Journal of Engineering and Applied Sciences 13 (Special Issue 1): 2301-2308, 2018
ISSN: 1816-949X
© Medwell Journals, 2018

Landslide Prediction Using Classifier Models

Sukriti Paul, Arpit Garg and S. Chethan
Department of Information and Communication Technology,
Manipal Institute of Technology, Manipal, India

Abstract: Prediction systems have been flooding the IT industry ever since the concept of information retrieval came into existence. Every organization uses one of these prediction systems which helps the users of the system to predict useful information on the basis of their past records. The landslide prediction system aims to use this concept for not just different places but different terrains as well. NASA maintains the record of the landslides that have happened in the past, this record includes details such as area, terrain, rainfall, population, vegetation cover, etc. If this data can be classified on the basis of the patterns mined, it becomes very easy for the authorities to inform the people living in a particular area about landslides or development of a colony on a landslide prone area can be avoided. Not only that, reasons for landslide occurrence can be taken note off and be avoided. For example, lack of vegetation is causing landslides in an area. If this data is provided and wisely used, lives lost can be greatly reduced. This study aims to find ways to obtain this data.

Key words: Landslide, prediction, Naive Bayes, random forest, population, vegetation

INTRODUCTION

The outburst of population in recent year has resulted in a dire need to have controlled development projects. Recent deforestation, global warming, etc. have resulted in landslides as there are no plants to hold the soil. Even the uneven rain patterns due to global warming has been a crucial factor that has resulted in landslides. In addition to the population serving as a factor contributing to landslides, the terrain of a region is sometimes not strong enough to sustain population more than the maximum threshold. Using predictive analysis on a database obtained from NASA if we can come up with a prediction system to predict and review certain patterns that cause landslides, a lot of lives can be saved. Apart from that the amount of money spent on reconstruction and debris removal on the landslide affected area can be avoided. We have analyzed landslides globally, scaling over two continents and various countries belonging to them.

Data set: We have accessed the database from NASA website (NASA., 2017). It basically has location, types, reasons of landslide, population, etc. The database has been collected from year (2007-2015) (Fig. 1).

MATERIALS AND METHODS

For each of the cases mentioned earlier, we will be suggesting ways to achieve the goal and apply them on the data set. Some of them just involve a better visualization of the data set. To start with we have done data preprocessing using scaling, normalization, Yeo-Johnson transform and independent component analysis. Now, we will be comparing the different ways used in each of these methods. For example for the first case to identify the places prone to landslides in general would require a simple count vs. landslide type graph. However, clustering the data would provide a better idea of the geographical boundaries prone to landslides. Pros and cons will be discussed in each of the cases in their respective study. To generate results we will be using R for the purpose of demonstration and RapidMiner Studio is used for predictive analysis and classification tree. If a landslide prediction system is to be developed, it can be developed using any of the programming technologies that the developer feels comfortable in.

Data preprocessing: Data needs to be cleaned owing to inconsistencies and missing values. Certain issues encountered on using raw data are missing attributes, missing attribute values, values which are out of

Corresponding Author: Sukriti Paul, Department of Information and communication technology, Manipal Institute of Technology, Manipal, India

