Understanding the factors influencing soil erosion using data science and remote sensing

D.Sakhri

Department of Civil Engineering, Laboratory of LEEGO, University of Sciences and Technology Houari Boumerdiene (USTHB), Algeria

Z. Matougui

Centre de Recherche en Aménagement du Territoire (CRAT), Campus Zouaghi Slimane, Route de Ain el Bey, 25000 Constantine, Algérie

A. Medjnoun

R. Bahar

Department of Civil Engineering, Laboratory of LEEGO, University of Sciences and Technology Houari Boumerdiene (USTHB), Algeria

ABSTRACT: One of the significant challenges facing modern Algerian society is the siltation of dams, a strategic issue for the sustainable water supply, primarily caused by soil erosion. This study investigates the factors influencing soil erosion by leveraging geographical information systems (GIS), data science, and machine learning techniques. By analyzing extensive datasets on climate, topography, and land use, the research identifies key predictors of erosion. Interpretable machine learning models, including Random Forest and Logistic Regression algorithms, are employed to model erosion patterns. The findings reveal that slope aspect, NDVI (Normalized Difference Vegetation Index), and aspect are the most significant factors influencing erosion. Additionally, a counterintuitive result shows that areas with lower precipitation areas are more affected by erosion, likely due to less cohesive soils and sparse vegetation. These insights are intended to enhance soil conservation strategies and promote sustainable land management practices, offering valuable guidance for mitigating the impact of soil erosion on water resources.

1 INTRODUCTION

Water security in Algeria faces significant challenges, particularly due to the silting of dams. Silting, the accumulation of sediments in dam reservoirs reduces water storage capacity and impedes efficient water management. This issue is exacerbated by Algeria's arid climate, which limits water availability. The Fergoug dam plays a crucial role in water storage and supply for the region, specifically serving the Oran – Arzew corridor and irrigating the Habra perimeter and is part of a triplex system that includes the Ouizert and Bouhanifa dams. The dam's foundations are 33 meters wide, and it was designed to hold a total reservoir capacity of 30 million cubic meters. Over the years, the dam has undergone several modifications and repairs due to damage caused by severe flooding. In response to these events, the Fergoug Dam was rebuilt in its current location and put into service in 1970. However, the sustainability and viability of this structure are threatened by siltation, which is mainly caused by soil erosion in a particular catchment area, and fills the dam's reservoir, reducing its size and capacity and thus its lifespan. The location of Fergoug watershed is a mountainous region subject to severe droughts followed by heavy rains and is naturally extremely vulnerable to water erosion (Bouderbala et al., 2018). Recent silting measures carried out by the national agency of dams and transfers (ANBT) reveals a low rate observed in the Ouizert and Bouhanifia dams (4%, 6%), a very high rate in the Fergoug dam (95%) (Gliz et al., 2015). Since then, the rate of siltation has accelerated as a result of the region's dry spells. Figure 1 shows the reduction in the dam's storage capacity between 2010 and 2021, based on Google Maps. The watershed can absorb an average of 156 million m³ of water annually. However, the dam is currently 100% silted up, making it unusable, as erosion progressively filled in the dam.

Remote sensing and geographic information systems (GIS) are increasingly used for the study of surface phenomena and form tools essential in interactive decision support systems operational for risk management operations (Aslam et al., 2021; Matougui et al., 2023). The implementation of effective measures for the conservation of soil must be preceded by an assessment of the erosion risk in the space. This study focuses on

Figure 1 Comparative aerial images of the Fergoug Dam in Algeria, showing significant siltation and loss of water capacity from 2010 to 2021

mapping the erosion sensitivity of a large area of the Fergoug dam catchment. In this work, we propose a methodology for mapping areas vulnerable to erosion as the source of solid materials extracted and transported by water based on remote sensing and climate data. The data analysis is done using a data science approach with the objectives of improving the estimation of the most affecting factors and establishing erosion susceptibility maps.

2 STUDY AREA

The study area is centred around the Fergoug dam and its watershed covering a total area of 550 km^2. Located in the northwest of Algeria in the province of Mascara, between latitude 35°15' N to 35°34' N and longitude 0°10' W to 0°15' E (Figure 2). The area has been characterized as weak lithology with marly series dating from the Cretaceous or the Neogene occupied almost all the land (Bouderbala et al., 2018). The region has a hilly relief and sparse vegetation cover, making it prone to erosion.

