**Literature Review for Landslide Analysis**

Aliyah Kabeer   
PES University   
Bangalore, India

Paul John  
PES University   
Bangalore, India

Serena A. Gomez  
PES University   
Bangalore, India

Prachi Sengar  
PES University  
Bangalore, India

*Abstract*—This literature review covers our research on previous work on Landslide Datasets and its Analysis using Machine Learning.

Keywords—Machine Learning, Landslide Prevention, Landslide Detection, Analysis

# Introduction

Landslides are the downward movement of rock mass/groundmass/rock blocks by gravity. Landslides are one of the most devastating natural hazards causing huge loss of life and damage to properties and infrastructures and adversely affecting the socioeconomic aspects of the country. Landslides occur in hilly and mountainous areas all over the world. Rainfall, earthquakes, and slope excavation are triggering factors for the occurrence of landslides. Some of the influencing factors of landslides include the topography, geology, hydrology, and land use pattern of the area. In the recent decade, landslide events have increased in both magnitude and frequency due to climate change effect reflected in rainfall patterns. Therefore, it is desirable to identify landslide-susceptible zones for better landslide management and disaster reduction. This will further help in aiding disaster prevention methods.

# Problem Statement

Our problem statement includes predicting the occurrence of different types of landslides induced by rainfall. This aims to answer questions such as - What are the regions most susceptible to landslides? How intense will the landslide be? What type of landslide is it? How fatal will the landslide be?

# Literature Survey

**Rainfall-Induced Landslide Prediction Using Machine Learning Models: The Case of Ngororero District, Rwanda**

In this research paper, two approaches, namely random forest (RF) and logistic regression (LR), were proposed to analyse the rainfall data along with other external and internal factors to develop a prediction model for landslide incidences for an early warning system. These models are evaluated using the receiver operating characteristics, area under the curve (ROC-AUC) and false negative rate (FNR) to measure the landslide cases that were not reported. The results revealed that landslides are triggered by current rainfall but very correlated with consecutive precipitation before the day of incidence (i.e., antecedent rainfall). This is because the soil strength reduces with water content depending on the type of soil material and gets strong again during the evapotranspiration process or the sunny season. It is also noted that among the various internal factors used for prediction, slope angle has the highest impact than other factors. On comparing both the models, the logistic regression proves to be the best approach to be used for landslide prediction and early warning. LR model’s incorrect prediction rate FNR is 9.61% without including antecedent precipitation data and is 3.84% after including the antecedent precipitation data. Therefore, LR model can be used for the early warning system.

**Pros:** Takes 5-days antecedent precipitation into account along with other external and internal factors resulting in a high accuracy and low error rate prediction model.

**Cons:** Does not evaluate based on soil moisture level. Presence of large number of trees in random forest can make the algorithm too slow and ineffective for real-time predictions.

**Machine learning for landslides prevention: a survey**

This research paper titled ‘Machine learning for landslides prevention: a survey’ focuses on landslide prevention by presenting a comprehensive survey of relevant research on machine learning applied in this domain with primary focus on(1) landslides detection based on images, (2) landslides susceptibility assessment, and (3) the development of landslide warning systems. Landslide prediction involves static and dynamic methods to predict landslide occurrences from spatial and temporal perspectives, respectively. Except for satellite images, other available datasets for detecting landslides include bitemporal aerial photographs , LiDAR, and InSAR. Supervised learning is by far the most widespread form of machine learning applied in landslide susceptibility assessment. The most frequently used supervised learning methods include LR ANN, SVM , NB , and DT. Unsupervised learning methods such as cluster sampling can evaluate factors by weighting the relative importance of each conditioning factor. All machine learning-based methods need to extract certain features regarding landslide and non-landslide data samples for analysis and then find a classification boundary to divide the training areas into two classes (i.e., landslides and non-landslides). However, keeping the limitations in check, the survey shows that machine learning methods have been widely used in landslides prevention and can achieve satisfactory performance.

**Pros:** Gives detailed information on how to use different ML and Deep Learning models for landslide prevention. It also points out the potential challenges and limitations of machine learning in landslides prevention and proffers several strategies that have been utilized in other research domains to overcome or circumvent them.

**Cons:** professional knowledge is needed, which can facilitate the selection of more appropriate variables and datasets when facing increasingly complex and massive data.

