Proceedings of the XVII ECSMGE-2019
Geotechnical Engineering foundation of the future
ISBN 978-9935-9436-1-3

doi: 10.32075/17ECSMGE-2019-0521

XVII European Conference
ECSMGE 2019
Reykjavik
on Soil Mechanics
and Geotechnical Engineering

A framework for predicting rainfall-induced landslides using machine learning methods

Un cadre pour prédire les glissements de terrain induits par les précipitations à l'aide d'un apprentissage automatique

F. S. Tehrani
Deltares & Delft University of Technology, Delft, The Netherlands

G. Santinelli
Deltares, Delft, The Netherlands

M. Herrera
Delft University of Technology, Delft, The Netherlands

ABSTRACT: Landslides are catastrophic geo-hazards that threaten urbanization. Growth in population besides construction of critical infrastructures such as roads and pipelines in landslide-prone areas elevates the risk associated with landslides. Therefore, a system that is able to predict landslides and issues warning in a timely manner is very appealing. Various factors influence the stability of natural and engineered slopes and cause landslides, including topography, geology of slopes, precipitation, temperature changes, snowmelt, seismic activities, volcanic activities, and human actions. It is widely accepted that precipitation is one of the most influential factors for triggering landslides. In this paper, we present the preliminary results of a practical research study that has been carried out in Deltares, The Netherlands. To that end, we have set up a framework that combines geo-engineering, remote sensing, hydrology with machine learning to predict the onset of landslides under the effect of precipitation. In this data-driven approach, Machine Learning (ML) methods are used to predict landslides by exploiting multiple Earth observation datasets, including rainfall data (e.g. TRMM 3B42) and Digital Elevation Models (e.g. SRTM1) and the NASA Global Landslide Catalogue. A detailed inventory of 10,988 landslides at a global level is built out of which 4,542 cases are used to train a supervised machine learning algorithm. The trained ML model is then fed by rainfall data, topography features such as slope and elevation relief, soil and bedrock data, and vegetation index of target regions to assess the stability of the studied area.

RÉSUMÉ: Les glissements de terrain sont des géo-aléas catastrophiques qui menacent l'urbanisation. La croissance de la population en plus de la construction d'infrastructures critiques telles que les routes et les pipelines dans les zones exposées aux glissements de terrain augmente le risque associé aux glissements de terrain. Par conséquent, un système capable de prédire les glissements de terrain et les alertes en temps opportun est très attrayant. Divers facteurs influent sur la stabilité des pentes naturelles et aménagées et provoquent des glissements de terrain, notamment la topographie, la géologie des pentes, les précipitations, les changements de température, la fonte des neiges, les activités sismiques, les activités volcaniques et les actions humaines. Il est largement admis que la précipitation est l'un des facteurs les plus influents du déclenchement des glissements de terrain. Dans cet article, nous présentons les résultats préliminaires d'une étude de recherche pratique réalisée à Deltares, aux Pays-Bas. À cette fin, nous avons mis en place un cadre combinant géo-ingénierie, télédétection, hydrologie et apprentissage automatique afin de prédire l'apparition de glissements de terrain sous l'effet des précipitations. Dans cette approche pilotée par les données, les méthodes Machine Learning (ML) sont utilisées pour prédire les

glissements de terrain en exploitant plusieurs jeux de données d'observation de la Terre, notamment les données pluviométriques (par exemple, TRMM 3B42), les modèles numériques d'altitude (par exemple, SRTM1) et le catalogue mondial des glissements de terrain de la NASA. Un inventaire détaillé de 10 988 glissements de terrain au niveau mondial est constitué de 4 542 cas utilisés pour former un algorithme d'apprentissage automatique supervisé. Le modèle ML formé est ensuite alimenté par les données pluviométriques, les caractéristiques topographiques telles que le relief de la pente et de l'élévation, les données sur le sol et le substratum rocheux et l'indice de végétation des régions cibles pour évaluer la stabilité de la zone étudiée.