The dam plays a crucial role in water storage and supply for the region, specifically serving the Oran – Arzew corridor and irrigating the Habra perimeter and is part of a triplex system that includes the Ouizert and Bouhanifa dams. The watershed can absorb an average of 156 million m³ of water annually. However, the dam is currently 100% silted up, making it unusable, as erosion progressively filled in the dam.

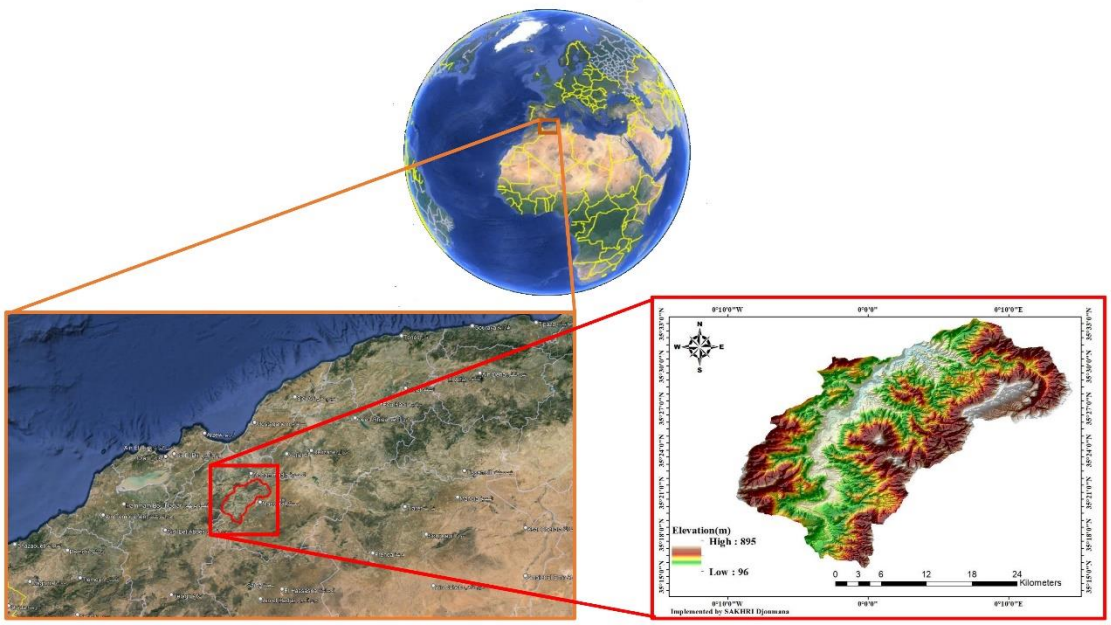

Figure 2 Localization of the watershed of Fergoug dam

3 MATERIAL AND METHODOLOGY

3.1 *Data used*

The inventory of eroded areas presented in Figure 3 was created using Sentinel imagery. This satellite imagery provides high-resolution data, allowing for detailed observation and analysis of the region's topographical and environmental features. The use of Sentinel data is integral to accurately mapping erosion patterns and assessing the extent of soil degradation in the study area. This data serves as a crucial resource for understanding the spatial distribution of erosion.

To create a comprehensive dataset of 12 factors, various data sources were exploited including Elevation, Aspect, Slope, Stream density, Topographic Wetness Index (TWI), maximum Land Surface Temperature (LST), Rainfall, Wind Speed, Normalized Difference Vegetation Index (NDVI) and Land use maps (Figure 4). Table 1 provides an overview of the various data sources used in the study to analyse the Fergoug Dam and its watershed. The climatic factors included in the study, such as rainfall wind speed, and LST are based on the mean values calculated over a 23-year period, from 2000 to 2023. This long-term dataset provides a comprehensive understanding of the climate patterns in the region, which is crucial for analysing the impacts on soil erosion and water management at the Fergoug Dam and its watershed.

