Integrating Multicriteria Environmental Factors and Spatially Cross-Validated Machine Learning for Synthetic Erosion Vulnerability Assessment in the Betsiboka Region, Madagascar

Rasoanaina Jacquis[1,1]; Rakotoson Andriatiana Tolontsoa[1,2];
Tovonirina Mamiharizo Jackie[1,3]; Razafiarisera Ralay Tiana[1,4];
Rasolomanana Eddy Harilala[1,5]

[1]Doctoral School of Engineering and Geosciences,
[2]Doctoral Research Team: Spatial Planning and Territorial
[3]Research Laboratory: Spatial Planning and Territorial Development Laboratory
[4]University of Antananarivo, Madagascar

Publication Date: 2026/07/29

Abstract: This study integrates multicriteria environmental analysis and spatially cross-validated machine learning to assess synthetic erosion vulnerability in the Betsiboka Region, Madagascar. The vulnerability index combined low NDVI, slope, hydrographic proximity, elevation, and available geological, soil, rainfall, and land-cover factors on a 1 km grid. Hydrographic proximity was modelled using an exponential distance-decay function with a characteristic distance of 3 km, while geology was represented through normalized susceptibility scores ranging from 0 to 1. Three regression algorithms—Random Forest, Extra Trees, and Histogram Gradient Boosting—were evaluated using five-fold spatial block cross-validation. Histogram Gradient Boosting achieved the best performance, reproducing the synthetic index with $R^2 = 0.999$, RMSE = 0.003, and MAE = 0.002. Residuals were tightly centered around zero, although slightly larger errors occurred at low predicted vulnerability values. The normal Q–Q plot showed strong overall agreement with normality (correlation = 0.987), with departures mainly in the distribution tails. Spatial patterns highlighted high geological susceptibility in western and northwestern Betsiboka and elevated hydrographic influence along the dense drainage network. These results demonstrate strong model emulation and spatial consistency, but not independent validation of observed erosion.

Keywords: *Erosion Vulnerability; Multicriteria Analysis; Spatial Cross-Validation; Machine Learning; Betsiboka Region.*

How to Cite: Rasoanaina Jacquis; Rakotoson Andriatiana Tolontsoa; Tovonirina Mamiharizo Jackie; Razafiarisera Ralay Tiana; Rasolomanana Eddy Harilala (2026) Integrating Multicriteria Environmental Factors and Spatially Cross-Validated Machine Learning for Synthetic Erosion Vulnerability Assessment in the Betsiboka Region, Madagascar. *International Journal of Innovative Science and Research Technology*, 11(7), 1750-1758. https://doi.org/10.38124/ijisrt/26jul913

I. INTRODUCTION

Soil erosion is a major form of land degradation that affects soil fertility, agricultural productivity, water quality, reservoir capacity, and river-system functioning. Its spatial distribution results from interactions among climate, topography, vegetation, geology, hydrology, soils, and human activities. At the global scale, land-use change and agricultural expansion are expected to intensify erosion in many tropical and subtropical regions, particularly in sub-Saharan Africa [1]. Identifying erosion-prone areas is therefore essential for soil conservation, watershed management, and the prioritization of restoration measures. Madagascar is especially vulnerable to erosion because of its deeply weathered soils, lateritic landscapes, seasonal rainfall, steep terrain, and widespread occurrence of lavaka gullies. Previous research has shown that these erosional landforms can supply substantial quantities of sediment to river systems, while erosion rates may vary considerably across spatial and temporal scales [2]. This variability highlights the need for regional approaches capable of distinguishing the relative influence of environmental conditions and land degradation processes. The Betsiboka River basin is one of Madagascar's most important sediment-transfer systems. It connects the central and northwestern parts of the island to the Betsiboka estuary and Bombetoka Bay, where high suspended-sediment concentrations and active sediment deposition have been documented [3]. Within the region, erosion susceptibility is

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul913

influenced by drainage connectivity, rainfall variability, vegetation condition, topographic gradients, and lithological contrasts. The spatial variability of lateritic weathering and geological materials across Madagascar further supports the inclusion of geology and elevation in regional erosion assessments [4]. Because erosion results from multiple interacting factors, it cannot be adequately represented by a single environmental variable. Steep slopes may increase runoff velocity, whereas vegetation cover can reduce rainfall impact, improve infiltration, and stabilize the soil. Proximity to rivers influences the connection between sediment-producing areas and drainage channels, while rainfall, lithology, soil properties, and land use affect the resistance of the landscape to erosion. Geographic information systems, remote sensing, and multicriteria analysis provide a practical framework for standardizing these variables and integrating them into spatial vulnerability models [5]. Multicriteria models are transparent and useful for regional screening, but their outputs depend on factor selection, normalization procedures, assigned scores, and weighting schemes. Machine-learning methods provide a complementary approach because they can represent nonlinear relationships and interactions among predictors. Algorithms such as Random Forest, Extra Trees, and gradient-boosting models have increasingly been applied to erosion-susceptibility mapping using terrain, vegetation, rainfall, drainage, soil, and geological variables [6]. However, when machine learning is trained on a synthetic index constructed from similar predictors, its performance should be interpreted as model emulation and spatial consistency rather than independent validation of observed erosion. Validation strategy is particularly important for spatial environmental modelling. Random cross-validation can lead to overly optimistic estimates when neighbouring observations occur in both the training and testing datasets. Spatial block cross-validation provides a more rigorous alternative by separating calibration and validation samples geographically and testing the model's ability to predict in withheld areas [7]. This approach is well suited to gridded erosion studies because adjacent cells often share similar environmental characteristics. Accordingly, this study integrates multicriteria environmental factors and spatially cross-validated machine learning to assess synthetic erosion vulnerability in the Betsiboka Region, Madagascar. The modelling framework combines NDVI, slope, hydrographic proximity, elevation, geology, soils, rainfall, and land-cover information on a regional analysis grid. Several machine-learning algorithms are compared using spatial block cross-validation, while residual analysis, predictor importance, and partial-dependence diagnostics are used to examine model behaviour. The objective is to provide a reproducible framework for regional vulnerability mapping and watershed prioritization, while recognizing that field observations and independent erosion measurements remain necessary for operational validation.

