International Conference on
Technology in Natural Disaster Prevention and Risk Reduction

ASSESSMENT OF LANDSCAPE DISTURBANCE IN TRUNG KHANH AREA, CAO BANG PROVINCE USING DIFFERENT DECISION TREE (C4.5, CART AND LMT) MODELS

Nguyen Quoc Phi[1*]
[1]Faculty of Environment, Hanoi University of Mining and Geology

Abstract

The efficiency of data mining approaches has gained immense popularity in various fields, including researches on natural hazards. The three decision tree models, namely C4.5, Classification and Regression Trees (CART) and Logistic Model Tree (LMT), were used for landscape disturbance assessment in the Trung Khanh region of Cao Bang province, and the results were compared. Factors for landscape disturbance were analysed and mapped as data inventory after gathering information from historical records, remote sensing detection and periodic field investigations. In the database, a total of 12 disturbance factors were considered as model inputs, and the results of each model were categorised under four disturbace classes. The receiver operating characteristics (ROC) curve and three statistical measures (Kappa statistic, Precision and F-Measure) were used to evaluate and prioritise the models. The CART model achieved the priority rank 1 with 88.6% corrected prediction and the area under the curve of 0.928, followed by LMT and C4.5 models. This research might be useful in sustainable studies in mountainous areas, especially in locations with comparable geophysical and climatological characteristics, to aid in decision making for land use planning.

Keywords: Data mining approaches, landscape disturbance, Trung Khanh, Cao Bang

1. Introduction

Landslides, debris flows, and human activities are the factors that affect the landscape disturbance the most. In Trung Khanh area, those natural hazards and mining activities usually cause loss of life and significant economic losses. How to assess the risk of landscape disturbance effectively has always been the focus and difficulty to reduce disaster risk. Risk is composed of the hazard of disaster, the vulnerability and exposure of victims and the disaster preparedness and mitigation capacity. For sustainable studies, the susceptibility assessment is the key point of the risk assessment.

There are various methods to map the sustainable of landscape including traditional mathematical and statistical models and advanced machine learning techniques. Traditional statistical analysis methods can be utilized, such as weight of evidence (Nguyen and Nguyen, 2004), frequency ratio (Wu, 2019) and other weighted index methods (Han et al., 2019; Yi et al., 2019). Various big data algorithms, such as logistic regression (Long and De Smedt, 2019; Yang et al., 2019), naive Bayes (Pham et al., 2017), decision tree (Mao et al., 2017), support vector machines (Aktas and San, 2019; Huang and Zhao, 2018), genetic algorithm (Dou, 2015), artificial neural network (Bragagnolo et al., 2020; Zhou, 2018), convolutional neural network (Wang, 2019) can also be used in landscape

disturbance mapping. However, due to the complex environmental factors system of disturbance in different study areas, the accuracy and scientific nature of landscape susceptibility drawn by each model is very important. Therefore, the evaluations of models, including their advantages and applicability, are very important to obtain a satisfactory susceptibility map of disturbance.

In this study, we collected multi-source data such as field survey data, precipitation data and remote sensing satellite data from Trung Khanh County, Cao Bang province and used the advanced big data models to construct susceptibility map of disturbance, and compared their performance and applicability in the study area.

2. Study Area and Data

2.1. Study Area

The area of Trung Khanh County is about 688.01km^2 and the population is about 70,424. The study area is one of the most landslide-prone areas in Cao Bang province, one of the reasons is because it is karst landscape with easily leaking surface water and high soil moisture content. The annual average precipitation of study area is about 1,500-2,000mm, precipitation is the major inducing factor for landscape disturbance. On 14 and 15 July 2019, many rainfall-induced landslides and debris flows occurred in the study area, which caused the destruction of many houses and roads and the death of villagers, 918 houses and 1.000ha of rice field were under the flood water. According to the Cao Bang Meteorological Bureau, between 14 and 16 July 2019, Trung Khanh Town experienced three periods of heavy rain shortly before the landslides. The cumulative rainfall at this site reached 189.1mm for 15-16 July and 98mm for 17-18 July.

2.2. Data

The monitoring of landscape disturbance in Trung Khanh County is done using field surveys, remote sensing images and combining the natural hazards' historical locations recorded by the Vietnam Geological Survey projects during 2012-2020. These areas are the centroid of landslide scarps and debris flows valleys, which has been proved the best sampling strategy. Combined with multi-source data, the data inventory was collected and 87 location fo landslides sites and debris flows valleys were obtained.