**Landslide identification using machine learning**

This research paper titled Landslide Identification using Machine Learning aims to propose an integrated landslide identification method which is able to identify both relict and recent landslides. It discusses a novel machine-learning and deep-learning method to identify natural-terrain landslides using integrated geodatabases. To establish the databases for learning, the approach followed includes three components: layer stacking, data extraction and establishment of databases. As this study aims to compare the performance difference of the same machine learning or deep learning algorithm on different types of landslides, three different landslide inventories are formed; namely, the Relict Landslide Inventory (RelLI), the Recent Landslide Inventory (RecLI) and the Joint Landslide Inventory (JLI). The raw data is processed in stages, starting from deriving landslide-related predictors from the raw data. Across the above mentioned three databases, due to its strengths in feature extraction and processing multi dimensional data, CNN proves to have the highest accuracy. Boosting methods come second in terms of accuracy, followed by RF, LR and SVM. Limitations of this paper are some inconsistencies in the terrain and landslide data. The relict landslide records of ENTLI are not completely accurate. Third, CNNs with more layers should be investigated in the future when higher computational power is available.

**Pros:** It can identify both relict and recent landslides using an integrated landslide identification method. It used logistic regression (LR), support vector machine (SVM), random forest (RF), boosting methods and convolutional neural network (CNN) and compared the accuracies to choose the best model.

**Cons:** Inconsistencies in the terrain and landslide data. The relict landslide records of ENTLI are not completely accurate, and CNNs with more layers should be investigated in the future when higher computational power is available.

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

In the research paper titled ‘A framework for predicting rainfall-induced landslides using machine learning methods’, the preliminary results of a practical research study that has been carried out in Deltares, Netherlands has been presented. It involves a framework that combines geoengineering, remote sensing, hydrology with machine learning to predict the onset of landslides under the effect of rain and precipitation. The main focus is on supervised learning methods for classification of landslide and non-landslide events. The trained ML model is then fed by rainfall data, topography features such as slope, elevation relief, soil and bedrock data, and vegetation index of target regions to assess the stability of the studied area. Using logistic regression as the ML model, the relationship between the probability p of the landslide and the triggering and controlling factors (features) are investigated and expressed mathematically and it performs binary classification to distinguish landslides and non-landslide cases. At the end of the paper, it is observed after comparing the AUC values that having rainfall data type fixed, elevation relief can be more effective than slope angle in producing a LR model with higher accuracy, through which the paper concluded that elevation relief, when dealing with large scale zones, can be more representative of the topography of the region than slope angle. Though in its initial stage in terms of its database and forecasting framework, this research 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.

**Pros:** It combines geoengineering, remote sensing, hydrology with machine learning to predict the onset of landslides under the effect of precipitation

**Cons:** the database and forecasting framework that were reported in this study are at its initial stage

**Landslide Susceptibility Mapping Using Single Machine Learning Models: A Case Study from Pithoragarh District, India**

Three single ML models, namely, Linear discriminant analysis (LDA), logistic regression (LR), and radial basis function network (RBFN), for landslide susceptibility mapping. LDA is a monitoring algorithm which calculates the linear differentiator by maximizing the interval between classes and minimizing the interval between the categories. LR is a supervised learning algorithm mainly used for binary classification. In landslide modeling using this method, the presence of landslides is considered as “1” and the absence of landslides is considered as “0”. RBFN is a simulator that consists of radial subordinates. RBFN can be applied to achieve performance approximation and pattern recognition. In this study, the receiver operating characteristic (ROC) curve and a set of statistical indices were used for model validation.

**Pros:** Twelve landslide-variables were generated for landslide susceptibility modeling, which include altitude, lithology, distance to faults, normalized difference vegetation index (NDVI), landuse/landcover (LULC), distance to roads, slope angle, distance to streams, profile curvature, plan curvature, slope length (LS), and slope-aspect

**Cons**: Precipitation and earthquakes are two landslide triggering factors that were not considered in this study.

**Mapping landslide susceptibility and types using Random Forest**

This paper presents a data mining approach to producing LSM for a large, heterogeneous region that is susceptible to multiple types of landslides. Using a case study of Piedmont, Italy, a Random Forest algorithm is applied to produce both susceptibility maps and classification maps. These maps are combined to give a highly accurate LSM (over 85% classification accuracy) which contains a large amount of information and is easy to interpret. Of the predictor variables for susceptibility, distance from rivers and TWI are most important. This novel method of mapping landslide susceptibility demonstrates the efficacy of Random Forest to produce highly accurate susceptibility maps for a large heterogeneous region without the need for multiple susceptibility assessments.