II. MATERIALS AND METHODS

The study was conducted in the Betsiboka Region of northwestern Madagascar using a combination of administrative, hydrographic, geological, soil, topographic, climatic, and satellite-derived datasets. Regional, district, and commune boundaries were obtained from available geographic information system layers and used to define the study extent and calculate zonal statistics. All spatial data were projected to WGS 84/UTM Zone 38S to ensure consistency in distance, area, slope, and spatial-overlay calculations. The hydrographic network was extracted from the Madagascar Waterways dataset distributed through the Humanitarian Data Exchange. This OpenStreetMap-derived database contains mapped rivers, streams, canals, and other linear water features [8]. Only watercourses intersecting the Betsiboka Region were retained, and invalid geometries were repaired before clipping and merging. Satellite-detected surface-water polygons produced by UNITAR–UNOSAT during the 2020 Madagascar floods were used as supplementary hydro-environmental information [9]. These data supported the interpretation of areas exposed to surface-water expansion but were not considered direct measurements of erosion. Vegetation conditions were represented using the USGS eVIIRS Global Normalized Difference Vegetation Index product. The 10-day NDVI composite, distributed at an approximate spatial resolution of 1 km, was converted to the conventional NDVI range after removing fill values and invalid observations [10]. Low NDVI values were interpreted as reduced vegetation protection and therefore as a higher potential contribution to erosion vulnerability. Elevation data were obtained from the USGS Shuttle Radar Topography Mission 1 Arc-Second Global digital elevation model, which provides terrain information at approximately 30 m resolution [11]. The elevation raster was resampled to the common analysis grid, and slope was derived from elevation differences between neighbouring cells. Rainfall, land cover, geology, and soil datasets were incorporated when valid regional layers were available. Continuous rasters, including elevation and rainfall, were resampled using bilinear interpolation, whereas categorical layers were processed using nearest-neighbour resampling to preserve class values. All datasets were harmonized to a 1 km grid, corresponding to the nominal resolution of the eVIIRS NDVI product. Vector processing, clipping, reprojection, geometry repair, and rasterization were performed using GeoPandas, Shapely, Pyogrio, PyProj, and Rasterio. NumPy and pandas supported numerical and tabular operations, SciPy was used for distance transformations and residual diagnostics, Matplotlib generated publication-quality figures, Folium produced the interactive map, and scikit-learn provided the machine-learning framework [12]. The synthetic erosion-vulnerability index combined low NDVI, slope, hydrographic proximity, elevation, rainfall, land cover, geology, and soils. Low NDVI, slope, and hydrographic proximity were treated as core predictors, while the remaining factors were included when suitable data were available. The NDVI contribution increased as vegetation cover decreased, and the slope contribution increased with terrain steepness. Hydrographic proximity was represented using an exponential distance-decay function with a characteristic distance of 3 km, assigning higher vulnerability values to cells located near watercourses. Elevation and rainfall were normalized using robust percentile limits to reduce the influence of extreme values. Land-cover classes were assigned relative susceptibility scores according to their expected influence on soil exposure and protection. Forested

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul913

surfaces received lower scores, agricultural and built-up areas received higher values, and bare or sparsely vegetated surfaces received the highest values. Geological and soil units were converted into relative susceptibility scores using predefined keyword-based classifications. These values were treated as heuristic indicators and should therefore be interpreted as preliminary representations requiring expert validation. The initial weighting scheme assigned the largest contributions to slope, low NDVI, and hydrographic proximity, followed by rainfall, elevation, land cover, geology, and soils. When optional factors were unavailable, the remaining weights were automatically renormalized. The continuous index was restricted to values between 0 and 1 and divided into five classes representing very low, low, moderate, high, and very high vulnerability. Regional, district, and commune statistics were then calculated from the valid raster cells, including the mean, median, minimum, maximum, class proportions, and the estimated area classified as high or very high vulnerability. Machine-learning analysis was used to evaluate the ability of nonlinear algorithms to reproduce the spatial structure of the synthetic index. Because no independent erosion-observation raster was available, the synthetic vulnerability index was used as the response variable. Consequently, the machine-learning results were interpreted as model emulation and spatial-consistency assessment rather than independent validation of observed erosion. Predictor variables included NDVI, slope, distance to watercourses, elevation, and, where available, rainfall, land-cover susceptibility, geological susceptibility, and soil susceptibility. Missing predictor values were replaced with the median, and the working sample was limited to a maximum of 100,000 valid grid cells to control computational demand. Three ensemble regression algorithms were compared. Random Forest combined predictions from multiple decision trees generated from randomized samples and predictor subsets [13]. Extra Trees introduced additional randomness in predictor and split selection to increase model diversity [14]. Histogram Gradient Boosting constructed trees sequentially, with each stage reducing the prediction error remaining from the previous stage [15]. The models were evaluated using five-fold spatial block cross-validation based on blocks of approximately 20 km. This procedure separated training and validation samples geographically and reduced the risk of overly optimistic accuracy caused by spatial autocorrelation. Model performance was assessed using the coefficient of determination, root mean squared error, and mean absolute error. The best-performing algorithm was selected according to the lowest mean cross-validated RMSE. Out-of-fold predictions were retained to produce target-versus-predicted plots, residual diagnostics, normal Q–Q plots, and spatial residual maps. Predictor importance was evaluated using permutation importance on a withheld spatial fold, with sequential processing and a restricted validation sample to reduce memory consumption. Partial-dependence plots were generated for the most influential predictors, while learning curves were used to compare training and spatial-validation errors across increasing sample sizes. The final selected model was applied to the full regional grid to produce a spatial prediction map, and relative uncertainty was represented using the dispersion among Random Forest tree predictions.