Keywords: Landslide; Rainfall; Precipitation; Machine Learning; Soil

1 INTRODUCTION

Landslides can pose serious threat to urban environment and to line infrastructures such as roads and pipelines. They can be triggered by natural factors such as precipitation and earthquake or by anthropogenic causes such as mining and slope cutting. Among multiple triggering factors of landslides, precipitation is one of the most common ones which has caused thousands of landslides in the past decade some of which are amongst the deadliest landslides (e.g. the debris flow occurred in August 2017 in and around Freetown in Sierra Leone with 1141 fatalities). Therefore, forecasting rainfall-induced landslides can be extremely helpful to minimize mortalities of landslides and planning mitigation and rescue measures. Forecasting rainfall-induced landslides is typically done using hydrology-based approaches that incorporate rainfall threshold (e.g. Guzzetti et al. 2007; Rossi et al. 2017), which is the basis of many landslide early warning systems in landslide prone areas across the world. Rainfall thresholds are based on rainfall conditions once exceeded landslide might be triggered. Although rainfall thresholds are widely used methods for predicting the occurrence of landslides, they suffer from certain limitations; one of them is that they have been mostly developed for regional and local landslide predictions (Segoni et al. 2018), therefore the outcome is partially biased by geography. Besides hydrological methods, geotechnical engineering methods have been also practiced for evaluating the probability of landslide occurrence. However, these methods bear a major limitation in that they are dependent on the quality of site investigation in the specific location of interest, which can be very costly. In the past years and by re-introduction of machine learning into natural hazard community, machine learning has gained popularity in supporting landslide analysis, especially for landslide susceptibility mapping (e.g. Marjanović et al. 2011; Goetz et al. 2015). There have been fewer works in landslide prediction using machine learning methods such as the work of Farahmand & AghaKouchak (2013), in which they used 581 landslide events, from landslide catalogue of National Aeronautics and Space Administration (NASA) of years 2003, 2007 to 2009, for global landslide prediction. One of the major challenges in using machine learning or any statistical-based methods in landslide prediction is concerned with availability of landslide data, namely, landslide inventories (location and date of landslides), landslide triggering factors (rainfall intensity and duration in this study) and landslide controlling factors (e.g. topography, geology, soil, land cover). In this study we have tried to overcome this limitation by integrating the most recent (updated up to the December 2017) Global Landslide Catalogue (GLC) of NASA (Kirschbaum et al. 2010 and 2015) with global rainfall datasets and publicly available datasets of landslide controlling factors. The predictive framework that we built is based on this quantitative landslide dataset and the use of machine learning algorithms for differentiating landslide and non-landslide

events. For this study, we present the results of logistic regression and decision tree algorithms as two popular machine learning classification algorithms to identify landslide and non-landslide events.

2 DATASETS

The landslide database built at Deltares is based on the landslide inventory (GLC) of NASA and on the associated triggering and controlling factors. Figure 1 shows the data used to create the landslide database.

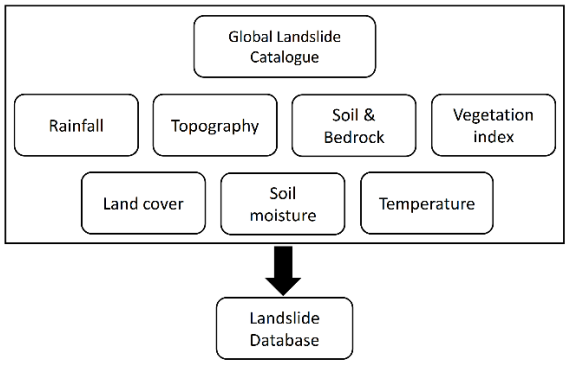

Figure 1. Elements of the Deltares landslide database.

In this paper, land cover, soil moisture and temperature are not considered for landslide prediction.

2.1 Global landslide inventory

The global landslide inventory is derived from the global landslide catalogue (GLC), which was developed by NASA Goddard Space Flight Center (GSFC). This catalogue provides initial insights into the spatio-temporal trends in landslide distribution and impact worldwide (Kirschbaum et al. 2010). The GLC is based on various online news media, scholarly articles, and existing hazard databases such as the International Consortium on Landslides (ICL), International Landslide Centre, University of Durham (ILC), and International Federation of Red Cross and Red Crescent Societies field reports, Reliefweb, humanitarian disaster information run by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) and other online regional and national newspaper articles and media sources. As of April 2018, the GLC consisted of 11,055 landslides with 10,988 landslides occurred after 2007. Figure 2 shows the type of data reported in GLC. The yellow cells indicate missing data.

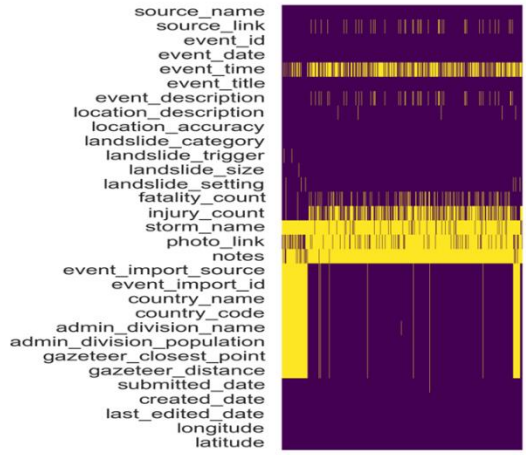

Figure 2. Data presented at NASA GLC.

