Assessing the Landslide susceptibility in Samdrupjongkhar Dzongkhag using Machine Learning Models.

Vasker Sharma*, Thongley, Phurba Tamang, Dechen Wangmo, Lobzang Dorji, Jigme Tenzin.

Department of Civil Engineering and Surveying, Jigme Namgyel Engineering College, Dewathang, Samdrupjongkhar, Royal University of Bhutan.

vaskersharma.jnec@rub.edu.bt *, thongley.jnec@rub.edu.bt, phurbatamang.jnec@rub.edu.bt, dechenwangmo.jnec@rub.edu.bt, lobzangdori.jnec@rub.edu.bt, jigmetenzin.jnec@rub.edu.bt

ABSTRACT

Natural hazards that are associated with landslides are predominant in a hilly and mountainous area owing to several causative factors that trigger the landslides in a region. In this study, Machine Learning Models such as Artificial Neural Network (ANN) and Logistic Regression (LR) and conventional methods like Weight of Evidence (WoE) were used to develop the Landslide susceptibility maps (LSM) for Samdrupjongkhar Dzongkhag in Bhutan, which could be used for land management in strategic areas. The causative factor was kept the same for all the methods. The landslide inventory was developed based on field observation in conjunction with Remote sensing data. The landslide data were divided into training and testing data with 70% for training the model and 30% for testing. The AUC success rate for the LSM was 64%, 93%, and 89% for WoE, ANN, and LR respectively. Similarly, the AUC prediction rate was 58%, 88%, 89%, for WoE, ANN, and LR respectively. From the validation, it was found that the LR method followed by ANN is better than WoE in predicting the landslide occurrence zones. From the study, it emerged out that, slopes facing South-East to South-West were more prone to landslide failures. This is mainly attributed to the monsoon which brings rain from the south and mostly precipitates on the south-facing slopes making it more vulnerable to landslides. Also slope value perse tends to influence the landslides predominantly. This perhaps offers preliminary ways to identify the landslide zones in spatially large areas where geotechnical studies become strenuous and costly.

Keywords: Landslides, Susceptibilty mapping, ANN, Logistic regression, WoE.

1. INTRODUCTION

The regions in the Himalayan area are ever more faced with natural hazards due to flooding and landslides which are mostly associated with an increased amount of rainfall. While flooding has been mostly associated with extreme rainfall events, landslides are attributed to a multitude of other causative factors besides rainfall events. Landslide-related risks have been rampant especially along the highways in the hill and mountains impacting the environment, infrastructures such as roads and bridges, and livelihood owing to intense and erratic rainfall patterns which are most likely induced by climate change and other anthropogenic activities. Further, as per climate projection (IPCC, 2014), South Asia is expected to receive a higher amount of rainfall leaving the region in a vulnerable state. However, rainfall alone is not the leading cause of landslides in these regions. The immediate factor after rainfall is the disturbances caused by construction per se along the highways which causes the slope instability. These constructions include new road construction, road widening works, and routine maintenance which makes the slope more susceptible to increased infiltration due to the removal of vegetation along the road corridor. The slope of the area per se can be a good indicator of the landslide susceptibility in the hilly region however the slopes with a firm rock material are less likely to fail than the loose soil with the same slope. Further, distance to stream and erosive power of the stream aggravate the problem. The geological structure of the Himalayas is such that there are several roughly east-west trending major thrust and faults with several weak and crushed zones. Such areas are susceptible to numerous landslides along with these linear structures. Further, the orientation of folds and bedding, foliation, and joints in the rocks also play a vital role in landslide events (Upreti & Dhital, 1996). Therefore, it necessitates identifying the causative factors and their contribution in triggering the landslides in the region. The southern foothill of Bhutan receives the highest rainfall than any part of the country and these regions report the highest number of landslides and roadblocks. The expenditure incurred by the Royal Government of Bhutan (RGoB) is enormous for the maintenance of roads due to landslides (Wangchuk, 2019; Nima, 2020; Wangdi, 2015).