III. RESULTS

➤ *Spatial Distribution of the Geology-Based Erosion Vulnerability Factor in the Betsiboka Region*

Figure 1 illustrates the spatial distribution of the normalized geology-based vulnerability factor across the Betsiboka Region. The factor ranges from 0 to 1, with values close to 0 indicating a relatively low geological contribution to erosion vulnerability and values close to 1 indicating a relatively high contribution. Commune, district, and regional boundaries are superimposed to facilitate the interpretation of spatial variations within the administrative units.

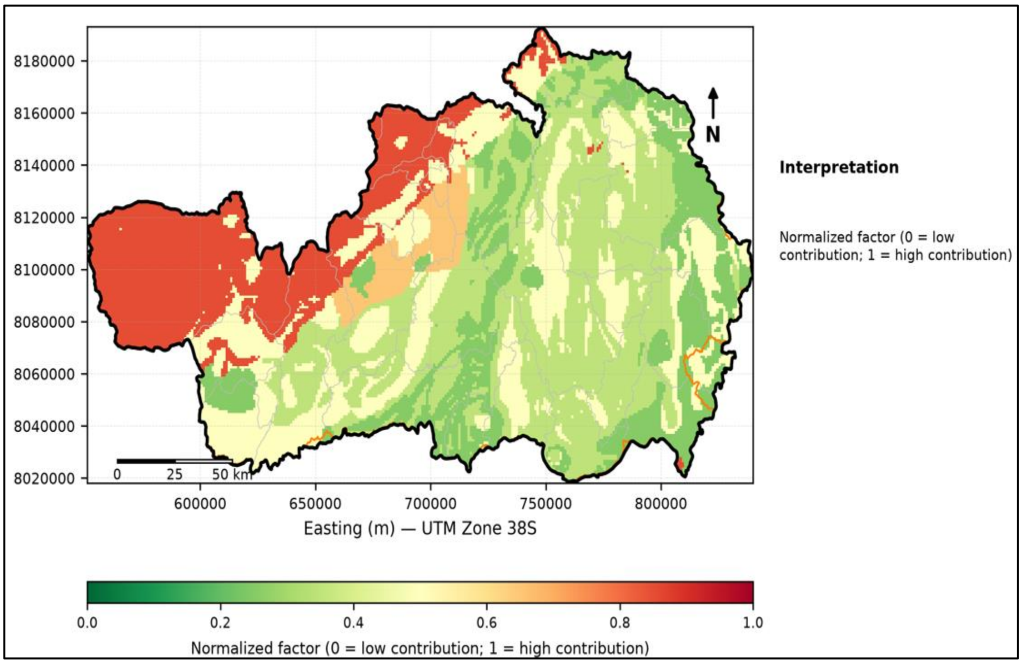

Fig 1 Spatial Distribution of the Geology-Based Erosion Vulnerability Factor in the Betsiboka Region.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul913

The highest geological vulnerability is concentrated mainly in the western and northwestern parts of the region, where extensive red and orange areas correspond approximately to normalized values between 0.75 and 1.00. A broad high-vulnerability belt extends from the western boundary toward the central-western sector, suggesting the predominance of geological formations considered comparatively susceptible to weathering, material detachment, or sediment production. Several smaller high-value zones are also visible along the northern margin and in isolated parts of the southeastern and eastern boundaries. The central portion of the region is characterized predominantly by intermediate geological vulnerability, with values generally ranging from approximately 0.40 to 0.70. These yellow-green and light-orange areas form transitional zones between the highly susceptible western formations and the less susceptible geological units observed farther east. The elongated spatial patterns visible across the central region may reflect the distribution and orientation of distinct lithological formations rather than a gradual environmental transition. Lower geological vulnerability dominates much of the eastern and northeastern sectors, where green shades indicate normalized values of approximately 0.10 to 0.40. These areas are interpreted as being associated with geological materials assigned lower susceptibility scores in the multicriteria model. Nevertheless, narrow linear and fragmented zones of moderate to high vulnerability occur within this generally low-risk sector, indicating substantial local geological heterogeneity. The marked west–east contrast suggests that geology may contribute unevenly to erosion susceptibility across Betsiboka. Western districts are likely to receive a stronger geological contribution to the final synthetic erosion index, whereas the geological influence is comparatively weaker over much of the eastern region. However, this factor represents relative susceptibility rather than observed erosion intensity. Actual erosion conditions also depend on slope, vegetation cover, rainfall, soil properties, drainage proximity, and land-use practices.

