
F. S. Tehrani
Deltares, Delft


G. Santinelli
Deltares, Delft


M. Herrera
Delft University of Technology, Delft

MACHINE LEARNING FOR FORECASTING RAINFALL-INDUCED LANDSLIDES

Figure 1 – Sierra Leone landslide in 2018.

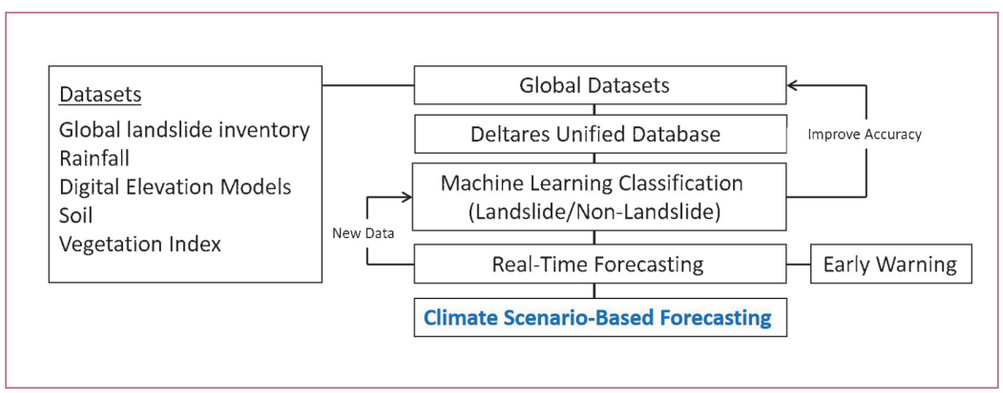

Figure 2 – Landslide forecasting framework of this study.

Introduction

Landslides can pose serious threat to urban environment and to line infrastructures such as roads and pipelines. Among multiple triggering factors of landslides, precipitation is one of the most common ones, causing thousands of landslides in the past decade, some of which are amongst the deadliest landslides. For instance, the debris flow occurred in August 2017 in and around Freetown in Sierra Leone caused 1141 fatalities (Figure 1). Therefore, forecasting rainfall-induced landslides can be extremely helpful to minimize mortalities due to landslides and plan mitigation and rescue measures.

Forecasting rainfall-induced landslides is typically done based on rainfall thresholds (e.g. Guzzetti et al., 2007; Rossi et al., 2017). Although rainfall thresholds are widely used for predicting the occurrence of landslides, they suffer from certain limitations; one of them is that they have been mostly developed for region-specific prediction of landslides (Segoni et al., 2018), hence the outcome suffers from geographical biases. To overcome the limitations of the conventional landslide forecasting methods, next to the rainfall intensity, duration and frequency, one needs to consider controlling factors, which include, among others, topography, lithology and geomorphology of slopes, soil type, ambient temperature, surface radiation, vegetation, soil moisture, land use and land cover. This was the subject of our study, for which we have set up a Machine Learning (ML) framework to better estimate the onset of rainfall-induced landslides. Figure 2 shows the forecasting framework that was adopted in this research project. We used the NASA Global Landslide Catalogue (Kirschbaum et al., 2010) to build the detailed database of landslides.

Datasets

GLOBAL LANDSLIDE INVENTORY

The global landslide inventory is derived from the global landslide catalogue (GLC), which was developed by NASA Goddard Space Flight Center. The GLC is based on various online news media, scholarly articles, and existing hazard databases. As of April 2018, the GLC consisted of 11,055 landslides with 10,988 landslides occurred after 2007. The GLC contains a limited number of landslides triggered by factors other than rainfall, such as earthquake and human action. In this study, we filtered these types of landslides out and focused only on rainfall-induced landslides.

With regard to location accuracy, Kirschbaum et al. (2010) reported large uncertainties when assigning geographic coordinates to a landslide event. 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 events only with nearly exact locations and with short-term rainfall (to be explained later) greater than 20 mm are considered in this study [this is because the focus of this study is on rainfall-induced landslides]. For training a ML algorithm, non-landslide cases are also needed. We sampled non-landslides from landslide events with radius of confidence greater

SUMMARY

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. It is widely accepted that precipitation is one of the most influential factors for triggering landslides. In this article, we present the preliminary results of a practical research study that has been carried out in Deltares. To that end, we have set up a framework that combines geo-engineering, remote sensing, hydrology with Machine Learning (ML) to predict the onset of landslides under the effect of precipitation. In this data-driven approach, 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. SRTM), and the NASA Global Landslide Catalogue. A detailed inventory of landslides at a global level is built out of which a supervised ML algorithm is trained with landslide/non-landslide events. 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.

than 25 km and short-term rainfall less than 60 mm. Since every landslide in the GLC has a coordinate, it can be suggested that a landslide event with radius of confidence greater than 25 km did not happen on the reported coordinate. This was further verified by visual inspection. Applying these filters, the final dataset consist of 235 landslides and 1696 non-landslide events.

RAINFALL DATA

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.