The GLC contains a limited number of landslides triggered by factors other than rainfall, such as earthquake and human action. We filtered these types of landslides out and focused only on rainfall-induced landslides in this study.

In terms of location accuracy, Kirschbaum et al. (2010) reported large uncertainties when assigning geographic coordinates to a landslide event, especially when multiple landslides on affected areas were referenced in the same report. To deal with this uncertainty, they assigned a radius of confidence (which spans from tens of meters to tens of kilometres) to the location, indicating the estimated radius of a circle over which the landslide may have occurred. To reduce the uncertainty in finding the triggering and controlling factors associated with landslides only landslides with 5 km radius of confidence were considered in this study. Applying the two filters (triggering

factor and radius of confidence), 4,542 landslides were resulted for further analysis.

2.2 Rainfall Data

There is a consensus in landslide community supported by concrete evidences that rainfall and subsequent phenomena such as infiltration, exfiltration and run-off trigger the majority of landslides worldwide. As such, in this study we focused only rainfall as the triggering factor.

As reported by Sun et al. (2018), currently there are approximately 30 available global precipitation datasets, including gauge-based, satellite-derived, and reanalysis datasets. These authors suggest that the reliability of precipitation datasets is mainly limited by the number and spatial coverage of surface stations, the accuracy of satellite algorithms, and the data assimilation models. For the scope of the current study, the maximum daily rainfall data from Tropical Rainfall Measurement Mission of NASA (TRMM 3B42) has been used for estimating the accumulated intensity of rainfall on the day of landslide event, the day before (short term rainfall) and nine days before these two days (long term rainfall) prior to the event. Figure 3 shows the frequency of the accumulated short term and long term rainfalls. It is seen in Figure 3 that for both short term and long term rainfall data, a remarkable number of landslide events are associated with accumulated rainfall less than 25 mm. This can be due to two major reasons: 1) the accuracy of landslide location is not sufficient, which is an inherent issue with GLC and landslide data based on media, 2) the rainfall data is not correctly estimated for the region where landslide events occurred. This is somehow a known issue with sattelite based rainfall data (Sun et al. 2018). Ideally, rainfall data should be measured by ground-based gauges. However, due to sparsity of these gauges and their delayed temporal coverage, use of satellite based data is the most optimal method for estimating the instensity and duration of percipitation.

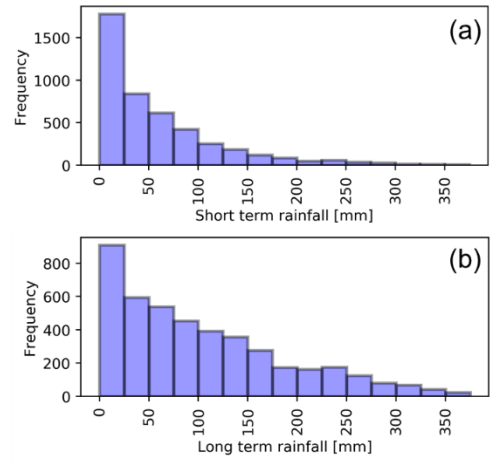

Figure 3. Accumulated rainfall for the filtered landslide events (4542 cases) based on TRMM3B42: (a) Short term and (b) long term.

2.3 Digital Elevation Model

Digital elevation models (DEMs) are considered as one of the main datasets for analysing the controlling factors involved in the landslide hazard assessments (van Westen et al. 2008). These three-dimensional representations of the terrain are useful for extracting key topographical and geomorphological parameters including elevation, slope, and aspect of the ground surface. DEM data normally consists of regular raster grids which are mostly organized as two-dimensional arrays with individual cells having a particular elevation value.

Global DEM datasets with different parameters such as horizontal grid spacing, spatial resolution and temporal coverage are nowadays available from a variety of open web portals. In this study, one of the most well-known satellite-derived DEMs, namely the NASA Shuttle Radar Topography Mission (SRTM, 2000) was used to obtain topographical features of the terrains where landslide occurred. SRTM1 is selected due to the high spatial resolution (30 m) and its temporal coverage with an acquisition date before the occurrence of all the landslides recorded in the database. Figure 4 shows the slope and elevation relief (difference between the maximum and

minimum elevation within the landslide confidence area) for the filtered landslide data.