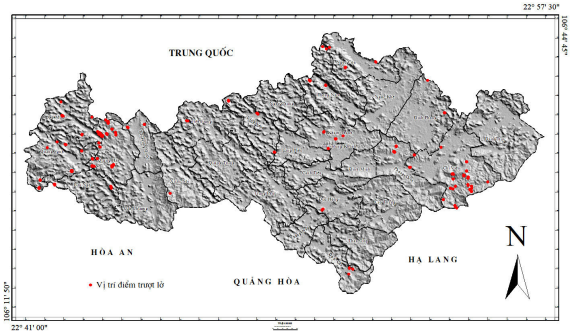


Figure 1. Location of landslides and debris flows in Trung Khanh County

The conditioning factors are extremely important for landscape disturbance assessment. A total of 12 conditioning factors were selected based on their impact on the disturbance and the data accessibility. To store these conditioning factors into a uniform attribute table, according to the DEM pixel size, all of the factors' pixel size was resampled to 20m×20m.

Topography is the most dominant factor in slopes stabilities. In this study, topography factors are calculated by the Digital Elevation Model (DEM) with 20m×20m pixel size.

Elevation, slope data were extracted from DEM. Elevation affects the degree of rock weathering in landslides assessment. The slope is another important factor that can reflect the steepness of the topography. Generally speaking, the greater the slope, the higher the possibility of landscape disturbance, when other conditions are the same.

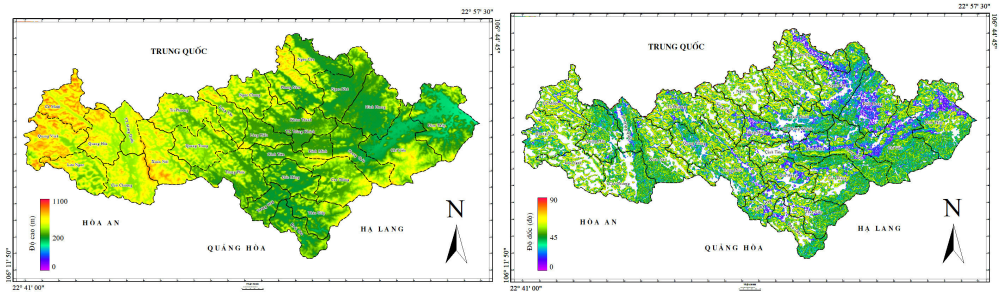


Figure 2. Maps of a) elevation and b) slope angle

Lithology affects the shear strength and permeability of slopes, which is another important conditioning factor for landslides and then, debris flows occurrences. Geological age can also characterize the development degree of regional lithology. Faults control the formation and development of natural hazards and geological processes are more active in the vicinity of faults. The lithology map were digitized from geological maps and we were able to calculate the distance to faults by spatial interpolation.

