Machine learning applications for modeling and mapping soil erosion in tropical regions

Francisco Hélter Fernandes do Amaral[1], Andrés Velastegui-Montoya[2,3], Eder MS de Paula[4]

[1] Postgraduate Program in Geography, Federal University of Pará, Belém, Brazil – francisco.amaral@ifch.ufpa.br
[2] Faculty of Engineering in Earth Sciences, ESPOL Polytechnic University, Guayaquil, Ecuador – dvelaste@espol.edu.ec
[3] Laboratory of Geoinformation and Remote Sensing, Faculty of Engineering in Earth Sciences, ESPOL Polytechnic University, Guayaquil, Ecuador
[4] Faculty of Geography, Federal University of Pará, Belém, Brazil – edermileno@ufpa.br

Keywords: Climate change, Soil loss, Environmental degradation, Machine learning

Abstract

Soil erosion represents a major environmental issue that threatens ecosystem integrity and land sustainability, making the development of reliable susceptibility models crucial for supporting mitigation and management policies. This study evaluates the potential of three machine learning algorithms Weighted Subspace Random Forest (WSRF), Regularized Random Forest (RRF), and Naive Bayes (NB) for soil erosion susceptibility mapping in the Pardo River watershed, situated between the states of São Paulo and Minas Gerais, Brazil. A total of 120 sampling locations, including erosion and non-erosion occurrences, were identified through field surveys and high-resolution imagery obtained from Google Earth Pro. Initially, fifteen conditioning factors related to erosion processes were considered; however, after applying multicollinearity and relevance analyses, thirteen variables were retained for the final modeling framework. To evaluate model robustness, the dataset was randomly partitioned into training (70%) and testing (30%) subsets. Model performance was assessed using statistical indicators, including accuracy and AUC-ROC metrics. The NB, RRF, and WSRF models achieved accuracy values of 0.87, 0.89, and 0.88, respectively, while the corresponding AUC-ROC values reached 0.93, 0.96, and 0.95. Among the evaluated approaches, RRF yielded the highest predictive performance, demonstrating the effectiveness of machine learning techniques for supporting sustainable land management and erosion-prone area conservation. In addition, the proposed methodological framework offers a transferable approach for future susceptibility studies and contributes to expanding geospatial modeling applications across different environmental settings.

1. Introduction

Soil erosion has become one of the major environmental challenges affecting watershed systems, as it compromises land productivity, disrupts basin stability, and contributes to environmental degradation in both upstream and downstream areas. In addition, erosion processes increase the likelihood of flooding and intensify ecosystem deterioration across landscapes (Velastegui-Montoya et al., 2023). The severity of this process is often amplified by anthropogenic activities, including infrastructure expansion, intensive agricultural practices, and the unsustainable exploitation of natural resources, which accelerate soil degradation and landscape instability (Khosravi et al., 2023). From a biophysical perspective, soil erosion involves the removal and degradation of the upper soil layer, resulting in the loss of organic matter and essential nutrients that are fundamental for soil fertility and agricultural productivity (Lu et al., 2023). The occurrence and intensity of erosion are controlled by multiple interacting factors, such as vegetation cover, land use patterns, topographic conditions, soil properties, and management practices, highlighting its complex and dynamic nature (Fenta et al., 2020).

In this context, soil erosion is one of the major challenges affecting ecosystems, especially in tropical regions such as Brazil, where deforestation and intensive agriculture exacerbate the problem, demanding more effective anticipatory actions. Traditionally, empirical models such as the Revised Universal Soil Loss Equation (RUSLE) and multicriteria decision-making methods, including the Analytical Hierarchy Process (AHP), have been widely used to perform preliminary assessments of areas susceptible to water erosion (Mosavi et al., 2020). However, the inherent limitations of these methods restrict their applicability on broader scales, highlighting the need for new methodological approaches (Mohammed et al., 2020).

Recent technological developments have significantly transformed environmental modeling by expanding the availability of computational resources and geospatial data processing techniques. The growing accessibility of programming environments such as R and Python, together with advances in remote sensing technologies and cloud-based data processing platforms, has facilitated the application of data-driven approaches based on machine learning (ML) and deep learning (DL) methods (Mosavi et al., 2020; Band et al., 2020). Compared with conventional techniques, these approaches have demonstrated greater capability in representing complex environmental interactions and producing more accurate and efficient physical–geographical models (Lu et al., 2023).

Within this context, the present study evaluates the predictive performance of the Weighted Subspace Random Forest (WSRF), Regularized Random Forest (RRF), and Naive Bayes (NB) algorithms for soil erosion susceptibility mapping in the Pardo River Basin. Furthermore, this research seeks to address existing gaps related to the application of ML techniques for soil erosion studies in Brazil by providing a robust analytical framework capable of supporting conservation planning and the development of sustainable land-use strategies.

Although the machine learning algorithms employed in this study, Weighted Subspace Random Forest (WSRF), Regularized Random Forest (RRF), and Naive Bayes (NB), are not novel per se, this research advances the current literature by systematically evaluating their comparative performance under data-constrained conditions typical of large tropical watersheds. Rather than

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-145-2026

proposing a new algorithm, the contribution of this study lies in (i) the integrated framework for factor selection, validation, and spatial interpretation; (ii) the explicit discussion of model behavior under limited and heterogeneous datasets; and (iii) the translation of susceptibility outcomes into a structured basis for defining erosion-prone management units.