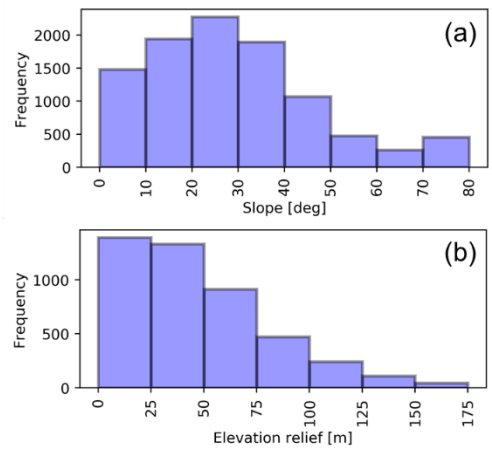

Figure 4. DEM properties for the filtered landslide events (4542 cases) based on SRTM1: (a) Slope and (b) Elevation relief.

2.4 Soil and bedrock

The comprising material of slopes and the depth of the bedrock can highly affect the hydro-geomechanical response of slopes to rainfall. Therefore, estimating the soil composition of hillslopes can potentially enhance the predictability of rainfall-induced landslides.

Soil composition was retrieved as raster data from the SoilGrids datasets (Hengl et al. 2014) at 250 m resolution with a global coverage. SoilGrids provides global predictions for standard numeric soil properties (organic carbon, bulk density, Cation Exchange Capacity (CEC), pH, soil texture fractions and coarse fragments) at seven standard depths (0, 5, 15, 30, 60, 100 and 200 cm), in addition to predictions of depth to bedrock and distribution of soils classes based on the World Reference Base (IUSS working group WRB 2006) and USDA classification systems (US Department of Agriculture 2010).

Soil content was reported as percentage of clay, sand, and silt. The estimated amount of each soil at each depth was normalized by dividing the soil estimated amount by the summation of amount of sand, clay and silt to ensure that all estimations are between 0 and 100 and that the fractions sum up to 100% (Hengl et al. 2014). Among the information available of SoilGrid, the estimated fraction of sand, clay and silt and depth to the bedrock are used in this study. The average sand, silt and clay fraction of the seven standard depths are calculated as features to be used later in the prediction stage. Figure 5 shows the fraction of these soil types for the filtered landslide events.

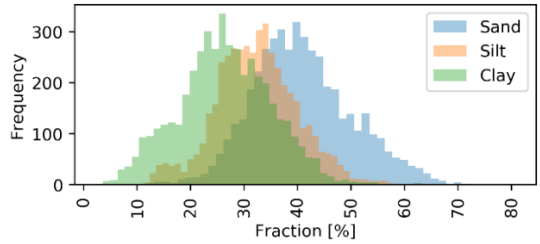

Figure 5. Soil fraction for the filtered landslides .

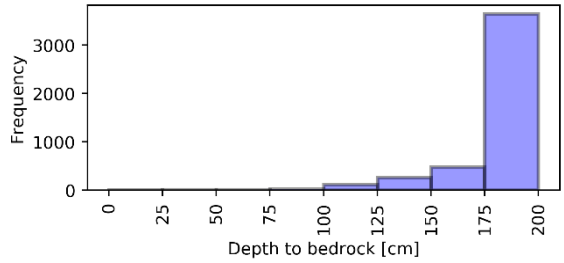

Figure 6. Depth to the bedrock for the filtered landslides.

2.5 Vegetation

Vegetation is another controlling factor that can highly influence the stability of natural slopes and therefore play a vital role in predicting landslides. Leaves control soil moisture through evapotranspiration and roots can mechanically reinforce soil particles and increase shear strength of soil compound by increasing the matric suction. Therefore, it is accepted that in general lack or shortage of vegetation can increase the susceptibility of slopes to landslides. One way of quantifying vegetation density is through calculating the Normalized Difference Vegetation Index (NDVI). NDVI quantifies vegetation by measuring the difference between near-infrared (NIR),

which is strongly reflected by vegetation, and red (visible) light (R), which is strongly absorbed by vegetation. NDVI is calculated per pixel as a function of the red and near infrared bands:

$$NDVI = \frac{NIR-R}{NIR+R} \quad (1)$$

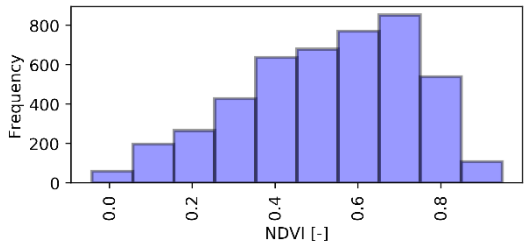

Figure 7. NDVI before landslide occurrence for the filtered landslides.