➤ *Spatial Pattern of the Hydrographic Proximity Factor in the Betsiboka Region*

Figure 2 presents the spatial distribution of the hydrographic proximity factor used in the multicriteria erosion-vulnerability model for the Betsiboka Region. This normalized factor ranges from 0 to 1, where values close to 0 represent areas located far from the hydrographic network and values close to 1 indicate areas situated very near a watercourse. The spatial pattern was derived using an exponential distance-decay function with a characteristic distance of 3 km, and the mapped values are shown together with commune boundaries, district boundaries, the regional boundary, and the hydrographic network.

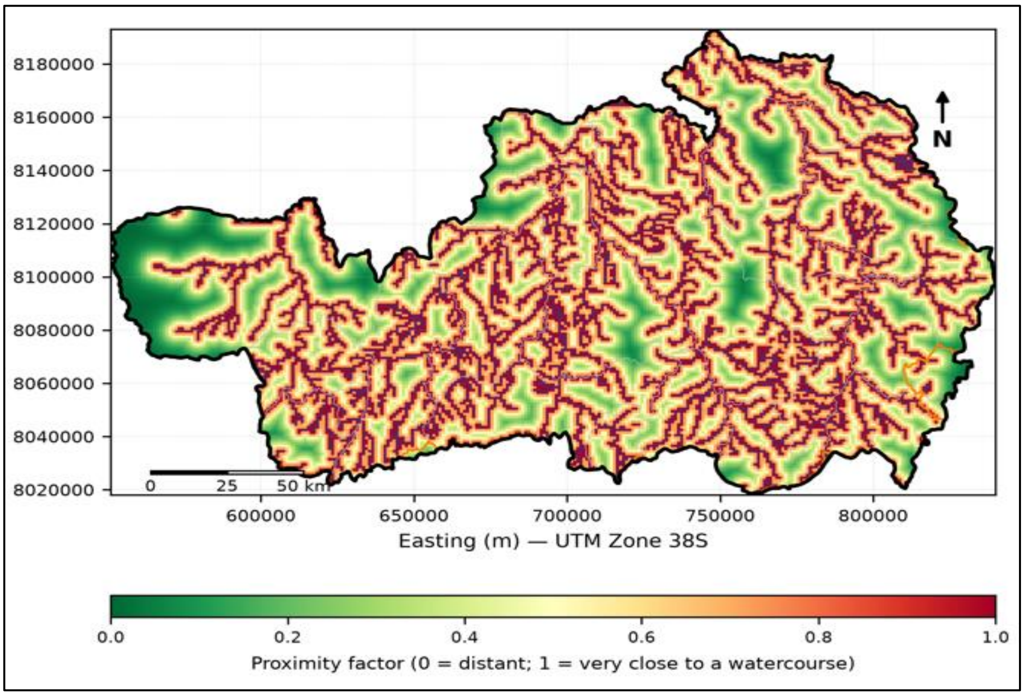

Fig 2 Spatial Pattern of the Hydrographic Proximity Factor in the Betsiboka Region

The map reveals a strongly linear and dendritic spatial organization that closely follows the structure of the drainage network. The highest factor values, generally between 0.80 and 1.00, occur along the main rivers and their tributaries and appear as narrow red to dark-orange corridors distributed throughout the region. Intermediate values, approximately between 0.40 and 0.80, form transitional belts around these channels, indicating areas that remain relatively close to the drainage system and may therefore experience stronger hydrological connectivity. In contrast, low values, commonly below 0.20, dominate the interfluves and the more distant upland sectors, particularly in parts of the western and southwestern region where the spacing between channels is greater. This spatial configuration indicates that hydrographic proximity contributes pervasively to the regional vulnerability pattern, but in a highly localized manner. Areas adjacent to rivers are assigned the highest contribution because their closeness to flowing water increases the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul913

likelihood of concentrated runoff, channel-bank instability, sediment transfer, and the rapid mobilization of eroded material. The widespread occurrence of high-value linear bands across the region also reflects the dense drainage organization of Betsiboka, implying that a substantial proportion of the territory lies within relatively short distances of a watercourse. Conversely, the green areas located farther from stream channels are expected to receive a weaker hydrographic contribution within the synthetic erosion-vulnerability index. The map also suggests that the influence of hydrographic proximity is not spatially uniform in terms of pattern density. In the central, eastern, and northeastern sectors, the river network is highly ramified and produces a dense mosaic of high- and intermediate-value zones, whereas some western sectors show broader low-value patches between channels. This contrast may indicate differences in drainage density and hydrological connectivity, which are important for understanding how sediment and runoff are routed across the landscape.

- *Normal Q–Q Plot of Spatially Cross-Validated Residuals*

Figure 3 presents the normal quantile–quantile plot of the out-of-fold residuals produced by the Histogram Gradient Boosting model. The ordered residuals are compared with the theoretical quantiles of a normal distribution, while the dashed line represents the expected pattern under perfect normality. The reported Q–Q correlation of 0.987 indicates a strong overall correspondence between the empirical residual distribution and a Gaussian distribution.