**Pros:** Uses two stage LSM - binary and multi-class prediction. Use of RF data mining algorithm - non-linear and non parametric, can deal with large datasets containing both categorical and numerical data. It can also handle missing values and maintain accuracy for missing data. It does not require much fine tuning of hyper-parameters like ANN and SVM thus increasing throughput. It can also rank predictor variables in order of their importance hence increasing performance speed. Bagging is used to reduce the variance of a decision tree which increases the performance of the model.

**Cons**:

Although RF ranks variables in order of importance, there is no way of knowing the process that the variables represent. Nor is there any way of determining a definitive ranking of variables as this will depend on the location of the case study, type of model, resolution of training data, and sampling scheme. Hence it is overfitted to the specific scenario and thus cannot be used in neighbouring regions.

**Landslide Susceptible Locations in Western Ghats: Prediction through openModeller**

In this paper, the probable landslide prone areas are predicted using two different algorithms – GARP (Genetic Algorithm for Rule-set Prediction) and Support Vector Machine (SVM) in a free and open source software package - openModeller. Several environmental layers such as aspect, digital elevation data, flow accumulation, flow direction, slope, land cover, compound topographic index, and precipitation data were used in modelling. A comparison of the simulated outputs, validated by overlaying the actual landslide occurrence points showed 92% accuracy with GARP and 96% accuracy with SVM in predicting landslide prone areas considering precipitation in the wettest month whereas 91% and 94% accuracy were obtained from GARP and SVM considering precipitation in the wettest quarter of the year.

**Pros:** Higher accuracy because the model uses two different precipitation layers were used to predict landslides – precipitation of wettest month and precipitation in the wettest quarter of the year along with the seven other layers; and also multiple parameters such as slope, land cover, compound topographic index, precipitation data etc.

**Cons**: A subject matter expert (SME) is required to ensure that the algorithm works efficiently.

**A Research on Deep Learning Advance for Landslide Classification using Convolutional Neural Networks**

In the past years, many researchers have shown their immense interest in the area of landslide. Landslide classification using the deep neural network is a newly pursued subject in which landslide is detected in the images which will be helpful in distinguishing landslide regions in maps and through images of sites prone to landslides. Truong et al. have used ensemble technique which is a fusion of Bagging Ensemble technique which is a fusion of Bagging Ensemble and Logistic Model Tree for refining the performance of landslide susceptibility model. Oh and Lee proposed data mining approaches such as Artificial Neural Network and Boosted Tree for landslide susceptibility modeling giving the validation result of 82.5% and 90.79% respectively. Ramesh discusses the design and deployment of a landslide detection system using a Wireless Sensor Network system to detect landslides. To detect landslides of five different types Martha, Kerle, Western, Hetten, and Kumar presents an intelligent classification approach with the thresholds of diagnostic parameters derived from if-means cluster analysis with an overall test accuracy of 77.7%.

**Pros:** Takes into consideration a wide variety of images from the dataset to improve the model.

**Cons**: Does not take into account many of the factors responsible for landslides.

**Mine landslide susceptibility using IVM, ANN and SVM models.**

The main objective of present study is to evaluate and compare the performance of feature selection arithmetic and three assessment methods, including two machine learning models: ANN, SVM and one conventional statistical model: IVM, for mine landslide susceptibility assessment. The uncertainty of the models is analyzed based on the resampling techniques and the rank probability score. For this reason, we extract evaluation factors from remote sensing images and spatial data, which are then represented by three methods, respectively. These models were evaluated using the landslide dataset of Shangli county, China. Analysis of landslide data and model construction have been carried out using ArcGIS 10.2 and Tensorflow 1.2 software. The area with high-prone landslide will be identified and the causes will be discussed in this study.

**Pros:** Large dataset and may features recorded. Also provides multicollinearity analysis.

**Cons**: Discards many less important factors that could play a role in determining the accuracy of the model.

**Machine Learning Based Early Prediction of Rainfall Induced Landslide**

In this system, Logistic Regression(LR) algorithm is used along with receiver operating characteristics, area under the curve(ROC-AUC) and false negative rate(FNR). Various factors triggering a landslide are considered as in input. The target variables or classes considered are:- landslide and non-landslide. Probability of landslide incidence is computed. The evaluation of explained variance of landslide triggering factors to landslide predictive model has bas been done. A high explained variance is achieved here. ROC curve is used to evaluate the model along with AUC.