Although several researchers (El Jazouli et al., 2020; Prakasam et al., 2020; Pasierb et al., 2019; Mondal, 2016), assessed the landslides from a geological, geotechnical, and geophysical standpoint, it can be very costly owing to the use of sophisticated equipment in analyzing the geology of the affected areas. The cost is further escalated when such assessment is carried out on large spatial areas. Therefore, simplistic morphological, and environmental factors can be considered to assess the landslides on a large spatial scale using GIS employing different mapping methods. Numerous studies (Thongley & Vansarochana, 2021; Arzu et al., 2010; Erener et al., 2012; Kavzoglu et al., 2014; Feizizadeh & Blaschke, 2013; Yilmaz, 2010) have studied the landslide considering myriads of factors in producing Landslide Susceptible Maps (LSM) where such maps were produced using different mapping methods such as Index of Entropy (IoE), Weight of Evidence (WoE), Frequency Ratio (FR), Analytical Hierarchical Process (AHP), Logistic Regression (LR), Support Vector Machine (SVM), Artificial Neural Networks (ANN).

Therefore, there is a need for landslide hazard zonation for the identification of potential landslide areas via remote sensing and GIS techniques. Thus, effective management and mitigation measures could be reciprocated and referred to these potential landslide areas. In this study, we aim to prepare the landslide inventory using field surveying, GIS, and remote sensing techniques and identify the most prominent causative factors triggering the landslide in the region. The study will also prepare a prediction map using the weight of evidence, artificial neural network, and logistic regression. The study can also put forward site-specific mitigation measures to prevent landslides and road subsidence.

1.1 STUDY AREA

The Kingdom of Bhutan is situated on the southern slopes of the Eastern Himalayas and being a part of the young fold-thrust Himalayan Mountain belt, more than 90 percent of the country's area is topographically rugged and geologically very fragile. Monsoonal winds are most intense between June to September, making Bhutan the wettest country within the Himalayan range (Sharma & Adhikari, 2020). Rainfall-induced slope failure is the most common geo-environment hazard in the country. Samdrupjongkhar Dzongkhag, known for its industrial and commercial trade routes in the country, connects the rest of the eastern part of the country to the neighboring country, India.

