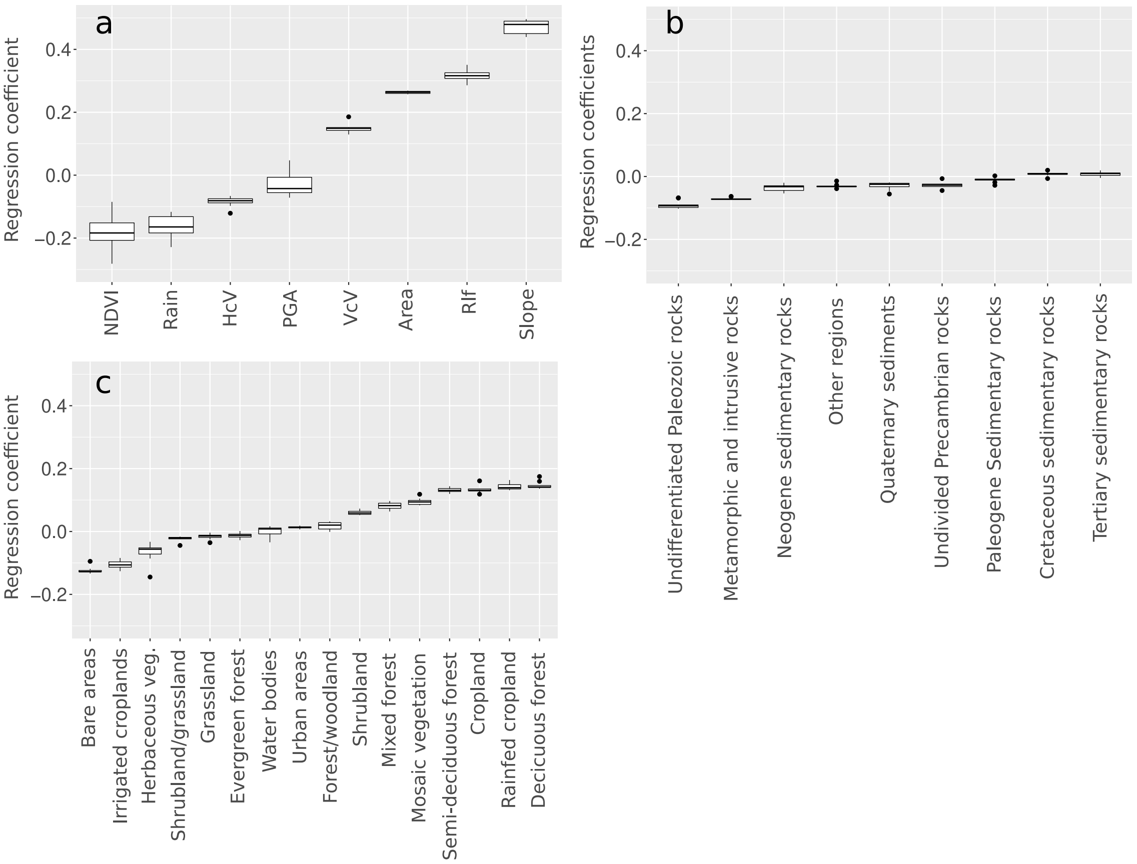


Figure 4.9: Boxplots from the LR 10-fold cross-validation regression coefficients of: a) continuous covariates, b) lithology, c) land cover.

To make the maps suitable for stakeholders and end users involved in land planning, the final maps have been classified into 4 classes by using the '01 Classify vector by ROC' (Fig. 4.11a, Fig. 4.11c) and '02 Classify vector by weighted ROC' functions (Fig. 4.11b, Fig. 4.11d). The latter classification has been weighted according to the slope units' area.

Figure 4.10 shows the result of the function 'Classify SI/03 True/False' which maps the distribution of the True Positive, True Negative, False Positive, False Negative mapping units respect to a cutoff value equal to the median of the LR susceptibility index.

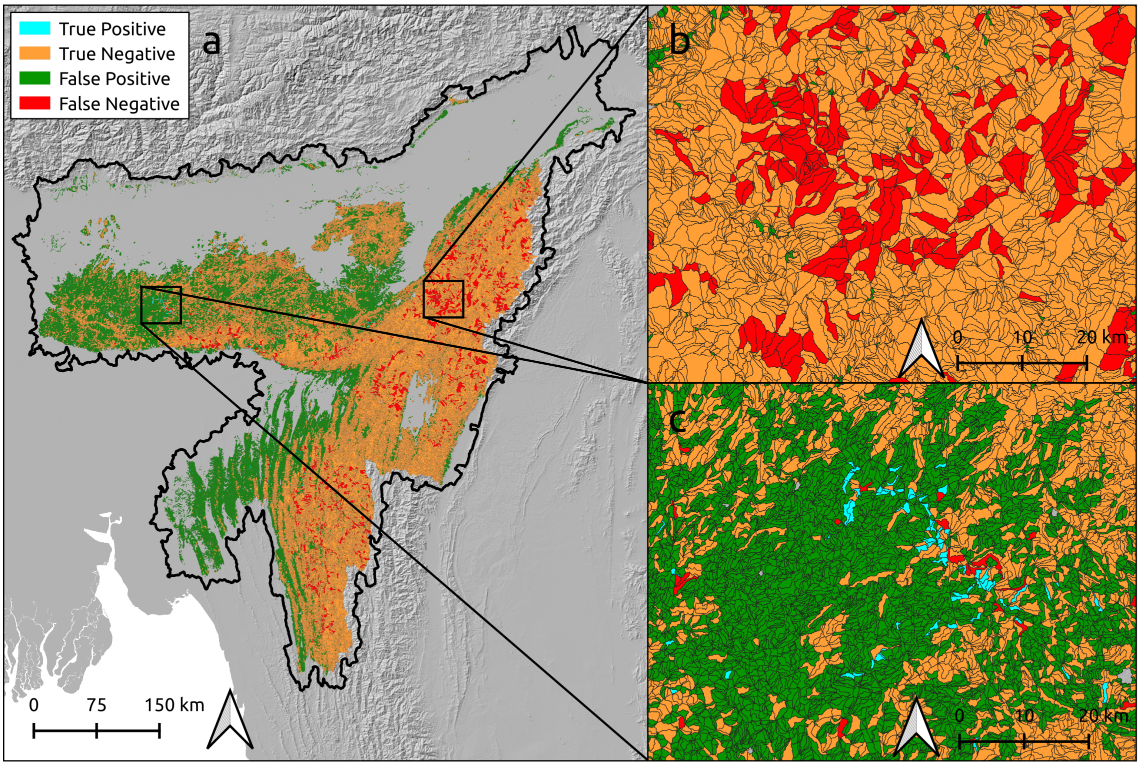


Figure 4.10: True Positive, True Negative, False Positive, False Negative mapping units respect to a cutoff value equal to the median of the LR susceptibility index.