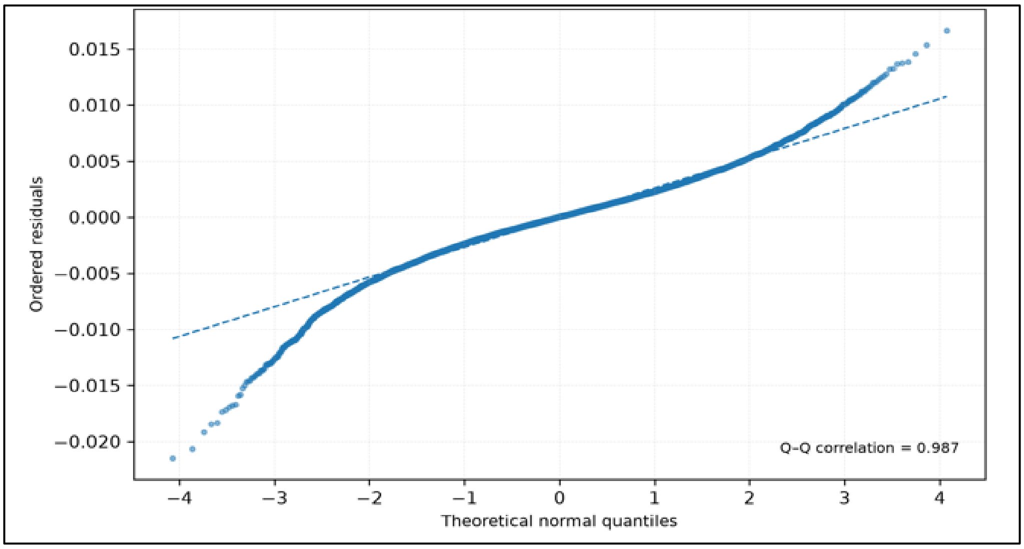

Fig 3 Normal Q–Q Plot of Spatially Cross-Validated Residuals

The central portion of the residual distribution follows the reference line closely, particularly between theoretical quantiles of approximately −2 and 2. This indicates that most residuals, especially those near zero, are approximately normally distributed. The ordered residuals in this central region range roughly from −0.006 to 0.006 and show only limited departure from the expected linear pattern. More pronounced deviations occur in both tails. At the lower end, residuals fall below the reference line and reach values close to −0.022, whereas the upper tail rises above the line and reaches approximately 0.017. This S-shaped pattern indicates heavier tails than expected under a strict normal distribution. The stronger departure in the lower tail suggests that a small number of grid cells exhibit relatively large negative residuals, corresponding to local overprediction by the model. Positive extreme residuals are also present, but their magnitude is slightly smaller than that of the most negative values. The high Q–Q correlation nevertheless confirms that non-normality is mainly restricted to a relatively small proportion of extreme observations. The bulk of the residual distribution remains well behaved, which is consistent with the very low overall RMSE and MAE values. For tree-based machine-learning models, exact residual normality is not a formal requirement for fitting or prediction. However, the tail deviations are scientifically relevant because they indicate that extreme local errors occur more frequently than would be expected under a Gaussian error structure. These departures may reflect spatial heterogeneity, unusual combinations of environmental predictors, or edge effects in areas with very low or very high vulnerability. They may also arise from the synthetic construction of the response index. Accordingly, the Q–Q plot supports the general stability of the model while highlighting the need to examine spatial residual maps and extreme-error locations before drawing conclusions about the model's applicability to observed erosion processes.

- *Spatially Cross-Validated Performance of the Histogram Gradient Boosting Model in Reproducing the Synthetic Erosion Vulnerability Index*

Figure 4 presents the relationship between the synthetic erosion vulnerability index and the corresponding out-of-fold predictions generated by the best-performing machine-learning algorithm. The predictions were obtained using a five-fold spatial block cross-validation procedure, which reduces the influence of spatial dependence between training and validation samples. The density of observations is represented by the hexagonal colour scale, while the dashed diagonal line indicates perfect agreement between the target and predicted values.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul913