Furthermore, this work emphasizes the role of erosion susceptibility mapping not merely as a predictive exercise, but as a decision-support framework capable of informing land-use planning and environmental policy at watershed and regional scales. By explicitly linking modeling outputs to practical implications, the study seeks to bridge the gap between data-driven modeling and policy-oriented environmental management, particularly in tropical regions where empirical erosion data remain scarce.

2. Materials and Methods

2.1 Study Area Characterization

The Figure 1, show Pardo River Basin (PRB) is located between the states of Minas Gerais and São Paulo, Brazil. It covers an area of 12,621.72 km² and is drained by the Pardo River and its main tributaries, including the Ribeirão das Antas (also known as the Lambari River), the Canoas River, the Araraquara River, and the Ribeirão da Figueira (or Ribeirão Tamanduá).

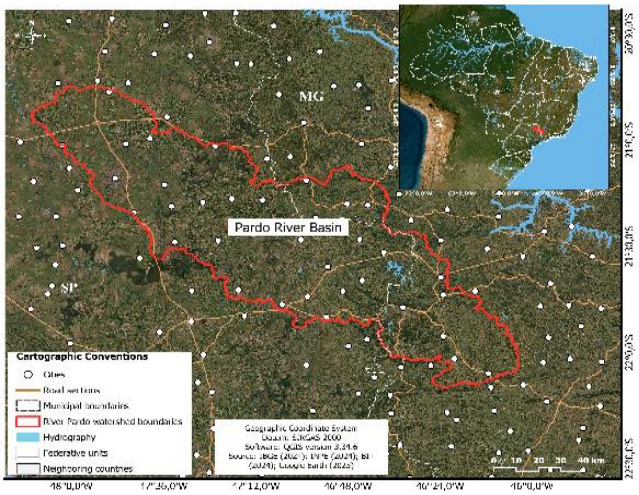

Figure 1. Location of the Pardo River Basin.

The basin is part of a larger hydrological network that includes sub-basins such as the Ribeirão das Antas Basin, the Canoas River Basin, and the Araraquara River Basin, which are tributaries of the Grande River and form part of the Paraná River system. The Pardo River, its main watercourse, has a total length of approximately 545 km, Originating in the state of Minas Gerais to the west and flowing into the Rio Grande in the state of São Paulo to the east.

2.2 Database Formulation

To map soil erosion susceptibility in the Pardo River Basin, areas with and without recorded erosion were first catalogued. A total of 120 points were sampled, including 60 locations exhibiting erosion and 60 locations without erosion. These samples were obtained through field surveys, during which the x and y coordinates of each site were recorded. Based on this information, soil erosion susceptibility was modeled using a binary scale, assigning a value of 1 to areas with erosion occurrence and 0 to areas without recorded erosion.

In this study, and guided by the literature review, 13 factors associated with soil erosion were identified. These factors were standardized to a spatial resolution of 30 m² per pixel and reprojected to UTM coordinates, Zone 22 South. Among the topographic conditioning factors, illustrated in Figure 2, are elevation, slope, aspect, slope length (SL-RUSLE), and profile curvature. Variations in elevation, aspect, and profile curvature influence soil temperature, evaporation conditions, soil moisture, and solar radiation, thereby conferring greater resistance or susceptibility to erosion depending on the degree of exposure to weathering processes (Fenta et al., 2020).

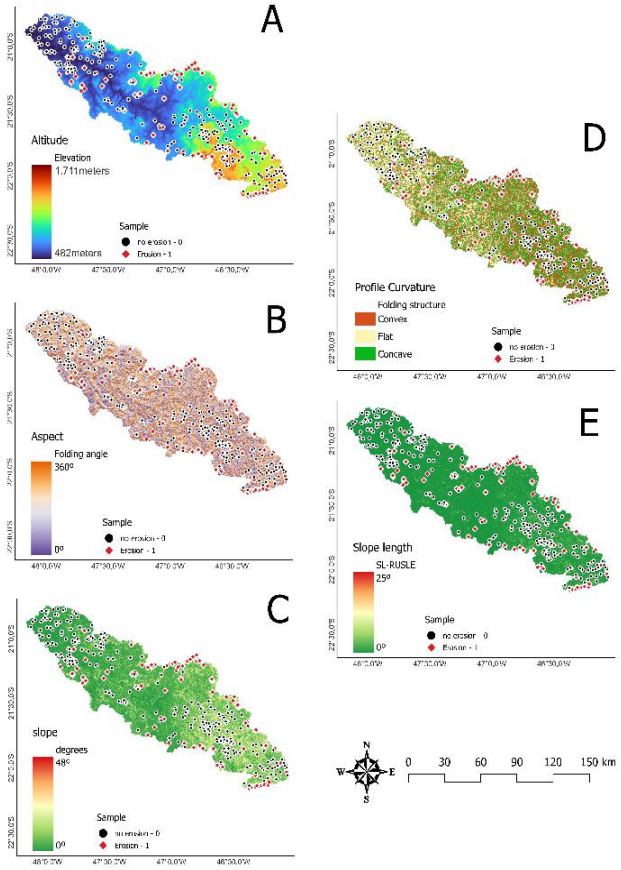

Figure 2. Topographic Conditioning Factors

Additionally, slope angle and slope length influence the velocity and volume of surface runoff, with steeper slopes generally increasing water-induced soil erosion (Salcedo-Sanz et al., 2020). The topographic factors were generated using the open-source software QuantumGIS (QGIS) version 3.34.6, based on a digital elevation model (DEM) derived from the SRTM project.