**Pros:** Various factors are considered including rainfall, slope, geology, geomorphology, distance from road, distance from stream, distance from lineaments and deforestation.

**Cons**: Does not consider factors like density of the forest, properties of soil(mositure content etc). The study is conducted specifically for Idukki district, Kerala.

**Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia**

To perform landslide susceptibility mapping, two datasets were considered. The first dataset represents the landslide inventory map. The second dataset is related to landslide conditioning factors. An inventory map of existing landslides was prepared by integration of the historical records, field investigations, and data extracted from satellite images interpretation. Seven advanced machine learning techniques that vary in their degree of complexity were applied to evaluate their efficacy in landslide susceptibility mapping. They include SVM, RF, MARS, ANN, QDA, LDA, and NB. Landslide susceptibility index (LSI) for every pixel in entire study area was calculated. These landslide susceptibility maps (LSMs) were reclassified into four susceptibility categories including low, moderate, high, and very high susceptible zones, using the natural breaks (Jenks) classification method. The receiver-operating characteristic-area under the curve (ROC-AUC), and the root mean square error (RMSE) were applied to evaluate and measure the predictive performance of the MLTs. This analysis revealed that MARS gives the best performance.

**Pros:** Twelve landslide-variables were generated for landslide susceptibility modeling, which include altitude, lithology, distance to faults, normalized difference vegetation index (NDVI), landuse/landcover (LULC), distance to roads, slope angle, distance to streams, profile curvature, plan curvature, slope length (LS), and slope-aspect

**Cons**: Precipitation and earthquakes are two landslide triggering factors that were not considered in this study.

**Novel GIS Based Machine Learning Algorithms for Shallow Landslide Susceptibility Mapping**

The main purpose of this study was to evaluate the efficiency of several ensemble techniques (MB, BA, RF, and RS) in improving the performance of a base classifier, namely Alternating Decision Trees (ADTree). The difference between this study and earlier studies is using two scenarios of the combination of sample size and raster resolution.

**Pros:** Various features recorded and taken into consideration. Model provides high accuracy for certain sets.

**Cons**: Various features recorded and taken into consideration. Model provides high accuracy for certain sets.

##### Conclusion

After surveying various different research papers on landslide detection, analysis and prevention, we conclude that the approach to our project is different from what’s already out there in the following ways: It gives information about how intense the landslide will be, about what type of landslide will occur and about how fatal the landslide will be. We also conclude from our study that we wish to leverage various machine learning models for classification to perform the above predictions as it has proven to be efficient and accurate in the past.

##### References

**Links to the research papers surveyed:**

<https://link.springer.com/article/10.1007/s00521-020-05529-8>

<https://www.sciencedirect.com/science/article/pii/S1674987120300542>

<https://www.ecsmge-2019.com/uploads/2/1/7/9/21790806/0521-ecsmge-2019_tehrani.pdf>

<https://www.mdpi.com/1660-4601/17/11/4147>

<https://www.tandfonline.com/doi/full/10.1080/20964471.2018.1472392>

<https://www.researchgate.net/publication/256838741_Landslide_Susceptible_Locations_in_Western_Ghats_Prediction_through_OpenModeller>

<https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F11840486S419.pdf>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0215134>

<https://www.xajzkjdx.cn/gallery/12-april2021.pdf>

<https://www.sciencedirect.com/science/article/pii/S1674987120301298>



**Data Analytics Project (UE19CS312)**

**Title:**

Predicting occurrence of different types of rainfall-induced landslides

**A Project By:**

|  |  |
| --- | --- |
| **Name** | **SRN** |
| Aliyah Kabeer | PES2UG19CS031 |
| Paul John | PES2UG19CS275 |
| Prachi Sengar | PES2UG19CS285 |
| Serena A. Gomez | PES2UG19CS372 |

**Introduction**

Landslides are the downward movement of rock mass/groundmass/rock blocks by gravity. Landslides are one of the most devastating natural hazards causing huge loss of life and damage to properties and infrastructures and adversely affecting the socioeconomic aspects of the country. Landslides occur in hilly and mountainous areas all over the world.  Rainfall, earthquakes, and slope excavation are triggering factors for the occurrence of landslides. Some of the influencing factors of landslides include the topography, geology, hydrology, and land use pattern of the area. In the recent decade, landslide events have increased in both magnitude and frequency due to climate change effect reflected in rainfall patterns. Therefore, it is desirable to identify landslide-susceptible zones for better landslide management and disaster reduction. This will further help in aiding disaster prevention methods.