In figure 4.11, it is evident that the susceptibility patterns produced by WoE and LR are quite similar. Differences become more visible when using the weighted and non-weighted classifiers. In both models (WoE and LR), the 'very high' class covers a larger area in the none weighted-classified map than in the weighted ones. The 'very high' class in the WoE non weighted-classified map covers the 31% of the total surface mapped (130,736 $km^2$) and the 14% in the weighted ones. The LR classified maps have given a similar response to the classification: 31% and 12% of the total surface mapped for the non weighted-classified map and weighted-classified map, respectively.

We report additional references about the predisposing factors and landslide inventory: lithology (https://catalog.data.gov accessed 2021-08-23), land cover (http://due.esrin.esa.int , accessed 2021-04-15), PGA (https://sedac.ciesin.columbia.edu , accessed 2021-04-15), annual precipitation from Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS), NDVI from Landsat 7 Collection 1 Tier 1 composites for 32-day period provided by U.S. Geological Survey, landslide inventory (https://bhukosh.gsi.gov.in , accessed 2021-11-15).

The effective reduced data and the outputs of the application described above are

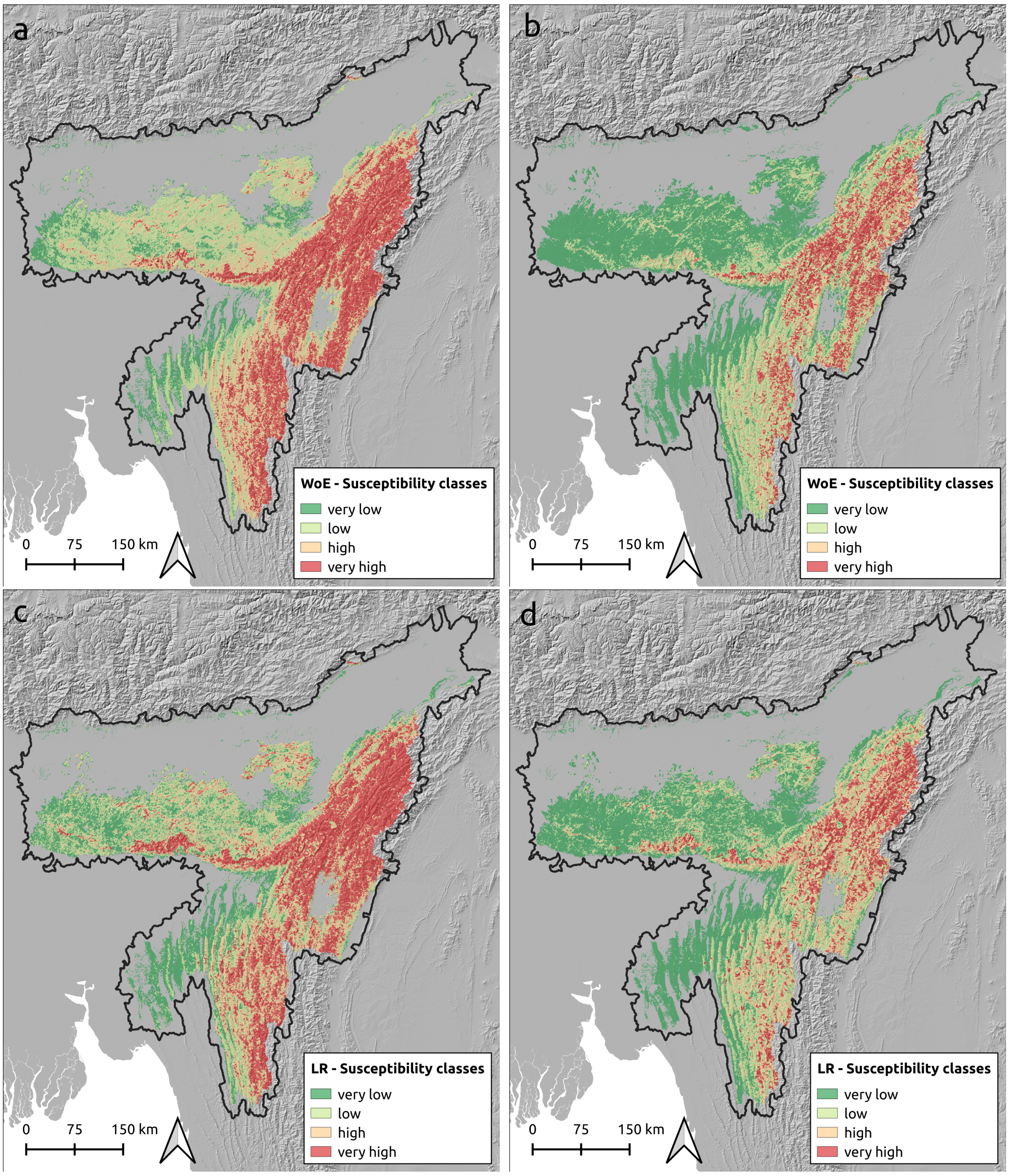


Figure 4.11: Landslide susceptibility maps classified using the WoE and the LR model. a) the classifier has maximized the AUC of the ROC curve derived from the WoE map; b) the classifier has maximized the AUC of the area-weighted ROC curve derived from the WoE map. c) the classifier has maximized the AUC of the ROC curve derived from the LR map; d) the classifier has maximized the AUC of the area-weighted ROC curve derived from the LR map.

downloadable from the GitHub repository CNR-IRPI-Padova/SZ.

## 4.5 Conclusions

A new plugin for landslide susceptibility zoning is introduced here. The SZ-plugin is a collection of processing scripts in Python language which runs as part of the QGIS platform. This framework has been chosen to maximize the accessibility to our plugin. QGIS is the most widely used open-source GIS environment.

The implemented functions have been organised into four groups: 'Data preparation', 'SI' (Susceptibility index), 'SI k-fold', 'Classify SI' which allow to carry out a complete susceptibility analysis from the data pre-processing, the prediction analysis, to the reclassification of the susceptibility index. The susceptibility can be generated using 6 different classifiers: Weight of Evidence, Frequency Ratio, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression. The statistical models may be applied to fit the input dataset or cross-validate the final map. In particular the latter can be done by a simple random split in test/train samples or by a k-fold method. Both are completed by a Receiving Operating Characteristic (ROC) analysis for performance assessment. Finally, a GA-based (Genetic Algorithms) classifier has been implemented to classify the susceptibility index.