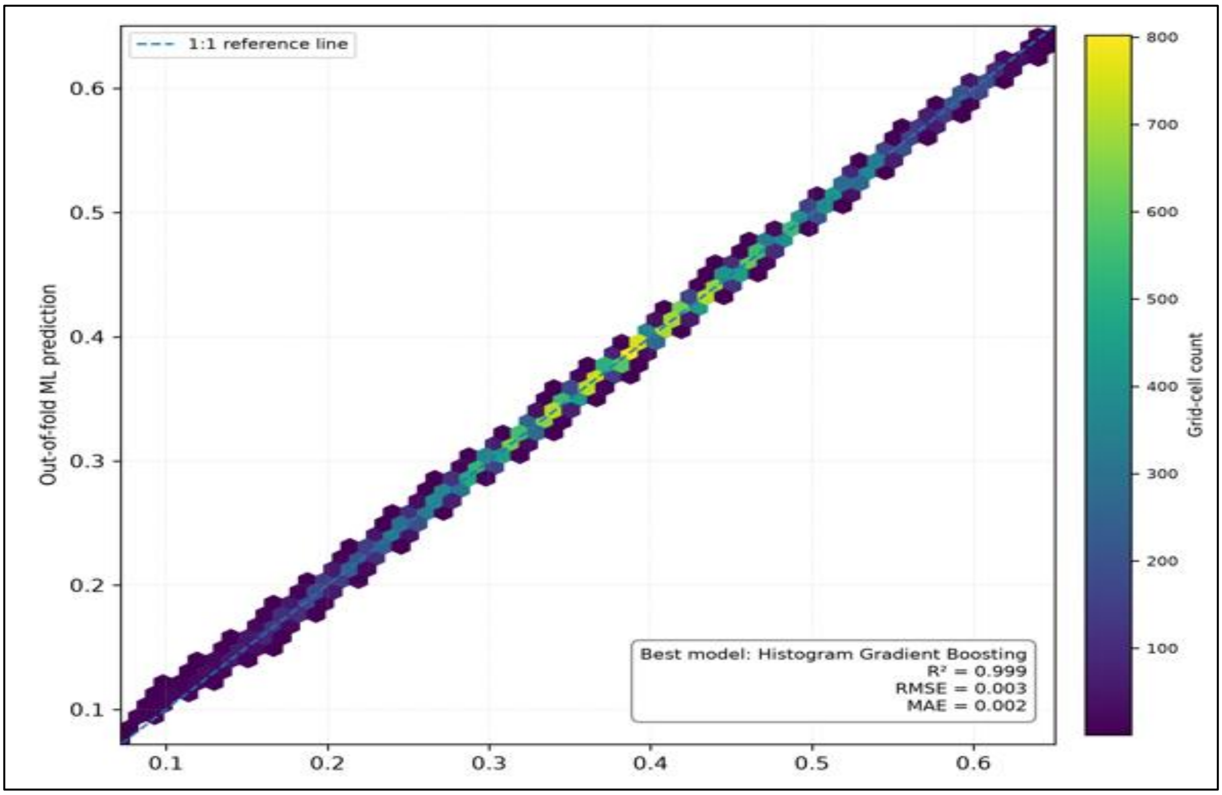

Fig 4 Spatially Cross-Validated Performance of the Histogram Gradient Boosting Model in Reproducing the Synthetic Erosion Vulnerability Index

The predicted values closely follow the 1:1 reference line throughout the entire vulnerability range, from approximately 0.08 to 0.65. The Histogram Gradient Boosting model achieved a coefficient of determination of $R^2 = 0.999$, indicating that approximately 99.9% of the spatial variation in the synthetic vulnerability index was reproduced by the model. The root mean squared error was 0.003, while the mean absolute error was 0.002, demonstrating that the average prediction deviations were extremely small relative to the full index range. The highest concentration of grid cells occurs within the intermediate vulnerability interval, particularly between approximately 0.25 and 0.45, where predicted and target values remain strongly aligned. The limited dispersion around the reference line indicates high numerical stability and strong spatial predictive consistency across the validation blocks. No substantial systematic overprediction or underprediction is apparent, although minor deviations occur near the lower and upper limits of the index. These deviations remain negligible compared with the overall variation in vulnerability values. This near-perfect agreement should nevertheless be interpreted cautiously because the response variable is the synthetic erosion vulnerability index rather than an independent measurement of observed erosion. The machine-learning model was trained using predictors that also contributed to the construction of the synthetic index, including NDVI, slope, hydrographic proximity, and elevation. Consequently, the reported performance primarily demonstrates the model's ability to emulate the multicriteria index and capture its nonlinear structure. It should not be interpreted as independent validation of actual erosion processes until the model is tested against field measurements, sediment-yield observations, mapped erosion inventories, or independently estimated soil-loss data.

➢ *Residual Distribution of the Spatially Cross-Validated Histogram Gradient Boosting Model*

Figure 5 shows the distribution of residuals as a function of the out-of-fold predictions generated by the selected Histogram Gradient Boosting model. Residuals were calculated as the difference between the synthetic erosion vulnerability index and the corresponding machine-learning prediction, such that positive values indicate underprediction and negative values indicate overprediction. The dashed horizontal line represents a residual value of zero, while the hexagonal colour scale highlights the concentration of grid cells across the prediction range.