**Problem Statement**

Predicting occurrence of different types of landslides induced by rainfall.

Questions to be answered:

* What are the regions most susceptible to landslides?
* How intense will the landslide be?
* What type of landslide is it?
* How fatal will the landslide be?

In our project we wish to leverage Support Vector Machine (SVM) and logistic regression (LR) for classification to perform the above predictions.

The approach to our project is different from what’s already out there in the following ways:

* Gives information about how intense the landslide will be.
* Gives information about what type of landslide will occur.
* Gives information about how fata the landslide will be.

**Exploratory Data Analysis**

**The Dataset:**

Name: Landslides after rainfall (2007- 2016)

Source: Kaggle

The dataset consists of 23 columns (features or variables) and 1694 rows(recordings).

Data types and variables:

Numeric variables: id, population, distance, latitude, longitude, fatalities, injuries.

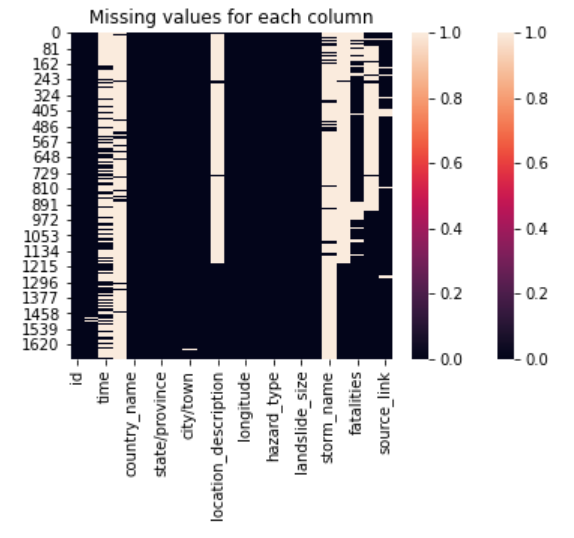
Categorical variables: continent\_code, country\_name, state/province, city/town, location\_description, hazard\_type, landslide\_type, landslide\_size, trigger, storm\_name, source\_name, source\_link.

Temporal data: date, time.

Spatial data: latitude, longitude, geolocation.

**Data Cleaning:**

Heatmap to detect presence of null values:



Performed the following operations for cleaning:

* Drop continent\_code column.
* Clean trigger column by changing all ‘unknown’, ‘other’, ‘downpour’ to ‘Unknown’, ‘Unknown’, ‘Downpour’; and by filling all null rows with ‘Unknown’.
* Clean location\_description column by changing all ‘Other’ to ‘Unknown’; and by filling all null rows with ‘Unknown’.
* Cleaning time column by changing all ‘evening’ to ‘Evening’; and by filling all null rows with ‘Unknown’.
* Cleaning Landslide\_type column by changing all ‘landslide’, ‘mudslide’, ‘Unknown’ to ‘Landslide’, ‘Mudslide’, ‘Other’ respectively; and by filling all null rows with ‘Other’.
* Cleaning Landslide\_size column by changing all ‘large’, ‘medium’, ‘small’ to ‘Large’, ‘Medium’, ‘Small’ respectively.
* Cleaning source\_name and source\_type column by filling all null rows with ‘Unknown’.

Parsing Dates:

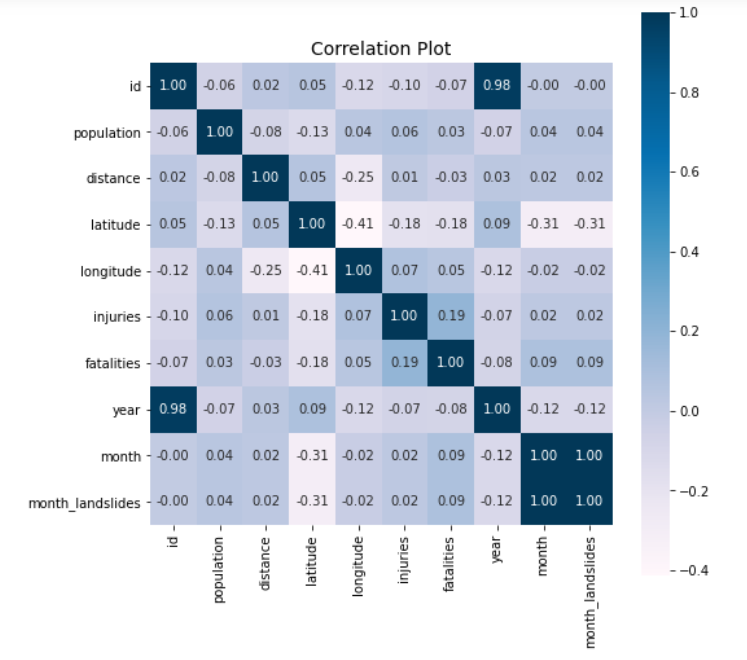
Drop all rows with null dates.