Hydrological conditioning factors considered in this study comprised Height Above the Nearest Drainage (HAND), distance to drainage networks, the Topographic Wetness Index (TWI), and the Stream Power Index (SPI) (Figure 3). Among these variables, the HAND model (Figure 4a) is particularly relevant because it adopts a topo-hydrological framework that adjusts elevation values according to the configuration of drainage systems and estimates the gravitational drainage potential within sub-basins. This methodology has been shown to exhibit a strong relationship with soil moisture saturation patterns, making it useful for identifying areas characterized by hydrological instability and

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-145-2026 | © Author(s) 2026. CC BY 4.0 License.

regions susceptible to sediment transport and deposition processes (Rennó et al., 2008).

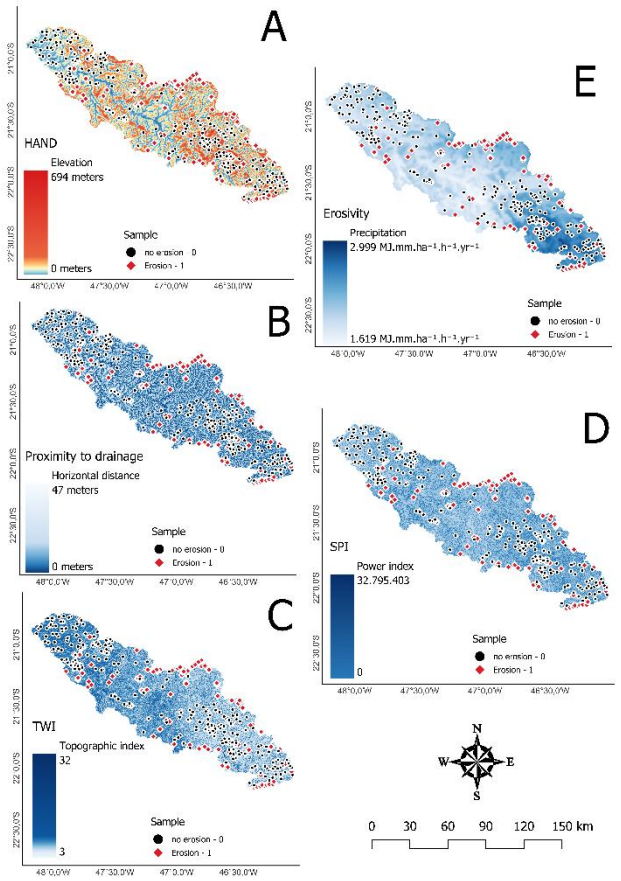

Figure 3. Hydroclimatic Conditioning Factors

Proximity to drainage networks (Figure 3-B) represents areas with an increased probability of soil erosion due to sediment transport processes (Sajedi-Hosseini et al., 2018). This layer was generated using the proximity tool available in the QGIS v. 3.34.6 geographic information system (GIS). The Topographic Wetness Index (TWI; Figure 3-C) describes the spatial distribution of moisture conditions across the landscape and identifies zones of soil water saturation (Sørensen et al., 2006).

The Stream Power Index (SPI; Figure 3-D) indicates the potential for erosion resulting from the gravitational acceleration of water flow, with higher values corresponding to greater erosive power (Moore et al., 1992). For both TWI and SPI, it was necessary to estimate flow accumulation (FA) and slope angle in degrees; these variables were calculated using tools from the System for Automated Geoscientific Analyses (SAGA GIS 2.0.7), accessed through the QGIS v. 3.34.6 interface (Mosavi et al., 2020).

The rainfall erosivity factor R (Figure 3-E) represents the precipitation erosivity index, which is strongly associated with soil erosion potential (Islam, 2022). Although the ideal method for calculating this factor involves direct measurement of soil erosion in experimental plots (Xu et al., 2013), in this study the monthly R factor was estimated using the method proposed by Wischmeier and Smith (1978). For this purpose, Google Earth Engine (GEE) was employed, along with precipitation data provided by the dataset “OpenLandMap/CLM/CLM_PRECIPITATION_SM2RAIN_M/v01”.

In this research, geological factors included proximity to fractures (PF) and lithology (Figure 4). PF influences infiltration and runoff processes, thereby affecting soil erosion. Additionally, the presence of faults may accelerate mass movement processes (Mosavi et al., 2020). The PF layer (Figure 4-C) was generated in QGIS using the proximity tool applied to the fault layer.

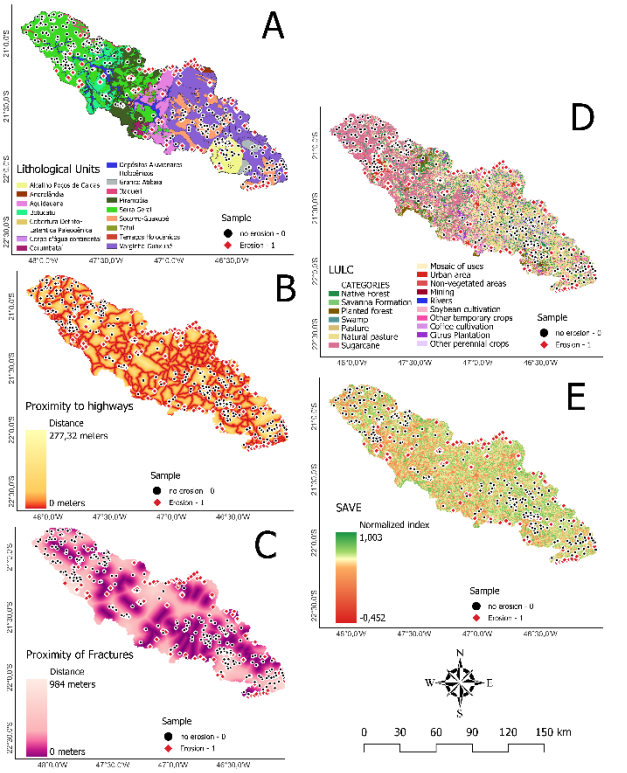

Figure 4. Geological and Anthropogenic Conditioning Factors

Lithology exerts one of the strongest controls on erosion, as it is conditioned by the weathering properties of exposed geological materials (Band et al., 2020). The lithological map (Figure 4-A) was obtained from the national geological survey provided by the Brazilian Geological Survey (CPRM), at a scale of 1:100,000.

