6th Workshop on Environmental Engineering
Programa de Engenharia Ambiental | UFRJ
September 20-22nd, 2023

Landslide Susceptibility Modelling by Machine Learning: Angra dos Reis City (RJ), Brazil

Amanda Alves da Silva[1], Marcos Barreto de Mendonça[1] and André de Souza Avelar[1]
[1] Programa de Engenharia Ambiental, UFRJ, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 21941-901, Brazil.

amandaalves@poli.ufrj.br; mbm@poli.ufrj.br; andreavelar@poli.ufrj.br

Abstract:
This work carried out a landslide susceptibility modelling in Angra dos Reis city, once is an area where recurrent mass movements occurred. After a review of the 98 references for last decade, we identified many studies related to susceptibility mapping, statistical methodologies and machine learning. The first work step we proceeded a combination of the documentary and geospatial data, in order to prepare the landslide inventory from 2010 to 2021. The present modelling involves a combination of the geospatial data preparation (geology, geomorphology and land use parameters), grid sampling and supervised algorithms such as Principal Component Analysis and Logistic Regression. Classification analysis with Support Vector Machine (SVM) and Random Forest (RF) algorithms further evaluates model performance. Results indicate that SVM outperforms RF have more significant performance in several metrics and, it can be considered the best choice for this classification task, contributing to the assessment of landslide susceptibility analysis.

Keywords: Landslide Susceptibility, Machine Learning, Random Forest, Support Vector Machines.

1 Introduction

Angra dos Reis City is located on the seashore of Rio de Janeiro State (Figure 1) and present recurrent landslides, usually resulting in deads and injuries. Susceptibility map is necessary to allow temporal analysis in landslide causes, however new methodologies are emergent in order to be able to improve this map. Currently, there are several approaches to prepare landslide susceptibility maps, including heuristic, deterministic and statistical methods (Zêzere et al., 2017). In statistical methods, the weights related to the landslides are found by landslide inventories, based on previous occurrences. The premise of this approach is that past landslide events can inform future occurrences (Guzzetti et al., 1999; Fell et al., 2008).

Considering advances in Remote Sensing and Machine Learning (ML), some studies have employed new approaches to susceptibility mapping (Hader et al., 2022) such as analysis of landslide occurrences and non-occurrences and morphometric terrain attributes, employing the ML algorithms Random Forest (RF) and Support Vector Machines (SVM) to identify susceptible slopes (Marjanovic et al., 2011). The present work we perform a statistical Landslide Susceptibility Model (LSM), as well as which Landslide Conditioning Factors (LCFs) most influence the variability of the model. The final result of this model will support the Landslide Susceptibility Chart.

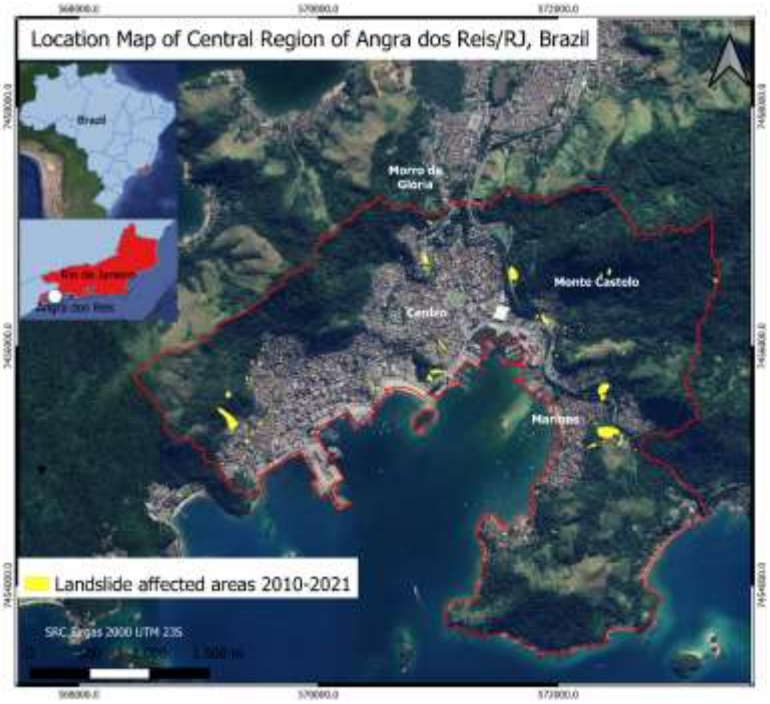

Figure 1: Study Area

2 Research Methodology

The research method began with an literature review focused on the RF and SVM algorithms in LSM. This review encompassed academic sources, such as articles, theses, and dissertations, resulting in the identification of 98 pertinent references. They were selected by themes, location, typology, and date, resulting in 68 papers, 12 textbooks, and 18 mixed (thesis, abstracts, meeting paper and reports). Inside this literature 23 focused on landslides, 12 on multivariate analisys, 27 on methodological and statistical aspects, and 36 on ML. Around 80% of references were international, reflecting a notary trend of increased in ML at the past decade. In addition to the literature review, data from the State Geological Survey (DRM/RJ) and Civil Defense of Angra dos Reis were analysed, focusing on mass movement at the city since 2010 to 2021.