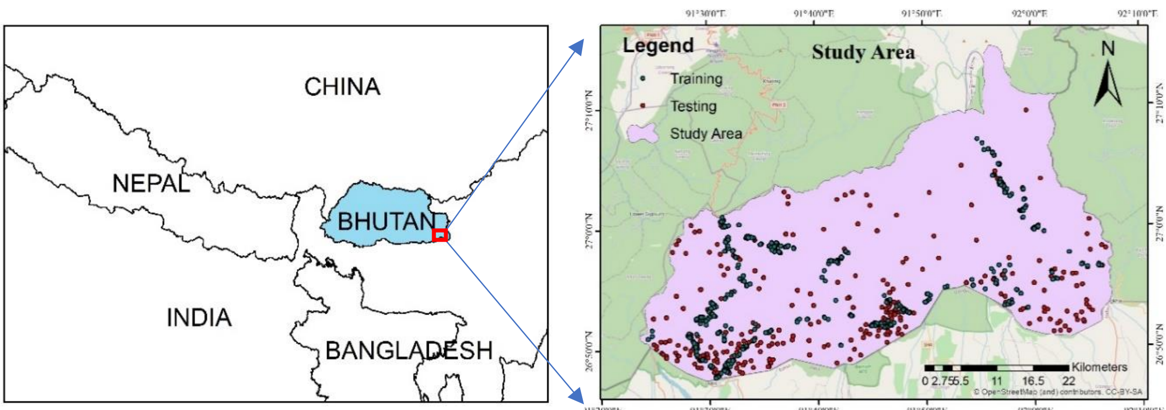

Figure 1 Study Area

2. DATA AND METHODS

The remotely sensed data such as Digital Elevation Model (DEM) and Landsat data (2020) has been obtained from www.usgs.gov.bt. Further locally available data like land use data and geological maps were also obtained for the analysis. Using DEM, various causative factors such as elevation, aspect, and slope were generated in the form of a raster layer. Further different raster like geological map, distance to road, distance to stream, FAO soil, and rainfall were also generated. Finally, Normalized Differential Vegetation Index (NDVI) was generated from Landsat data. A total of nine factors were considered in this study. However, the aspect data were further segregated into eight parts considering different aspect angles. Therefore, for the ANN and logistic regression model, a total of sixteen factors were considered. The factors are well represented in Fig. 2.

2.1 Data Preparation

The landslides inventory for training and testing were digitized using GIS from Google Earth which were further confirmed by field surveying. A total of 1675 points were digitized of which 1172 (70%) points were used for training the model while 503 (30%) points were used for testing the model. Within the training data, there are 812 landslide points while there are 274 landslides points in testing data.

The data from the raster layers mentioned above were extracted based on the location of training and testing data. The extracted data were then normalized using min-max normalization given by equation 1. The training and testing data was used in ANN and logistic regression model.

$$y = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Where y is the normalized data, x is the original extracted data.

2.2 Weight of Evidence (WoE) Method

The Weight of Evidence (WoE) is a data-driven quantitative statistical method (Bonham-Carter et al., 1988). Initially, the weight of evidence is developed for the mineral potential study. However, several authors have applied for the prediction of landslides, floods, and groundwater potential (Dahal et al., 2008). For landslide susceptibility assessment, the landslide probability is determined using the weight of the factors (evidence). In this method, a series of mathematical calculations are required and the detail of mathematical procedures are explained by (Van Westen, 2002). Two types of data required for the WoE were landslide inventory and landslide factors (Roy et al., 2019). The weight describes the probability of landslide occurrence in the case of the presence as well as the absence of the evidence (Kayastha et al., 2012). In this method, the positive and negative weights *(W+ and W-)* are assigned to each pixel of the factor classes and calculated using Equation 2 and Equation 3. *W+* expresses the chances of occurrence of a landslide in the case of the evidence being present and its magnitude indicates the positive association between landslide event and the factor class (Neuhäuser et al., 2012). On the other hand, *W-* describes the chances of landslide in case of absence of the evidence and its magnitude indicates a negative association between the landslide event and the factor class (Neuhäuser et al., 2012).

$$W^{+} = \log_e \frac{P\{B \mid S\}}{P\{B \mid \bar{S}\}} \tag{2}$$

$$W^{-} = \log_e \frac{P\{\bar{B} \mid S\}}{P\{\bar{B} \mid \bar{S}\}} \tag{3}$$

Where P is the probability, B is the presence of potential landslide factor, while $\bar{B}$ is the absence of a potential landslide factor. S is the presence of landslide and $\bar{S}$ is the absence of landslide.

The difference between $W+$ and $W-$ is known as weight contrast C, which indicates the spatial association between classes of the factor and landslides events. The $C > 0$, if the spatial association is positive while $C < 0$, if the spatial association is negative and $C = 0$ if the spatial association is lacking (Carranza, 2004).

$$C = W^{+} - W^{-} \tag{4}$$

Where C is the weight contrast of $W+$ and $W-$.

To develop LSM, all the classes of the factors are reclassified using the contrast value (C Then, add all the reclassified factors. Equation 5 is used for the development of the LSM.

$$LSM_{WOE} = \sum_{i=1}^{n} C_{ij} \tag{5}$$

Where LSM is a landslide susceptibility map, and C_{ij} is the contrast for class i of conditioning factor j, n is the number of factors.

2.3. Artificial Neural Network (ANN)

ANN model was developed using the R-programme, where a backpropagation algorithm was employed to adjust the weight and bias. The stopping criteria for the model was set by the error threshold of *0.01* and logistic function was used as an activation function in the model. The learning rate for the model was set at *0.0001*.