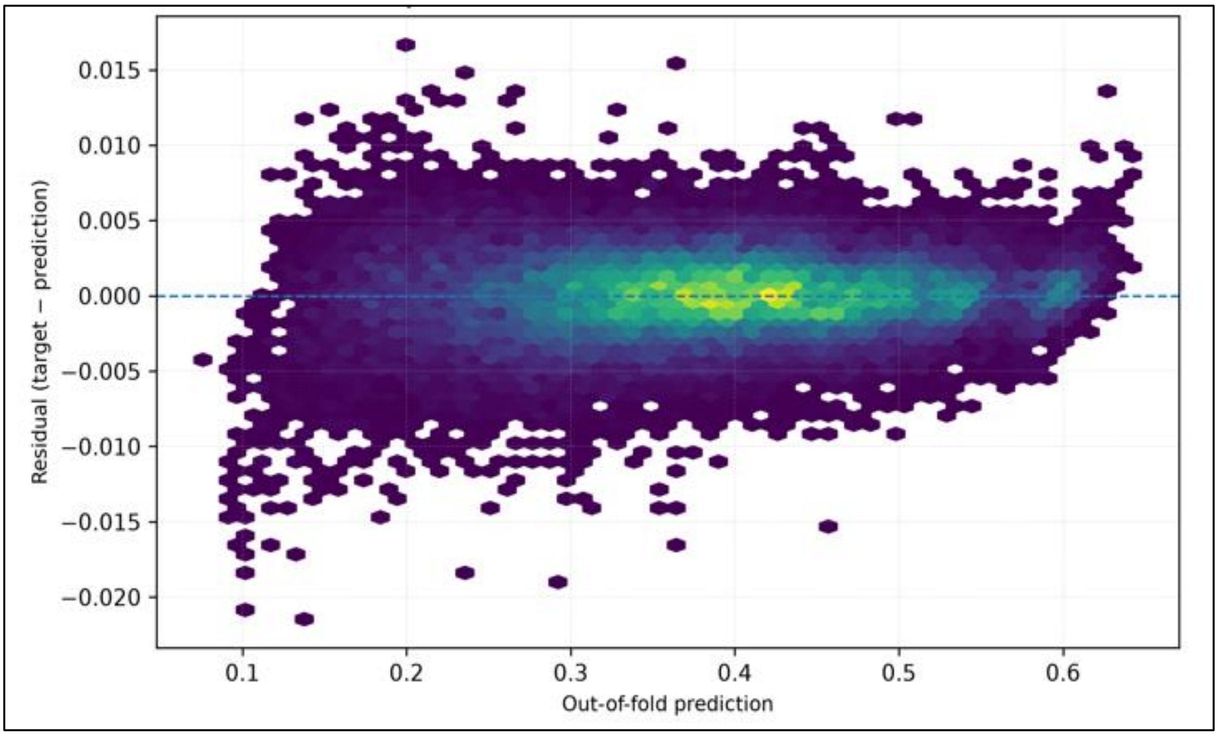

Fig 5 Residual Distribution of the Spatially Cross-Validated Histogram Gradient Boosting Model

Most residuals are tightly concentrated around zero, particularly for predicted vulnerability values between approximately 0.30 and 0.50. The highest density of observations occurs near predicted values of about 0.40, where residuals are generally confined to a narrow interval of roughly −0.003 to 0.003. This pattern confirms the very small prediction errors previously reflected by the overall RMSE of 0.003 and MAE of 0.002. The residual cloud remains broadly centred on the zero line across the full prediction range, from approximately 0.08 to 0.65, indicating the absence of a strong systematic bias. However, the dispersion is slightly wider at the lower end of the prediction range, especially below about 0.20, where residuals extend to nearly −0.022 and 0.015. This suggests that the model exhibits somewhat greater variability when reproducing the lowest vulnerability values. At higher predicted values, above approximately 0.55, the residual spread narrows again, although a few isolated positive and negative deviations remain visible. The overall residual pattern does not show a pronounced curvature, which indicates that the model adequately captures the main nonlinear structure of the synthetic erosion index. Nevertheless, the mild change in residual variance across the prediction range suggests limited heteroscedasticity, with larger errors concentrated mainly in low-vulnerability areas. These deviations remain small relative to the total index range and do not materially affect the overall predictive agreement. As with the previous validation figure, this result should be interpreted as evidence of strong model emulation rather than independent validation of observed erosion. Because the response variable is a synthetic index derived from environmental predictors also used by the machine-learning model, the near-zero residuals primarily demonstrate the model's capacity to reproduce the multicriteria formulation. Independent field-based erosion observations would still be required to assess the model's real-world predictive validity.

IV. DISCUSSION

The spatial assessment indicates that erosion vulnerability in the Betsiboka Region results from the combined influence of vegetation condition, terrain, hydrographic connectivity, elevation, geology, soils, rainfall, and land cover. The geological factor shows a marked west–east contrast, with higher susceptibility concentrated mainly in the western and northwestern sectors. This pattern suggests that lithological and weathering conditions may contribute more strongly to erosion vulnerability in these areas, whereas the eastern part of the region generally presents lower geological susceptibility. The hydrographic proximity factor displays a dense dendritic pattern closely associated with the drainage network. High values occur along rivers and tributaries, while lower values dominate the interfluves and areas located farther from watercourses. These results emphasize the role of hydrological connectivity in controlling sediment transfer. Areas near channels are more likely to facilitate the rapid transport of detached material, although stream proximity alone does not determine erosion severity. Among the tested machine-learning algorithms, Histogram Gradient Boosting provided the strongest spatially cross-validated performance. The model reproduced the synthetic vulnerability index with an $R^2$of 0.999, an RMSE of 0.003, and an MAE of 0.002. The close alignment between predicted and target values indicates that the model effectively captured the nonlinear relationships embedded in the multicriteria index. Residuals were generally centred around zero, with slightly greater dispersion at low vulnerability values, suggesting limited local variability but no major systematic bias. The normal Q–Q plot showed a high correlation of 0.987, indicating that most residuals followed an approximately normal distribution. However, deviations in both tails reveal the presence of a small number of extreme

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul913

prediction errors. These departures may be associated with unusual combinations of environmental conditions, spatial heterogeneity, boundary effects, or limitations in the original susceptibility scores. Despite the very high predictive accuracy, the machine-learning results should be interpreted cautiously. The response variable was a synthetic vulnerability index derived partly from the same predictors used in the models. Therefore, the analysis demonstrates strong model emulation and spatial consistency rather than independent validation of actual erosion. In addition, the geological and soil susceptibility values were based on heuristic classifications and require expert and field validation. Overall, the combined multicriteria and machine-learning framework provides a consistent basis for identifying spatial vulnerability patterns and prioritizing areas for watershed management. Western and northwestern sectors, together with river-adjacent zones, appear particularly important for erosion-control planning. Future work should incorporate measured soil loss, sediment yield, gully inventories, or field observations to evaluate the model against independent erosion data and improve its operational reliability.