Table 1 Data sources of the factors used in this study

Data	Sources
Elevation Aspect Slope Stream density TWI	https://doi.org/10.5067/Z97HFCNKR6VA)
LST	https://doi.org/10.5067/MODIS/MOD11A2.061
Rainfall	https://doi.org/10.1038/sdata.2015.66
Wind speed	https://doi.org/10.1038/sdata.2017.191
NDVI	https://doi.org/10.5067/MODIS/MOD13Q1.061
Land use	https://doi.org/10.5281/zenodo.5571936
Landform	https://doi.org/10.1371/journal.pone.0143619

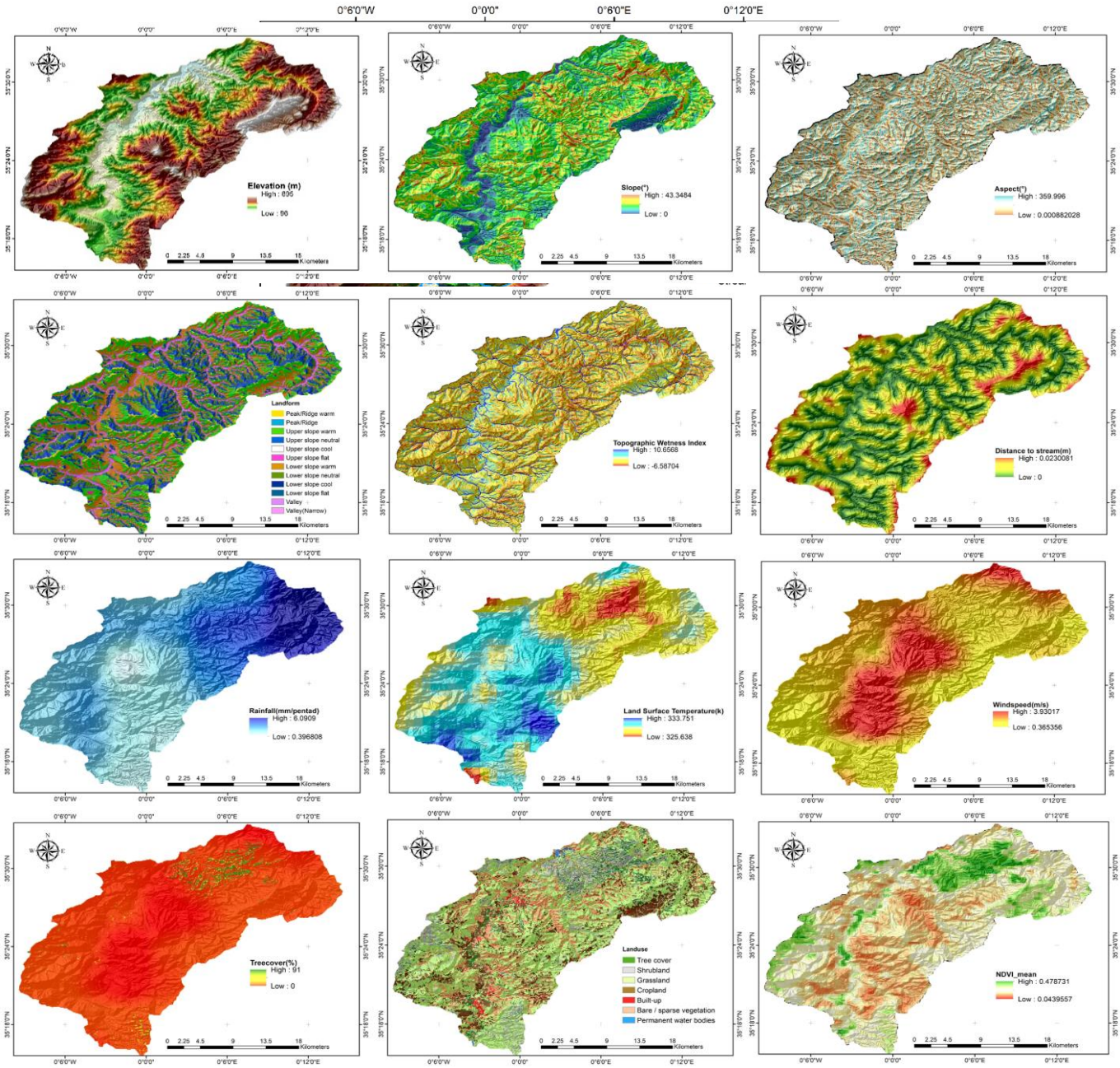

Figure 4. Set of raster variables of both environmental and geographical factors obtained from an eroded area map extracted from remote sensing imagery.

