Geo-Resilience 2023 Conference, Cardiff, Wales
D.G. Toll & M.G. Winter (Eds)
British Geotechnical Association, London. 2024
https://doi.org/10.53243/Geo-Resilience-2023-5-4

Random Forest and Frequency Ratio: A Comparison of Methods for Landslide Susceptibility Mapping

José Maria DOS SANTOS RODRIGUES NETO[1], Netra Prakash BHANDARY[2]

[1]Ehime University Graduate School of Science and Engineering, Matsuyama, Japan
[2]Affiliation Ehime University Faculty of Collaborative Regional Innovation, Matsuyama, Japan
Corresponding author: José Maria dos Santos Rodrigues Neto (rodrigues_neto94@hotmail.com)

Abstract

In this study, we perform an efficiency comparison between two methods for landslide susceptibility mapping (LSM): the machine learning technique random forest (RF), which is evidenced to be the most efficient for landslide susceptibility assessment tasks and the statistics-based frequency ratio (FR) method. We have selected an area comprising Kure City and its neighbourhood in Southern Hiroshima Prefecture in Japan. This area is known to suffer from frequent rain-induced landslide disasters and the most recent one occurred in July 2018. Both the above LSM methods require a collection of landslide conditioning factors (LCFs), which in this study are: (1) geology; (2) altitude; (3) slope angle; (4) slope aspect; (5) drainage density; (6) soil profile; (7) land use; (8) distance from lineaments; and (9) mean annual precipitation. The rainfall LCF data comprise of XRAIN (eXtended RAdar Information Network) radar records, which are novel in the task of LSM production. The accuracy of the produced LSMs was calculated with the receiver operating characteristic's (ROC) area under curve (AUC), giving a result of 0.84 for the FR method and 0.92 for the RF method. It is also noteworthy that the RF method is substantially swifter and more practical than the FR method and allows for multiple and automatic experimentations with different parameters, providing fine and accurate outcomes with the given data. The results evidence that machine learning techniques such as the RF method are most advisable for dealing with hazard assessment problems such as the one exemplified in this study, and that XRAIN radar-acquired mean annual precipitation data is effective when used as a LCF in the production of LSMs.

Keywords: landslide, machine learning, random forest, frequency ratio, susceptibility map, XRAIN

1. Introduction

Landslides are common natural disasters that kill a few thousand people worldwide every year. If landslide prevention methods are not developed further, the damage caused by such disasters is expected to increase in the next few years due to urbanization, deforestation and climate change. A recent case of wide-area landslide disasters in Japan was recorded in July 2018. The landslides together with massive flooding in a large part of Southwest Japan were caused by heavy rains. During the course of about 10 days from 28th June until 8th July, the rainfall reached as much as 1800 mm on the island of Shikoku and 1200 mm in Tokai region. Many cities recorded more than 400 mm of rainfall over 72 hours (Japan Meteorological Agency, 2018).

Property loss caused by the July 2018 disasters is estimated to be ¥1.09 trillion (i.e., about US$ 10 billion), including damage to industries and public infrastructures. Although emergency warnings were issued for eight prefectures, the death toll caused by the landslides and floods during the July 2018 disasters was above 225 people. In Hiroshima Prefecture, one of the most affected areas was Kure, with 24 deceased due to landslides. Additionally, most transportation lines into the city (except maritime ways) were cut off and 760 houses were damaged.

One of the strategies for minimizing the damage of landslide disasters is the production and use of landslide susceptibility maps, which assess the probability of landslide occurrence in an area considering slope failure-related factors and the actual occurrence of past landslides in a GIS platform. Currently, there are various methods of calculating spatial landslide probability and producing landslide susceptibility maps. Yilmaz et al. (2009) defend that the Frequency Ratio (FR) method is one of the most practical and efficient methods for landslide susceptibility calculation in GIS platform. However, advancements in programming and computation technology in recent years have put through the extensive use of machine learning (ML) methods in myriads of areas of application, including the domain of natural hazards and landslide susceptibility assessment, as advocated by Goetz et al. (2015), Youssef & Pourghasemi (2021).