Mathematically, the neural network is expressed as:

$$y = f \sum_{i=1}^{N} w_i \times x_i + b \tag{6}$$

Where y is the output vector and xi is the input vector in the neural network, N is the number of neurons, wi is the connection weight between input and output, f is the activation function, and b is the bias term.

Weights and bias are adjusted using the ANN's back-propagation algorithm, where the objective function (also known as loss function) is the error between the network's output and the observed output. The error is minimized using the optimization algorithm known as "Gradient descent" which minimizes the error value by taking steps from an initial guess until it reaches the best value. This makes Gradient descent useful when it is not possible to solve where the derivative of the objective function is equal to zero. The step size is usually calculated by providing the learning rate and is expressed as follows:

$$step\,size = slope\ of\ objective\ function \times learningrate \tag{7}$$

2.4 Logistic Regression (LR)

Logistic regression is employed when the response variable is a categorical variable involving 0/1. It models non-normal distribution and used the logit link function given by (4)

$$\log\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1.x_1 + ... + \beta_n.x_n + \varepsilon \tag{8}$$

Where y is the probability of occurrence of landslide event while *(y/1-y)* is the odds ratio and *log(y/1-y)* is the log odds ratio. β_0 is the intercept and $\beta_1, \beta_2 ..... \beta_n$ are coefficients that measure the contribution of each independent variable $(x_1, x_2 .... x_n)$ which are landslide influencing factors and ε is the error.

3. RESULTS AND DISCUSSION

The landslide susceptibility maps have been produced using the methods described above and are shown in Fig 2.