3.2 *Methodology*

The methodology employed in this study integrates remote sensing data and climatic data. Soil degradation, specifically erosion, is influenced by numerous interacting factors, which often combine in complex ways. By leveraging data science techniques, we aim to understand the influence of each factor on erosion.

The target variable, indicating the presence of erosion, was classified into binary categories: '1' for areas affected by erosion and '0' for non-affected areas. Exploratory data analysis was performed using density plots for continuous variables and bar plots for categorical variables, allowing for a visual comparison of the

distribution of features between eroded and non-eroded areas. This analysis highlighted patterns and differences in feature distributions, helping to identify factors strongly associated with erosion.

These factors are then utilized to implement interpretable machine learning models to elucidate how each factor affects the susceptibility to erosion. The models are trained using an erosion map and the associated factors, enabling it to make predictions about new areas based on the characteristics observed in regions already affected by erosion. The spatial dataset was divided into two regions to evaluate the model's ability to generalize the phenomenon. Following the scheme outlined in Figure 5, areas impacted by erosion will be identified through modelling, and the most significant factors influencing erosion will be determined.

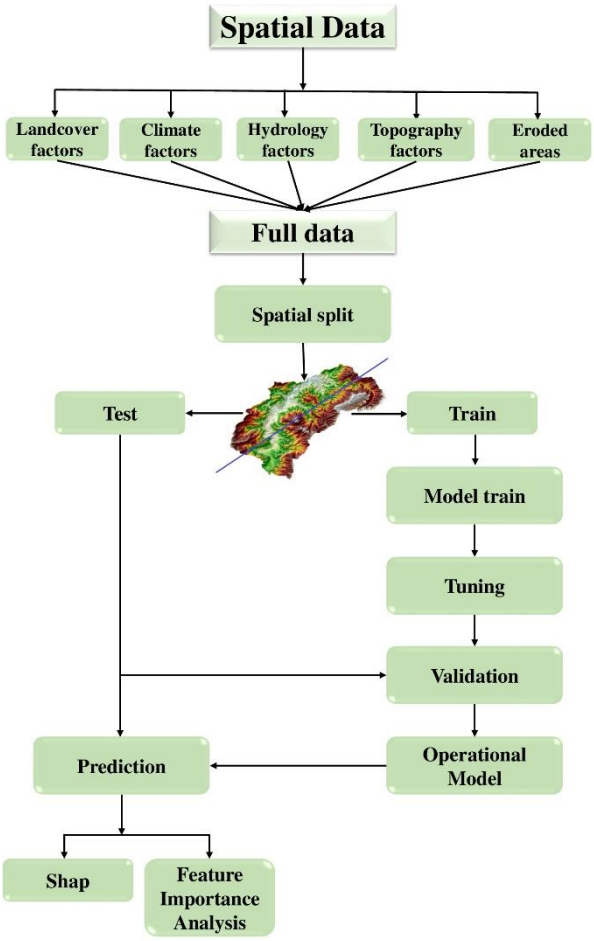

Figure 5 Methodology used in this study.

4 RESULTS AND DISCUSSIONS

4.1 *Binary analysis of the factors influencing the erosion*

The figure 6 presents a binary analysis of the factors related to soil erosion in the study area. This analysis distinguishes between areas affected by erosion (denoted as '1') and those not affected (denoted as '0'). The figure comprises several subplots, each examining a different variable's impact on erosion.

The figure underscores the complex interplay between multiple factors in determining erosion susceptibility. For instance, areas with elevations between 200 and 600 m and steeper slopes are more prone to erosion, likely due to increased runoff and the potential accumulation of water in these mid-elevation zones. The aspect distribution suggests that north-facing slopes, which receive less sunlight and may retain moisture longer, could be less vulnerable to erosion. The correlation between TWI and erosion indicates that drier areas are at higher risk, supporting the theory that soil moisture contributes significantly to soil stability.

The land use data provide critical insights into how human activities and natural land features contribute to soil degradation. Grassland and spares vegetation areas, particularly those lacking adequate vegetation cover, are more susceptible to erosion, highlighting the need for sustainable land management practices.