Many libraries in R and other libraries such as Scikit-learn in Python are probably the best solutions to perform susceptibility with statistical models, but they require a good knowledge in coding. Overall, the tool proposed can simplify the access to statistical assessments of susceptibility for users who are not familiar in coding or for those who wish to achieve results rapidly.

In this paper the version 1.0 of the SZ-plugin has been presented to the geoscientific community. Several improvements and new functionalities are under development which will be uploaded with the further versions of the plugin.

# 5. Conclusions

This text presents a multi-scale approach to the landslide susceptibility in the countries involved in the Silk Road Economic Belt project. A first landslide susceptibility map at continental scale has been produced in south-Asia. The evaluation has been conducted with a pixel-based approach using 8 predisposing factors within a Weight of Evidence model. This method allows us to make a quantitative, and thus objective, zonation through the use of a simple, and easy to replicate, model. The results of the validation reveals a good prediction capacity of the model with an AUC of validation equal to 0.9. The resulting landslide susceptibility map of south-Asia shows the most susceptible administrative units crossed by the new silk road which have then been analysed in detail at national scale. Figure 5.1 shows the areas which are moderately, highly and very highly susceptible to landslides. In our opinion those regions require a second investigation at national scale. In total, they cover the 28% of China, the 17% of Pakistan, the 28% of India, the 83% of Tajikistan, the 15% of Bangladesh, the 92% of Nepal, the 18% of Afghanistan, the 98% of Bhutan, the 84% of Myanmar, the 35% of Cambodia, the 71% of Kyrgyzstan, the 94% of Laos, the 61% of Thailand and the 81% of Viet Nam. In this study, we focused our efforts at national scale in the areas of west Tajikistan and North-east India.

The detailed dataset available in Tajikistan allows us to compile a good-quality explanatory susceptibility map based on 9 predisposing factors spatially reduced by Unique Condition Units. The analysis has been processed using a Generalized Additive Model. The landslide susceptibility map proposed in Figure 3.6c can support the design of infrastructures at national scale indicating which area is affected by high and very high probability of landslide occurrence and thus which requires solutions to avoid or mitigate the effects induced by probable landslides. For sure, this type of maps can support the decision process, even though, the final choice is of the local authorities. Moreover, unfortunately, we cannot indicate which section of the Silk Road Economic Belt infrastructures in Tajikistan may require mitigation measures because the executive design of the path is not available.

In addition, to investigate the possibility to replicate the analysis in the rest of the high-susceptible countries which are characterized by data scarcity we repeated the evaluation of the landslide susceptibility in 2000 different scenarios by varying the number of the landslide available from the original dataset. The result shows that a dataset of at least 220 unstable UCUs, on average, which corresponds to the 32% of the total unstable UCUs (1.5% of all UCUs), could be enough to produce a reliable landslide susceptibility map of west Tajikistan. The relative valiation AUC mean of the scenarios with 220 unstable UCUs is equal to 0.85.

Despite the fact that the result is extremely site-specific and mapping unit-specific the protocol proposed reveals a new perspective which would be investigated in further analysis.

In the end of the PhD project, thousands rows of code have been written. Most of

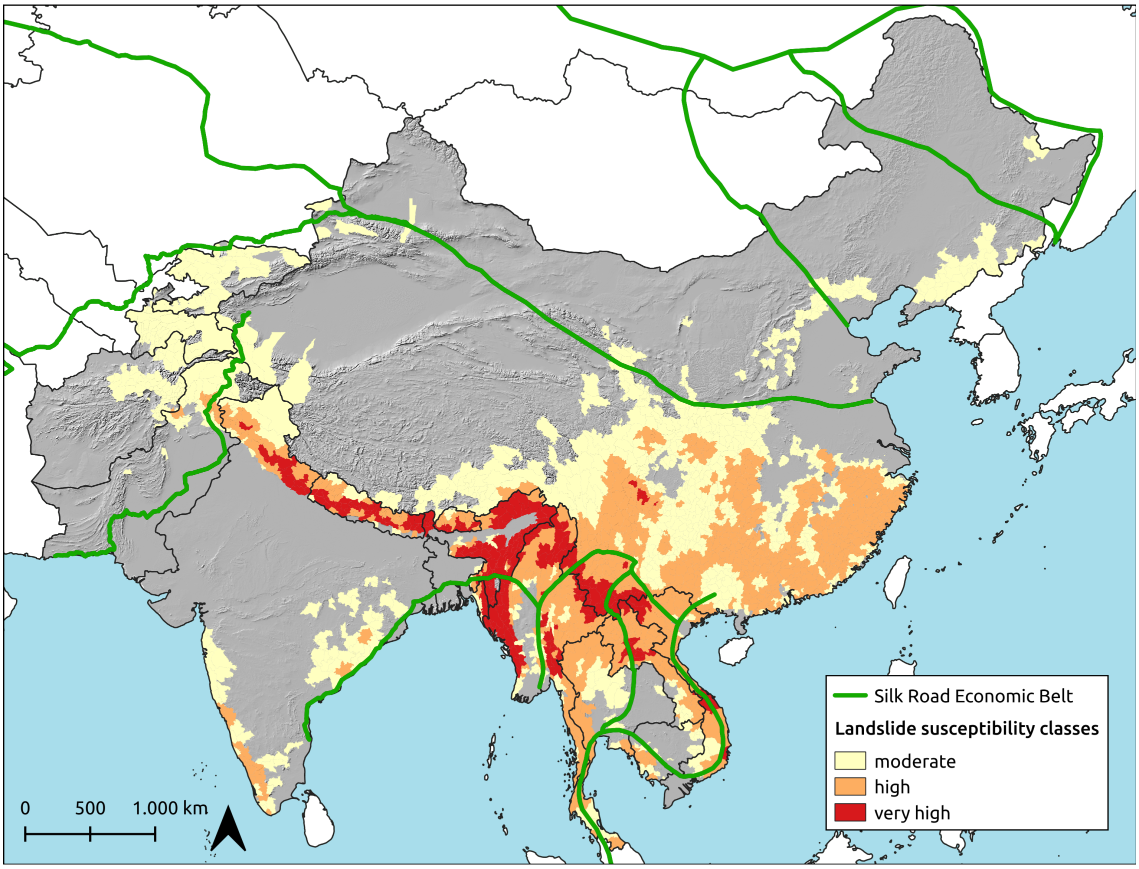


Figure 5.1: From moderate to very high susceptible administrative units of south-Asia.

these have been collected in two open-source tools: the SZ-plugin and the SRT. We tested both in the study area of North-east India to carry out a complete landslide susceptibility assessment from data collection to interpretation. The results are two slope units-based susceptibility maps modeled using the Logistic Regression and Weight of Evidence models. The analysis confirms the user-friendly properties of the published tools and their effective capacity to support the user during the result interpretation.