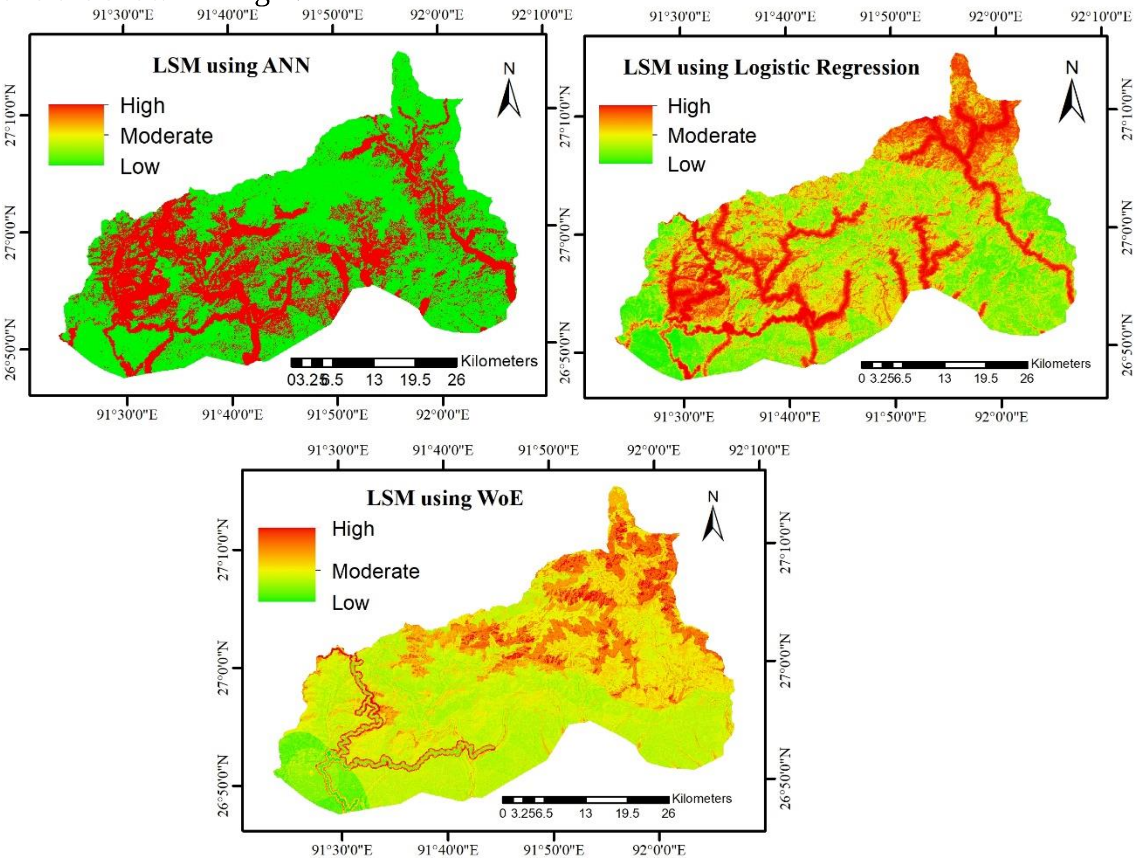

Figure 2 Landslide Susceptibility Maps

Although the model like ANN and LR yields better predictivity (Fig 3) of the landslide zones, a few clarifications are indispensable considering the overall interpretation of the weights generated by the model. From the model weights it is found that slope and aspect with south, southeast, and southwest have been the most influencing factor causing landslides in the region. It is because the majority of the monsoon rainfall falls on the slopes facing south which always remain in the grip of landslides. However, the factor like rain perse has relatively low weight for all the models indicating their low influence on the landslides. But it is an obvious fact that the intensity of rainfall invariably causes landslides and such ambiguity in the model weights is mainly due to the non-availability of adequate rainfall stations in the region. In this study only two rainfall stations were available and the rainfall intensity from these two stations has been interpolated over the whole region where actual rainfall in the region is either over or under-represented. This explains the non-influence of the rain on landslides in this study. It is also observed that the majority of the factors in the study do not tend to influence the landslides in the region.

From the Receiver Operating Characteristics (ROC) curve (Fig 4), it is observed that the success rate for ANN and LR is 93% and 89.48% respectively while for WOE is 64.15%. Similarly, the prediction rate for ANN and LR is 87.92% and 89.88% respectively while for WOE is only 58.10%. ANN and Logistic regression prove to be the better predictability models for assessing the landslide susceptibility zones. However, models like ANN takes an enormous amount of computation time to train and test the model and therefore LR can be a preferred method to preliminary assess the landslide zones given the instant computation time. Although predictability from the WOE method is strikingly low in this study, generally they have a better edge in explaining the influence of subfactors on the landslides which ANN and Logistic regression cannot.