The analysis of LST, wind speed, and rainfall reveals significant insights into their roles in soil erosion. Higher LST values can contribute to soil desiccation, reducing soil cohesion and increasing susceptibility to erosion. Wind speed is another critical factor, with higher velocities exacerbating wind erosion. Rainfall reveals a counterintuitive finding, areas with lower precipitation levels are more affected by erosion. This phenomenon could be attributed to several factors, including the type of soil and vegetation cover present in

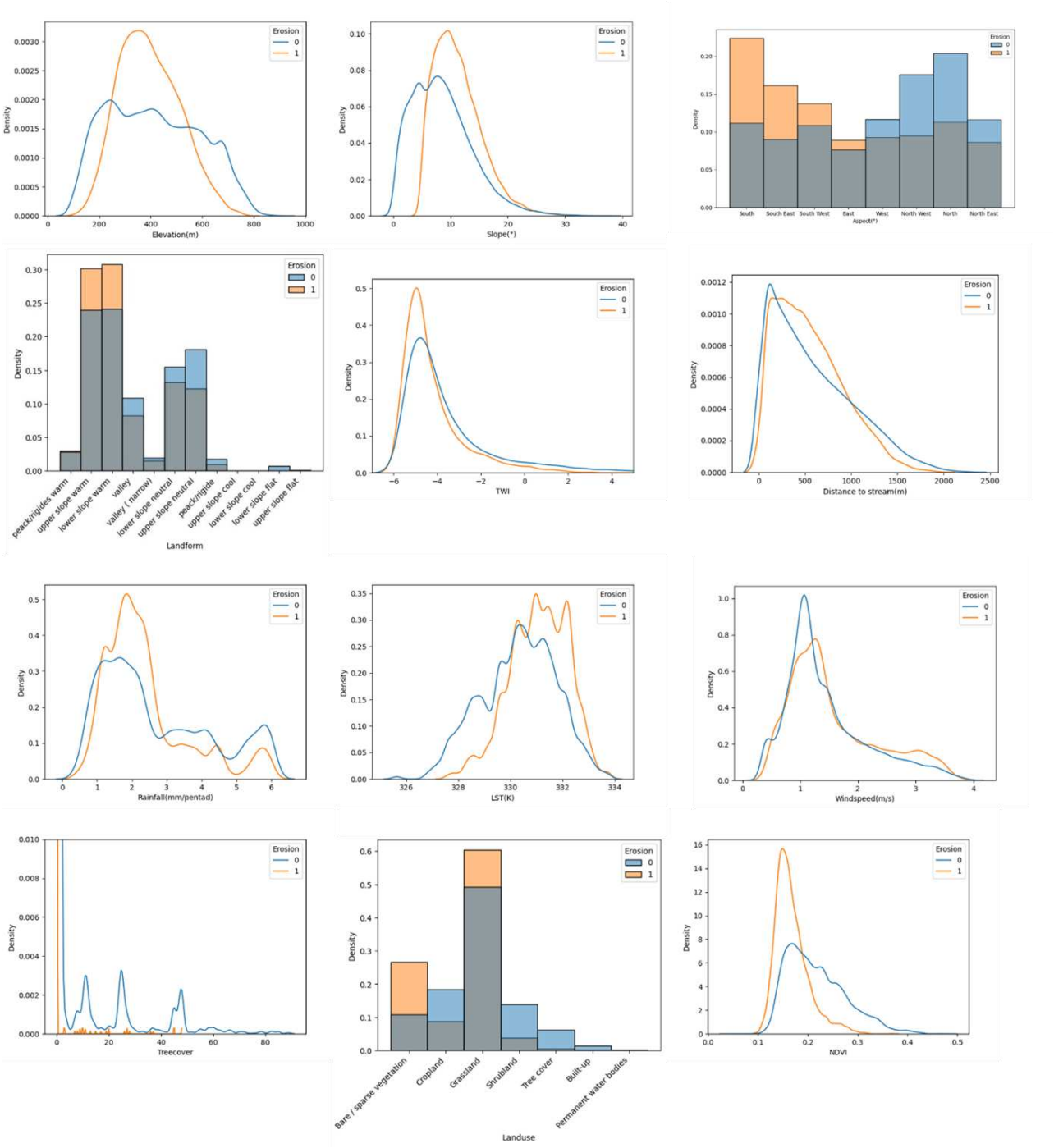

Figure 6. Binary analysis of various factors influencing soil erosion

these regions. In low rainfall areas, soils may be less cohesive and vegetation cover sparse, which reduces the protection against both wind and water erosion.