V. CONCLUSION

This study developed an integrated framework combining multicriteria environmental analysis and spatially cross-validated machine learning to assess synthetic erosion vulnerability in the Betsiboka Region. The results highlighted strong spatial contrasts associated with geological conditions and hydrographic proximity, with higher susceptibility concentrated mainly in western, northwestern, and river-adjacent areas. Histogram Gradient Boosting produced the best predictive performance, with $R^2 = 0.999$, RMSE = 0.003, and MAE = 0.002. Residual analyses confirmed high spatial consistency, although small deviations remained at the extremes of the vulnerability range. These findings demonstrate that machine learning can effectively reproduce the nonlinear structure of a multicriteria erosion index. However, because the response variable was synthetic, the results should be interpreted as model emulation rather than independent validation of observed erosion. The proposed approach provides a useful basis for regional screening and watershed prioritization, but future applications should incorporate field observations, sediment measurements, or independently estimated soil-loss data to improve validation and operational reliability.

REFERENCES

[1]. Borrelli, P., Robinson, D.A., Fleischer, L.R., Lugato, E., Ballabio, C., Alewell, C., Meusburger, K., Modugno, S., Schütt, B., Ferro, V., Bagarello, V., Van Oost, K., Montanarella, L. and Panagos, P. (2017) 'An assessment of the global impact of 21st century land use change on soil erosion', Nature Communications, 8, Article 2013. https://doi.org/10.1038/s41467-017-02142-7.

[2]. Cox, R., Bierman, P., Jungers, M.C. and Rakotondrazafy, A.F.M. (2009) 'Erosion rates and sediment sources in Madagascar inferred from ^{10}Be analysis of lavaka, slope, and river sediment', The Journal of Geology, 117(4), pp. 363–376. https://doi.org/10.1086/598945.

[3]. Ralison, O.H., Borges, A.V., Dehairs, F., Middelburg, J.J. and Bouillon, S. (2008) 'Carbon biogeochemistry of the Betsiboka estuary, north-western Madagascar', Organic Geochemistry, 39, pp. 1649–1658. https://doi.org/10.1016/j.orggeochem.2008.01.010.

[4]. Paul, J.D., Radimilahy, A., Randrianalijaona, R. and Mulyakova, T. (2022) 'Lateritic processes in Madagascar and the link with agricultural and socioeconomic conditions', Journal of African Earth Sciences, 196, Article 104681. https://doi.org/10.1016/j.jafrearsci.2022.104681.

[5]. Elbadaoui, K., Mansour, S., Ikirri, M., Abdelrahman, K., Abu-Alam, T. and Abioui, M. (2023) 'Integrating Erosion Potential Model (EPM) and PAP/RAC guidelines for water erosion mapping and detection of vulnerable areas in the Toudgha River watershed of the Central High Atlas, Morocco', Land, 12(4), Article 837. https://doi.org/10.3390/land12040837.

[6]. Mosavi, A., Sajedi-Hosseini, F., Choubin, B., Taromideh, F., Rahi, G. and Dineva, A.A. (2020) 'Susceptibility mapping of soil water erosion using machine learning models', Water, 12(7), Article 1995. https://doi.org/10.3390/w12071995.

[7]. Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F. and Dormann, C.F. (2017) 'Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure', Ecography, 40(8), pp. 913–929. https://doi.org/10.1111/ecog.02881.

[8]. Humanitarian OpenStreetMap Team (2026) Madagascar: Waterways (OpenStreetMap Export). Humanitarian Data Exchange. Dataset accessed 16 July 2026.

[9]. UNITAR–UNOSAT (2020) Satellite-detected surface waters in the Republic of Madagascar, 9–13 February 2020. Product FL20200128MDG, Product ID 2807. Geneva: United Nations Institute for Training and Research.

[10]. U.S. Geological Survey, Earth Resources Observation and Science Center (2023) USGS EROS Archive—Vegetation Monitoring—eVIIRS Global Normalized Difference Vegetation Index. Sioux Falls, SD: USGS EROS Center.

[11]. U.S. Geological Survey, Earth Resources Observation and Science Center (2018) USGS EROS Archive—Digital Elevation—Shuttle Radar Topography Mission 1 Arc-Second Global. Sioux Falls, SD: USGS. doi: 10.5066/F7PR7TFT.

[12]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011) 'Scikit-learn: Machine learning in Python', Journal of Machine Learning Research, 12, pp. 2825–2830.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul913

[13]. Breiman, L. (2001) 'Random forests', Machine Learning, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.
[14]. Geurts, P., Ernst, D. and Wehenkel, L. (2006) 'Extremely randomized trees', Machine Learning, 63(1), pp. 3–42. doi: 10.1007/s10994-006-6226-1.
[15]. Friedman, J.H. (2001) 'Greedy function approximation: A gradient boosting machine', The Annals of Statistics, 29(5), pp. 1189–1232. doi: 10.1214/aos/1013203451.