One of the landslide conditioning factors (LCFs) commonly used in the production of landslide susceptibility maps is precipitation volume, which may be measured with various methods, one of the most effective ones for spatial-related tasks being radar-based methods, such as XRAIN (eXtended Radar Information Network). XRAIN data is operated by the Ministry of Land, Infrastructure, Transport and Tourism (MEXT) and made available through University of Tokyo's Data Integration & Analysis System (DIAS) platform. Starting operation in 2014, XRAIN comprises of a real time rainfall measuring system based on Multi-factor (MP) radars, which allow for more spatially accurate measurements of rainfall volume (Data Integration & Analysis System, 2022).

This study aims to produce landslide susceptibility maps through both the frequency ratio (FR) method and the random forest (RF) machine learning method for the area of Kure City, Southern Hiroshima, and compare the efficiencies of both methods. The current study has the particularity of utilizing long-range radar acquired rainfall XRAIN data as one of the factors for landslide probability mapping. Advancements in the landslide susceptibility assessment methods (such as the use of ML or radar-acquired rainfall data) may lead to more efficient strategies in minimizing damage caused by landslide disasters.

2. Methodology

2.1 Study area

For this study, a rectangular area of 390.5 sq. km. (approximately 28 km x 14 km) covering Kure City, south of Hiroshima Prefecture, was used (Figure 1). A geographically small port town adjacent to the Seto Inland Sea, Kure started as a shipbuilding facility in the end of 19th century. Soon made into a major dockyard and military base, the port and the city around it grew quickly due to the Imperial Japanese Navy's and its facilities' rapid development until the end of World War 2. However, the area's flat terrains are cramped and limited by mountains (a common scenario in Japan), which forced the town's rapid expansion into and near adjacent hills. The most predominant bedrock lithology of the area, the Hiroshima Granitic Rocks group, is very easily weathered, changing into a soil commonly referred to as Masado. Masado granitic soil is known to have good permeability and be very brittle when wet, which causes it to be prone to lose its structure and stability when infiltrated, and thus be a very supscetible soil for landslide occurrence in heavy rainfall events.

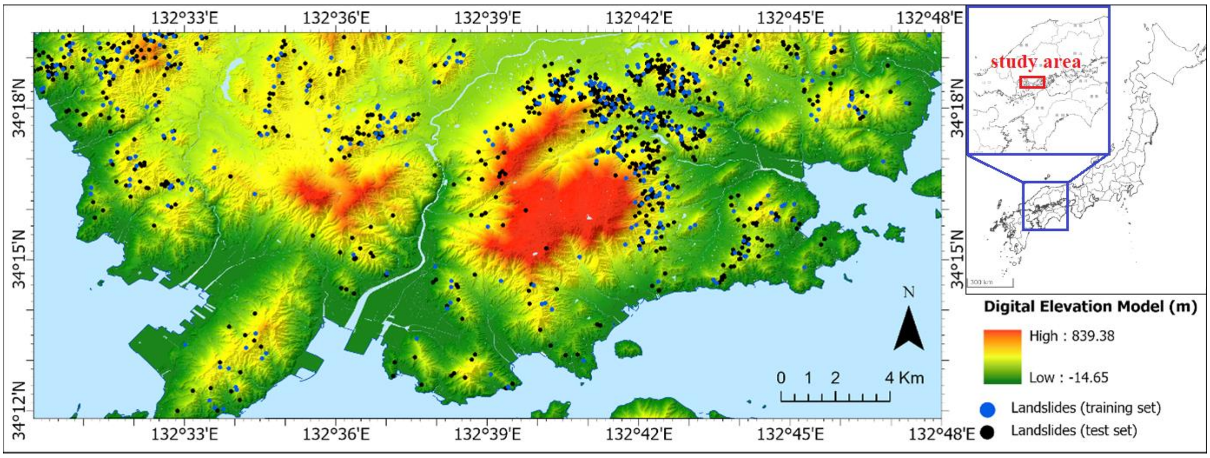

Figure 1: Localization map of the study area, in Kure City, Southern Hiroshima, along with landslide points referent to the July 2018 disasters.

2.2 Landslide Susceptibility Map (LSM) production

In this study, two different methods for production of the LSM were experimented, one being a statistical approach (frequency ratio) and the other a machine learning technique (random forest), in order to find which is most efficient for the intended objectives. Both LSMs contain 432,258 20-meters pixels. For both methods, final results were separated in 5 susceptibility zones (from very low to very high) with the natural breaks (Jenks) method.