The analysis of tree cover and NDVI highlights the importance of vegetation in mitigating soil erosion. Sparse vegetation cover and low NDVI values are strongly associated with higher erosion. This is likely due to the lack of root systems to hold the soil in place and the reduced canopy cover to protect the soil surface from the impact of raindrops and wind. Conversely, areas with higher tree cover and NDVI values benefit from better soil stability and reduced erosion risk due to the protective and stabilizing effects of vegetation.

4.2 *Implementation of predictive models results*

Through the analysis of each factor, most of the factors have an impact on the erosion of the dam watershed and this analysis will be used as full datasets with the eroded area map in modelling the phenomenon of erosion using machine learning algorithms. Two prominent algorithms used in this analysis are Random Forest and Logistic Regression, each offering unique advantages in modelling erosion susceptibility.

4.2.1 *Random forest map*

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes of the individual trees. This approach helps to improve the model's accuracy and control over-fitting (Ho, Kam, 1998). In the context of erosion prediction, the Random Forest model integrates the various factors, to predict erosion-prone areas. It is particularly valued for its robustness and ability to handle large datasets with numerous input variables.

The majority of the locations indicated on the earlier erosion map were found by the model. The model also indicated regions that it anticipated would be at risk of erosion. Nonetheless, the error rate is quite low, indicating that the model correctly identified regions devoid of erosion (Figure 6). This was confirmed by determining if the region is a water surface, an agricultural area, or a construction site.