Perform the following operations after paring the date:

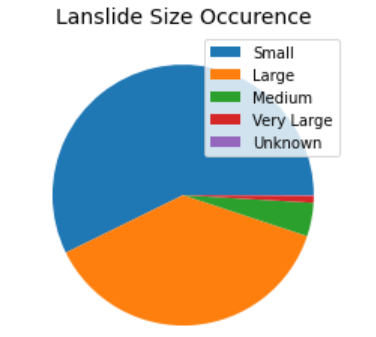
* Create a new year and month column
* Add a column with abbreviated months
* Create month\_landslides column
* Create grouped\_df dataframe (grouped by month\_lanslides and year)

**Data Visualisation:**

Correlation plot:

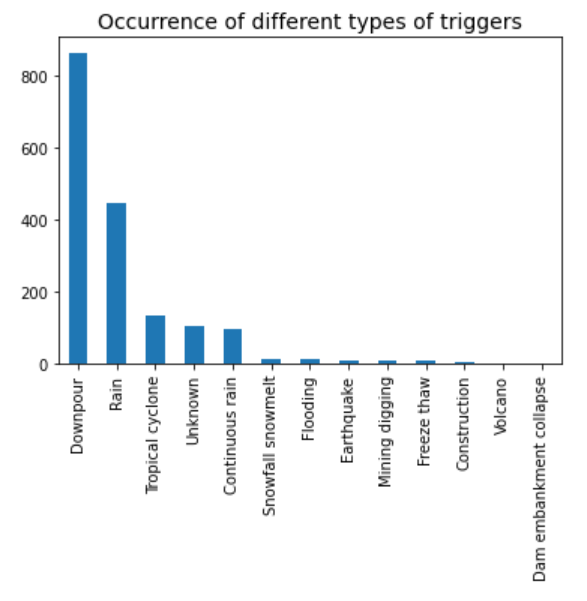


What size of landslides are most likely to occur?



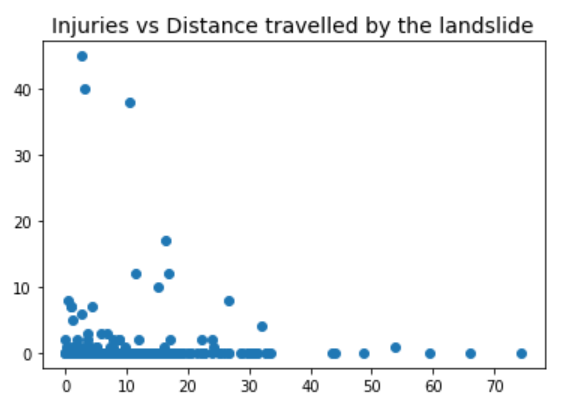
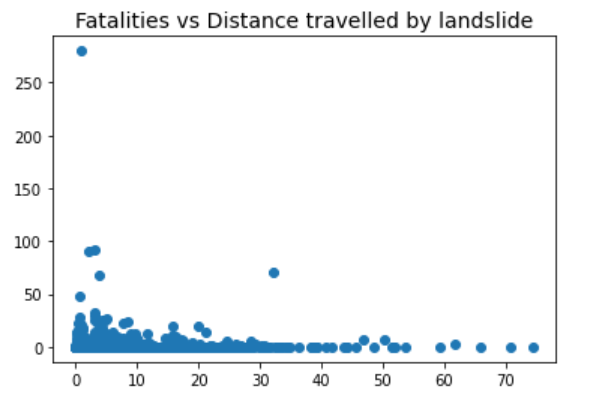
Inference: From the above chart we can infer that medium sized landslide are most likely to occur and very large landslides are least likely to occur.