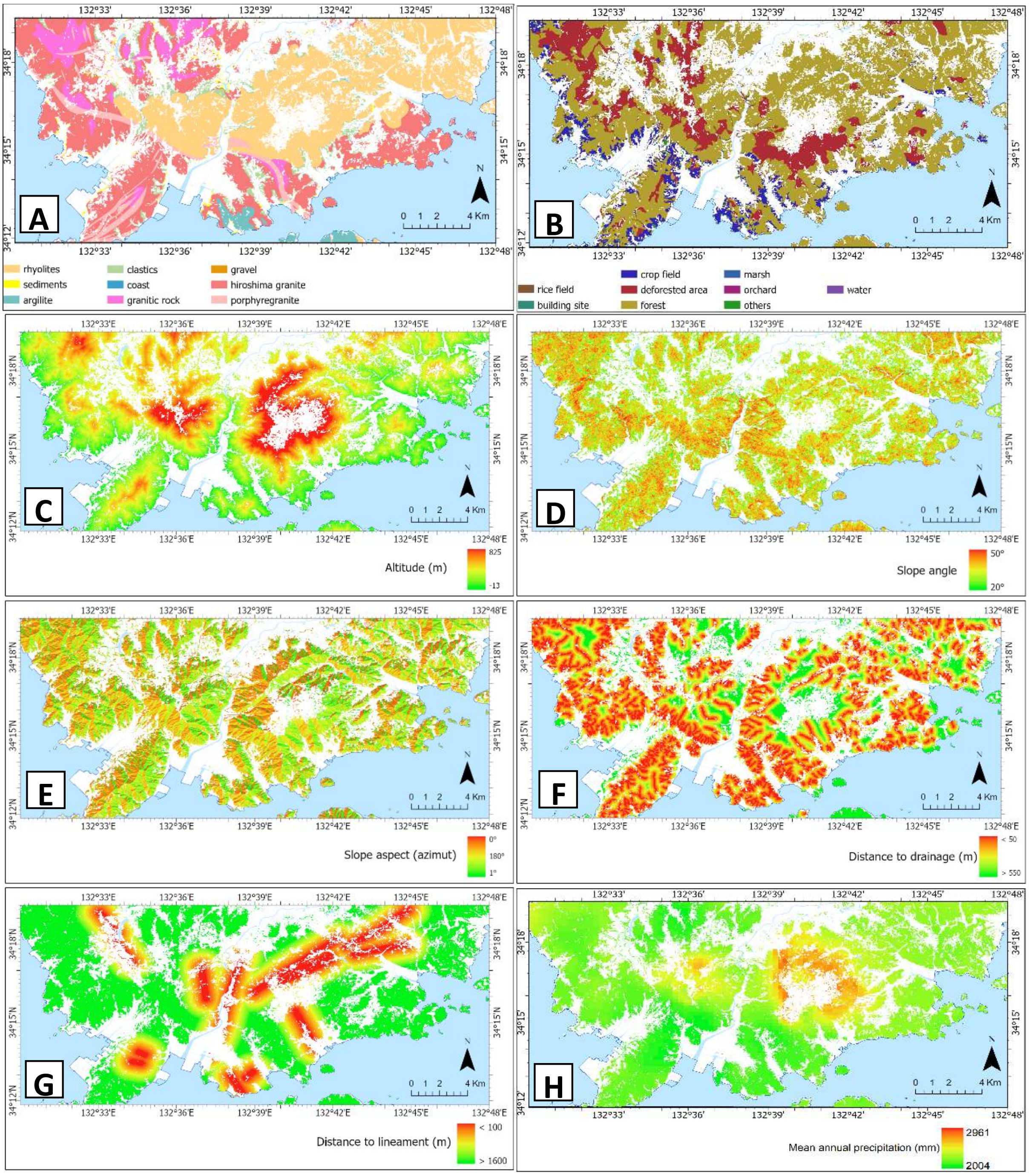

Figure 2: LCFs used in the LSM production: (a) geology, (b) land use, (c) altitude, (d) slope angle, (e) slope aspect, (f) distance to drainage, (g) distance to linemanet, (h) mean annual precipitation. White areas comprise of slopes lower than 20° or higher than 50°, which were left out of the analysis for being considered not prone to landslides.