The results of the last three years of research have been summarized in three open access articles published by international journals. All the data used are open and the tools used are open-source. Moreover, all the data and tools produced are freely available in open and open-source forms.

In conclusion, in this research we faced a very challenging project with the practical goal to support the development of infrastructures through the prevention of damages induced by some geological hazards. The dimension of the study area and the resulting induced obstacles make difficult the evaluation of a detailed landslide susceptibility map of the entire study area, in particular in a limited time of three years. Therefore, we focused all our resources to promote a way to approach the problem and to provide tools in order to make reproducible the solutions carried out in the tested study area of south-Asia. Therefore we encourage the use of the tools provided to analyse the regions where a landslide susceptibility map is necessary.

# References

Acharya, G., Cochrane, T., Davies, T., and Bowman, E. (2009). The influence of shallow landslides on sediment supply: a flume-based investigation using sandy soil. *Engineering Geology*, 109(3-4):161–169.

Agterberg, F. P., Bonham-Carter, G. F., and Wright, D. F. (1989). Weights of evidence modelling: a new approach to mapping mineral potential. In Agterberg, F. P. and Bonham-Carter, G. F., editors, *Statistical applications in the earth sciences*. Geological Survey Canada Paper, Windsor.

Alvioli, M., Guzzetti, F., and Marchesini, I. (2020). Parameter-free delineation of slope units and terrain subdivision of italy. *GEOMORPHOLOGY*, 358.

Alvioli, M., Marchesini, I., Reichenbach, P., Rossi, M., Ardizzone, F., Fiorucci, F., and Guzzetti, F. (2016). Automatic delineation of geomorphological slope units with r.slopeunits v1.0 and their optimization for landslide susceptibility modeling. *Geoscientific Model Development*, 9(11):3975–3991.

Amato, G., Eisank, C., Castro-Camilo, D., and Lombardo, L. (2019). Accounting for covariate distributions in slope-unit-based landslide susceptibility models. a case study in the alpine environment. *Engineering Geology*, 260:In print.

Arabameri, A., Chen, W., Loche, M., Zhao, X., Li, Y., Lombardo, L., Cerda, A., Pradhan, B., and Bui, D. T. (2020). Comparison of machine learning models for gully erosion susceptibility mapping. *Geoscience Frontiers*, 11(5):1609–1620.

Arabameri, A., Pradhan, B., and Lombardo, L. (2019). Comparative assessment using boosted regression trees, binary logistic regression, frequency ratio and numerical risk factor for gully erosion susceptibility modelling. *Catena*, 183:104223.

Arup (2020). The global landslide hazard MapFinal project report. the world bank. https://development-data-hub-s3-public.s3.amazonaws.com/ddhfiles/1191621/global-landslide-hazard-map-report.pdf. accessed 15 apr 2021.

Axing, Z. H. U., Tao, P. E. I., Ping, Q. J., Yongbo, C., Chenghu, Z., Qiangguo, C. A. I., Axing, Z. H. U., Tao, P. E. I., Ping, Q. J., Yongbo, C., Chenghu, Z., and Qiangguo, C. A. I. (2010). A landslide susceptibility mapping approach using expert knowledge and fuzzy logic under GIS, a landslide susceptibility mapping approach using expert knowledge and fuzzy logic under GIS. *Progress Geogr*, 25.

Ayalew, L., Yamagishi, H., and Ugawa, N. (2004). Landslide susceptibility mapping using GIS-based weighted linear combination, the case in tsugawa area of agano river, niigata prefecture, japan. *Landslides*, 1.

Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., and Lindgren, F. (2018). Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6):e1443.

Barredo, J. I., Benavides, A., Hervás, J., and Westen, C. J. (2000). Comparing heuristic landslide hazard assessment techniques using GIS in the tirajana basin, gran canaria island, spain. *Int J Appl Earth Obs Geoinf*, 2.

Bonham-Carter, G. F. (1989). Weights of evidence modeling: a new approach to mapping mineral potential. *Statistical applications in the earth sciences*, pages 171–183.

Bonham-Carter, G. F., Agterberg, F. P., and Wright, D. F. (1988). Integration of geological datasets for gold exploration in nova scotia. *Photogrammetric Engineering and Remote Sensing*, 54(11):1585–1592.

Bontemps, S., Defourny, P., Van Bogaert, E., Arino, O., Kalogirou, V., and Perez, J. (2011). GLOBCOVER 2009, product description and validation report.

Brabb, E. E. (1985). Innovative approaches to landslide hazard and risk mapping. In *International Landslide Symposium Proceedings, Toronto, Canada*, volume 1, pages 17–22.

Brenning, A. (2005). Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Science*, 5(6):853–862.

Broeckx, J., Vanmaercke, M., and Duchateau, R. (2018). A data-based landslide susceptibility map of africa. *Earth Sci Rev*, 185.

Bühlmann, E., Wolfgramm, B., Maselli, D., Hurni, H., Sanginov, S., and Liniger, H. (2010). Geographic information system–based decision support for soil conservation planning in Tajikistan. *journal of soil and water conservation*, 65(3):151–159.

Bui, D. T., Tuan, T. A., Klempe, H., Pradhan, B., and Revhaug, I. (2016). Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 13.

Carrara, A., Cardinali, M., Guzzetti, F., and Reichenbach, P. (1995). GIS technology in mapping landslide hazard. In *Geographical information systems in assessing natural hazards*, pages 135–175. Springer.

Carrara, A., Crosta, G., and Frattini, P. (2008). Comparing models of debris-flow susceptibility in the alpine environment. *Geomorphol GIS Technol Models Assess Landslide Hazard Risk*, 94.

Catani, F., Lagomarsino, D., Segoni, S., and Tofani, V. (2013). Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Natural Hazards and Earth System Sciences*, 13(11):2815–2831.

Cawsey, D. C. and Mellon, P. (1983). A review of experimental weathering of basic igneous rocks. *Geological Society, London, Special Publications*, 11(1):19–24.

Chacón, J., Irigaray, C., Fernández, T., and Hamdouni, R. E. (2006). Engineering geology maps: landslides and geographical information systems. *Bull Eng Geol Environ*, 65.