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. In this study, the NASA Shuttle Radar Topography Mission (SRTM, 2000) was used to obtain topo-graphical features of the terrains where landslide occurred. SRTM 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. 4 shows the mean slope and elevation relief (difference between the maximum and minimum elevation within the landslide confidence area) for the filtered landslide data.

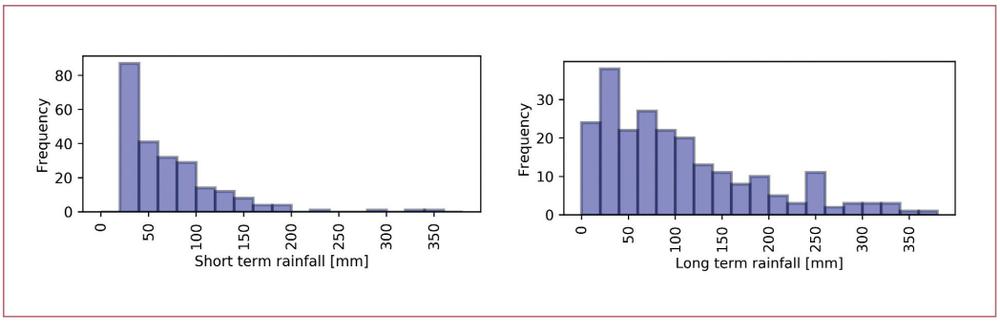

Figure 3 – Accumulated rainfall for the filtered landslide events based on TRMM3B42: (a) Short term and (b) long term.

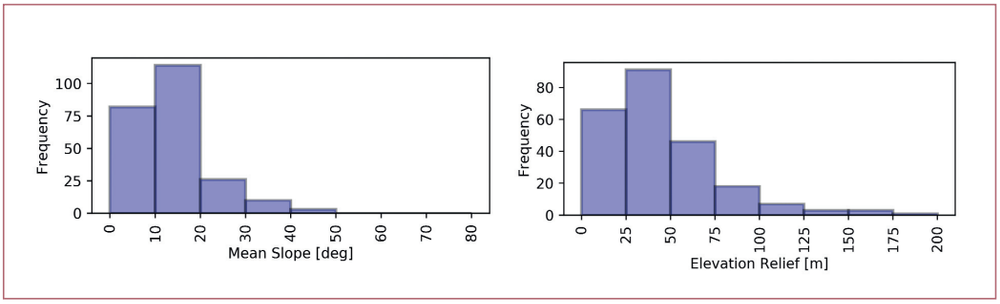

Figure 4 – DEM properties for the filtered landslide events based on SRTM: (a) Slope and (b) Elevation relief.

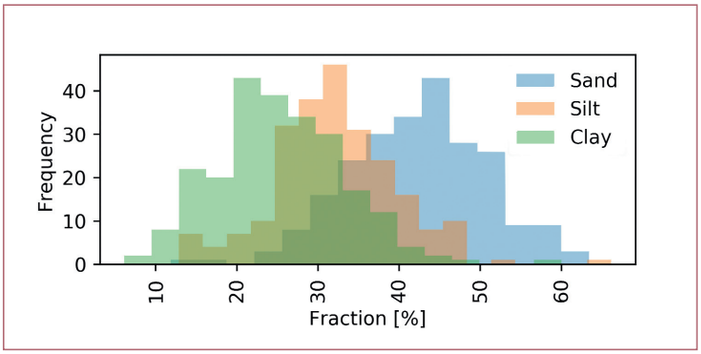

Figure 5 – Soil fraction for the filtered landslides.

SOIL AND BEDROCK

The comprising material of slopes and the depth of the bedrock can highly affect the hydro-geo-mechanical 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. 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.

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:

$$\mathrm{NDVI}=\frac{NIR-R}{NIR+R}$$

MACHINE LEARNING

For the current work, we used the Logistic Regression (LR) algorithm as a supervised ML method for classification of landslide and non-landslide events (binary classification). The LR algorithm is trained with training sub-sets, which include controlling and triggering factors as predictors (X) and labeled (landslide or non-landslide) output (Y). The performance of LR models is measured on test sub-sets to evaluate the accuracy of predicting outputs. LR algorithm calculates the pro-bability 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 (predictors or features) can be expressed using the sigmoid function:

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

where $z = w0 + w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$ is a linear combination of predictors x_1 to x_n, 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 actual output. If the probability of occurrence is greater than 50%, the model classifies the output as 1 (landslide), otherwise 0 (non-landslide).

As mentioned earlier 235 landslides and 1696 non-landslide events are used to build the ML dataset. However, the dataset suffers from imbalanced landslide and non-landslide events which can influence the performance of any ML algorithm. To overcome this issue, the non-landslide events are undersampled by random removal of 1000 non-landslides. The outcome is a dataset that consists of 235 landslides and 696 non-landslides.