An identical collection of landslide conditioning factors (LCFs) were used in shapefile form for both FR and RF attempts: (a) geology; (b) altitude; (c) slope angle; (d) slope aspect; (e) drainage density; (f) soil profile; (g) land use; (h) distance from lineaments; and (i) mean annual precipitation (Figure 2). The Digital Elevation Model as well as other LCF shapefiles such as drainage, land use and soil profile maps were provided by the Geospatial Information Authority of Japan (2018). The geological and lineament data was extracted from the Kure Geological map by Higashimoto et al. (1985). The mean annual precipitation factor comprises of recorded rainfall with XRAIN technology. Most LSM attempts take use of interpolated rain gauge station measurement data, which is not as spatially accurate. It is expected that XRAIN data, used as an LCF, may provide good results in the LSM production. Each pixel of the XRAIN data mesh used as a LCF has dimensions 280x230 meters, accounting for 4999 pixels in the study area in a 96x67 grid. The collection of landslides used as output data in both methods comprises of 1177 landslide points referent to the July 2018 disasters, which were mapped and provided by the Geospatial Information Authority of Japan (2022) with aerial photography (Figure 1). The landslide points were randomly separated into training and test sets, with a ratio of 30% for the training set (353 landslide points) and 70% for the test set (824 landslide points).

2.3 Frequency ratio (FR) method

The frequency ratio method uses the assumption that landslide occurrence is determined by factors that are related to the event, and thus new landslides will generally occur under the same conditions of past landslides (Lee & Talib, 2005; Yilmaz, 2009; Rasyid et al., 2016). In the frequency ratio method, FR values represent the ratio between the landslide occurrence and the area of specified factor for a given factor class. After the frequency ratio is calculated for each LCF class, the values are summed in each pixel of the susceptibility map to calculate the landslide susceptibility index (LSI) for that specific pixel (Lee and Talib, 2005; Yilmaz, 2009). Once the LSI values are established for each pixel of the map, a final landslide susceptibility map is produced, where higher LSI values represent higher risk of landslide occurrence in that location.

2.4 Machine learning random forest (RF) method

Machine learning (ML) is a form of artificial intelligence comprising of methods where a system "learns" based on a set of data, looking for patterns in it and how they affect a certain result relative to a problem. It was found that this technology is successful in completing varied specialized tasks, when set up with a sufficient dataset and adequate parameters. The utilization of ML methods in the domain of natural hazards (including landslide susceptibility assessment) is supported by Goetz et al. (2015), Youssef & Pourghasemi (2021). Some of the advantages compassed by the use of MLs include adjusting its internal structure to the experimented data, as well as efficiency and practicability even in large areas (Youssef & Pourghasemi, 2021). Bibliographical research suggests that the ML technique judged most effective for the specific case of landslide risk assessment mapping is the random forest (RF) technique (Youssef & Pourghasemi, 2021, Yilmaz et al. 2009).

The RF technique is actually an expansion of another ML method, the decision tree. The decision tree technique is a supervised ML technique where the algorithm observes a provided dataset and looks for patterns that creates results, observed in the test dataset. The algorithm then recreates these patterns in a "tree", where each decision (variance in the data) creates a new "branch", until these finish at the results, or a "leaf". Decision trees learners, however, are prone to create over-complex trees, which do not reflect an accurate representation of the data in which is known as overfitting. This is solved with the random forest ML technique, which, as the name suggests, is an ensemble model comprising of a "forest" with many decision trees. Each tree is a completely random and independent experiment, which prevents overfitting by outputting a result comprising of an ensemble averaged prediction for all the decision trees in the random forest (Youssef & Pourghasemi, 2021). In this study, the RF algorithm utilized was the one provided in the scikit-learn ML library. Once the optimal parameter values for the algorithm were designated through automatic testing and evaluation, a final prediction model was executed and the resulting predictions were then laid in map form using ArcGIS Pro, providing the RF-based landslide susceptibility map.

