Random Forest Method Utilization For Landslide Hazard Zonation In Lima Puluh Kota Regency

Ahyuni[1], Rizki Atthoriq Hidayat[2], Endah Purwaningsih[3], Yurni Suasti[4]
{ahyuniaziz@fis.unp.ac.id[1], rizkiatthoriq99@gmail.com[2], endahgeo@fis.unp.ac.id[3], yurnisuasti@fis.unp.ac.id[4]}

Geography Department, Universitas Negeri Padang, Padang, Indonesia

Abstract. There are various statistical methods that are applicable for landslide hazard zonation. Random Forest classification method is deployed in determining landslide hazard in Lima Puluh Kota Regency. Landslide records are obtained from field measurement for 145 points, 80% of total points used as train data and 20% for test data. Predictor data obtained from related agencies, namely rainfall, geology, LULC, soil type, aspect, elevation, slope, curvature, road distance, and river distance. Random Forest analysis applied 1000 trees to get probability values. ROC was obtained for 0.8137 which represents the model is qualified to predict landslide hazard zones at very high (6.1%), high (11.106%), moderate (19.036%), and low (63.748%).

Keywords: Landslide; Random Forest; Hazard zonation.

1 Introduction

Indonesia is an archipelago in the tropical area which has many areas prone to landslides. Landslides are found almost in all provinces in Indonesia throughout the year. Lima Puluh Kota District in West Sumatra Province has frequent landslide records. The whole world regards landslides as one of the most formidable and dangerous natural disasters [1]. Landslides can destroy man-made structures, dramatically alter landscapes, and endanger human lives (Khaled Taalab, Tao Cheng, Yang Zhang). Large-scale movements during landslides are caused by specific strata, rugged and rough ground, and various climate extremes, resulting in a high degree of instability. Moreover, the dynamics of hydrological processes and the specific elevation patterns due to large changes in elevation mean that rather large differences in environmental properties in mountainous areas can trigger landslides. In addition, human activities such as road construction and deforestation may contribute to this hazard [2,3].It needs to avoid road construction by creating routes that can traverse landslide-prone areas with minimal damage (Ender Budai, Abdullah Emin Akai). Landslides have negative short-term and long-term economic consequences and high costs [4].

A number of statistical methods can be used to identify landslide-prone areas. Even some experts compare results between methods. For example, compared results between the logistic regression method and the Wu-Yaung random forest model, random forest results outperform logistic regression [5-8]. As an example of the benefits of this model, the study used the random forest method to identify landslide-prone areas in Lima Puluh Kota Regency. There are two methods commonly used to assess landslide vulnerability: data exploration and assessing expert site-specific experience knowledge approaches. A data-driven approach

ICGEO 2022, November 19-20, Padang, Indonesia

DOI 10.4108/eai.19-11-2022.2332288

combines statistical methods with probabilistic analysis. [7-9]. Implementation of this method has recently been supported by geographic information system technology (GIS) [6].

A random forest is a decision tree or collection of decision trees. This algorithm is a combination of each tree of decision trees and is combined into a model. A random forest is an ensemble learning algorithm built from decision trees. The algorithm builds multiple decision trees using bootstrap data and randomly selects a subset of variables in each decision tree. A decision tree is a machine learning algorithm that uses a set of rules to make decisions in a tree-like structure that models expected outcomes, resource costs, benefits, and possible outcomes or risks. Random Forest works in two phases. In the first phase, a number of N decision trees are combined to create a random forest. Then, in the second stage, we make predictions for each tree created in the first stage. The study used R in Google Colab and QGIS 3.22.6 to perform the computational process and compile the data to create a landslide susceptibility map.