3 MACHINE LEARNING

Machine learning (ML) is a sub-field of artificial intelligence that uses data and statistical techniques to give a machine the ability to learn and improve its learning process without being explicitly programmed. One of the main advantage of machine learning over classic prediction methods is its capability in dealing with complex datasets that include both quantitative and qualitative data. In this study, our focus is on supervised learning methods for classification of landslide and non-landslide events (binary classification). The machine learning models are trained with training datasets which includes features or inputs (X) and target output (Y). Then the performance of models is measured on test inputs to evaluate the accuracy of predicting outputs.

3.1 Logistic regression

Logistic regression, applied as classification, calculates the probability that the predicted output belongs to a particular category or class (landslide and non-landslide in this study). Mathematically, the relationship between the probability p of landslide and the triggering and controlling factors (features) can be expressed using the sigmoid function:

$$p(z) = \frac{1}{1+e^{-z}} \quad (2)$$

where z is a linear combination of features x_1 to x_n as:

$$z = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \quad (3)$$

where w_0 is the intercept or bias of the model, and w_i (i =1, 2, ..., n) are the weights (fitting coeffects) of the features. These weights are derived by optimizing the cost function which measures the difference of predicted output and target output. If the probability of occurrence is greater than 50%, the model classifies the output as 1 (landslide), otherwise 0 (non-landslide). Figure 8 shows the probability function and the threshold line (p = 0.5) for identifying landslide and non-landslide events.

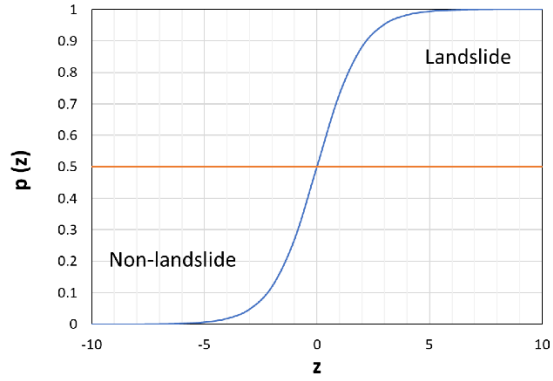

Figure 8. Probability (sigmoid) function for logistic regression.

3.2 Training examples

In order to make a training example, non-landslide cases must be added to the example set. To that end, artificial non-landslide cases were created by applying random fraction (less than 0.5) to the rain and topographic features of selected landslide cases and adding the results to landslide

example set. This resulted in an example set of 9,084 landslide/non-landslide cases. This approach in creating non-landslide events might impose a bias to the training and test sets. Therefore, it would be better to create non-landslide events through finding actual cases where no slope instability took place. This will be done in our future works.

4 RESULTS AND DISCUSSION

At this stage, logistic regression (LR) algorithm as a binary classification method was used to distinguish landslides and non-landslide cases. To train the ML model, eleven example sets (E0 to E10) with different combination of triggering and controlling factors (model features) were built. Figure 9 shows the combination of controlling and triggering features (x1 to x6) used for training ML in this study.

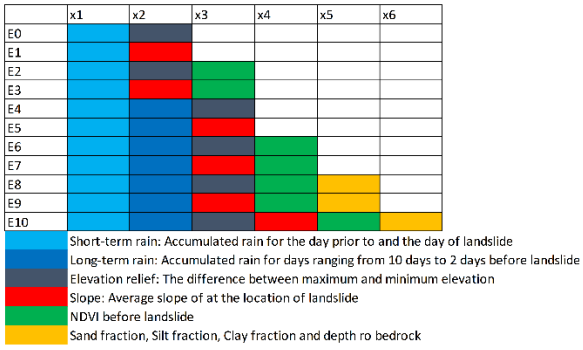

Figure 9. Features used for each example sets.