Land cover factors include the Soil-Adjusted Vegetation Index (SAVI), proposed by Huete et al. (1988), the land-use/land-cover map, and proximity to roads (Figure 4). The SAVI layer (Figure 4-E) was derived from Sentinel-2 satellite imagery, based on a stacked composite representing the meaning of all scenes acquired in 2023. The land-use and land-cover map was obtained from the MapBiomas project, which provides 30 m resolution data derived from Landsat imagery (Figure 4-D). Proximity to roads (Figure 4-B) is considered an important factor because roads tend to increase the availability of energy for erosion processes and sediment production within the watershed. The road proximity layer was generated using the proximity tool in QGIS v. 3.34.6.

Although the total number of samples (n = 120) may appear limited relative to the size of the study area (12,621.72 km²), this sampling strategy reflects a common constraint in erosion studies conducted in tropical environments, where systematic field-based erosion inventories are scarce, costly, and spatially uneven. To mitigate potential overfitting and sampling bias, the dataset was carefully balanced between erosion and non-erosion classes and combined with a rigorous factor selection process based on multicollinearity (VIF/TOL) and relevance (IGR) analysis.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-145-2026 | © Author(s) 2026. CC BY 4.0 License.

Additionally, the inclusion of ensemble-based algorithms such as RRF and WSRF, which are known for their robustness to small and noisy datasets, further reduces the risk of model instability. Nonetheless, the limitations imposed by sample size are explicitly acknowledged, and future research should prioritize the expansion of erosion inventories through higher-resolution field surveys, crowd-sourced data, or long-term monitoring programs.

2.3 Evaluation of the Suitability of Conditioning Factors

Assessing multicollinearity is essential in binary modeling studies, as it allows the identification of linear dependence between two or more conditioning factors within a predictive model (Dormann et al., 2013). To evaluate and mitigate the effects of multicollinearity, the Variance Inflation Factor (VIF) and Tolerance (TOL) metrics are commonly employed. VIF measures the extent to which the variance of a regression coefficient is inflated due to collinearity, with values greater than 5 indicating significant multicollinearity (Band et al., 2020). TOL, the inverse of VIF, reflects the proportion of variability in a variable that is not explained by other variables in the model; thus, TOL values below 0.1 or 0.2 also indicate high multicollinearity (Zhu & Huang, 2006). In such cases, the removal of conditioning factors exhibiting high multicollinearity is recommended (Whang et al., 2020).

Additionally, to evaluate the relevance of the conditioning factors in relation to the erosion process, the Information Gain Ratio (IGR) was used. IGR is a widely employed technique due to its simplicity, quantifying the information gain contributed by each conditioning factor to the model, measured by the reduction of uncertainty in erosion prediction (Shahzad et al., 2022). If a factor yields an IGR value below 0.05, it indicates a lack of correlation with erosion. Including such a factor may be detrimental, introducing noise and reducing model accuracy (Yu et al., 2019).

2.4 Machine Learning Algorithms Employed

Three Machine Learning models were used in this study. The first was the Weighted Subspace Random Forest (WSRF) model, introduced by Xu et al. (2012), which is an extension of the traditional Random Forest (RF) algorithm designed to more efficiently handle high-dimensional and sparse datasets. Unlike the conventional random variable sampling strategy, WSRF incorporates a weighting scheme to select subspaces of variables that are more informative for model construction. This approach ensures that the selected subspaces contain attributes that significantly contribute to model performance, thereby improving predictive accuracy (Lu et al., 2023).

The second model employed was the Regularized Random Forest (RRF), another variation of the Random Forest algorithm. RRF differs by incorporating a regularization mechanism that selects a more compact and relevant subset of features during the construction of decision trees (Deng & Runger, 2012).

The third model, Naive Bayes (NB), is an algorithm used for statistical regression and classification tasks based on Bayes’ Theorem. It assumes conditional independence among the atributes, that is, each feature used for prediction is considered independent from the others. This assumption allows NB to calculate posterior probabilities for a given class based on prior probabilities using a labeled dataset (Mosavi et al., 2020). Despite the independence assumption, which may not always reflect real-world conditions, Naive Bayes is computationally efficient and can be trained with relatively small datasets.

2.5 Validation of Susceptibility Modeling

Evaluating the accuracy of soil erosion susceptibility maps generated by Machine Learning algorithms is essential to assess their predictive performance. In this study, both threshold-dependent and threshold-independent statistical metrics were applied to the validation dataset (30% of the samples), following recommendations by Mousavi et al. (2017). The threshold-dependent metrics used, derived from the binary contingency matrices of each model, include sensitivity (SST), specificity (SPF), positive predictive value (PPV), negative predictive value (NPV), true skill statistic (TSS), Matthew’s correlation coefficient (MCC), and misclassification rate (MR). Except for MR, higher values for these metrics indicate better model performance, as discussed in Darabi et al. (2021).