date	landslide_type	trigger	landslide_size	distance	population	changeset_id	latitude	longitude	geolocation
09/25/2015	Landslide	Rain	Small	1.16705	335007	3456375012	9.9402	-84.0771	(9.9402000000000008, -84.077100000000002)
9/9/2007	Landslide	Rain	Medium	0.26208	21947	1	10	-84.1167	(10, -84.116699999999994)
9/9/2007	Landslide	Rain	Medium	2.59849	16571	1	15.3055	-61.3642	(15.3055, -61.364199999999997)
2/3/2016	Rock_Fall	Unknown	Small	1.78429	2066	4281763033	43.4771	-72.4066	(43.4771, -72.406599999999997)
02/27/2016	Rock_Fall	Unknown	Small	5.07093	2184	580117165	37.3287	-81.3134	(37.328699999999998, -81.313400000000001)
09/27/2015	Landslide	Rain	Medium	1.83863	994938	3858422560	14.6217	-90.4956	(14.621700000000001, -90.495599999999996)
06/23/2014	Landslide	Continuous_rain	Small	31.14242	5827	1	12.3535	-84.8095	(12.3535, -84.8095)
09/27/2015	Landslide	Rain	Medium	2.08425	994938	1893779508	14.6219	-90.5119	(14.6219, -90.511899999999997)
06/23/2014	Landslide	Continuous_rain	Medium	32.77401	5827	1	12.352	-84.7932	(12.352, -84.793199999999999)
12/9/2013	Landslide	Rain	Medium	6.81843	2400	1	39.8839	-105.3033	(39.883899999999997, -105.30329999999999)
12/9/2013	Landslide	Rain	Medium	3.762	2400	1	39.8958	-105.3357	(39.895800000000001, -105.3357)
02/27/2016	Rock_Fall	Rain	Small	24.16064	485	4112533001	44.104	-116.0033	(44.103999999999999, -116.0033)
02/27/2016	Rock_Fall	Unknown	Small	22.78728	939	3766295569	44.3127	-116.076	(44.3127, -116.07599999999999)
02/26/2016	Rock_Fall	Unknown	Small	12.00678	1048	3659565428	37.5011	-81.1093	(37.501100000000001, -81.109300000000005)
1/12/2015	Landslide	Rain	Medium	4.68732	14400	2310518521	18.134	-76.4551	(18.134, -76.455100000000002)
06/23/2014	Landslide	Continuous_rain	Medium	28.90294	5827	1	12.3129	-84.8199	(12.312900000000001, -84.819900000000004)
06/23/2014	Landslide	Continuous_rain	Medium	29.95253	5827	1	12.349	-84.8195	(12.349, -84.819500000000005)
07/30/2011	Mudslide	Rain	Small	10.57117	9614	1	39.4977	-107.2225	(39.497700000000002, -107.2225)
07/30/2011	Mudslide	Rain	Small	4.90954	3801	1	39.4274	-107.1291	(39.427399999999999, -107.12909999999999)
07/27/2011	Debris_Flow	Downpour	Small	17.95414	1297	1	32.6209	-107.9645	(32.620899999999999, -107.9645)

Fig. 1: NASA database

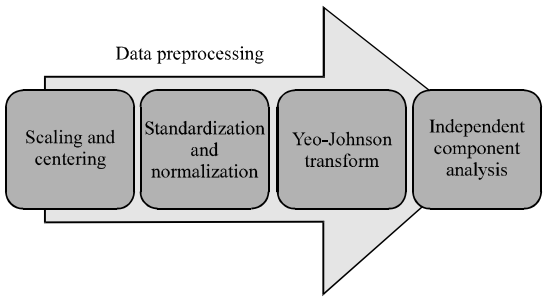

Fig. 2: Standard preprocessing technique

range or improper data types (numeric/factor). Certain methodologies by Famili *et al.* (1997) have been implemented via. R for data preprocessing using the carat package. Figure 2 explains the standard preprocessing technique flow followed. Independent component analysis is essential while using the Naive Bayes classifier. The Yeo-Johnson transform (Yeo and Johnson, 2000), a power transform taking zero and negative data values into consideration is used while normalization is used when the attribute range is very large.

Landslide analysis methods: We have used R to find the relation between:

- Landslide count and latitude
- Landslide count and longitude
- Population and latitude
- Population and longitude
- Trigger and latitude

Count vs. latitude: The graph in Fig. 3 shows count of different landslide types at particular latitude. Using this graph, we can find out what kind of landslide is more frequent on which latitude.

Count vs. longitude: The graph in Fig. 4 shows count of different landslide types at a particular longitude. Using this graph, we can find out what kind of landslide is more frequent on which longitude. Using the information gathered in section A and this study, we can find the area which is more prone to a particular area. The same observation can be used to deploy steps to avoid that kind of landslide.

Population vs. latitude: The graph in Fig. 5 shows landslide size depending on the population at particular latitude. Using this graph, we can find out what intensity of landslide is dependent on population number on a particular latitude.

Population vs. longitude: The graph in Fig. 6 shows landslide size depending on the population at a particular longitude. Using this graph, we can find out what intensity of landslide is dependent on population number on a particular longitude. And using the data collected in section C and this study, we can find out the area that is densely populated and is prone to large landslides. This will help us to warn people living in that area about the situation, this will also help in future development plans. For example, if an area with low population is prone to large landslides, major development projects are of no use in that area as it can't sustain large population.