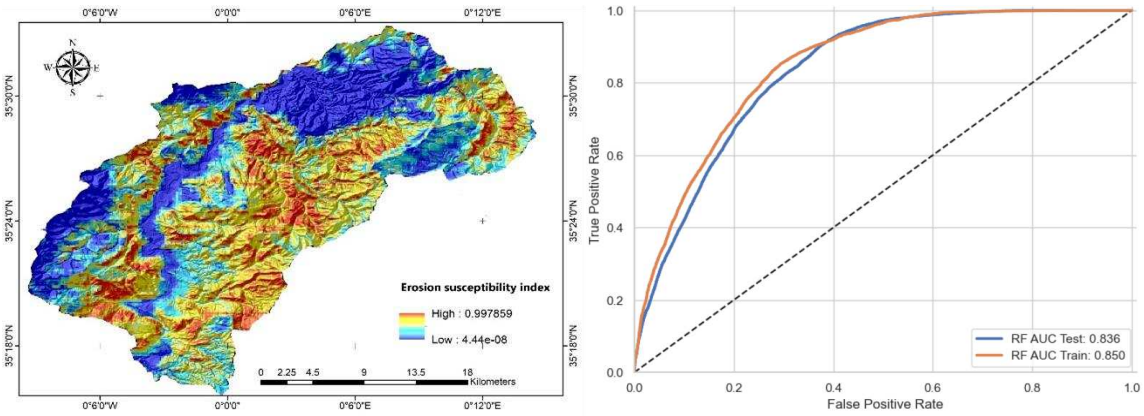

Figure 6. Random Forest model and corresponding ROC AUC

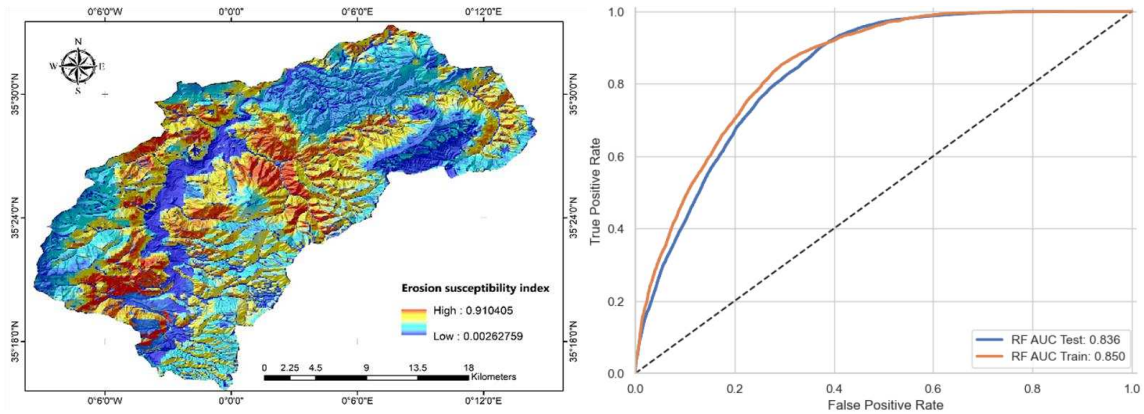

Figure 7. Logistic Regression model and corresponding ROC AUC

4.2.2 *Logistic regression map*

Logistic Regression is a statistical model commonly used for binary classification problems, such as predicting the presence or absence of erosion. This algorithm estimates the probability of a binary outcome based on one or more predictor variables, using a logistic function. Logistic Regression is useful in situations where the relationship between the dependent variable and the independent variables is not linear. In erosion prediction, it helps identify key factors contributing to erosion and provides a probability score indicating the likelihood of erosion occurring in a specific area. In addition to identifying areas that were predicted to be at risk of erosion, the model identified an important percentage of the same erosion areas as the previous erosion map. However, this model has fewer errors in the Logistic Regression map of the eroded area, which is confirmed by the (ROC AUC) result (Figure 7).

The AUC- ROC curve measures the classification issues' performance at different threshold settings. AUC is a metric or degree of separability, whereas ROC is a probability curve. It indicates the degree to which the model can discriminate between classes. These ROC curves indicate that the curve for both training and testing data is above the diagonal dashed-line area which corresponds to an AUC of 0.5, typical for a random classifier. Therefore, it is possible to note that compared to random prediction, the accuracy of the random forest model is still much higher. For the training set the values of AUC for the random forest were rather high and made 0.850 which speaks about good discrimination performance. A similar result has been observed on the test set, where the value of AUC is 0.836 which is slightly less than the value observed in the training set but still shows the promising ability to distinguish between different classes

4.2.3 *Feature importance in the model*

SHAP values offer an objective and consistent explanation of how each feature impacts a model's predictions. Rooted in game theory, SHAP (SHapley Additive exPlanations) values assign a numerical value to each feature (Shapley, 2020), indicating its contribution to the model's output. Positive SHAP values suggest that a feature influences the prediction positively, while negative SHAP values indicate a negative influence. The magnitude of these values reflects the strength of the feature's impact on the prediction.

As the Random Forest model provided the best performance in predicting erosion risk, the SHAP method was applied to this model to interpret and quantify the contribution of each feature to the model's predictions. By using SHAP values, we can objectively assess the importance of various factors.

Figure 8 presents the mean SHAP values for various factors affecting the output of a Random Forest model, with the factors ranked by their impact on the model's predictions. The SHAP values are divided into two categories, "Class 1" and "Class 0," indicating their contribution towards predicting the presence or absence of erosion, respectively. The figure revealed that slope is the most influential factor in predicting erosion, as