The Area Under the Receiver Operating Characteristic Curve (AUC–ROC) is one of the primary threshold–independent metrics. It is generated by plotting sensitivity (true positive rate, TPR) on the Y–axis against the false positive rate (1 – specificity, FPR) on the X–axis, providing a precise evaluation of model performance by measuring the ability to correctly distinguish between positive and negative classes across different decision thresholds, while minimizing errors and maximizing correct classifications (Pham et al., 2019). An AUC–ROC value approaching 1 indicates excellent predictive performance, whereas values near 0.5 suggest performance equivalent to random chance, thereby indicating inefficiency or lack of reliability (Band et al., 2020).

In addition to conventional accuracy-based metrics, the use of multiple complementary validation indicators allows for a more nuanced interpretation of model behavior, particularly under imbalanced or limited datasets. This multi-metric framework helps distinguish between models that maximize overall accuracy and those that prioritize the detection of erosion-prone areas, which is often more relevant for preventive environmental management.

Although cross-validation techniques such as k-fold validation could further enhance robustness, the holdout approach adopted here (70/30 split) remains consistent with similar regional-scale susceptibility studies and provides a transparent basis for inter-model comparison.

3 Results and Discussion

3.1 Analysis of the Suitability of Conditioning Factors

The multicollinearity assessment results presented in Table 1 indicate that the analyzed conditioning factors exhibited TOL values ranging from 0.137 to 0.944 and VIF values varying between 1.060 and 7.314. Among the evaluated variables, Altitude showed the highest VIF value (7.314), revealing a substantial degree of multicollinearity and indicating redundancy with other explanatory factors. Following the criteria proposed by Whang et al. (2020), this variable was excluded from the subsequent modeling procedures. In contrast, Aspect demonstrated the lowest VIF value (1.060) together with the highest TOL value (0.944), suggesting minimal correlation with

the remaining predictors and indicating strong independence within the dataset. Although Slope and Rainfall Erosivity presented comparatively elevated VIF values of 2.609 and 4.663, with corresponding TOL values of 0.383 and 0.214, these variables remained within the acceptable thresholds adopted in this study (VIF < 5 and TOL > 0.1) and were therefore retained in the modeling process.

Variáveis	TOL	VIF
Altitude	0.136	7.314
Aspect	0.943	1.060
Profile Curvature	0.775	1.289
Slope	0.383	2.609
Rainfall Erosivity	0.214	4.663
Distance to Faults	0.776	1.288
Geology	0.489	2.042
HAND	0.453	2.205
LS-EUSP	0.595	1.678
LULC	0.556	1.797
Distance to Rivers	0.746	1.339
Distance to Roads	0.927	1.077
SAVI	0.495	2.019
SPI	0.885	1.129
TWI	0.554	1.804

Table 1. Results of the collinearity analysis among the conditioning factors.

Complementarily, the assessment of factor relevance prior to the modeling process was estimated using the Information Gain Ratio (IGR) (Table 2).

Variáveis	IGR
Altitude	0,1167
Aspect	0,1171
Profile Curvature	0,1170
Slope	0,1171
Rainfall Erosivity	0,1232
Distance to Faults	0,1163
Geology	0,0489
HAND	0,1152
LS-EUSP	0,1055
LULC	0,2180
Distance to Rivers	0,0569
Distance to Roads	0,0939
SAVI	0,1163
SPI	0,1162
TWI	0,1162

Table 2. IGR values for the conditioning factors of erosion susceptibility.

The results indicate that the LULC factor was the most influential, presenting the highest IGR value (0.2181), followed by rainfall erosivity with an IGR of 0.1233. Other factors, such as slope, aspect, and profile curvature, exhibited similar IGR values, ranging between 0.1170 and 0.1172, indicating a certain degree of influence in characterizing areas that are more or less prone to erosion.

Among the evaluated conditioning factors, geology exhibited the lowest IGR value (0.04899), remaining below the predefined acceptance threshold of 0.05 and indicating an insignificant contribution to soil erosion susceptibility characterization. Following the criteria proposed by Whang et al. (2020), this variable was excluded from the subsequent modeling procedures. In contrast, variables such as altitude, HAND, distance to fractures, SAVI, SPI, and TWI presented slightly reduced but still relevant IGR values, ranging between 0.1152 and 0.1167, suggesting a moderate contribution to the predictive framework. Likewise, distance to rivers and distance to roads yielded comparatively lower IGR values (0.05692 and 0.09399, respectively), indicating a weaker influence on model performance and erosion prediction.

3.2 Accuracy Assessment of the Machine Learning Models

In the comparative analysis of the Naive Bayes (NB), Regularized Random Forest (RRF), and Weighted Subspace Random Forest (WSRF) models, several performance indicators were employed to evaluate the effectiveness of each model in predicting soil erosion susceptibility (Table 3).

Metricas	NB	RRF	WSRF
Accuracy	0.877	0.898	0.883
Kappa	0.755	0.797	0.766
Sensitivity	0.866	0.922	0.933
Specificity	0.888	0.875	0.833
Pos Pred Value	0.886	0.883	0.848
Neg Pred Value	0.869	0.916	0.925
Area under the curve	0.932	0.969	0.954
True Skill Statistic (TSS)	0.755	0.799	0.774
Matthews Correlation Coefficient (MCC)	0.755	0.798	0.770
Misclassification Rate (MR)	0.122	0.101	0.116
F1 Score	0.879	0.895	0.877

Table 3. Model validation using 30% of the sample dataset.