2 Method

2.1 Research Location

This Research Located Lima Puluh Kota Regency, West Sumatra, Indonesia which located between 00°25`28.71"N - 00°22'14.52" South Latitude and 100°15'44.10"E - 100°50'47.80" East Longitude. This region consists of 13 sub-districts with approximately 3,354.30 Km2 as the total area. The geological conditions of Lima Puluh Kota Regency are varied, influenced by the dynamic Sumatra Fault System (SFS). which is evidenced by the discovery of igneous rocks such as basalt, tuff, andesite, pumice, sedimentary rocks such as sandstone, limestone and metamorphic rocks such as slate, quartz, and philite. This geological formation is important to be studied because it is one of the main factors that cause landslides. The topography of the area varies in the range of flat, undulating, and hilly. It has a height between 110 meters and 2,261 meters from sea level. In this area there are 3 inactive volcanoes namely Mount Sago (2,261 m), Mount Bungsu (1,253 m), and Mount Sanggul (1,495 m), and 13 large and small rivers. According to the 2020 national census, the regency population is 383.525.

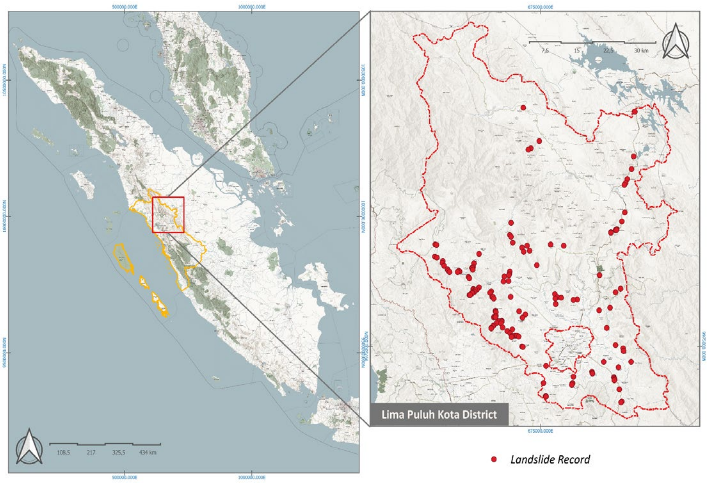

Fig. 1. Map of esearch location.

2.1 Random Forest (RF)

Random forest is one of the ensemble machine learning algorithms based on decision trees which is starting to be widely used in various data processing. This method is widely used for classification and regression [10]. The decision tree is a prediction model that uses a set of binary rules to determine the target variable [11]. Random forest is a machine learning model that is capable of producing a large number of decision trees to obtain spatial relationship information based on landslide events[12].

This algorithm randomly exploits binary trees using observations with the bootstrap technique by collecting original data, then a random selection of data is sampled and used to build a model [13]. RF requires two parameters to produce a classification model, namely ntree (number of decision trees) and mtry (number of factors used in the nodes of each decision tree) [10]. RF works by constructing decision trees during training and generating classes as a result of classification or regression of individual trees [13]. In the modeling in this study using the results of regression by extracting the probability value, so that the resulting output is continuous. The estimated dependent variable is obtained using the average results for regression [14]. The random forest works with the following procedure [15]: 1) determines the mtry value, 2) the model uses the bootstrap method to randomly select ntree samples in the original data to make an ntree decision tree, 3) ntree model random decisions forest and samples are predicted or classified based on random forest results.

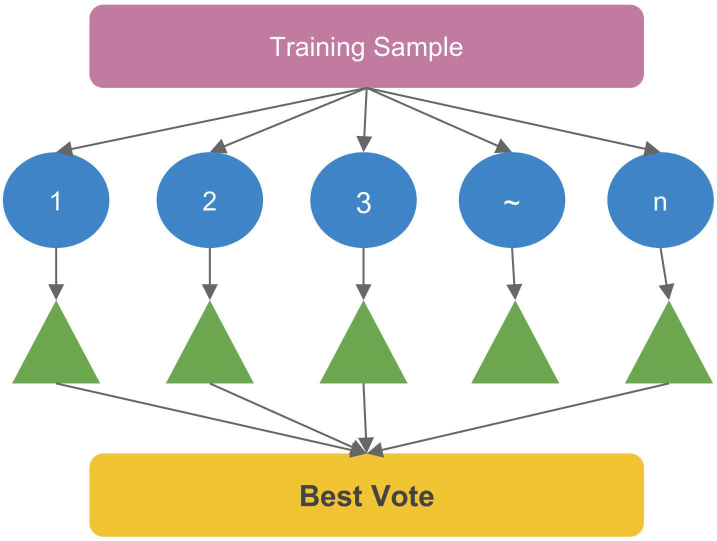

Fig. 2. Random forest algorithm scheme.