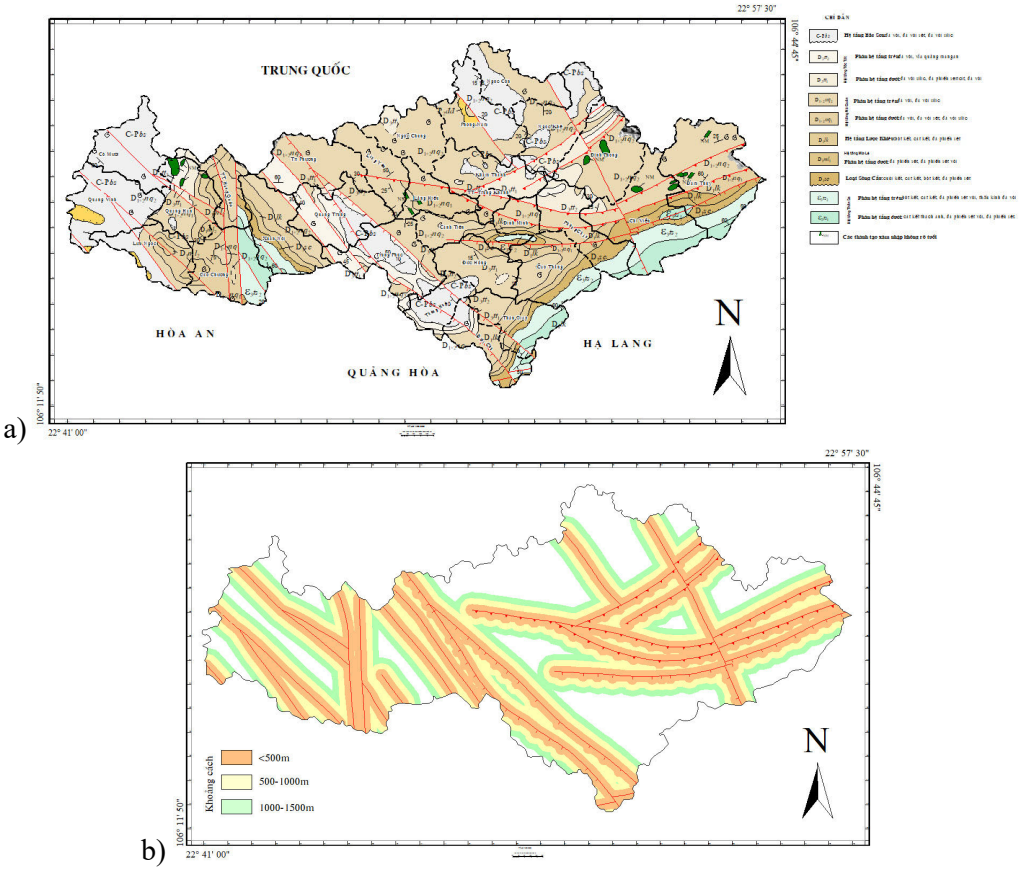
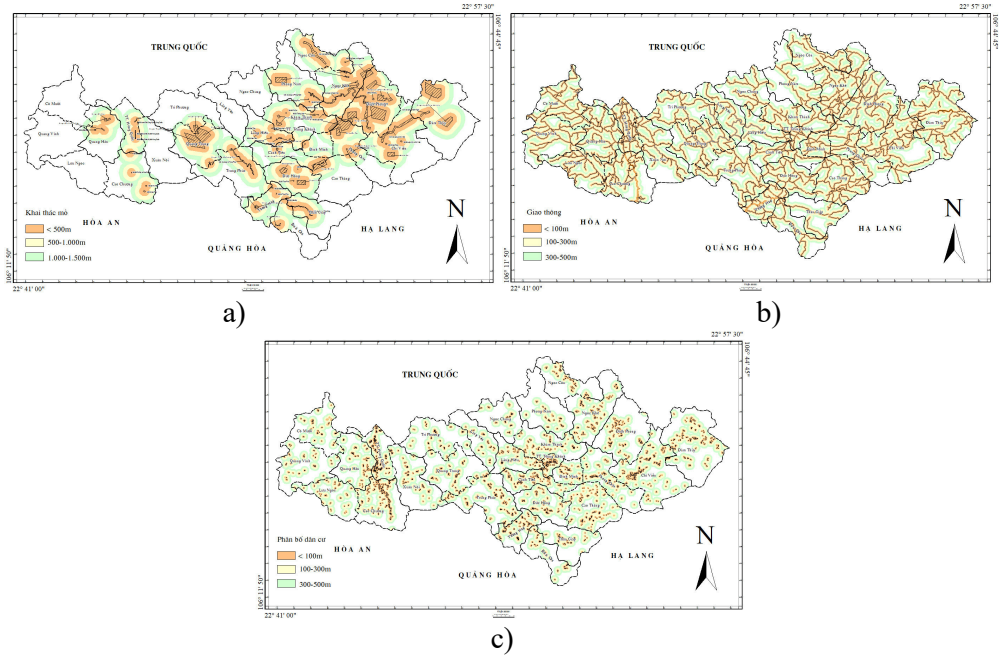


Figure 3. Maps of a) geology and b) distance to faults

In the study area, mining, road and house construction can also reflect the possible influence of human activities on the natural hazards to a certain extent. Meanwhile, the

mining and traffic on roads can cause vibration to destabilize rock material. The closer the distance to the mines and roads, the higher is the possibility that geo-hydrological hazards will occur. Therefore, we selected the distance to mines, roads and houese as conditioning factors.



a) b)

c)