Chatterjee, S., Carrera, C., and Lynch, L. A. (1996). Genetic algorithms and traveling salesman problems. *European journal of operational research*, 93(3):490–510.

Chen, W., Chai, H., Sun, X., Wang, Q., Ding, X., and Hong, H. (2016). A GIS-based comparative study of frequency ratio, statistical index and weights-of-evidence models in landslide susceptibility mapping. *Arab J Geosci*, 9.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., Duan, Z., and Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *CATENA*, 151.

Chiessi, V., Toti, S., and Vitale, V. (2016). Landslide Susceptibility Assessment Using Conditional Analysis and Rare Events Logistics Regression: A Case-Study in the Antrodoco Area (Rieti, Italy). *Journal of Geoscience and Environment Protection*, 4(12):1–21. Number: 12 Publisher: Scientific Research Publishing.

CHRR and CIESIN (2005). Global earthquake hazard distribution, peak ground acceleration. nasa socioeconomic data and applications center (sedac). Technical report, Center for Hazards and Risk Research, Columbia University, and Center for International Earth Science Information Network, Columbia University, Palisades, NY.

Chung, C.-J. F. and Fabbri, A. G. (2003). Validation of Spatial Prediction Models for Landslide Hazard Mapping. *Natural Hazards*, 30(3):451–472.

Ciurleo, M., Cascini, L., and Calvello, M. (2017). A comparison of statistical and deterministic methods for shallow landslide susceptibility zoning in clayey soils. *Engineering Geology*, 223:71–81.

Conoscenti, C., Rotigliano, E., Cama, M., Caraballo-Arias, N. A., Lombardo, L., and Agnesi, V. (2016). Exploring the effect of absence selection on landslide susceptibility models: A case study in Sicily, Italy. *Geomorphology*, 261:222–235.

Constantin, M., Bednarik, M., Jurchescu, M. C., and Vlaicu, M. (2011). Landslide susceptibility assessment using the bivariate statistical analysis and the index of entropy in the sibiciu basin (romania). *Environ Earth Sci*, 63.

Corominas, J., van Westen, C., Frattini, P., Cascini, L., Malet, J.-P., Fotopoulou, S., Catani, F., Van Den Eeckhaut, M., Mavrouli, O., Agliardi, F., et al. (2014). Recommendations for the quantitative analysis of landslide risk. *Bulletin of engineering geology and the environment*, 73(2):209–263.

Cui, P., Amar, D. R., Zou, Q., Lei, Y., Chen, X., and Cheng, D. (2017). Natural hazards and disaster risk in one belt one road corridors. In *Advancing culture of living with landslides-volume 2 advances in landslide science*. Springer, Berlin.

Dahal, R., Hasegawa, S., and Nonomura, A. (2008). Gis-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping. *Environmental geology*, 54:311–324.

Dang, V.-H., Dieu, T. B., Tran, X.-L., and Hoang, N.-D. (2019). Enhancing the accuracy of rainfall-induced landslide prediction along mountain roads with a GIS-based random forest classifier. *Bulletin of Engineering Geology and the Environment*, 78(4):2835–2849.

Depicker, A., Govers, G., Jacobs, L., Campforts, B., Uwihirwe, J., and Dewitte, O. (2021). Interactions between deforestation, landscape rejuvenation, and shallow landslides in the North Tanganyika–Kivu rift region, Africa. *Earth Surface Dynamics*, 9(3):445–462.

Dewitte, O., Dille, A., Depicker, A., Kubwimana, D., Mateso, J.-C. M., Bibentyo, T. M., Uwihirwe, J., and Monsieurs, E. (2021). Constraining landslide timing in a data-scarce context: from recent to very old processes in the tropical environment of the North Tanganyika-Kivu Rift region. *Landslides*, 18(1):161–177.

Di Gregorio, A. (2016). *Land cover classification system: classification concepts.* FAO, Rome.

Dilley, M., Chen, R. S., Deichmann, U., Lerner-Lam, A. L., and Arnold, M. (2005). *Natural disaster hotspots: a global risk analysis.* World Bank, Washington, D.C.

Dávid, G. (2021). frmod, frequency ratio modeller. Accessed: 2021-10-15, `https://github.com/gerzsd/frmod`.

EC (2006). Thematic strategy for soil protection. COM(2006)231 final.

Eckelmann, W., Baritz, R., Bialousz, S., Bielek, P., Carre, F., Houšková, B., Jones, R., Kibblewhite, M., Kozak, J., Bas, C., Tóth, G., Tóth, T., Várallyay, G., Yli, H., and Zupan, M. (2006). Common criteria for risk area identification according to soil threats. european soil bureau research report no. 20, EUR 22185 EN. office for official publications of the european communities, luxembourg.

Eeckhaut, M., Reichenbach, P., Guzzetti, F., Rossi, M., and Poesen, J. (2009). Combined landslide inventory and susceptibility assessment based on different mapping units: an example from the Flemish Ardennes, Belgium. *Natural Hazards and Earth System Sciences*, 9(2):507–521.

Eeckhaut, M. V. D., Hervás, J., Jaedicke, C., Malet, J. P., Montanarella, L., and Nadim, F. (2012). Statistical modelling of europe-wide landslide susceptibility using limited landslide inventory data. *Landslides*, 9.

Ermini, L., Catani, F., and Casagli, N. (2005). Artificial neural networks applied to landslide susceptibility assessment. *Geomorphology*, 66(1):327–343. Geomorphological hazard and human impact in mountain environments.

Evans, S. G., Roberts, N. J., Ischuk, A., Delaney, K. B., Morozova, G. S., and Tutubalina, O. (2009). Landslides triggered by the 1949 khait earthquake, tajikistan, and associated loss of life. *Engineering Geology*, 109(3):195–212.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D. (2007). The Shuttle Radar Topography Mission. *Reviews of Geophysics*, 45(2). _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2005RG000183.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874. ROC Analysis in Pattern Recognition.

Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., Savage, W. Z., et al. (2008). Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. *Engineering Geology*, 102(3-4):99–111.

Florinsky, V., Skrypitsyna, T. N., Trevisani, S., and Romaikin, S. V. (2019). Statistical and visual quality assessment of nearly-global and continental digital elevation models of trentino, italy. *Remote Sens Lett*, 10.

Frattini, P., Crosta, G., and Carrara, A. (2010). Techniques for evaluating the performance of landslide susceptibility models. *Engineering Geology*, 111(1):62–72.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., et al. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific data*, 2(1):1–21.

Glade, T. and Crozier, M. J. (2012). *A review of scale dependency in landslide hazard and risk analysis. Landslide hazard and risk.* Wiley Online Books, New Jersey.