Ensemble methods such as Random Forest have better accuracy than single membership if only the data appears random and varies [16]. Random forest It has been used in various studies which reveal that random forests have a high tolerance in terms of algorithms, noise, and outliers [17]. As in [18], who compared the performance of logistic regression models and random forests in determining the potential for landslides in Wuyaun, China, found that random forests proved to be more accurate as evidenced from the data training, validation, and area under the curve phases. Decision trees improve the accuracy and stability of the model better than single decision trees using randomly generated methods to select samples and features [15]. The random forest classification model uses the following equation [15]:

$$Y(X) \;=\; arg_{Z}^{max} \sum_{i=1}^{k} \quad I(y_i(X) = Z) \tag{1}$$

3 Result and Discussion

There are 290 points considered in this model with two categories, landslide point as recorded landslide data encoded as 1 and non-landslide points (randomly generated out of landslide) encoded as 0. Each category has 145 points. These points thus randomly splitted into training and test dataset with proportion 80% for training and 20% for test. The points are then used to processed with all variables to generate RF model.

3.1 Model Validation

To make sure the model can be accepted or proper for further analysis, it needs to assess the accuracy. ROC is useful to explain the accuracy with its area under curve (AUC) by comparing sensitivity and specificity of model result. AUC value ranges between 0 and 1. The model is better if the AUC approaches 1, greater than 0.6 is considered as a good model.

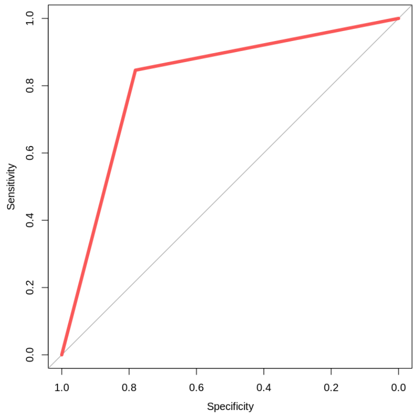

Fig. 3. ROC chart.

The AUC obtained in this analysis is 0.8137 and the area under the red line in Figure 1 is large, meaning that this model is good and worth predicting the landslide susceptibility. The information of the points is either classified or misclassified is generated by confusion matrix in the table below

Table 1. Confusion matrix.

	0	1	Class Error
0	93	23	0.1982759
1	14	102	0.1206897

Horizontal axis is the actual classes of the points and the vertical axis is the predicted classes. Non-landslides points are correctly predicted as 93 of 145 points with 0.98 class error, while for recorded landslides are correctly predicted as 102 of 145 points with 0.12 class error.

3.2 Random Forest Model

The result of RF used for predicting all the study area. All pixel values in the table are converted into raster format for further spatial analysis. The value chosen is probability rather than class value because from probability value, we can divide it into multiple levels of susceptibility. The susceptibility levels are low, moderate, high, and very high as the classification from probability value.

Table 2. Landslide susceptibility area in Lima Puluh Kota Regency

Susceptibility	Area (Ha)	Area (%)
Low	206440.138	63.748
Moderate	61647.415	19.036
High	35964.491	11.106
Very High	19787.474	6.11