2.5. LSM validation method

To assess the performance of the produced LSMs, the receiver operating characteristic (ROC) analysis was employed in this study. Initially developed for radar accuracy tests, the ROC method is recommended for landslide zoning validation tasks due to its threshold-independent nature (Beguería, 2006, Corominas et al., 2014), that is, it doesn't require a fixed value to determine that either negates or requires a landslide activation, since LSI is a probability assessment, not a deterministic one. Thus, ROC analysis uses multiple thresholds with different interspaces, and the points in the ROC curve represent these possible cutoff thresholds given by the multiple cases in a model (i.e., LSM). The area under curve (AUC) of the ROC curve value is used as a metric to assess the quality of the LSM, where a larger area (ranging from 0.5 to 1) represents better prediction performance, that is, how well the model separates the validations landslides throughout the susceptibility zones of the LSM. For that reason, AUC value is used as the primary meter for LSM accuracy in this study (Beguería, 2006; Corominas et al., 2014). As an additional method for LSM validation, landslide density is checked for each one of the 5 LSI zones attributed in the LSM. It expected that in an efficient LSM, the landslide density distribution will follow a proportionally direct growth with each zone change, from very low to very high.

3. Results

3.1 Frequency ratio method LSM

The calculation of FR values for each of the LCF classes used in the LSM production are presented in Figure 3. Calculation of the ROC curve's AUC for this LSM (Figure 5) resulted in a score of 0.84, considered "good" for landslide susceptibility assessment methods (Rasyid et al., 2016). The distribution of landslide density for each one of the 5 LSI zones is shown in Figure 6. It is observed that the FR method presents good distribution of landslide density throughout the 5 susceptibility zones, showing good application for risk assessment tasks in the field of disaster prevention.

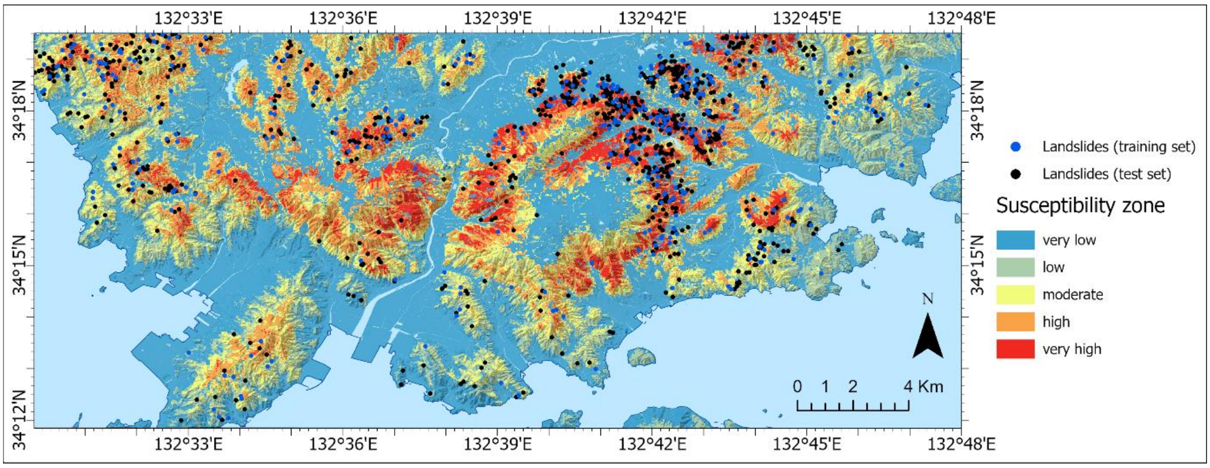

Figure 3: Landslide susceptibility map (LSM) produced with the frequency ratio (FR) statistical method, along with landslide points from the July 2018 disasters (both training and test sets).

3.2 Machine learning random forest method LSM

Execution of the random forest algorithm using the LCFs provided LSI predictions which were then inserted into map view, producing the final RF LSM (Figure 4). This map showed an AUC value of 0.92, a rating considered "excellent" for susceptibility assessment (Figure 5). Figure 6 shows the distribution of landslide density throughout the 5 susceptibility zones of the LSM. Compared to the FR LSM, there is a more regular distribution on density of landslides on the high and very high zones, which accounts to the higher accuracy calculated in the ROC AUC validation method.