Differences in the predictive performance of Machine Learning models can arise from several factors, among which the descriptive capability and quality of the selected conditioning variables play a fundamental role (Kulimushi et al., 2023). Inadequate parameter selection may introduce noise into the modeling process rather than improve prediction accuracy, ultimately reducing model reliability and predictive effectiveness (Yunkai et al., 2010). Within this context, the comparative evaluation performed in the present study demonstrated distinct behavioral patterns among the tested algorithms. The RRF model exhibited superior overall performance, achieving the best balance between classification accuracy and discriminative capacity. In contrast, WSRF showed a greater ability to correctly identify positive events, which may be advantageous in applications where maximizing the detection of susceptible areas is a priority (Khosravi et al., 2023). The NB model displayed a comparatively stable but more constrained performance, likely associated with its assumption of predictor independence, which may limit its capacity to represent more complex interactions within environmental datasets.

Although direct comparisons should be interpreted cautiously due to differences in environmental processes and datasets, the performance pattern observed in this study is consistent with findings reported in susceptibility modeling literature. For example, Pham and Prakash (2019) compared LogitBoost Ensemble (LBE), Support Vector Machine (SVM), Logistic Regression (LR), and Fisher’s Linear Discriminant Analysis (FLDA) for landslide susceptibility mapping and reported superior predictive performance for ensemble-based approaches, with LBE achieving the highest AUC value (0.972), followed by SVM (0.945). Similarly, the present study demonstrated that ensemble tree-based models, particularly RRF (AUC = 0.969) and WSRF (AUC = 0.954), outperformed the probabilistic NB model (AUC = 0.932), reinforcing the ability of ensemble strategies to capture nonlinear interactions and improve

predictive accuracy in susceptibility assessments across complex environmental systems (Pham; Prakash, 2019).

Among the evaluated algorithms, the Rotation Random Forest (RRF) demonstrated the strongest predictive performance across the analyzed metrics. The model achieved the highest AUC value (0.96), highlighting its superior capability to discriminate between erosion-susceptible and non-susceptible areas. Likewise, RRF obtained the best accuracy (0.89) and Kappa coefficient (0.79), indicating a high level of agreement between predicted and observed outcomes. The superior performance of this approach may be associated with its methodological structure, which integrates the strengths of random forest algorithms with feature-space rotation mechanisms, enabling a more efficient representation of data variability and complex relationships among predictors (Mohammed et al., 2020).

In comparison, the Weighted Subspace Random Forest (WSRF) also showed robustness, with a slightly lower accuracy (0.8833) and a Kappa index of 0.7667. However, the WSRF exhibited the highest sensitivity (0.9333), indicating a greater capability to correctly identify positive events. On the other hand, its specificity was the lowest among the models (0.8333), reflecting a higher rate of false positives. This behavior may result from the WSRF’s weighted-subspace approach, which, by emphasizing different combinations of features, increased true-positive detection but also amplified confusion between classes (Lu et al., 2023). The Naive Bayes (NB) model, despite delivering solid performance, ranked below the forest-based models. With an accuracy of 0.8778 and a Kappa index of 0.7556, NB showed good classification capability but did not reach the precision levels achieved by RRF and WSRF. Its AUC of 0.9327, although high, was still lower than that of the other two models. The NB’s simpler structure, which assumes independence among predictor variables, likely contributed to this difference, as such an assumption may not adequately capture the complex interactions inherent in the dataset (Mosavi et al., 2020).

In terms of other performance metrics, such as the True Skill Statistic (TSS), the Matthews Correlation Coefficient (MCC), and the F1 Score, the RRF again ranked highest, confirming its robustness and reliability in susceptibility modeling. The misclassification rate (MR) was lowest for the RRF (0.1011), reinforcing its predictive precision. Although competitive, the WSRF showed a slightly higher error rate (0.1166), while the NB presented the highest error rate among the three models (0.1222). Although the performance differences among the evaluated models are relatively small, they are consistent across multiple validation metrics, suggesting systematic rather than random variation. The superior performance of the RRF model across accuracy, AUC, Kappa, TSS, and MCC indicates a more stable balance between sensitivity and specificity.

While formal statistical significance tests (e.g., DeLong test for AUC comparison) were not applied in this study, the convergence of results across independent metrics supports the practical relevance of the observed differences. From an applied perspective, these differences are meaningful when selecting models for specific management objectives, such as minimizing false negatives in high-risk erosion zones.

The comparative performance of the models highlights important trade-offs relevant to erosion susceptibility mapping. The RRF model demonstrated the most balanced performance, making it particularly suitable for regional planning and policy applications where both false positives and false negatives carry significant implications. In contrast, the WSRF model exhibited higher sensitivity, suggesting its applicability in early-warning or precautionary frameworks where the primary goal is to avoid overlooking erosion-prone areas. The Naive Bayes model, despite its comparatively lower performance, remains relevant in contexts characterized by severe data limitations or computational constraints. Its simplicity and interpretability may be advantageous in preliminary assessments or in regions lacking extensive geospatial infrastructure.

3.2 Mapping Soil Erosion Susceptibility

Based on the results obtained using the Naive Bayes (NB), Regularized Random Forest (RRF), and Weighted Subspace Random Forest (WSRF) models for estimating soil erosion susceptibility, the maps were classified into five susceptibility levels: very low, low, moderate, high, and very high. The spatial distribution of susceptible areas within these categories reveals the extent of erosion risk across each susceptibility class. Overall, the results show consistent patterns, indicating that areas of high and very high susceptibility are concentrated in the western region and along steep floodplains, while low-susceptibility areas predominate in the eastern portion of the watershed.