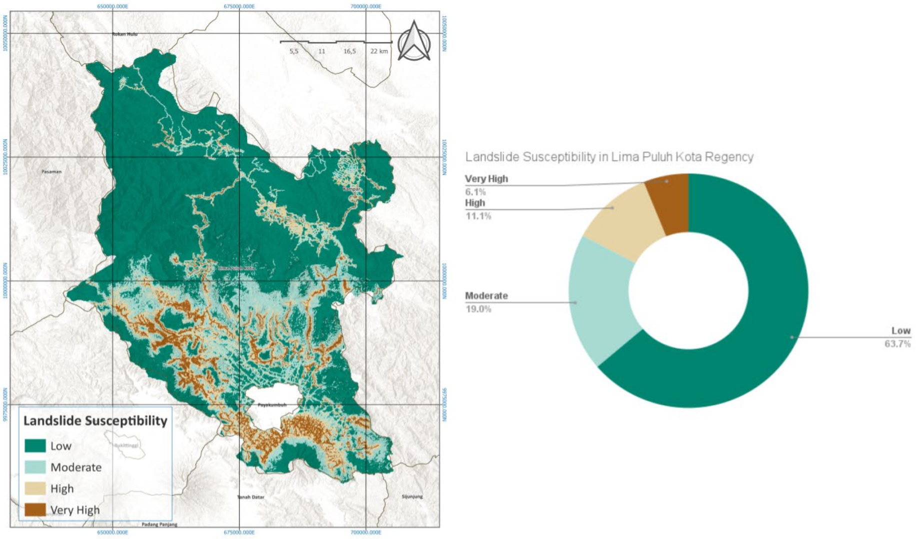

Fig. 4. Landslide susceptibility map and chart area

From the area of susceptibility level, it is found that research location is mostly covered by low susceptibility (67.75%) then followed by moderate (19.04%). Regarding the map, high (11.10%) and very high level (6.11%) are grouped and elongated in any specific region.

3.3 Variable Contribution

It is necessary to assess the importance of variables to discover the characteristics of the landslide itself. randomForest package in R provides the feature to do so. Importance variable means that if any variable is removed, it will decrease the accuracy. As the result of random forest assessment, variable importance is represented by mean decrease accuracy and mean decrease gini.

Table 3. Variable Importance

Variable	Mean Decrease Accuracy	Mean Decrease Gini
Aspect	0.65	4.88
Curvature	-1.92	1.19
Elevation	24.97	12.83
Geology	15.25	6.71
LULC	8.19	4.66
Rainfall	7.62	2.33
River Distance	9.3	4.12
Road Distance	50.75	27.13
Slope	6.55	6.54
Soil	3.72	2.87

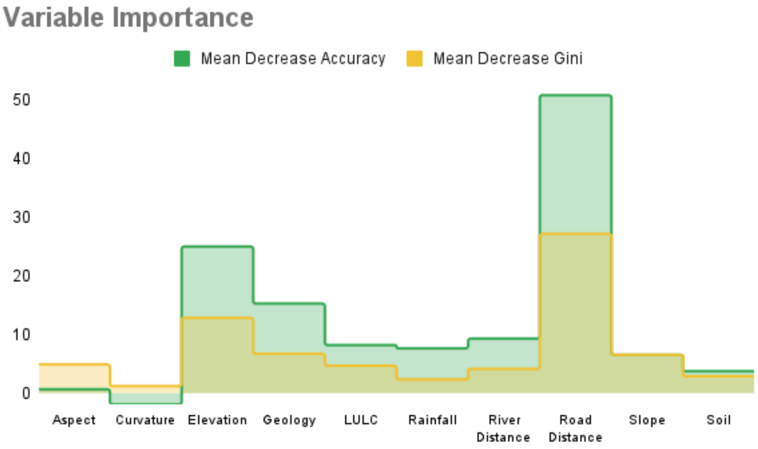

Fig. 5. Variable importance chart

Road distance is the most important variable in this analysis (MDA = 50.75, MG = 27.13), then followed by elevation (MDA = 24.97, MDG = 12.83) and geology type (MDA = 15.25, MDG = 6.71) because they are high both in mean decrease accuracy and mean decrease gini. While the weakest variable is curvature (MDA = -1.92, MDG = 1.19).

3.4 Landslide Susceptibility Area in Important Variable

Road distance, elevation, and geology are used for further spatial analysis to discover the characteristic of landslide susceptibility.