Figure 4. Human activities of a) mining, b) road and c) house

Hydrological factors are the factors that must be considered in natural hazards in the study area. The surface water, such as streams and rivers is one of the most active factors in external dynamic geological processes, distance to streams can clearly express the influence of surface water on the landscape disturbance. Also, this study selected two kinds of hydrological indexes which are mainly used in the study of landscape disturbance, including TWI (Terrain Wetness Index) and MBI (Mass Balance Index). TWI represents the effect of different terrains on saturation degree and surface runoff location and MBI represents slope failure and deposition. The calculation equations of these two hydrological indices are as follows:

$$TWI = \ln(A_s / \tan\beta) \tag{1}$$

$$MBI = \begin{cases} f(TC)*(1-f(\beta))*(1-f(A_s)) & f(TC) \le 0 \\ f(TC)*(1+f(\beta))*(1+f(A_s)) & f(TC) > 0 \end{cases} \tag{2}$$

where A_S represents the catchment area (m^2/m), β is the slope of each grid, TC is the total curvature of eight grids around each grid, respectively.

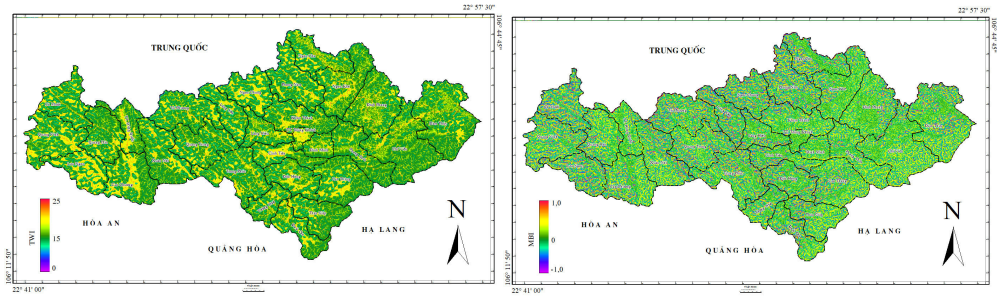
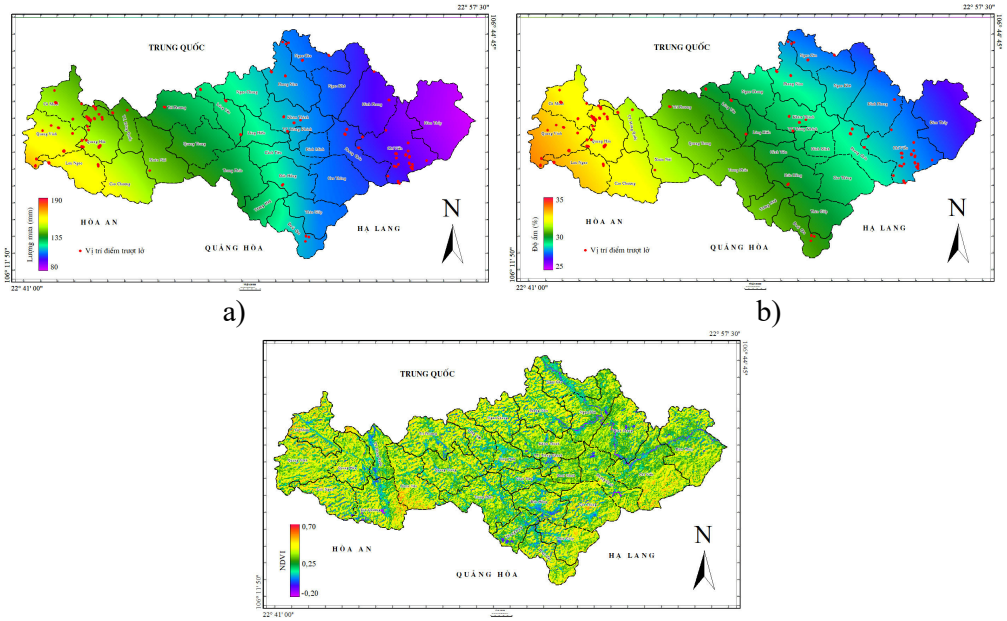


Figure 5. Maps of a) TWI and b) MBI

Land cover, especially vegetation cover, is often used as a factor in determining slope stability. Hence, we chose the normalized difference vegetation index (NDVI) as conditioning factor. The NDVI data was calculated through Landsat 8 OLI satellite remote sensing digital products using the band algebra tool of the Environment for Visualizing Images (ENVI) software.