The example sets were split into training (67%) and test (33%) sets which then were used for training and assessing the logistic regression model. The accuracy of the logistic regression model in form of Receiver Operating Characteristic (ROC) curves and the associated Area Under Curve (AUC) is illustrated in Figure 10. The ROC curve is a measure for evaluating a diagnostic test. In a ROC curve the true positive rate (Sensitivity) is plotted against the false positive rate (100 - Specificity) for various cut-off points of a parameter. Every point on the ROC curve represents a sensitivity/specificity pair that corresponds to a certain decision threshold. The area under the ROC curve (AUC) quantifies how well a group of features can be used to distinguish between two diagnostic groups (landslide / non-landslide). Higher AUC values (maximum = 1) indicate a more accurate classification.

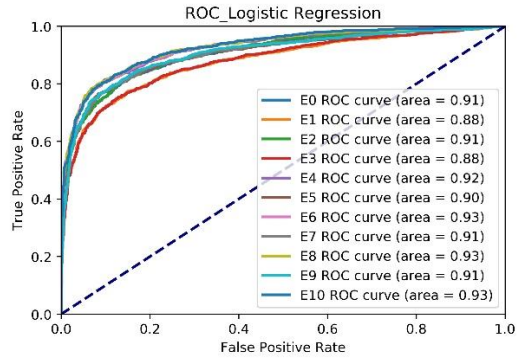

Figure 10. Accuracy of logistic regression model in classifying landslides and non-landslides.

It can be suggested from Figure 10 that in general the logistic-regression model can perform well in distinguishing landslide and non-landslide cases.

By comparing the results of E0 (AUC = 0.91) and E1 (AUC = 0.88), it is observed that having rainfall data type fixed, elevation relief can be more effective than slope angle in producing a LR model with higher accuracy, which indicates that elevation relief, when dealing with large scale zones, can be more representative of the topography of the region than slope angle. This has been suggested by other authors such as Lin et al. (2017).

Looking into example sets with highest accuracy (AUC = 0.93), namely E6, E8 and E10 it can be suggested that adding features to a training set of a machine learning model might not necessarily result in better prediction. In this case, E6 prediction is as good as E8 and E10. This observation emphasizes the role of feature analysis when dealing with machine learning problems. Such analysis can reduce the cost of prediction as

less number of features may result in highly accurate ML models.

5 SUMMARY CONCLUSIONS

In this paper, we presented the preliminary results of a practical research study on developing a data-driven framework for predicting rainfall-induced landslides. Logistic Regression as a Machine Learning algorithm was used to predict landslides by exploiting multiple Earth Observation datasets. Although the database and forecasting framework that were reported in this study are at its initial stage, the results of the study (AUC greater than 0.88) showed that such a framework, with enhanced datasets, can be used for forecasting rainfall-induced landslides and landslide early warning systems at a global and regional level.

6 REFERENCES

Farahmand A., AghaKouchak A., 2013, A Satellite-Based Global Landslide Model, Natural Hazards and Earth System Sciences, 13, 1259-1267, doi:10.5194/nhess-13-1259-2013.

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

Guzzetti, F., Peruccacci, S., Rossi, M., & Stark, C. P. (2007). Rainfall thresholds for the ini-tiation of landslides in central and southern Europe. Meteorology and atmospheric phys-ics, 98(3-4), 239-267.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., ... & Guevara, M. A. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLoS one, 12(2), e0169748.

IUSS working group WRB. (2006). World reference base for soil resources 2006. World soil resources reports, 132--pp.

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

Kirschbaum, D., Stanley, T., & Zhou, Y. (2015). Spatial and temporal analysis of a global landslide catalog. Geomorphology, 249, 4-15.

Lin, L., Lin, Q., & Wang, Y. (2017). Landslide susceptibility mapping on a global scale using the method of logistic regression. Natural Hazards and Earth System Sciences, 17(8), 1411-1424.

Marjanović, M., Kovačević, M., Bajat, B., & Voženílek, V. (2011). Landslide susceptibility assessment using SVM machine learning algorithm. Engineering Geology, 123(3), 225-234.

Rossi, M., Luciani, S., Valigi, D., Kirschbaum, D., Brunetti, M. T., Peruccacci, S., & Guzzetti, F. (2017). Statistical approaches for the definition of landslide rainfall thresholds and their uncertainty using rain gauge and satellite data. Geomorphology, 285, 16-27-267.

Segoni, S., Piciullo, L., & Gariano, S. L. (2018). A review of the recent literature on rainfall thresholds for landslide occurrence. Landslides, 1-19.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K. L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. Reviews of Geophysics, 56(1), 79-107.

US Department of Agriculture. (2010). Keys to Soil Taxonomy. U.S. Government Printing Office.

Van Westen, C. J., Castellanos, E., & Kuriakose, S. L. (2008). Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. Engineering geology, 102(3-4), 112-131.