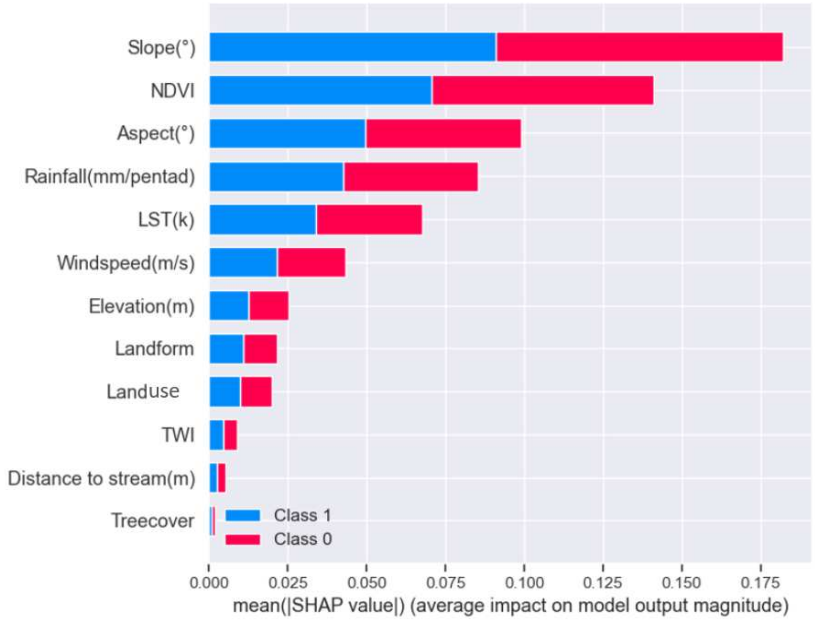

Figure 8. Mean SHAP values of factors on Random Forest

steeper slopes increase the risk due to faster surface runoff. NDVI is the second most significant factor, with denser vegetation generally reducing erosion susceptibility. The aspect also plays a critical role by affecting soil moisture and sunlight exposure, thereby influencing vegetation growth and erosion susceptibility. Rainfall intensity and frequency are crucial determinants of erosion. LST and wind speed moderately influence erosion, and wind contributing to erosion in areas with sparse vegetation. Other factors such as elevation, landform, land use, Topographic Wetness Index (TWI), distance to streams, and tree cover have varying but generally lower impacts on erosion risk, with tree cover showing minimal influence in this dataset, potentially due to limited variability in the area studied.

5 CONCLUSION

This study highlights the critical challenges facing water security in Algeria, particularly due to the significant siltation in dams, which reduces water storage capacity and impacts water management. The Fergoug Dam, a vital water source for the region, is severely affected by soil erosion and siltation, rendering it unusable. Through the use of remote sensing and GIS technologies, we have mapped areas vulnerable to erosion within the Fergoug Dam catchment, integrating data from various environmental and climatic factors over a 23-year period. The study revealed a counterintuitive finding regarding rainfall: areas with lower precipitation levels were more affected by erosion. This unexpected result suggests that in low rainfall areas, the lack of cohesive soil and sparse vegetation cover may reduce protection against erosion, emphasizing the complex interplay of environmental factors.

The application of machine learning models, specifically Random Forest and Logistic Regression, has provided valuable insights into the factors influencing erosion. The Random Forest model, enhanced by SHAP values, identified slope, NDVI, and aspect as the most significant contributors to erosion risk. These findings underscore the importance of targeted soil conservation strategies, focusing on maintaining vegetation cover and managing land use to mitigate erosion. The study demonstrates the efficacy of combining advanced data analytics and machine learning techniques to address environmental challenges, offering a pathway to improve water management and soil conservation in arid regions.

REFERENCE

Aslam, B., Maqsoom, A., Salah Alaloul, W., Ali Musarat, M., Jabbar, T., & Zafar, A. (2021). Soil erosion susceptibility mapping using a GIS-based multi-criteria decision approach: Case of district Chitral, Pakistan. *Ain Shams Engineering Journal*, *12*(2), 1637–1649. https://doi.org/10.1016/j.asej.2020.09.015

Bouderbala, D., Souidi, Z., Donze, F., Chikhaoui, M., & Nehal, L. (2018). Mapping and monitoring soil erosion in a watershed in western Algeria. *Arabian Journal of Geosciences*, *11*(23). https://doi.org/10.1007/s12517-018-4092-3

Gliz, M., Remini, B., Anteur, D., & Makhlouf, M. (2015). Vulnerability of soils in the watershed of Wadi El Hammam to water erosion (Algeria). *Journal of Water and Land Development*, *24*(1-3), 3–10. https://doi.org/10.1515/jwld-2015-0001

Ho, Kam, T. (1998). The Random Subspace Method for Constructing Decision Forests. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Matougui, Z., Djerbal, L., & Bahar, R. (2023). A comparative study of heterogeneous and homogeneous ensemble approaches for landslide susceptibility assessment in the Djebahia region, Algeria. *Environmental Science and Pollution Research*, *0123456789*. https://doi.org/10.1007/s11356-023-26247-3

Shapley, L. S. (2020). A VALUE FOR n-PERSON GAMES. In *Classics in Game Theory* (pp. 69–79). Princeton University Press. https://doi.org/10.2307/j.ctv173f1fh.12