For landscape disturbance in the area, rainfall is one of the most important triggering factors, especially short-term and instantaneous extreme rainfall. Meanwhile, rainfall also reflects the soil moisture. We chose the maximum daily rainfall as a conditioning factor to reflect the effect of rainfall. The rainfall data was collected from GSMaP of Japan Aerospace Exploration Agency (JAXA) and the soil moisture SMOPS (Soil Moisture Products) of the US National Environmental Satellite, Data, and Information Service (NESDIS).



a) b)

c)

Figure 6. Maps of a) rainfall (precipitation), b) soil moisture and c) NDVI

3. Methods

3.1. C4.5 decision tree (C4.5)

C4.5 decision tree is a non-parametric supervised learning method used for classifcation and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from data features. Creation of a C4.5

involves a multistage or hierarchical decision-making scheme. The tree features a root node, internal nodes, and terminal nodes (leaves) (Nguyen et al., 2018).

Each node of the tree makes a binary decision separating one or more classes from the remaining classes. Processing involves a gradual descent until the terminal node is attained (Gama, 2004). The figure below shows the sample of architecture of the decision tree model which consists of the three following elements: nodes, conditions, and productions.

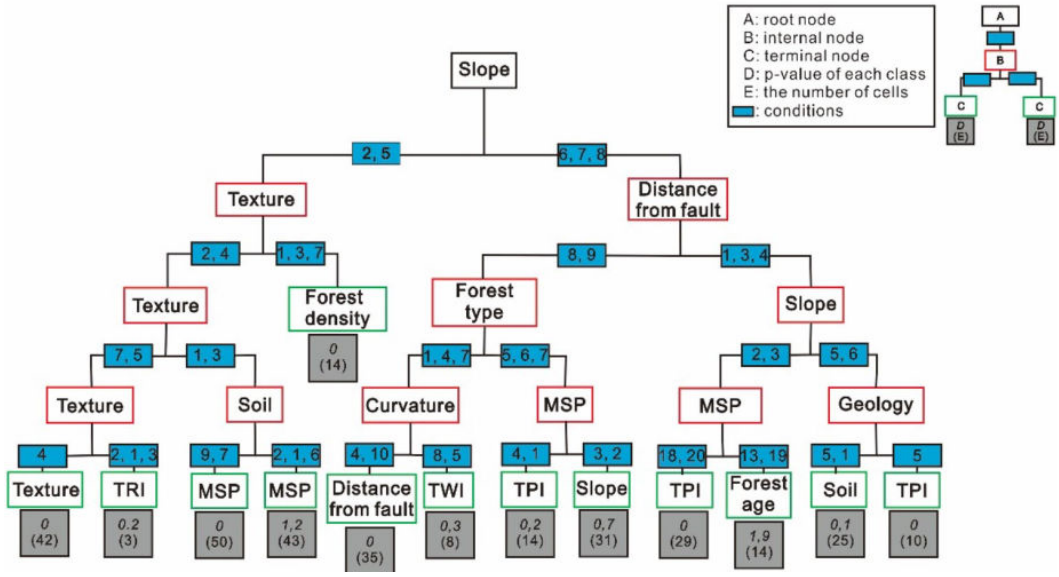


Figure 7. Architecture of a decision tree model

In this study, landscape disturbance was analyzed using J4.8 algorithm of open source Weka package, which is a modification on Java of original C4.5 algorithm.

3.2. Classification and Regression Trees (CART)

CART is a rule-based algorithm that constructs a binary tree by binary recursive partitioning. Binary recursive partitioning is a method that partitions a node into a yes/no response. The heterogeneity within each resultant subset is reduced on the basis of a single factor and the rule generated for each phase, which divides them depending on the various relationships of each division. Landslide susceptibility mapping using the CART technique has been used in several studies (Nefeslioglu et al., 2010). A "terminal" node's expected value is considered the average of the answer values in that node (Breiman et al., 1984). The predictor variables are extremely simple and can be comprised of different types: numeric, binary and categorical types. The model's results are not affected by monotonous transformations and different measurement scales between predictors. In regression trees, independent variables are insensitive to outliers and use surrogates to manage missing data (Breiman et al., 1984). The hierarchical structure of a regression tree indicates that the response to one input vector relies on higher input variables in the tree to model relationships between predictors automatically.