In the NB model, the “very low” susceptibility class occupies the largest area, covering 4,731.914 km² of the basin, representing a substantial proportion of the total. The “low” class covers 1,356.796 km², followed by the “moderate” class with 1,268.922 km². Areas of high and very high susceptibility encompass 1,476.718 km² and 3,762.076 km², respectively (Figure 5). This indicates that, while a considerable portion of the watershed is at “very high” risk of erosion, a similarly significant area falls under the “very low” risk category, demonstrating a varied distribution of susceptibility across the region.

The RRF model also presents an interesting variation in the distribution of susceptibility classes. The “very low” class covers 2,389.675 km², whereas the “low” class occupies 3,224.789 km². The “moderate” class spans 2,173.056 km², and the “high” and “very high” susceptibility classes correspond to 2,006.227 km² and 2,833.476 km², respectively. This model suggests a more balanced distribution of areas among the classes, with a slight emphasis on low-susceptibility zones, which may indicate a general resistance of the basin to more severe erosive events.

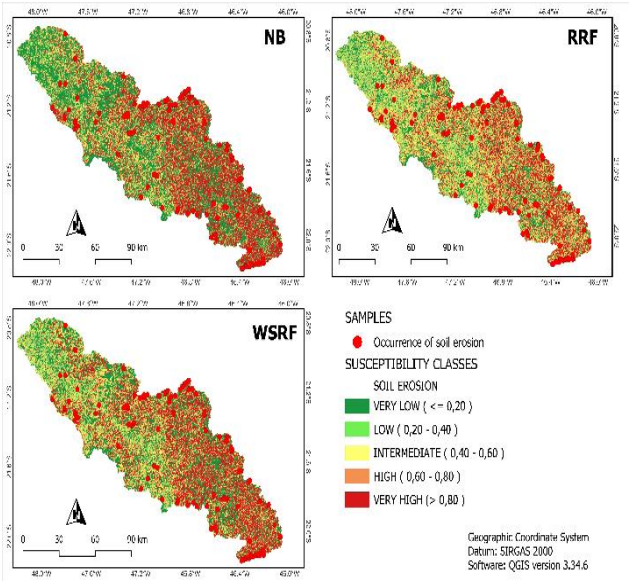

Figure 5. Soil erosion susceptibility maps generated using the NB, RRF, and WSRF models.

When applying the WSRF model, it is observed that the *very low* susceptibility class covers 3,162.841 km² of the watershed, representing the largest area among all classes. The *low* class covers 2,634.427 km², followed by the *moderate* class with 2,398.732 km². The areas classified as *high* and *very high* susceptibility to erosion total 2,444.692 km² and 1,955.733 km², respectively. The WSRF model indicates a predominance of areas with very low and low susceptibility, reflecting a potentially greater resistance of the basin to erosive processes, with the distribution of risk gradually decreasing toward the higher susceptibility classes.

4. Conclusion

The research was conducted under constraints related to data availability and compatibility, as datasets with different spatial scales, spectral resolutions, and temporal coverage were employed. This heterogeneity may introduce inconsistencies in the spatial representation of the analyzed phenomena, potentially affecting the accuracy of the results. Therefore, future studies should prioritize the selection and adaptation of datasets according to the specific characteristics of the study area, as well as the incorporation of more advanced approaches, such as deep learning techniques, which are better suited to handle the variability and complexity of geospatial data.

Among the evaluated models, RRF showed the best overall performance, combining high accuracy with a strong ability to discriminate between erosion and non-erosion areas, particularly in structurally heterogeneous environments. The WSRF model, although effective in identifying susceptible areas, exhibited a greater tendency toward overestimation, while the NB model proved more suitable for applications requiring simplicity and lower computational cost, despite its limitations in capturing complex patterns. In this context, the susceptibility generated maps can support territorial planning, the identification of priority areas for soil conservation, and the implementation of environmental monitoring strategies.

References

Band et al., 2020. Flash flood susceptibility modeling using new approaches of hybrid and ensemble tree-based machine learning algorithms. *Remote Sensing, 12*(21), 3568. https://doi.org/10.3390/rs1221356

Darabi, H., Rahmati, O., Naghibi, S.A., Mohammadi, F., Ahmadisharaf, E., Kalantari, Z., Torabi Haghighi, A., Soleimanpour, S.M., Tiefenbacher, J.P. and Tien Bui, D., 2021. Development of a novel hybrid multi-boosting neural network model for spatial prediction of urban flood. *Geocarto International, 37*(19), 5716–5741. https://doi.org/10.1080/10106049.2021.1920629

Deng, H., & Runger, G., 2012. Feature selection via regularized trees. In *The 2012 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., et al., 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography (Cop.), 36*(1), 27–46.

Fenta, A. A., Tsunekawa, A., Haregeweyn, N., Poesen, J., Tsubo, M., Borrelli, P., et al., 2020. Land susceptibility to water and wind erosion risks in the East Africa region. *Science of the Total Environment, 703*, 135016.

Huete, A. R., 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment, 25*, 295–309. https://www.researchgate.net/publication/220040775_Huete_A_R_A_soiladjusted_vegetation_index_SAVI_Remote_Sensing_of_Environment

Islam, Z., 2022. Soil loss assessment by RUSLE in the cloud-based platform (GEE) in Nigeria. *Modeling Earth Systems and Environment, 8*, 4579–4591. https://doi.org/10.1007/s40808-022-01467-7

Khosravi, K., Rezaie, F., Cooper, J. R., Kalantari, Z., Abolfathi, S., Hatamiafkoueieh, J., 2023. Soil water erosion susceptibility assessment using deep learning algorithms. *Journal of Hydrology, 618*, 129229.