Occurrence of different types of triggers of landslides:



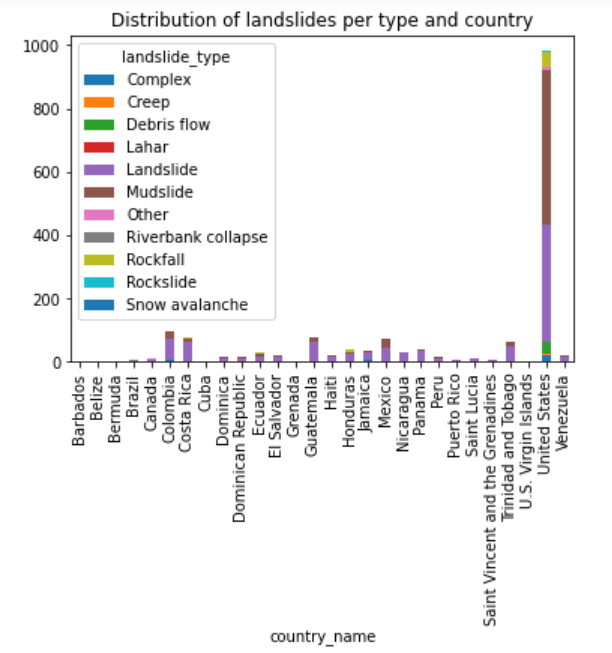
Inference: From the above plot we can infer that landslide are most likely to be caused by downpour followed by rain, and it is least likely to be caused by a volcano and embarkment collapse

Number of fatalities and injuries that occurred vs the distance travelled by the landslide:



Inference: From the above plots we can infer that landslide that happen over a shorter distance are more fatal and injurious than the ones that have been recorded over longer distances.

Distribution of landslide per type and country:

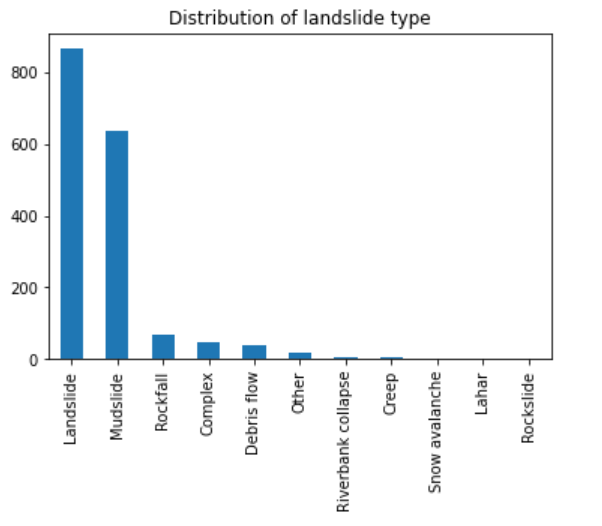


Inference:

From the above plot we can infer:

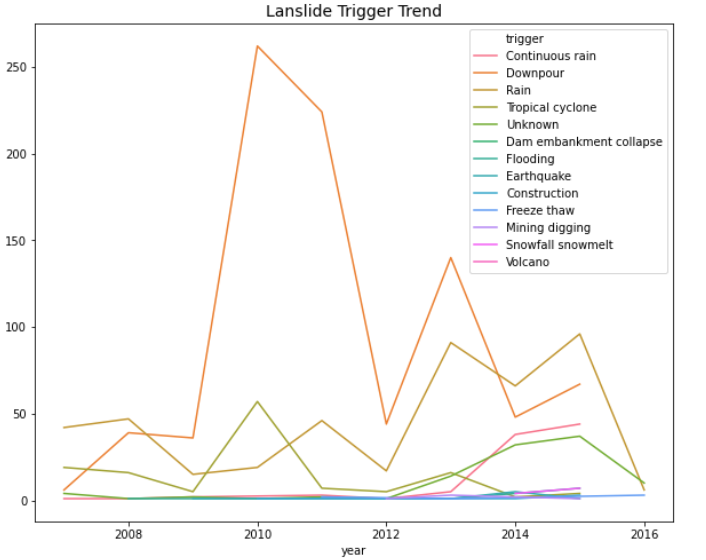
* Landslides most commonly occur in the United States.
* Landslides and mudslides are the types of landslides that are most likely to occur.