Regression trees typically lead to an overcomplex decision tree where only the most relevant knowledge, that is, the nodes that illustrate the largest amount of deviance, needs to be 'pruned' to communicate (Nefeslioglu et al., 2010). CART, similar to other DT algorithms, does not need the identification of independent variables in advance because the most relevant variables are discovered during the selection of the optimal splitting characteristic in each node (Breiman et al., 1984). Thus, CART is appropriate for issues in

which the correlation between input and output parameters is unknown in advance, making the CART model's outputs interpretable (Nefeslioglu et al., 2010).

Depending on whether the output is qualitative or quantitative, CART can be used to solve classification and regression issues. CART is used in this study as a classifier for landscape disturbance. "tree" package in open source Weka package was used for preparing the CART model.

3.3. Logistic Model Tree (LMT)

LMT is a classification model in computer science that integrates logistic regression with DT learning, with the related supervised training algorithm (Landwehr, 2005). The earlier concept of a model tree is used in logistic LMT. A DT on its leaves uses linear regression (LR) models where a piecewise constant model is generated by ordinary DTs with constants on their leaves (Landwehr et al., 2005). This process is performed to obtain a piecewise linear regression model. The LogitBoost algorithm is used in the logistic version to generate a logistic regression model at each tree node. The model uses cross-validation for searching the multiple LogitBoost iterations to control overfitting of training data. For each Mi class, the LogitBoost model utilises least-squares fitting additive logistic regression, and the later likelihood of leaf nodes is measured by LR (Wang et al., 2015).

$$L_M(x) = \sum_{i=1}^{n} \beta_i x_i + \beta_0 \tag{3}$$

where β_i is the coefficient of the ith element of vector x, n is the total factors, and D is the total classes.

In the LMT model, the posterior probabilities of leaf nodes were computed by using the linear logistic regression technique [64].

$$P(M|x) = \frac{\exp(L_M(x))}{\sum_{M=1}^{D} \exp(L'_M(x))} \tag{4}$$

3.4. Model evaluation

The Precision, Recall, Accuracy and the area under curve (AUC) of receiver operating characteristic (ROC) were selected to verify the models' accuracy. Those methods are based on the statistics of true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

4. Results and discussions

4.1. Feature importance

Ridge regression (RR) was applied to verify the importance of the selected input factors. A total of 12 factors were tested through RR to analyse their importance for disturbance modelling. The outcome of RR revealed that rainfall (RR = 0.377) had the highest predictive capability in this study. Comparatively, other factors, such as soil moisture (RR = 0.282), elevation (RR = 0.256), lithology (RR = 0.214), NDVI (RR = 0.247), mining (RR = 0.176), slope angle (RR = 0.161), house construction (RR = 0.153), MBI (RR = 0.156), road construction (RR = 0.135), fault (RR = 0.131), and TWI (RR = 0.057) played positive and significant roles for the modelling of landscape disturbance in this research. Among the human induced factors, mining activities has the greatest damage to the surface disturbance.