Goetz, J., Brenning, A., Petschko, H., and Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & geosciences*, 81:1–11.

Goetz, J. N., Guthrie, R. H., and Brenning, A. (2011). Integrating physical and empirical landslide susceptibility models using generalized additive models. *Geomorphology*, 129(3-4):376–386.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

Gorsevski, P., Gessler, P., and Foltz, B. (2000). Spatial prediction of landslide hazard using discriminant analysis and GIS.

Gorsevski, P. V., Gessler, P. E., Boll, J., Elliot, W. J., and Foltz, R. B. (2006). Spatially and temporally distributed modeling of landslide susceptibility. *Geomorphology*, 80.

Guha-Sapir, D., Below, R., and Hoyois, P. (2016). The cred/ofda international disaster database. *Nature*, 2:250–261.

Günther, A., Eeckhaut, M. V. D., Reichenbach, P., Hervás, J., Malet, J. P., Foster, C., and Guzzetti, F. (2013a). New developments in harmonized landslide susceptibility mapping over europe in the framework of the european soil thematic strategy. In *Landslide science and practice*. springer, Berlin.

Günther, A., Hervás, J., Eeckhaut, M. V. D., Malet, J. P., and Reichenbach, P. (2014). Synoptic pan-european landslide susceptibility assessment: the ELSUS 1000 v1 map. In *Landslide science for a safer geoenvironment*. Springer, Cham.

Günther, A., Reichenbach, P., Guzzetti, F., and Richter, A. (2007). Criteria for the identification of landslide risk areas in europe: the tier 1 approach. In *Guidelines for mapping areas at risk of landslides in Europe*. Institute for Environment and Sustainability Joint Research Centre, Rome.

Günther, A., Reichenbach, P., Malet, J. P., Eeckhaut, M. V. D., Hervás, J., Dashwood, C., and Guzzetti, F. (2013b). Tier-based approaches for landslide susceptibility assessment in europe. *Landslides*, 10.

Guzzetti, F., Mondini, A. C., Cardinali, M., Fiorucci, F., Santangelo, M., and Chang, K.-T. (2012). Landslide inventory maps: New tools for an old problem. *Earth-Science Reviews*, 112(1-2):42–66.

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., and Galli, M. (2006). Estimating the quality of landslide susceptibility models. *Geomorphology*, 81(1–2):166–184.

Hanisch, J. (2018). Usoi landslide dam in Tajikistan–the world's highest dam. First stability assessment of the rock slopes at Lake Sarez. In *Landslides*, pages 189–192. Routledge.

Hartmann, J. and Moosdorf, N. (2012). The new global lithological map database GLiM: a representation of rock properties at the earth surface. *Geochem Geophys Geosyst.*

Havenith, H., Strom, A., Torgoev, I., Torgoev, A., Lamair, L., Ischuk, A., and Abdrakhmatov, K. (2015). Tien shan geohazards database: Earthquakes and landslides. *Geomorphology*, 249:16–31. Geohazard Databases: Concepts, Development, Applications.

Havenith, H.-B., Strom, A., Jongmans, D., Abdrakhmatov, A., Delvaux, D., and Tréfois, P. (2003). Seismic triggering of landslides, Part A: Field evidence from the Northern Tien Shan. *Natural Hazards and Earth System Sciences*, 3(1/2):135–149.

Heckmann, T., Gegg, K., Gegg, A., and Becht, M. (2014). Sample size matters: investigating the effect of sample size on a logistic regression susceptibility model for debris flows. *Natural Hazards and Earth System Sciences*, 14(2):259–278.

Hervás, J. (2007). Guidelines for mapping areas at risk of landslides in europe. *Inst Environ Sustain Jt Res Centre.*

Hong, Y., Adler, R., and Huffman, G. (2007). Use of satellite remote sensing data in the mapping of global landslide susceptibility. *Nat Hazards*, 43.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression.* Wiley, New York, 2nd ed edition.

Huffman, G., Stocker, E., Bolvin, D., Nelkin, E., and Jackson, T. (2019a). Gpm imerg final precipitation l3 1 day 0.1 degree x 0.1 degree v06, edited by andrey savtchenko. Technical report, Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD.

Huffman, G., Stocker, E., Bolvin, D., Nelkin, E., and Jackson, T. (2019b). Gpm imerg final precipitation l3 1 day 0.1 degree x 0.1 degree v06, edited by andrey savtchenko, greenbelt, md, goddard earth sciences data and information services center (ges disc). Accessed: 2021-04-15, 10.5067/GPM/IMERGDF/DAY/06.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Hussin, H. Y., Zumpano, V., Reichenbach, P., Sterlacchini, S., Micu, M., van Westen, C., and Bălteanu, D. (2016). Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model. *Geomorphology*, 253:508–523.

Ischuk, A., Bjerrum, L. W., Kamchybekov, M., Abdrakhmatov, K., and Lindholm, C. (2017). Probabilistic Seismic Hazard Assessment for the Area of Kyrgyzstan, Tajikistan, and Eastern Uzbekistan, Central Asia. *Bulletin of the Seismological Society of America*, 108(1):130–144.

Jacobs, L., Dewitte, O., Poesen, J., Delvaux, D., Thiery, W., and Kervyn, M. (2016). The Rwenzori Mountains, a landslide-prone region? *Landslides*, 13(3):519–536.

Jacobs, L., Dewitte, O., Poesen, J., Maes, J., Mertens, K., Sekajugo, J., and Kervyn, M. (2017). Landslide characteristics and spatial distribution in the rwenzori mountains, uganda. *Journal of African Earth Sciences*, 134:917–930.

Jacobs, L., Dewitte, O., Poesen, J., Sekajugo, J., Nobile, A., Rossi, M., Thiery, W., and Kervyn, M. (2018). Field-based landslide susceptibility assessment in a data-scarce environment: the populated areas of the Rwenzori Mountains. *Natural Hazards and Earth System Sciences*, 18(1):105–124.

Jebur, M., Pradhan, B., Shafri, H., Yusoff, Z., and Tehrany, M. S. (2015). An integrated user-friendly arcmap tool for bivariate statistical modelling in geoscience applications. *Geoscientific Model Development*, 8(3):881–891.

Juang, C. S., Stanley, T. A., and Kirschbaum, D. B. (2019). Using citizen science to expand the global map of landslides: Introducing the cooperative open online landslide repository (COOLR). *PLoS ONE*, 14.

Kirschbaum, D. and Stanley, T. (2018). Satellite-Based Assessment of Rainfall-Triggered Landslide Hazard for Situational Awareness. *Earth's Future*, 6(3):505–523.