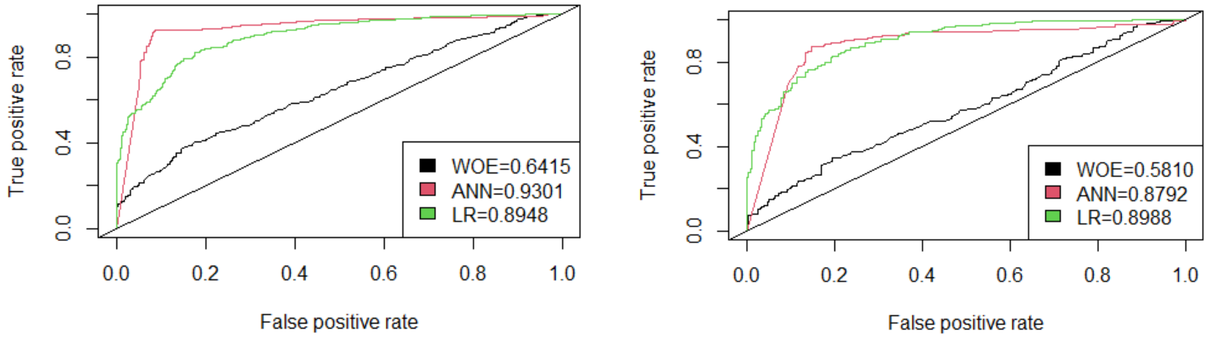

Figure 3 ROC-AUC curve (a) Success rate (b) prediction rate

4. CONCLUSION AND RECOMMENDATIONS

Natural hazards in the form of landslides are a common sight in the hilly and mountainous areas especially during the event of extreme rainfall and occasional earthquakes. Such hazards necessitate identifying the most prominent causative factor responsible for landslides in the region. Although various geotechnical and geophysical investigations can identify the causative factors, such investigation is enormous and costly especially if the investigation is carried out at a large spatial scale. Therefore, simple spatial analysis using Machine learning and other conventional techniques is a viable option in identifying the causative factors at a large spatial scale. Here in this study, ANN, LR, and WoE have been used to identify the causative factors causing the landslides in the region.From the models, it was observed that ANN and logistic regression yielded better predictability, while the WOE has low predictability. The prediction rate for ANN and Logistic regression is 87.9% and 89.88%, while for WOE the prediction rate is only 58%. Although ANN has a better predictability rate, it takes an enormous amount of computation time and therefore LR can be a preferred method for assessing the landslide susceptibility zones. From the model weights, it also emerged that slope and Aspect (South, Southwest, and South East) are the most influencing factors for the landslides in the region. Although rain is widely known as the cause of the landslide, the model weights did not justify the cause which was mainly attributed to the unavailability of an adequate rainfall station in the region. Nevertheless, from the visual observation in the region rainfall has been the major factor influencing the landslides in the region.

Further, the hills and mountains formed in the region are geologically young and weak, therefore the soils and slopes in these regions are fractured and weathered making them prone to slope failures(DGM, 2016). From the field, it was also observed that the soil along the highway and in the localized region was sandy-granular. This perhaps has a great influence on the stability of the slopes in the region as the rainwater infiltrating into such soil renders them unstable as they are not able to withstand the pore pressure. Furthermore, the landslide is more prominent in anthropologically active areas where road construction and widening are predominant. To reduce the landslide impact, detailed geo-technical investigations are deemed necessary on strategic location whereby site-specific mitigation measures could be employed. The mitigation measures range from constructing bio-engineering slopes and facilitating drainage especially along the highways where there is a higher chance of landslide failures. It is also believed that the non-frame method is an innovative slope stabilization technique and a disaster prevention technology widely adopted in Japan (Pokhrel, 2015) which could be reciprocated in Bhutan as well.

5. REFERENCES

Bonham-Carter, G., Agterberg, F., & Wright, D., 1988. Integration of geological datasets for gold exploration in Nova Scotia. *Photogrammetric Engineering and Remote Sensing*, 54(11), 1585-1592.

Carranza, E. J. M., 2004. Weights of evidence modeling of mineral potential: a case study using small number of prospects, Abra, Philippines. *Natural Resources Research*, 13(3), 173-187.

Dahal, R. K., Hasegawa, S., Nonomura, A., Yamanaka, M., Dhakal, S., & Paudyal, P., 2008. Predictive modelling of rainfall-induced landslide hazard in the Lesser Himalaya of Nepal based on weights-of-evidence. *Geomorphology*, 102(3-4), 496-510.

DGM (Department of Geology and Mines)., 2016. *Integrated Geo-Hazard Risk Assessment of Critical Landslide at Arong / Lamsorong , Samdrupjongkhar-Trashigang Highway , under Samdrupjongkhar Dzongkhag.* Thimphu, Bhutan: Ministry of Economic Affairs.