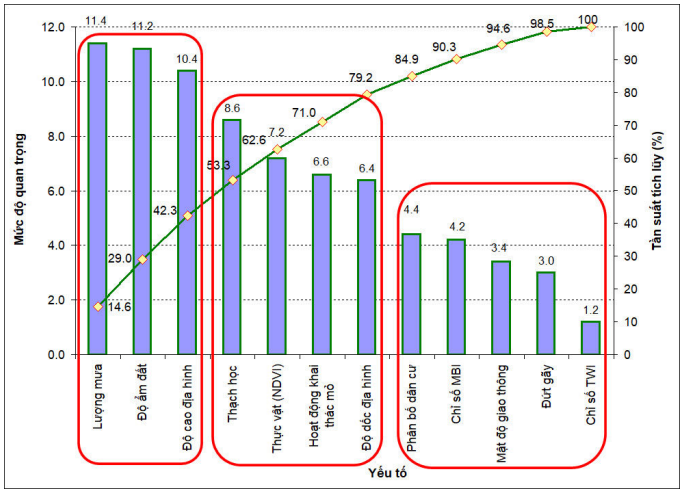


Figure 8. The importance of input factors

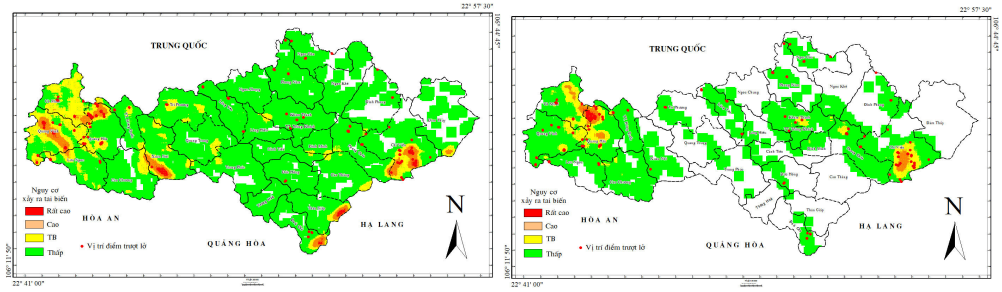
4.2. Model accuracy evaluation

For the use of the three models, the training and testing data are all determined as 70/30. In the training process of the models, the redundant nodes of the tree are pruned along with the tree growth, and the child nodes are created 10 times and the other model parameters are given as default. The model will stop building when the accuracy is no longer improved. The average accuracy rates of CART, LMT, and C4.5 are 88.6%, 87.9% and 83,6%, respectively. CART model has the highest accuracy in both the training and testing stages.

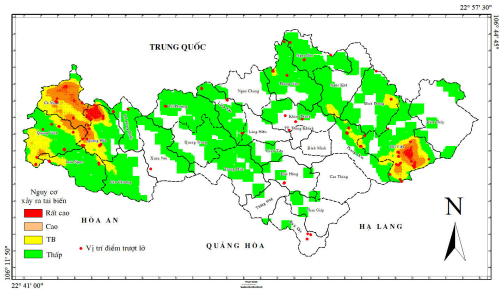
4.3. Susceptibility map analysis

The percentage distributions of the susceptibility classes of the CART model for very high, high, moderate and low classes were 18.59, 28.38, 18.17 and 3.23%, respectively. The LMT model predicted 12.18% as low, 20.70% as moderate, 27.44% as high, and 34.90% as very high landscape disturbance. The C4.5 model categorised 6.54, 16.71, 26.05, 26.65 and 24.05% of the area under low, moderate, high and very high disturbance zones.

The study area contains 1,755,012 pixels, converted into point type and mapped by PG-Steamer platform. Using Jenks Natural Breaks algorithm, the three models were re-classified into four susceptibility levels of low, medium, high, and very high.



a) b)



c)

Figure 9. Landscape disturbance from a) CART, b) LMT and c) C4.5 tree model

The predictability of the models was validated by using the AUC of ROC and threee other statistical measures. The success rate curve (using the training dataset) was drawn for each model. The result showed that CART was the best fit with an AUC of 0.928, followed by LMT (AUC = 0.901) and C4.5 (AUC = 0.843). The predictive capability of the models was assessed by using the prediction rate curve, which provided a similar result to the success rate curve. The table below contains the summary statistics of other measures.

Table 1. Statistical measures of the models

Model	Kappa statistic	Precision	F-Measure
CART	0,763	0.887	0.886
LMT	0,755	0.891	0.880
C4.5	0.664	0.841	0.837

The above outcome showed the relative priority ranking of models by considering all the exactness metrics calculated using the training and validation data sets. Using the training dataset, most priority was assigned to the CART model because it ranked 1, followed by LMT (rank 2) and C4.5 (rank 3) model. In the validation dataset, the result was the same as the training set.

5. Conclusion

This current research has contributed to comparison and evaluation of three decision tree models (CART, LMT, and C4.5) for landscape disturbance in Trung Khanh County of Cao Bang province. The results showed that these data mining approaches have been considered as robust and efficient tools and have been used in different fields of geographical research, geotechnical application, and natural hazards, including landscape disturbance mapping. All factors in a region are not equally responsible for causing of the disturbance. In this study, RR was adopted, confirming factors such as rainfall, soil moisture, and elevation, as the most important driving factors for the disturbance in the study area. The outcome showed that the most vulnerable zones of the disturbance are found in the northwesten and south east portions of the district where soil condition, weak geology, torrent runoff, high altitude, steep sloping, rugged topography and heavy rainfall are the chief reasons for the disturbance.

Of all the 12 condition factors, the three factors including rainfall, soil moisture and elevation are the most suitable condition factors for landscape disturbance. The lithology, land cover (NDVI), mining activities and slope angle has a medium contribution to the three algorithms and plays an obvious role in the disturbance in the area.