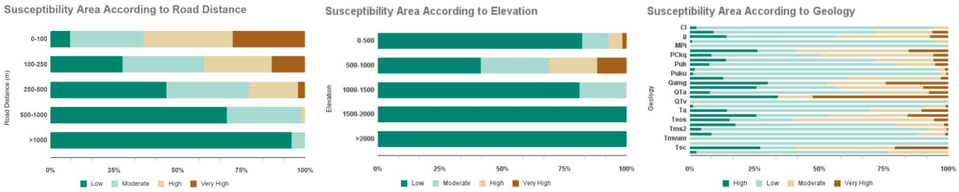

Fig. 6. Susceptibility proportion in important variables

Road distance has a gradual characteristic that the closer to the road, the susceptibility will be figured out as bigger. In the range of 0-100 m, low susceptibility was only found 7.7%, but high and very high susceptibility are 35.06% and 28.40%. Higher susceptibility will decrease if farther from the road and vice versa, such as in the range of >1000, there is no longer very high susceptibility and low susceptibility covers almost all its class (94.66%). Landslide susceptibility levels are vary and higher in the 0 until 1000 masl and increase following the elevation class. But elevation classes greater than 1000 are already covered by low susceptibility level. In the geological aspect, Qta and Tsc are the highest susceptibility, while Tmvam and Tpc are the lowest susceptibility.

3.5 Landslide Records in Important Variable

To figure out the characteristic of the landslide occurrence regarding the important variables, landside points were overlaid with landslide susceptibility and important variables. This analysis is useful to discover the landslide characteristic spatially.

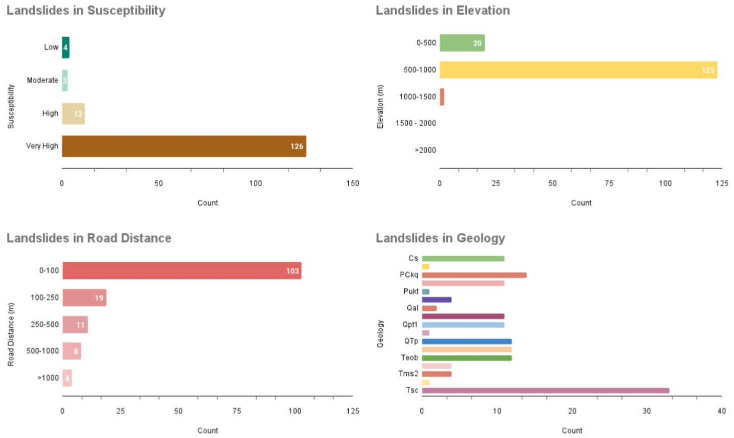

Fig. 7. Landslide records in susceptibility class and important variables

Landslide records are mostly found in very high susceptibility level (126 points), then followed by high level. According to the road distance, 103 landslides are found in a range of 0-100m, then the number decreases following the distance. The relation between susceptibility and road distance point that the higher susceptibility is closer to the road, the map also shows the distribution of landslide points and higher susceptibility elongated following the road pattern. Regarding the elevation, it is related to the area with the higher susceptibility found between 500 - 100 masl (123 points). Tsc and Pckq are the geology type which contain landslide point the most.

4 Conclussion

Random forest is a powerful algorithm because of its accuracy. Based on landslide records and predictor variables analyzed using Random forest, the model resulted 0.8137 AUC value which means the model is good for predicting landslide susceptibility. Road distance, elevation, and geology are the most important variables to model the result. Considering the area of susceptibility levels and the number of landslide records, the closer to the road, the wider of higher susceptibility. Plateau or mountainous area (500 - 1000 masl) is the elevation class that has higher susceptibility and landslide records oiver other elevation levels. While in geological aspect, Qta and Tsc are the highest susceptibility and the landslide records found mostly in Tsc.

Acknowledge

The authors would like to thank Lembaga Penelitian dan Pengabdian Masyarakat Universitas Negeri Padang for funding this work with a contract number: 914/UN35.13/LT/2022.