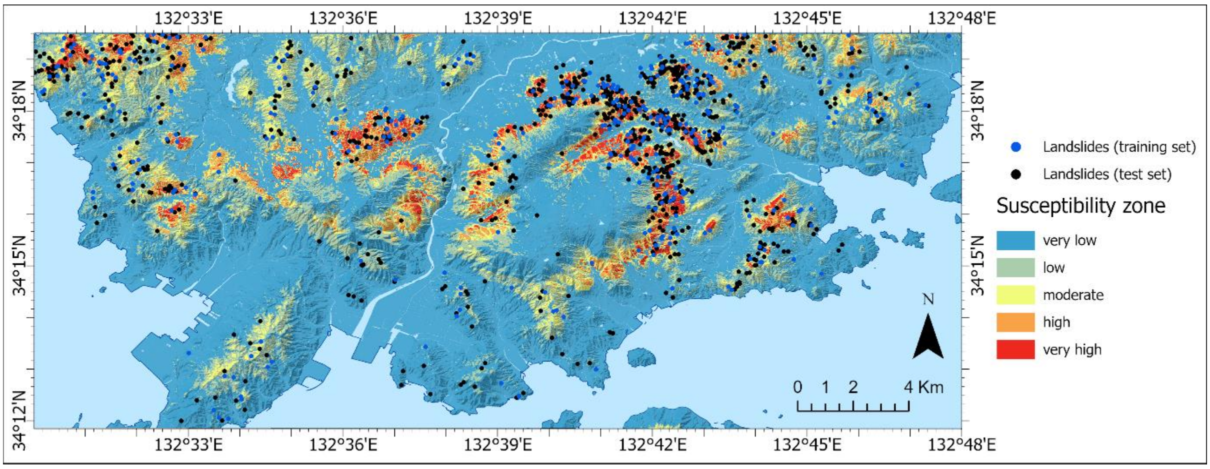

Figure 4: Landslide susceptibility map (LSM) produced with the random forest (RF) machine learning method, along with landslide points from the July 2018 disasters (both training and test sets).

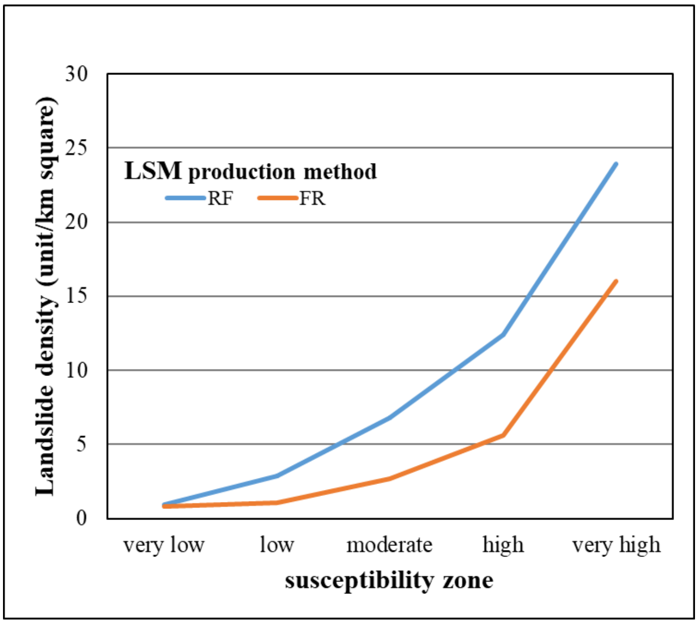

Figure 5: Receiver operating characteristic (ROC) area under curve (AUC) graphs for maps produced with FR (orange line) and RF (blue line) methods.

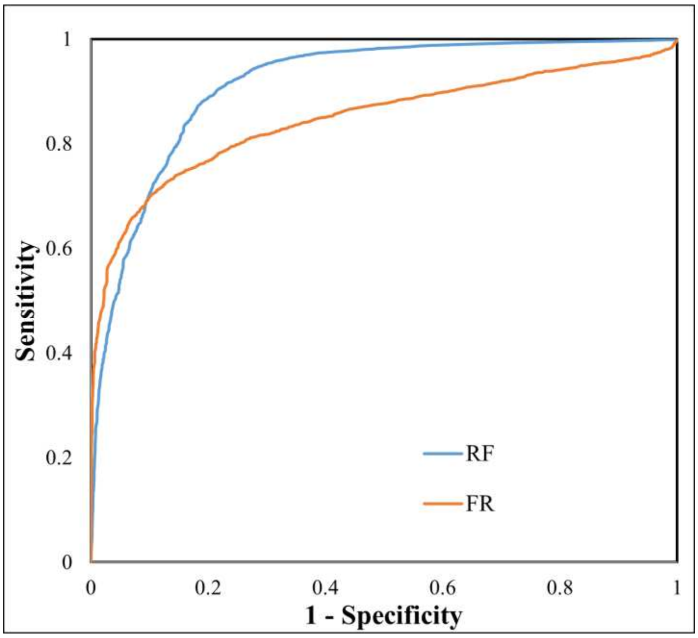

Figure 6: Graphs representing landslide density distribution in units per square kilometer in the susceptibility zones for maps produced with FR (orange line) and RF (blue line) methods.