Kulimushi, L.C., Bashagaluke, J.B., Prasad, P., Heri-Kazi, A.B., Kushwaha, N.L., Masroor, M.D., Choudhari, P., Elbeltagi, A., Sajjad, H. and Mohammed, S., 2023. Soil erosion susceptibility mapping using ensemble machine learning models: A case study of the upper Congo river sub-basin. *Catena, 222*, 106858.

Lu, Q.O., Ahmadi, K., Mahmoodi, S., Karami, A., Elkhrachy, I., Mondal, I., Arshad, A., Nguyen, T.T., Chi, N.T.L. and Thai, V.N., 2023. Hybrid regularization and weighted subspace algorithms with random forest model for assessing piping erosion in semi-arid ecosystem. *Environmental Earth Sciences*, *82*(22), 527.

Mohammed, S., Al-Ebraheem, A., Holb, I.J., Alsafadi, K., Dikkeh, M., Pham, Q.B., Linh, N.T.T. and Szabo, S., 2020. Soil management effects on soil water erosion and runoff in central Syria - A comparative evaluation of general linear model and random forest regression. *Water, 12*(9), 2529.

Moore, I. D., & Wilson, J. P., 1992. Length-slope factors for the Revised Universal Soil Loss Equation: Simplified method of estimation. *Journal of Soil and Water Conservation, 47*, 423–428.
Mosavi, A., Sajedi-Hosseini, F., Choubin, B., Taromideh, F., Rahi, G., & Dineva, A. A. (2020). Susceptibility mapping of soil water erosion using machine learning models. *Water, 12*(7), 1995. https://doi.org/10.3390/w12071995

Mousavi, S. M., Golkarian, A., Amir Naghibi, S., Kalantar, B., Pradhan, B., 2017. GIS-based groundwater spring potential mapping using data mining boosted regression tree and probabilistic frequency ratio models in Iran. *AIMS Geosciences, 3*(1), 91–115.

Pham, B. T., Prakash, I., 2019. Evaluation and comparison of LogitBoost Ensemble, Fisher’s Linear Discriminant Analysis, logistic regression and support vector machines methods for landslide susceptibility mapping. *Geocarto International, 34*, 316–333.

Pham, B.T., Prakash, I., Khosravi, K., Chapi, K., Trinh, P.T., Ngo, T.Q., Hosseini, S.V. and Bui, D.T., 2019. A comparison of Support Vector Machines and Bayesian algorithms for landslide susceptibility modelling. *Geocarto International*, *34*(13), 1385-1407.

Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J., 2008. HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. *Remote Sensing of Environment*, *112*(9), 3469-3481.

Sajedi-Hosseini, F., Choubin, B., Solaimani, K., Cerdà, A., Kavian, A., 2018. Spatial prediction of soil erosion susceptibility using a fuzzy analytical network process: Application of the fuzzy decision making trial and evaluation laboratory approach. *Land Degradation & Development, 29*, 3092–3103.

Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., Izquierdo-Verdiguier, E., Muñoz-Marí, J., Mosavi, A., Camps-Valls, G., 2020. Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion, 22*, 480–545.

Shahzad, N., Ding, X., Abbas, S., 2022. A comparative assessment of machine learning models for landslide susceptibility mapping in the rugged terrain of northern Pakistan. *Applied Sciences, 12*(5), 2280.

Sörensen, R., Zinko, U., Seibert, J., 2006. On the calculation of the topographic wetness index: Evaluation of different methods based on field observations. *Hydrology and Earth System Sciences, 10*, 101–112.

Wang, G., Chen, X., Chen, W., 2020. Spatial prediction of landslide susceptibility based on GIS and discriminant functions. *ISPRS International Journal of Geo-Information, 9*, 144.

Wischmeier, W. H., Smith, D. D., 1978. *Predicting rainfall erosion losses: A guide to conservation planning* (No. 537). Department of Agriculture, Science and Education Administration.

Velastegui-Montoya, A., Montalván-Burbano, N., Carrión-Mero, P., Rivera-Torres, H., Sadeck, L., Adami, M., 2023. Google Earth Engine: a global analysis and future trends. *Remote Sensing*, *15*(14), 3675.

Xu, B., Huang, J. Z., Williams, G., Ye, Y., 2012. Hybrid weighted random forests for classifying very high-dimensional data. *International Journal of Data Warehousing and Mining, 8*(2), 44–63.

Xu, L., Xu, X., & Meng, X., 2013. Risk assessment of soil erosion in different rainfall scenarios by RUSLE model coupled with information diffusion model: A case study of Bohai Rim, China. *CATENA, 100*, 74–82. https://doi.org/10.1016/j.catena.2012.08.012

Yu, L., Cao, Y., Zhou, C., Wang, Y., Huo, Z., 2019. Landslide susceptibility mapping combining information gain ratio and support vector machines: A case study from Wushan Segment in the Three Gorges Reservoir Area, China. *Applied Sciences, 9*(22), 4756. https://doi.org/10.3390/app9224756

Yunkai, L., Yingjie, T., Zhiyun, O., Lingyan, W., Tingwu, X., Peiling, Y., Huanxun, Z., 2010. Analysis of soil erosion characteristics in small watersheds with particle swarm optimization, support vector machine, and artificial neuronal networks. *Environmental earth sciences*, *60*(7), 1559-1568.

Zhu, L., Huang, J., 2006. GIS-based logistic regression method for landslide susceptibility mapping in regional scale. *Journal of Zhejiang University - Science A, 7*(12), 2007–2017.