Distribution of landslide type:



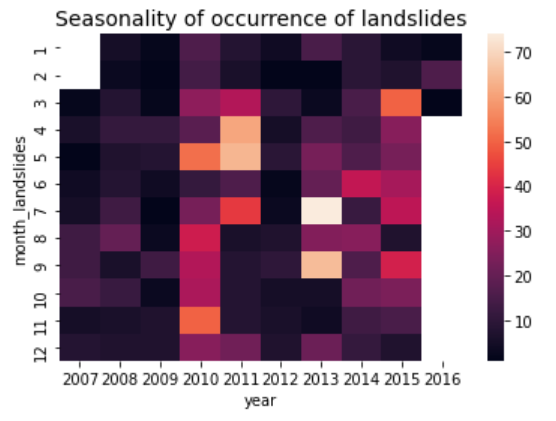
Inference: From the above plot we can infer that landslide are the most common type of landslide to occur followed by mudslides

Distribution of landslides per trigger per year:



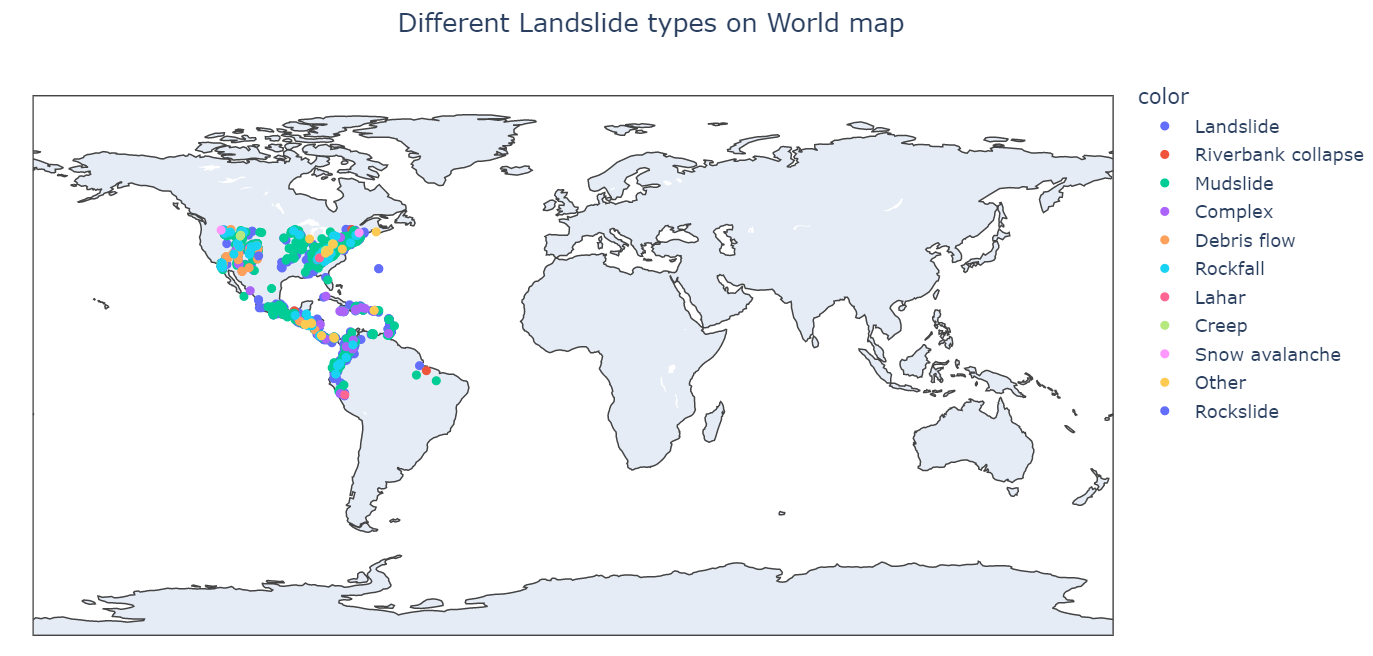
Inference: From the above plot we can infer that landslide are most commonly triggered by downpour followed by rain.

To check if there is a seasonality of occurrence of landslides:



Inference: There is no clear seasonality that can be visualized in the heatmap. However, in the boxplots there is a seasonality where in the beginning and end of the year there are less landslides than during the mid of the year.

To analyse the region of occurrences of various types of landslides and the number of fatalities they caused:



Inference:

From the above plot we can infer:

* Most common landslide types are landslides and mudslides.
* Most number of fatalities are caused by mudslides.
* Landslides are most likely to occur in regions closer to waterbodies.

**Literature Survey**

Google sheets link: <https://docs.google.com/spreadsheets/d/1_3AHfY87QwdvS5wZymjp05ifsFGr6D29kXPp_PP4KPo/edit#gid=0>

**References**

Github link for code: <https://github.com/Killawog/DAProject>