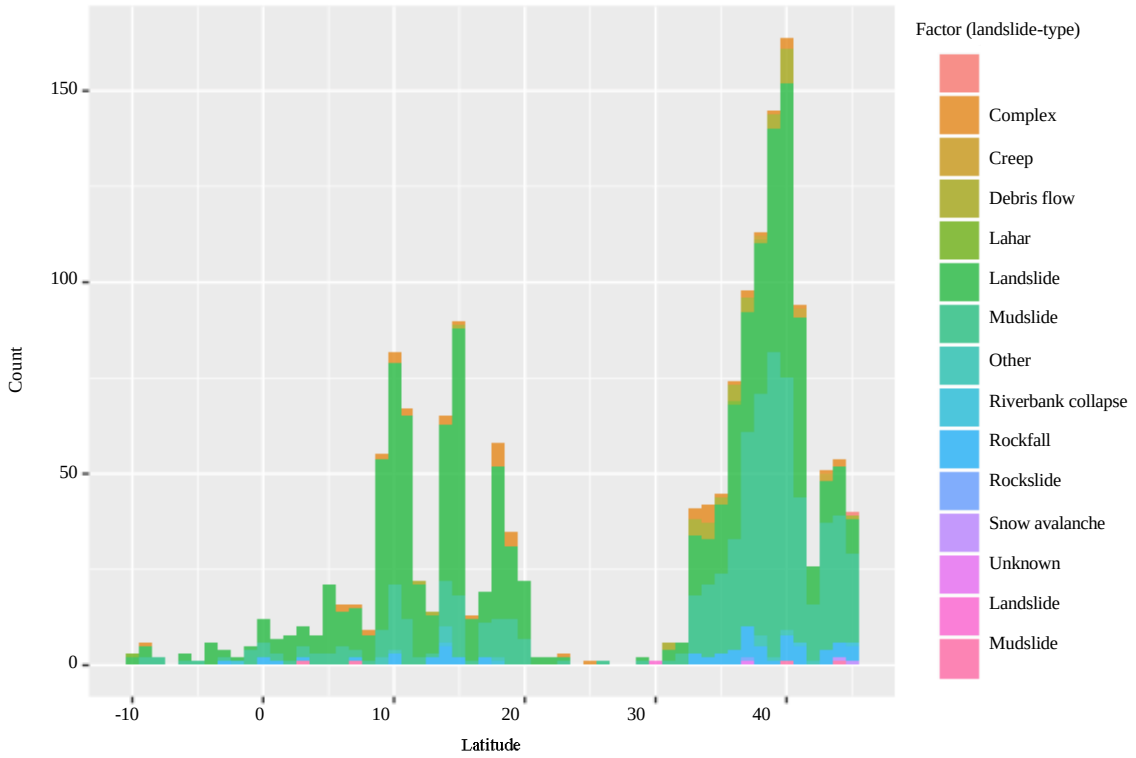

Fig. 3: Count vs. latitude

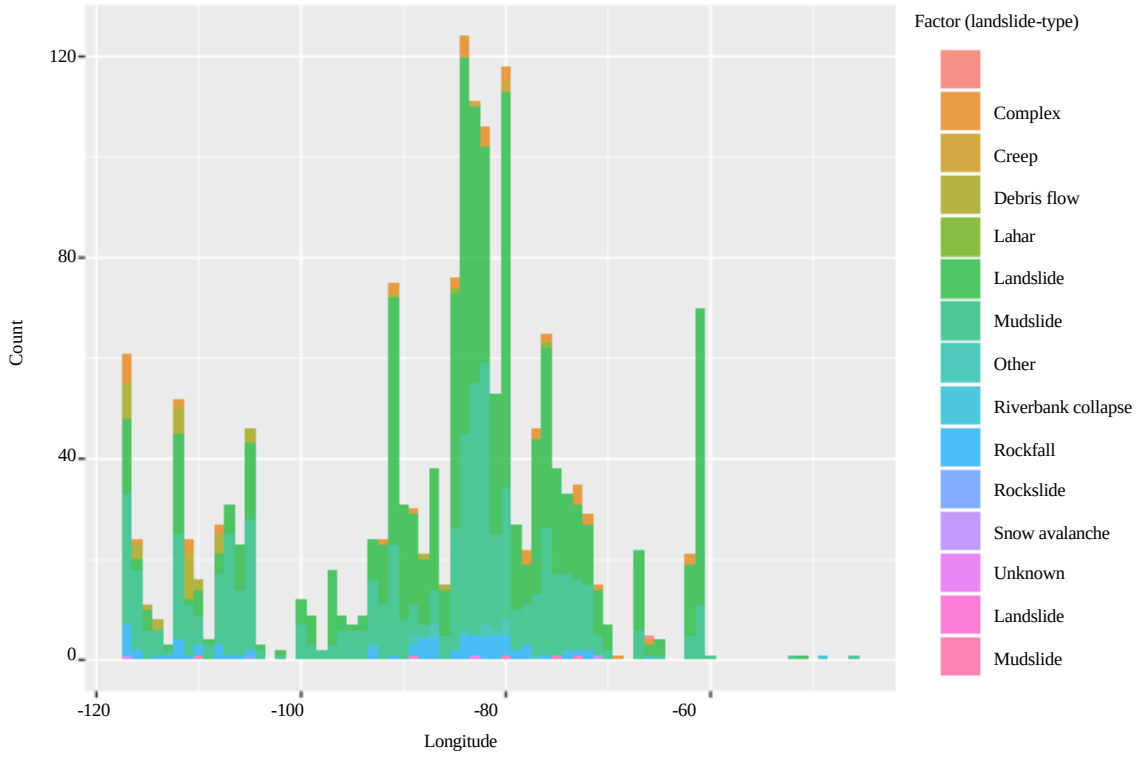

Fig. 4: Count vs. longitude

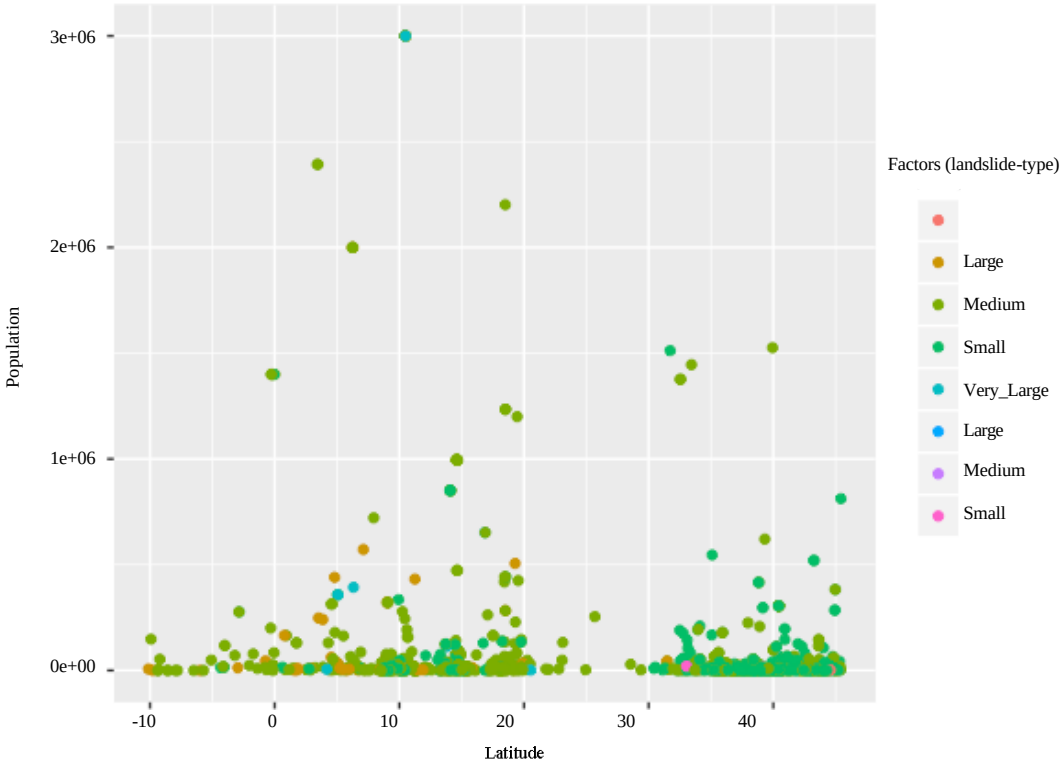

Fig. 5: Population vs. latitude

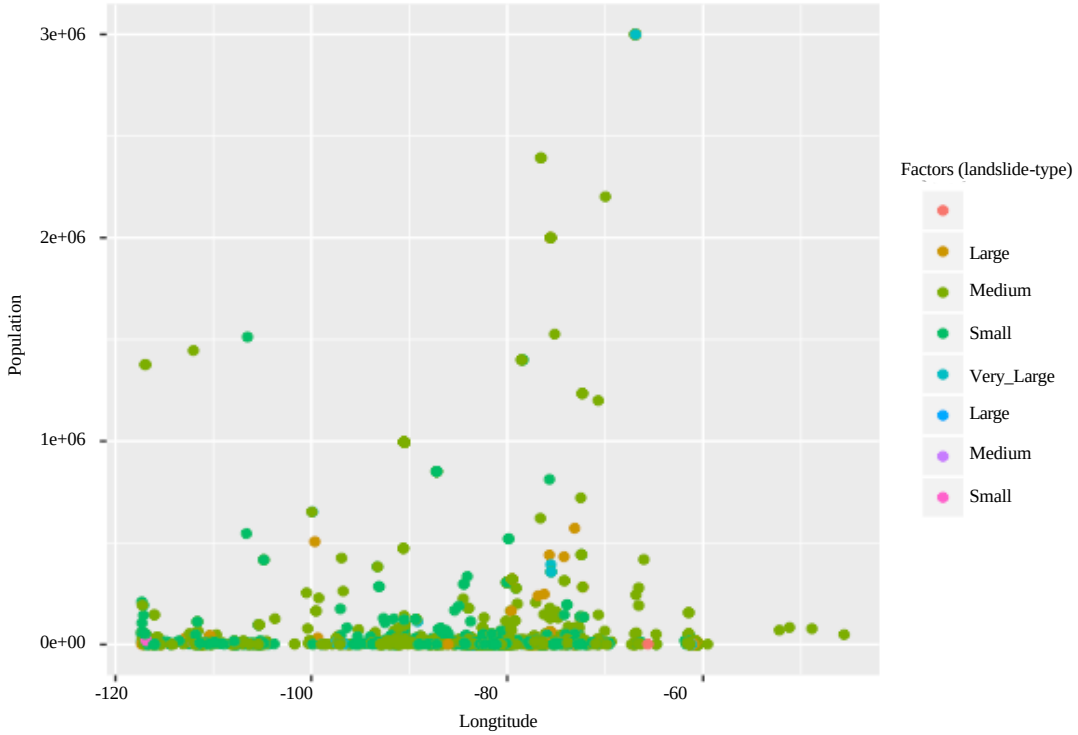

Fig. 6: Population vs. longitude

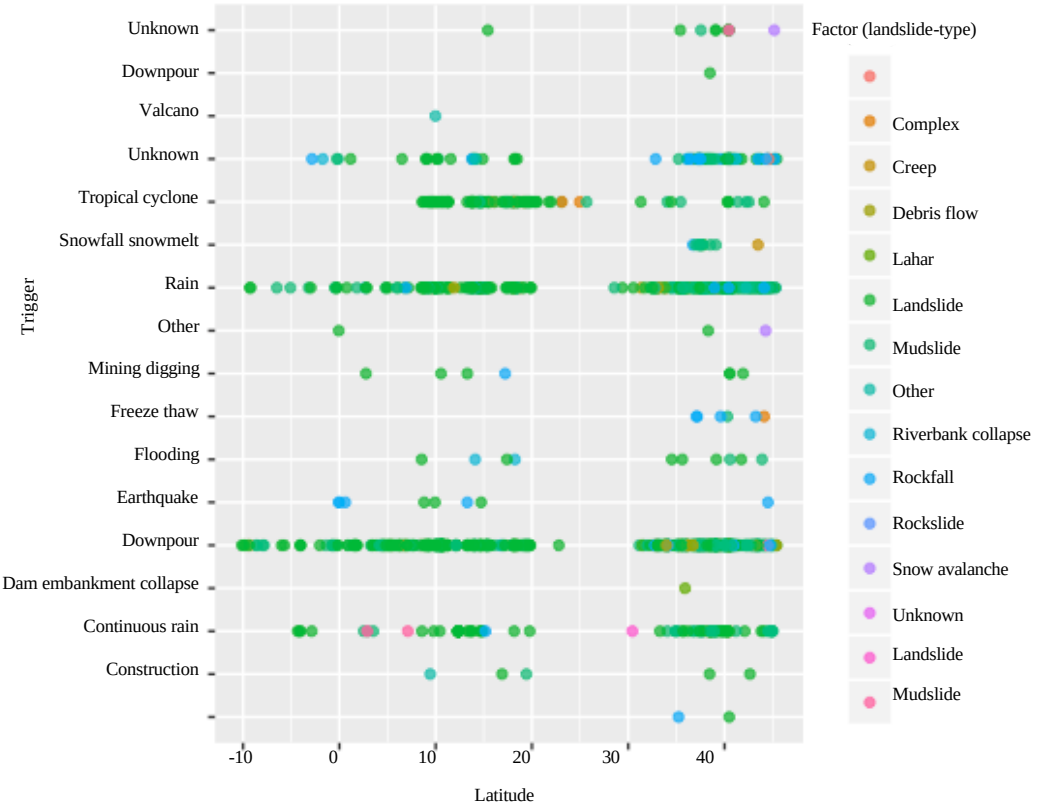

Fig. 7. Trigger vs. latitude

Trigger vs. latitude: The graph in Fig. 7 shows the triggering factor of a particular landslide on a latitude. Triggering factor could be rain, earthquake, construction, etc.

RESULTS AND DISCUSSION