However, every research work has a certain limitation. The limitation of the present

study is the absence of some geological properties, such as joint, foliation and bedding. Only surface geology data were used. However, despite this limitation, the current study has good scope to accurately demarcate the landslide-susceptible area for future planning and management.

References

1. Aktas H., San B. T., 2019. Landslide susceptibility mapping using an automatic sampling algorithm based on two level random sampling. Comput. Geosci. 133, 104329.
2. Bragagnolo L., da Silva R. and Grzybowski J., 2020. Artificial neural network ensembles applied to the mapping of landslide susceptibility. Catena 184, 104240.
3. Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984. Classification and Regression Trees; Wadsworth International Group: Belmont, CA, USA.
4. Dou J., Chang K. T., Chen S., Yunus A. P., Liu J. K., Xia H. and Zhu Z., 2015 Automatic Case-Based Reasoning Approach for Landslide Detection: Integration of Object-Oriented Image Analysis and a Genetic Algorithm. Remote Sens. 7, 4318-4342.
5. Gama J., 2004. Functional trees. Mach. Learn. 55, 219-250.
6. Han L., Zhang J., Zhang Y. and Lang Q., 201. Applying a Series and Parallel Model and a Bayesian Networks Model to Produce Disaster Chain Susceptibility Maps in the Changbai Mountain area, China. Water 11, 2144.
7. Huang Y., Zhao L., 2018. Review on landslide susceptibility mapping using support vector machines. Catena 165, 520-529.
8. Landwehr N., Hall M., Frank E., 2005. Logistic model trees. Mach. Learn. 59, 161-205.
9. Long N. and De Smedt F., 2019. Analysis and Mapping of Rainfall-Induced Landslide Susceptibility in A Luoi District, Thua Thien Hue Province, Vietnam. Water 11, 51.
10. Mao Y., Zhang M., Sun P. and Wang G., 2017. Landslide susceptibility assessment using uncertain decision tree model in loess areas. Environ. Earth Sci. 76, 752.
11. Nefeslioglu H. A., Sezer E. A., Gokceoglu C.; ,Bozkir A. S.; ,Duman T. Y., 2010 Assessment of Landslide Susceptibility by Decision Trees in the Metropolitan Area of Istanbul, Turkey. Math. Probl. Eng. 2010, 901095.
12. Pham B. T., Bui D. T., Pourghasemi H. R., Indra P., Dholakia M. B., 2017. Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: A comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. Theor. Appl. Clim. 128, 255-273.
13. Quoc Phi Nguyen, Du Duong Bui, SangGi Hwang, Khac Uan Do, Thi Hoa Nguyen, 2018. Rainfall-triggered landslide and debris flow hazard assessment using data mining techniques: A comparison of Decision Trees, Artificial Neural Network and Support Vector Machines. Proceedings of the 2018 Vietnam Water Cooperation Initiative (VACI 2018) - Highlights. Science and Technics Publishing House, Hanoi, Vietnam, p.138-141. ISBN: 978-604-67-1059-2.
14. Wang L. J., Guo M., Sawada K., Lin J., Zhang J., 2015. A comparative study of landslide susceptibility maps using logistic regression, frequency ratio, decision tree, weights of evidence and artificial neural network. Geosci. J. 20, 117-136.
15. Wang Y., Fang Z. and Hong H., 2019. Comparison of convolutional neural networks

for landslide susceptibility mapping in Yanshan County, China. Sci. Total Environ. 666, 975-993.
16. Wu C., 2019. Landslide Susceptibility Based on Extreme Rainfall-Induced Landslide Inventories and the Following Landslide Evolution. Water 11, 2609.
17. Yang J., Song C., Yang Y., Xu C., Guo F. and Xie L., 2019. New method for landslide susceptibility mapping supported by spatial logistic regression and GeoDetector: A case study of Duwen Highway Basin, Sichuan Province, China. Geomorphology 324, 62-71.
18. Yi Y., Zhang Z., Zhang W., Xu Q., Deng C., Li Q., 2019. GIS-based earthquake-triggered-landslide susceptibility mapping with an integrated weighted index model in Jiuzhaigou region of Sichuan Province, China. Nat. Hazards Earth Syst. Sci. 19, 1973-1988.
19. Zhou C., Yin K., Cao Y., Ahmed B., Li Y., Catani F. and Pourghasemi H. R., 2018. Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir Area, China. Comput. Geosci. 112, 23-37.