El Jazouli, A., Barakat, A., & Khellouk, R., 2020. Geotechnical studies for Landslide susceptibility in the high basin of the Oum Er Rbia river (Morocco. *Geology, Ecology, and Landscapes*, 1–8. https://doi.org/10.1080/24749508.2020.1743527

Erener, A, & Duzgun, H. S. B., 2012. Landslide susceptibility assessment : what are the effects of mapping unit and mapping method ? *Environ Earth Sci*, *66*, 859–877. https://doi.org/10.1007/s12665-011-1297-0

Erener, Arzu, & Duzgun, H. S. B., 2010. Improvement of statistical landslide susceptibility mapping by using spatial and global regression methods in the case of More and Romsdal (Norway . *Landslides*, *7*, 55–68. https://doi.org/10.1007/s10346-009-0188-x

Feizizadeh, B., & Blaschke, T., 2013. GIS-multicriteria decision analysis for landslide susceptibility mapping : comparing three methods. *Nat Hazards*, *65*, 2105–2128. https://doi.org/10.1007/s11069-012-0463-3

IPCC (Intergovernmental Panel on Climate Change., 2014. *Climate Change 2014 Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]*. IPCC, Geneva, Switzerland.

Kavzoglu, T., Sahin, E. K., & Colkesen, I., 2014. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis , support vector machines , and logistic regression. *Landslides*, *11*, 425–439. https://doi.org/10.1007/s10346-013-0391-7

Mondal, S., 2016. Geomorphic Threshold and Landsliding in Paglajhora Sinking Zone, Darjiling Himalaya. *International Journal of Research in Geography*, *2*(1), 1–13.

Nima., 2020, July 15. The cost of keeping the Gelephu-Trongsa highway open. *Kuensel*. Retrieved from https://kuenselonline.com/the-cost-of-keeping-the-gelephu-trongsa-highway-open/

Pasierb, B., Grodecki, M., & Gwóźdź, R., 2019. Geophysical and geotechnical approach to a landslide stability assessment: a case study. *Acta Geophysica*, *67*, 1823–1834. https://doi.org/10.1007/s11600-019-00338-7

Pokhrel, N., 2015, October 28. Non-frame method to stabilize slopes. *Kuensel*. Retrieved from https://kuenselonline.com/non-frame-method-to-stabilize-slopes/

Prakasam, C., Aravinth, R., Nagarajan, B., & Kanwar, V. S., 2020. Site-specific geological and geotechnical investigation of a debris landslide along unstable road cut slopes in the Himalayan region, India. *Geomatics, Natural Hazards and Risk*, *11*(1), 1827–1848. https://doi.org/10.1080/19475705.2020.1813812

Sharma, V., & Adhikari, K., 2020. Rainfall and Rainy days trend and ENSO Phenomena in Himalayan Kingdom of Bhutan. *International Journal of Spatilal Temporal and Multimedia Information Systems*. https://doi.org/10.1504/IJSTMIS.2020.10035484

Thongley, & Vansarochana, C., 2021. Spatial Zonation of Landslide Prone Area Using Information Value in the Geologically Fragile Region of Samdrup Jongkhar-Tashigang National Highway in Bhutan. *Environment and Natural Resources Journal*, *19*. https://doi.org/10.32526/ennrj/19/2020171

Upreti, B. N., & Dhital, M. R., 1996. *Landslide Studies and Management in Nepal. International Center for Integrated Mountain Development (ICIMOD.* Kathmandu, Nepal: ICIMOD.

Wangchuk, K., 2019. More than Nu 4 M worth of roads damaged in Samdrup Jongkhar by the onset of monsoon. Retrieved April 22, 2021, from http://www.bbs.bt/news/?p=117939

Wangdi, T., 2015, September 15. Mitigating slides at Reutala and Box Cutting roads. *Kuensel*. Retrieved from https://kuenselonline.com/mitigating-slides-at-reutala-and-box-cutting-roads/

Yilmaz, I., 2010. Comparison of landslide susceptibility mapping methodologies for Koyulhisar , Turkey : conditional probability , logistic regression , artificial neural networks , and support vector machine. *Environ Earth Sci*, *61*, 821–836. https://doi.org/10.1007/s12665-009-0394-9