Table 1 - Example sets used in training the LR algorithm

Example set/Features	E0	E1	E2	E3	E4	E5	E6	E7	E8
x1 Short-term rain	1	1	1	1	1	1	1	1	1
x2 Long-term rain	0	0	0	1	1	1	1	1	1
x3 Mean slope	0	1	0	0	0	1	1	1	1
x4 Elevation relief	0	0	1	0	1	0	1	1	1
x5 NDVI	0	0	0	0	1	1	0	1	1
x6 Soil and bedrock	0	0	0	0	1	1	1	0	1

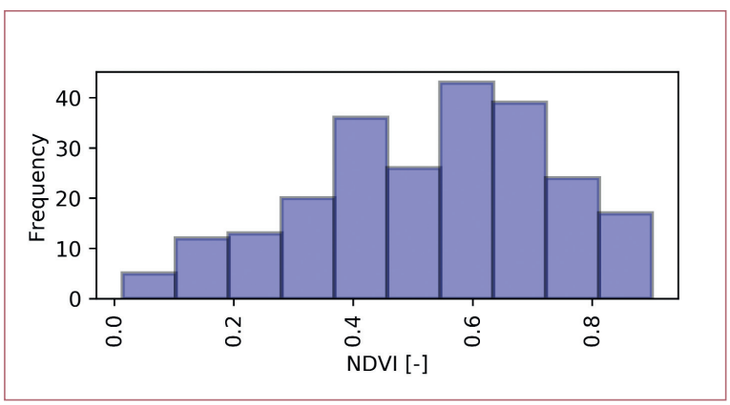

Figure 6 – NDVI before landslide occurrence for the filtered landslides.

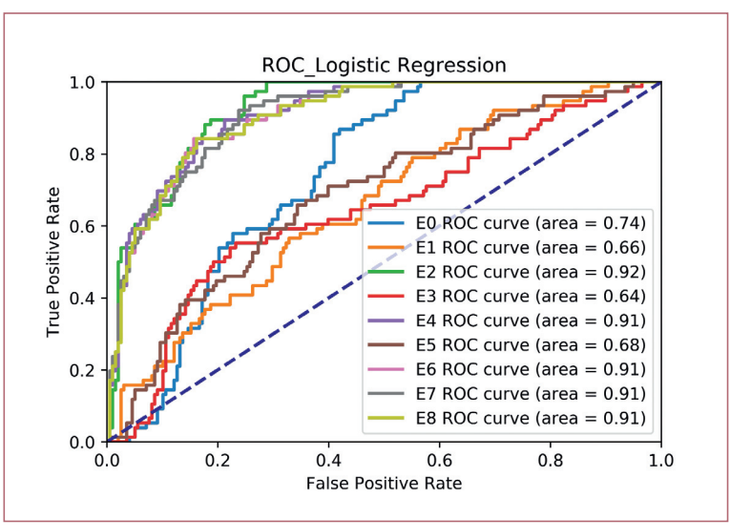

Figure 7 - Accuracy of logistic regression model in classifying landslides and non-landslides.

Research outcomes

LR algorithm was used to distinguish landslides and non-landslide cases. To train the LR model, nine example sets (E0 to E8) with different combination of triggering and controlling factors (model features) were built. Table 1 shows the combination of controlling and triggering features (x1 to x6) used for training the LR model, where 1 shows if the feature is used and 0 means otherwise.

The sample sets are split into training (70%) and test (30%) sets which then are used for training and assessing the LR model. The accuracy of the LR model in form of Receiver Operating Characteristic (ROC) curves and the associated Area Under Curve (AUC) is illustrated in 7. The ROC curve is a measure for evaluating a diagnostic test, where true positive rate (Sensitivity) is plotted against false positive rate (100 - Specificity) for various decision thresholds (between 0 to 100%). 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). In general, higher AUC values (maximum = 1) indicate a more accurate classification. However, other metrics such as number of true positives and negatives

and false positives and negatives should be also checked for further verifications.

7 shows that, in general, the LR model can perform well in distinguishing landslide and non-landslide cases. By comparing the results of E1 (AUC = 0.66) and E2 (AUC = 0.92), it can be suggested that having the short-term rainfall fixed, elevation relief can be more effective than slope angle for landslide/non-landslide classification. This indicates that elevation relief on regional scales 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 other cases with high accuracy (AUC = 0.91), namely E6, E8 and E10 it can be suggested that adding more features to a training set of an ML model might not necessarily result in better prediction. In this case, E2 prediction is as good as E6, E7 and E8. However, looking into number of true negatives (correctly predicted non-landslides), it seems E8 slightly performs better than the rest of example sets.

This observation emphasizes the role of feature engineering in ML problems. Feature engineering can reduce the cost of prediction as less number of features may result in highly accurate ML models.

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. LR as an MR 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 their initial stage, the results of the study (AUC greater than 90%) showed that such a framework, with enhanced datasets and perhaps more advanced ML algorithms, can be used for forecasting rainfall-induced landslides and landslide early warning systems at a global and regional scales.

References

- Guzzetti, F., Peruccacci, S., Rossi, M., & Stark, C. P. (2007). Rainfall thresholds for the initiation of landslides in central and southern Europe. Meteorology and atmospheric physics, 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.
- 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.
- 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.
- 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.
- 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.