References

[1] Tsangaratos, P., Ilia, I.: Landslides using a modified decision tree classifier in the Greek 10.1007/s10346-015-0565-6Xanthi perfection Susceptibility mapping. Landslide 13, 305-320.Doi: 10.1007/s10346-015-0565-6 (2016)

[2] Poudyal, C.P.: Landslide susceptibility analysis using decision tree method, Phidim, Eastern Nepal. Bull. Dep. Geol. Vol 15 2012. [3] Yeon, Y.-K., Han, J.-G., Ryu, K.H., 2010. Landslide susceptibility mapping in Injae, Korea, using a decision tree. Eng. Geol. 116, 274–283. doi: 10.1016/j.enggeo.2010.09.009. (2013)

[3] Yeon, Y.-K., Han, J.-G., Ryu, K.H.: Landslide susceptibility mapping in Injae, Korea, using a decision tree. Eng. Geol. 116, 274–283. doi: 10.1016/j.enggeo.2010.09.009. (2010)

[4] Klose, M., Damm, B., & Terhorst, B.: Landslide cost modeling for transportation infrastructures: a methodological approach. Landslides, 12(2), 321-334. (2015)

[5] Hong H, Naghibi SA, Pourghasemi H, Pradhan B GIS-based landslide spatial modeling in Ganzhou City, China. Arab J Geosci 9:112. doi:10.1007/s12517-015-2094-y(2016)

[6] Tsangaratos P, Ilia I.: Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: the influence of models complexity and training dataset size. CATENA 145:164–179 (2016)

[7] Ilia I, Tsangaratos P.: Applying weight of evidence method and sensitivity analysis to produce a landslide susceptibility map. Landslides 13(2):379–397 (2016)

[8] Chen W, Chai H, Zhao Z, Wang Q, Hong H.: Landslide susceptibility mapping based on GIS and support vector machine models for the Qianyang County, China. Environ Earth Sci 75:474. doi:10.1007/s12665-015-5093-0 (2016)

[9] Pourghasemi HR, Mohammady M, Pradhan B.: Landslide susceptibility mapping using index of entropy and conditional probability models in GIS: Safarood Basin, Iran. CATENA 97:71–84 (2012)

[10] Akinci, H., Kilicoglu, C., & Dogan, S.: Random forest-based landslide susceptibility mapping in coastal regions of Artvin, Turkey. ISPRS International Journal of Geo-Information, 9(9), 553. (2020)

[11] Taalab, K., Cheng, T., & Zhang, Y.: Mapping landslide susceptibility and types using Random Forest. Big Earth Data, 2(2), 159-178. (2018)

[12] Darminto, M. R., & Chu, H. J.: Mapping landslide release area using Random Forest Model. In IOP Conference Series: Earth and Environmental Science (Vol. 389, No. 1, p. 012038). IOP Publishing. (2019)

[13] Breiman L: Random forests. Mach Learn 45:5–32 (2001)

[14] Kim, J. C., Lee, S., Jung, H. S., & Lee, S.: Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. Geocarto international, 33(9), 1000-1015. (2018)

[15] Wang, Y., Sun, D., Wen, H., Zhang, H., & Zhang, F.: Comparison of random forest model and frequency ratio model for landslide susceptibility mapping (LSM) in Yunyang County (Chongqing, China). International journal of environmental research and public health, 17(12), 4206. (2020)

[16] Hansen L, Salamon P..: Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 12:993–1001(1990)

[17] Lu, W., Zhou, Z., Liu, T., & Liu, Y. H.: Discrete element simulation analysis of rock slope stability based on UDEC. In Advanced Materials Research (Vol. 461, pp. 384-388). Trans Tech Publications Ltd. (2012)

[18] Hong, H., Tsangaratos, P., Ilia, I., Chen, W., & Xu, C. Comparing the performance of a logistic regression and a random forest model in landslide susceptibility assessments. The Case of Wuyaun Area, China. In Workshop on world landslide forum (pp. 1043-1050). Springer, Cham. (2017)