2.1 Landslide Susceptibility Modelling (LSM) in Angra dos Reis

The study conducted in Angra dos Reis commenced by compiling landslide data, which was used to create a Landslide Inventory (LI). This inventory was established based on the geographical coordinates, temporal information, and geometrical characteristics of previous landslides identified in Google Earth images spanning from 2010 to 2021. These landslide polygons were superimposed onto a topographic map generated from a Copernicus DGED 30 digital elevation model and Landsat-8 imagery in the green, red, and infrared bands. Simultaneously, 11 LCFs were incorporated into the topographic map. These LCFs included lithology, geomorphology, elevation (hypsometry), slope, aspect, curvature, plane curvature, curvature profile, Normalized Difference Vegetation Index (NDVI), Normalized Difference Building Index (NDBI), and Topographic Wetness Index (TWI).

The 2nd phase of the study consisted of grid sampling with 30m resolution across the entire study area, aligning with the pixel resolution of the orbital images. Subsequently, all collected data was normalized and conducted for Weight Analysis between LCFs. In the third step, the analysis of LCF weights was performed using a combination of supervised algorithms, specifically Principal Component Analysis (PCA) and Logistic Regression (LR). LR, recognized as one of the widely used methods in multivariate analysis, was used due to its ability to solve non-linear problems, incorporating a regression function suitable for the case of binary variables. On the other hand, PCA algorithms helped determine the essential components among the 11 conditioning factors necessary to explain data variability with a probability of at least 95% and identify interactions between them.

After that, Classification Analysis was conducted using SVM and RF supervised classification algorithms. Both were employed to predict landslide susceptibility and calculate associated probabilities. Both were tested to determine their respective performance on the model. Subsequently, the LSM was validated using a confusion matrix, allowing a comparison between actual and predicted landslides. Consequently, a curve was established plotting the true positive rate against the false positive rate, producing an area under the curve (AUC), which served as a parameter to evaluate the model's ability to distinguish between susceptible and non-susceptible areas. Ideal models have an AUC close to 1.

3 Findings and Discussion

PCA analysis revealed a strong positive correlation between the slope, elevation, and NDVI parameters, as well as a strong inverse correlation between NDVI, NDBI, and TWI. Furthermore, among the LCFs inputted in the LSM, only 7 parameters are sufficient to explain 95% of the data variability. The LR summary reinforced the considerations raised by the PCA analysis, indicating that the 7 LCFs are the most influence the model (by importance order): elevation, slope, NDVI, Curvature Profile, Lithology, NDBI and Curvature. The application of RF method showed an average accuracy of 76% for both classes “No occurrence of landslides” and “Occurrence of landslides”. The AUC resulted in 0.85.

The SVM model (Figure 2a) showed an average accuracy of 89% for both classes, highlighting the algorithm's robustness in classification. The ability of the algorithm correctly identify positive cases in 61% for “No occurrence of landslides” and a surprising 99% for “Occurrence of landslides”. The AUC of the SVM surpassed that of the RF, reaching a value of 0.90. The analysis of the false negative results obtained by applying the RF and the SVM, brought to light different results for both algorithms. In a new round of analysis focused only on occurrence values, it was evident that the SVM algorithm excelled in predicting more true positives compared to false negatives, while the RF (Figure 2b) leaned toward predicting more false negatives than true positives.

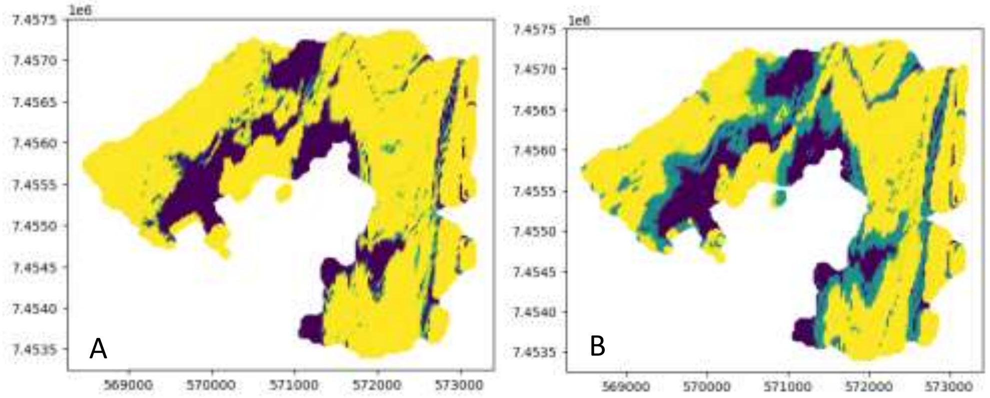

Figure 2: LSM Charts generated with SVM and RF

4 Conclusion

Weight analysis through PCA and LR revealed that 7 out of 11 LCFs are the most influential on LSM variability. The SVM algorithm showed better performance in LSM compared to RF. The false negative check found that the RF method misclassified more false negative values, resulting in poor performance.

5 Acknowledgment

We thank to the CAPES/Brazil - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior do Brasil (finance code 001) for their funding support.

6 References

Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., & Savage, W. Z. (2008). Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Engineering geology*, *102*(3-4), 85-98.

Guzzetti, F., Reichenbach, P., Cardinali, M., Galli, M., & Ardizzone, F. (2005). Probabilistic landslide hazard assessment at the basin scale. *Geomorphology*, *72*(1-4), 272-299.

Hader, P. R. P., Kaiser, I. M., Manzato, G. G., & Peixoto, A. S. P. (2019). Hazard Assessment of Landslides Disasters in the City of Cubatão, State of São Paulo, Brazil. In *International Congress on Engineering and Sustainability in the XXI Century* (pp. 1087-1101). Cham: Springer International Publishing.

Zêzere, J. L., Pereira, S., Melo, R., Oliveira, S. C., & Garcia, R. A. (2017). Mapping landslide susceptibility using data-driven methods. *Science of the total environment*, *589*, 250-267.