Kirschbaum, D., Stanley, T., and Zhou, Y. (2015). Spatial and temporal analysis of a global landslide catalog. *Geomorphol Geohazard Databases*, 249.

Kirschbaum, D. B., Adler, R., Hong, Y., Hill, S., and Lerner-Lam, A. (2010). A global landslide catalog for hazard applications: method, results, and limitations. *Natural Hazards*, 52(3):561–575.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC.

Lei, Y., Peng, C., Regmi, A. D., Murray, V., Pasuto, A., Titti, G., Shafique, M., and Priyadarshana, D. G. T. (2018). An international program on Silk Road Disaster Risk Reduction–a Belt and Road initiative (2016–2020). *Journal of Mountain Science*, 15(7):1383–1396.

Lima, P., Steger, S., and Glade, T. (2021). Counteracting flawed landslide data in statistically based landslide susceptibility modelling for very large areas: a national-scale assessment for austria. *Landslides*.

Lin, G.-F., Chang, M.-J., Huang, Y.-C., and Ho, J.-Y. (2017). Assessment of susceptibility to rainfall-induced landslides using improved self-organizing linear output map, support vector machine, and logistic regression. *Engineering geology*, 224:62–74.

Lin, Q., Lima, P., Steger, S., Glade, T., Jiang, T., Zhang, J., Liu, T., and Wang, Y. (2021). National-scale data-driven rainfall induced landslide susceptibility mapping for China by accounting for incomplete landslide data. *Geoscience Frontiers*, page 101248.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of statistical software*, 63(1):1–25.

Liu, C., Li, W., Wu, H., Lu, P., Sang, K., Sun, W., Chen, W., Hong, Y., and Li, R. (2013). Susceptibility evaluation and mapping of china's landslides based on multi-source data. *Nat Hazards*, 69.

Liu, W. and Dunford, M. (2016). Inclusive globalization: unpacking china's belt and road initiative. *Area Dev Policy*, 1.

Lombardo, L., Bakka, H., Tanyas, H., van Westen, C., Mai, P. M., and Huser, R. (2019). Geostatistical modeling to capture seismic-shaking patterns from earthquake-induced landslides. *Journal of Geophysical Research: Earth Surface*, 124(7):1958–1980.

Lombardo, L. and Mai, P. M. (2018). Presenting logistic regression-based landslide susceptibility results. *Engineering geology*, 244:14–24.

Lombardo, L., Opitz, T., Ardizzone, F., Guzzetti, F., and Huser, R. (2020a). Space-time landslide predictive modelling. *Earth-Science Reviews*, page 103318.

Lombardo, L., Saia, S., Schillaci, C., Mai, P. M., and Huser, R. (2018). Modeling soil organic carbon with Quantile Regression: Dissecting predictors' effects on carbon stocks. *Geoderma*, 318:148–159.

Lombardo, L. and Tanyas, H. (2020). Chrono-validation of near-real-time landslide susceptibility models via plug-in statistical simulations. *Engineering Geology*, 278:105818.

Lombardo, L. and Tanyas, H. (2021). From scenario-based seismic hazard to scenario-based landslide hazard: fast-forwarding to the future via statistical simulations. *Stochastic Environmental Research and Risk Assessment*, pages 1–14.

Lombardo, L., Tanyas, H., and Nicu, I. C. (2020b). Spatial modeling of multi-hazard threat to cultural heritage sites. *Engineering Geology*, page 105776.

Margottini, C., Canuti, P., and Sassa, K. (2013). *Landslide science and practice*, volume 1. Springer.

Menard, S. (2002). *Applied logistic regression analysis*, volume 106. Sage.

Mergili, M., Kopf, C., Müllebner, B., and Schneider, J. F. (2012). Changes of the cryosphere and related geohazards in the high-mountain areas of Tajikistan and Austria: a comparison. *Geografiska Annaler: Series A, Physical Geography*, 94(1):79–96.

Mitchell, M. (1995). Genetic algorithms: An overview. *Complexity*, 1(1):31–39.

Nazirova, D. and Saidov, S. (2015). Features of the development of landslides in tajikistan in various engineering-geological conditions (central tajikistan). *Science and new technologies*, 2:60–63.

Osna, T., Sezer, E. A., and Akgun, A. (2014). Geofis: an integrated tool for the assessment of landslide susceptibility. *Computers & Geosciences*, 66:20–30.

Pasuto, A. and Tagliavini, F. (2007). Landslide susceptibility and hazard mapping in high mountain regions: application in the italian alps. In *Guidelines for mapping areas at risk of landslides in Europe*. Institute for Environment and Sustainability Joint Research Centre, Rome.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Petschko, H., Brenning, A., Bell, R., Goetz, J., and Glade, T. (2014). Assessing the quality of landslide susceptibility maps–case study Lower Austria. *Natural Hazards and Earth System Sciences*, 14(1):95–118.

Pham, B. T., Pradhan, B., Tien Bui, D., Prakash, I., and Dholakia, M. B. (2016). A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of uttarakhand area (india). *Environ Model Softw*, 84.

Plotly Technologies Inc. (2015). Collaborative data science.

Polat, A. (2021). An innovative, fast method for landslide susceptibility mapping using gis-based lsat toolbox. *Environmental Earth Sciences*, 80(6):1–18.

QGIS.org (2021). QGIS Geographic Information System.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahmati, O., Kornejady, A., Samadi, M., Deo, R. C., Conoscenti, C., Lombardo, L., Dayal, K., Taghizadeh-Mehrjardi, R., Pourghasemi, H. R., Kumar, S., et al. (2019). PMT: New analytical framework for automated evaluation of geo-environmental modelling approaches. *Science of the total environment*, 664:296–311.

Razali, N. M. (2015). An Efficient Genetic Algorithm for Large Scale Vehicle Routing Problem Subject to Precedence Constraints. In *Procedia - Social and Behavioral Sciences*, volume 195, pages 1922 – 1931.

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

Remondo, J., González, A., Terán, J. R. D., Cendrero, A., Fabbri, A., and Chung, C. J. F. (2003). Validation of landslide susceptibility maps; examples and applications from a case study in northern spain. *Nat Hazards*, 30.

Reuter, H. I., Nelson, A., and Jarvis, A. (2007). An evaluation of void-filling interpolation methods for SRTM data. *Int J Geogr Inf Sci*, 21.

Rossi, M. and Reichenbach, P. (2016). Land-se: a software for statistically based landslide susceptibility zonation, version 1.0. *Geoscientific Model Development*, 9(10):3533–3543.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *EJOR*, 48.