Landslide prediction models: We have used RapidMiner for predictive analysis. A classifier consists of a function which helps in categorizing an object with feature vector $X = (x_1, x_2, ..., x_n)$ to its corresponding class or state $Y = \{1, 2, 3, ..., C\text{-}1, C\}$, where n is the number of features and C is the total number of classes. Labeling the training data is crucial in supervised training and classification is based on the same. We have considered two classifiers for prediction. In RapidMiner, we have used Naive Bayes and random forest. In Naive Bayes prediction model, we have used Naive Bayes theorem to predict certain landslides. This is a probabilistic model, centered around conditional probability. The posterior probability $Pr(C_k|x)$ is predicted based on prior probability $Pr(C_k)$, likelihood $Pr(x|C_k)$ and evidence $Pr(x)$ as shown:

$$Pr(C_k|x) = \frac{Pr(C_k)\, Pr(x|C_k)}{Pr(x)}$$

On assuming that the feature vectors which are used to determine the class of the object are not dependent, the equation or decision rule representing the model can be written as follows:

$$Pr(C_k|x_1, ..., x_n) = \frac{1}{Z} Pr(C_k) \prod Pr(x_i|Ck)$$

where, Z is a scaling factor equal to Pr (x). According to Brownlee (2016), Naive Bayes research well with both completely independent features and functionally independent features while it has its worst performance between these limits. While predicting landslide, we have considered the trigger, population and latitude as our features for classification as shown in Fig. 8. Most of the landslides are due to rainfall and downpour in averagely populated areas.

The landslide size is average for most of the cases in rainfall and downpour while for others it's scattered but size is more as shown in Fig. 9.