4. Conclusions

This study attempted to produce two LSMs with identical sets of LCFs and landslide test and training data, one with the statistical frequency ratio (FR) method, and another with the machine learning random forest (RF) method. Verification with ROC AUC method showed that both methods presented satisfactory results, with a rating of 0.84 for the FR method, and 0.92 for the RF method. Both maps also showed good performance with the use of XRAIN radar-acquired rainfall data as an LCF.

Although the overall quality is verified to be higher in the LSM produced with the RF method, the FR method provides a visualization of FR values for each class of the LCFs, which provides a good opportunity for readily understanding of how each LCF and its patterns may more or less influence activation of landslides. Since the random forest method takes use of an ensemble of decision trees, it's not possible to examine the specific decision-taking process and the intricacies of each LCF and its influence in a final prediction. However, it was verified by the ROC AUC values that the RF LSM shows a better fit and thus higher efficiency in risk assessment tasks. Moreover, the automated process of the RF method provides the possibility of multiple experimentations with varied parameters, as well as presents a much more practical process for the whole approach, once the mechanisms of putting the algorithm into use are properly grasped. For these reasons, it is evidenced that the use of RF and other ML techniques are advisable for tasks in the natural disaster risk assessment area, such as the in the case of landslides susceptibility mapping. Moreover, the utilization of XRAIN radar-acquired annual precipitation data as an LCF to produce LSMs was attempted and verified as effective.

It is recommended that future studies in the field include experimentations of other ML techniques other than the RF method, as well as the attempt of similar mappings in different areas, since the intricate qualities of various region and terrains may greatly affect the landslide process.

Acknowledgments

This research was made possible through the grant for foreign students provided by Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT). XRAIN data is made available by the Data Integration & Analysis System (DIAS), The University of Tokyo, and is sponsored by MEXT.

References

Beguería, S. (2006). Validation and evaluation of predictive models in hazard assessment and risk management. *Nat Hazards,* 37(3):315–329.

Corominas, J., van Westen, C., Frattini, P., Cascini, L., Malet, J.-P., Fotopoulou, S., Catani, F., Van Den Eeckhaut, M., Mavrouli, O., Agliardi, F., Pitilakis, K., Winter, M G., Pastor, M., Ferlisi, S., Tofani, V., Hervás, J., Smith, J. T. (2014). Recommendations for the quantitative analysis of landslide risk. *Bull Eng Geol Environ,* 73 (2):209–263.

Data Integration & Analysis System (2020). XRAIN Realtime Precipitation Data. The University of Tokyo, sponsored by the Ministry of Education, Culture, Sports, Science and Technology. https://diasjp.net/. Accessed June 2022.

Geospatial Information Authority of Japan (2022). Database. www.gsi.go.jp. Accessed June 2022.

Goetz, J.N., Brenning, A., Petschko, H., Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput. Geosci,* 81, 1–11.

Higashimoto, S., Hirohisa, M., Mizuno, K., Kawada, K. (1985). Kure 1:50,000 geological map 13-14, NI-53-33-7. Geological Survey of Japan.

Lee, S., Talib, J.A. (2005). Probabilistic landslide susceptibility and factor effect analysis. *Environmental Geology*, 47(7), 982–990.

Rasyid, A.R., Bhandary, N.P., Yatabe, R. (2016). Performance of frequency ratio and logistic regression model in creating GIS based landslides susceptibility map at Lompobattang Mountain, Indonesia. *Geoenvironmental Disasters*, 3(1).

Yilmaz, I. (2009). Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: A case study from Kat landslides (Tokat—Turkey). *Computers & Geosciences*, 35(6), 1125–1138.

Youssef, A.H., Pourghasemi, H.R. (2021). Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. *Geoscience Frontiers*, V. 12, I. 2.