Safanelli, J. L., Poppiel, R. R., Ruiz, L. F. C., Bonfatti, B. R., Mello, F. A. d. O., Rizzo, R., and Demattê, J. A. M. (2020). Terrain analysis in google earth engine: A method adapted for high-performance global-scale analysis. *ISPRS International Journal of Geo-Information*, 9(6).

Said, G., Mahmoud, A., and El Horbaty, E. (2014). A Comparative Study of Meta-heuristic Algorithms for Solving Quadratic Assignment Problem. *International Journal of Advanced Computer Science and Applications*, 5(1):1–6.

Saponaro, A., Pilz, M., Bindi, D., and Parolai, S. (2015). The contribution of EMCA to landslide susceptibility mapping in central asia. *Ann Geophys.*

Schneider, J. F., Mergili, M., and Schneider, D. (2010). Analysis and mitigation of remote geohazards in high mountain areas of Tajikistan with special emphasis on glacial lake outburst floods. In *EGU General Assembly Conference Abstracts*, page 4905.

Soeters, R. and van Westen, C. J. (1996). Slope instability recognition, analysis and zonation. In *Landslide Investigation and Mitigation.*, number 247 in Transportation Research Board Special Report, pages 129–177. National Research Council, Transportation Research Board.

Stanley, T. and Kirschbaum, D. B. (2017). A heuristic approach to global landslide susceptibility mapping. *Nat Hazards*, 87.

Steger, S., Brenning, A., Bell, R., and Glade, T. (2016). The propagation of inventory-based positional errors into statistical landslide susceptibility models. *Natural Hazards and Earth System Sciences*, 16(12):2729–2745.

Steger, S., Brenning, A., Bell, R., and Glade, T. (2017). The influence of systematically incomplete shallow landslide inventories on statistical susceptibility models and suggestions for improvements. *Landslides*, 14(5):1767–1781.

Steger, S., Mair, V., Kofler, C., Pittore, M., Zebisch, M., and Schneiderbauer, S. (2021). Correlation does not imply geomorphic causation in data-driven landslide susceptibility modelling–benefits of exploring landslide data collection effects. *Science of the total environment*, 776:145935.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227.

Tajikistan General Office of Geology (1974). Geological map of tajikistan. `https://geoportal-tj.org/maps/`.

The pandas development team (2019). pandas-dev/pandas: v0.25.3.

Titti, G., Borgatti, L., Qiang, Z., Peng, C., and Pasuto, A. (2021a). Landslide susceptibility in the Belt and Road Countries: continental step of a multi-scale approach. *Environ Earth Sciences*, 80(630).

Titti, G. and Sarretta, A. (2020). Cnr-irpi-padova/sz: Sz plugin.

Titti, G., Sarretta, A., and Lombardo, L. (2021b). Cnr-irpi-padova/sz: Sz plugin.

Titti, G., van Westen, C., Borgatti, L., Pasuto, A., and Lombardo, L. (2021c). When enough is really enough? on the minimum number of landslides to build reliable susceptibility models. *Geosciences*, 11(11):469.

Torgoev, A., Lamair, L., Torgoev, I., and Havenith, H.-B. (2013). A review of recent case studies of landslides investigated in the Tien Shan using microseismic and other geophysical methods. *Earthquake-Induced Landslides*, pages 285–294.

Torizin, J. (2012). Landslide susceptibility assessment tools for arcgis 10 and their application. *Proceedings of 34th IGC, Brisbane*, pages 5–10.

Ubaidulloev, A., Kaiheng, H., Rustamov, M., and Kurbanova, M. (2021). Landslide Inventory along a National Highway Corridor in the Hissar-Allay mountains, Central Tajikistan. *GeoHazards*, 2(3):212–227.

Van Den Bout, B., Lombardo, L., Chiyang, M., van Westen, C., and Jetten, V. (2021). Physically-based catchment-scale prediction of slope failure volume and geometry. *Engineering Geology*, 284:105942.

Van Den Bout, B., Lombardo, L., van Westen, C., and Jetten, V. (2018). Integration of two-phase solid fluid equations in a catchment model for flashfloods, debris flows and shallow slope failures. *Environmental Modelling & Software*, 105:1–16.

Van Den Eeckhaut, M., Vanwalleghem, T., Poesen, J., Govers, G., Verstraeten, G., and Vandekerckhove, L. (2006). Prediction of landslide susceptibility using rare events logistic regression: a case-study in the flemish ardennes (belgium). *Geomorphology*, 76(3-4):392–410.

van Westen, C., Castellanos, E., and Kuriakose, S. (2008). Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. *Engineering Geology*, 102(3-4):112–131.

Van Westen, C. J., Rengers, N., Terlien, M., and Soeters, R. (1997). Prediction of the occurrence of slope instability phenomenal through GIS-based hazard zonation. *Geologische Rundschau*, 86(2):404–414.

van Westen, C. J., Soeters, R., and Sijmons, K. (2000). Digital geomorphological landslide hazard mapping of the Alpago area, Italy. *International Journal of Applied Earth Observation and Geoinformation*, 2(1):51–60.

Wang, X., Otto, M., and Scherer, D. (2021). Atmospheric triggering conditions and climatic disposition of landslides in Kyrgyzstan and Tajikistan at the beginning of the 21st century. *Natural Hazards and Earth System Sciences Discussions*, pages 1–24.

Wilde, M., Günther, A., Reichenbach, P., Malet, J. P., and Hervás, J. (2018). Pan-European landslide susceptibility mapping: ELSUS version 2. *J Maps*, 14.

Wolfgramm, B., Liniger, H., and Nazarmavloev, F. (2014). Integrated watershed management in tajikistan.

Yablokov, A. (2001). The Tragedy of Khait: A Natural Disaster in Tajikistan. *Mountain Research and Development*, 21(1):91 – 93.

Yao, X., Tham, L. G., and Dai, F. C. (2008). Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of hong kong, china. *Geomorphology*, 101.

Yeon, Y.-K., Han, J.-G., and Ryu, K. H. (2010). Landslide susceptibility mapping in Injae, Korea, using a decision tree. *Engineering Geology*, 116(3-4):274–283.

Zevenbergen, L. W. and Thorne, C. R. (1987). Quantitative analysis of land surface topography. *Earth surface processes and landforms*, 12(1):47–56.

Zêzere, J., Pereira, S., Melo, R., Oliveira, S., and Garcia, R. A. (2017). Mapping landslide susceptibility using data-driven methods. *Science of the total environment*, 589:250–267.