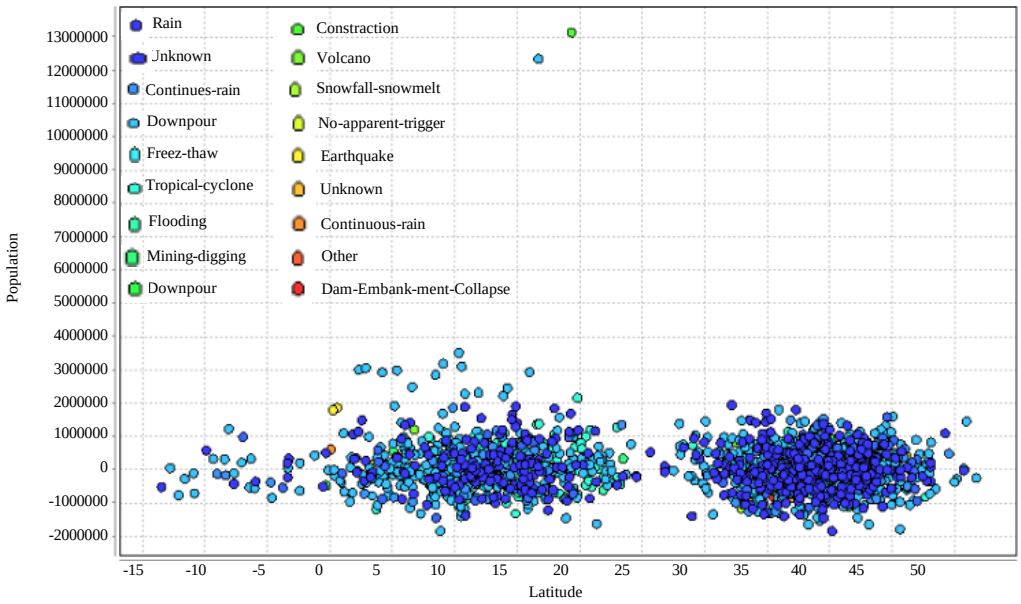

Fig. 8: Trigger, population and latitude

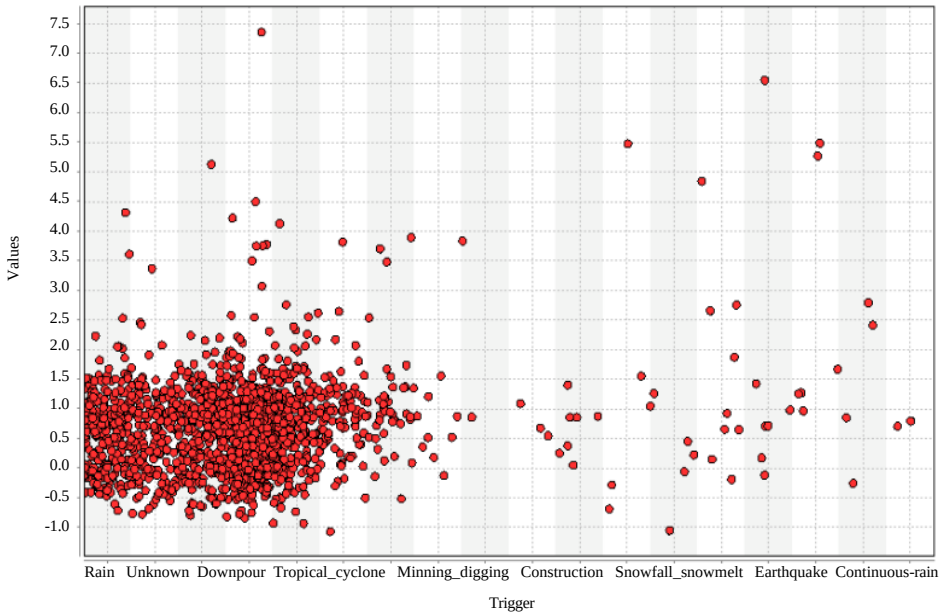

Fig. 9: Trigger and landslide size

We have also used random forest algorithm as showing in Fig. 10 for tree based classification. This is an ensemble learning algorithm which uses tree-type classifiers {h (x, ⊖ k), k = 1, ...,} where the {⊖ k} are independent identically distributed random vectors and x is an input pattern (Pal, 2003). Several trees which are synonymous with Classification And Regression Trees (CART) are generated and trained based on data samples. The split point of each node is an arbitrarily selected subset of data and the classification is based on a majority vote casted by the trees for the most popular and probable state of the object at input x. A single vote is casted by each tree. Gini index is observed to be the most optimal splitting criteria (Rodriguez *et al*., 2012). The parameters considered are the number of decision trees used and the number of attributes taken for classification and prediction.

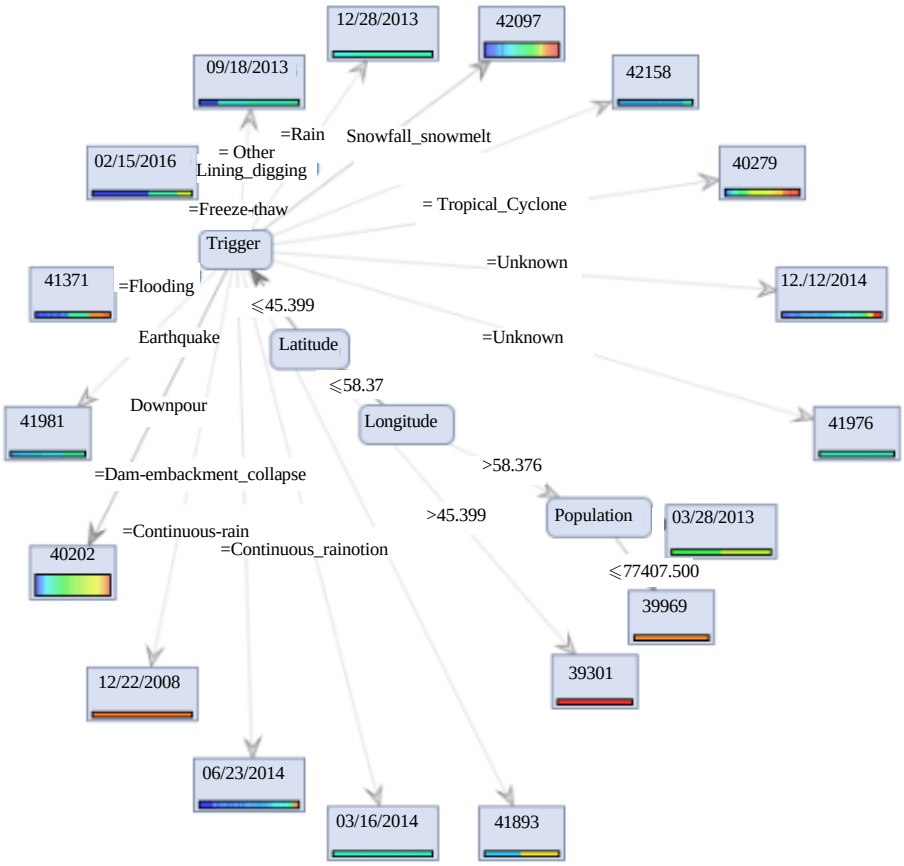

Fig. 10: Random forest

CONCLUSION

In this study, we discussed different ways a landslide predictive system can be modeled. We also discussed scenarios where this system will be used. Most of the prior research carried out in this domain focuses on the impact of landslides in a localized, smaller region of study like Selangor (Lee and Pradhan, 2007) and the Damrei Romel area in Cambodia (Lee and Sambath, 2006) by using logistic regression models and frequency ratio. Certain landslide analysis methods are limited to a single landslide as a case study such as the Vallcebre landslide (Corominas *et al*., 2005). Moreover, most of these classifiers have used satellite remote sensing data comprising topological maps and satellite images. In addition to geological features being our primary focus, we have also considered parameters like population. Very few researchers have ventured into analyzing landslides at a global level on a large-scale basis by using classifiers like the random forest classifier and Naive Bayes classifier. Although, these classifiers are very efficient, the following highlight limitations of our implementation. We have not taken factors like soil content and vegetation cover into account. The random forest classifier may yield non-reliable results if there are several categories under categorical variables and may overfit a noisy dataset that has not been pre-processed properly while the Naive Bayes classifier supposes the features to be independent. It can be concluded that the number of people along with geographical and climatic conditions can be used to predict landslide prone areas.

REFERENCES

Brownlee, J., 2016. Get your data ready for machine learning in R with pre-processing. Machine Learning Mastery, Toronto, Ontario, Canada.http://machinelearningmastery.com/pre-process-your-dataset-in-r/.

Corominas, J., J. Moya, A. Ledesma, A. Lloret and J.A. Gili, 2005. Prediction of ground displacements and velocities from groundwater level changes at the Vallcebre landslide (Eastern Pyrenees, Spain). Landslides, 2: 83-96.

Famili, A., W.M. Shen, R. Weber and E. Simoudis, 1997. Data preprocessing and intelligent data analysis. Intell. Data Anal., 1: 3-23.

Kipnis, I., 2017. Testing the hierarchical risk parity algorithm. R bloggers, Oakland California. https://www.r-bloggers.com/

Lee, S. and B. Pradhan, 2007. Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. Landslides, 4: 33-41.

Lee, S. and T. Sambath, 2006. Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. Environ. Geol., 50: 847-855.

NASA., 2017. Global landslide catalog export. NASA's Open Data Portal, USA. https://data.nasa.gov/dataset/Global-Landslide-Catalog-Export/dd9e-wu2v?category=dataset&%20view%20name=Global-Landslide-Catalog-Export.

Pal, M., 2003. Random forests for land cover classification. Proceedings of the 2003 IEEE International Symposium on Geoscience and Remote Sensing (IGARSS '03), July 21-25, 2003, IEEE, Kurukshetra, India, ISBN:0-7803-7929-2, pp: 3510-3512.

Rodriguez, G.V.F., M.C. Olmo, F.A. Hernandez, P.M. Atkinson and C. Jeganathan, 2012. Random forest classification of mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sens. Environ., 121: 93-107.

Yeo, I.K. and R.A. Johnson, 2000. A new family of power transformations to improve normality or symmetry. Biometrika, 87: 954-959.