



Mapping of Water-Induced Soil Erosion Using Machine Learning Models: A Case Study of Oum Er Rbia Basin (Morocco)

Ahmed Barakat¹ · Mouadh Rafai¹ · Hassan Mosaid¹ · Mohammad Shakiul Islam² · Sajjad Saeed^{3,4}

Received: 21 December 2021 / Revised: 6 May 2022 / Accepted: 9 May 2022 / Published online: 12 June 2022
© King Abdulaziz University and Springer Nature Switzerland AG 2022

Abstract

The basin of Oum Er Rbia River (Morocco) has been greatly affected by water-related erosion leading to loss of soils, land degradation, and deposits of sediment in dams. With this motivation, we estimated the soil erosion vulnerability using three machine learning (ML) techniques, namely random forest (RF), k-nearest neighbor (kNN), and extreme gradient boosting (XGBoost). From a total of 3034 known soil erosion locations, identified from google earth and other data archives and published works, 80% were used for soil erosion model training, with the remaining 20% used for model testing. The Boruta algorithm identified 17 most relevant environmental and geological factors, selected as the main contributors for modeling the soil erosion by water. The performance of the ML models was evaluated based on sensitivity, specificity, precision, and the Kappa coefficient. This evaluation revealed that RF, kNN and XGBoost are very good to excellent models for water-based soil erosion prediction in the study area. Soil erosion susceptibility (SES) maps were generated for all models, compared, and subsequently validated using the receiver-operating characteristic (ROC) curves and area under the curve (AUC). According to ROC results, all derived maps are reliably good predictors of potential soil erosion rates by water. The AUC values attest that all models performed comparably well, with very high accuracies, although RF had a better predictive performance (AUC = 92%) than the others (kNN AUC = 90%, XGBoost AUC = 91%). Hence, the methodology adopted in this study, based on ML algorithms, can be a helpful tool for soil erosion modeling and mapping in similar settings elsewhere. Moreover, our results provide beneficial information for decision-makers to propose appropriate measures to avoid soil loss in the Oum Er Rbia Basin.

Keywords Soil erosion modeling · Geo-environmental factors · Machine learning · Accuracy analysis · Susceptibility mapping · Oum Er Rbia Basin

1 Introduction

Soil erosion is one of the main causes of soil degradation worldwide, principally in mountainous regions, and poses a major threat to global food security and environmental sustainability (Rodrigo Comino et al. 2016). Soil erosion is a natural process caused by weathering and precipitation (Ionita et al. 2015; Pal et al. 2020; Poesen et al. 1996). However, it is also accelerated by human activities such as urbanization, agriculture practices, pasturage, and deforestation (El Jazouli et al. 2019b; Esa et al. 2018; Ionita et al. 2015). Therefore, implementing efficient preventive measures and strategies for managing and reducing the soil erosion hazard needs an appropriate assessment of its potential causes. Thus, it is crucially important to assess soil erosion status, evaluate the potential risks to soil and ecosystems safety

✉ Ahmed Barakat
a.barakat@usms.ma

¹ Geomatic, Georesources and Environment Laboratory, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Béni-Mellal, Morocco

² Department of Geosciences, Mississippi State University, Starkville MS-39762, USA

³ Earth System Physics Section, The Abdus Salam International Center for Theoretical Physics (ICTP), Trieste, Italy

⁴ Department of Earth and Environmental Sciences, University of Leuven (KU Leuven), Louvain, Belgium

and human health, and identify the factors that control soil erosion.

Numerous studies have been conducted to model and map erosion susceptibility through various techniques in the last 2 decades. Some hazard studies demonstrated the efficiency of remote sensing and Geographic Information System (GIS) (Barakat et al. 2019; El Jazouli et al. 2017, 2019a; Mohan et al. 2021; Parajuli et al. 2020; Sansare and Mhaske 2020). Combining these two techniques became widely used to assess, control, and predict soil erosion; they facilitate the extraction of enormous quantities of information about the factors that favor water-based erosion of soil. These factors can then be compiled and analyzed to map the soil vulnerability risk of a given region (Sansare and Mhaske 2020; Sarkar et al. 2020; Senanayake et al. 2020).

Many researchers also developed various models for evaluating soil erosion vulnerability and mapping those areas with high erosion risks. When integrated with geo-informatics, these models have been successively implemented to determine the various sets of conditions that control soil erosion and land degradation (Cabral et al. 2018; Jarrah et al. 2020; Puente et al. 2019). They are classified into empirical, physical, and conceptual categories with regional prediction capabilities. The most commonly used empirical soil erosion models are the Universal Soil Loss Equation (USLE) and its updated version, the Revised Universal Soil Loss Equation (Wischmeier and Smith 1978) developed for sheet and rill erosion in specific lands. Among the physically based models, the Water Erosion Prediction Project (WEPP) and Soil and Water Assessment Tool (SWAT) were developed to evaluate the impact of land use in a given watershed, predict soil erosion, and simulate sediment delivery and runoff processes (De Jong et al. 1999; Flanagan and Nearing 1995; Laffin et al. 1997). Taking a position between the physical and empirical models, the conceptual models are used to evaluate the qualitative and quantitative impacts of land use dynamics on erosion and sediment yield without the detailed information provided by high-resolution spatially and temporally distributed input data (Merritt et al. 2003). Nevertheless, most studies report that the utility of this model, or any other erosion model type, is constrained by the limits in our understanding of soil erosion processes and causes (Croke and Mockler 2001; Merritt et al. 2003). Furthermore, most of these models suffer from low predictive accuracy of gully erosion susceptibility (Conforti et al. 2011; Rahmati et al. 2016).

In recent years, several machine learning (ML) approaches have been adopted as an alternative tool to deal with the multivariate and complex nature of soil erosion hazards (Arabameri et al. 2020b; Gayen et al. 2019; Pourghasemi et al. 2017; Rahmati et al. 2017; Vu Dinh et al. 2021). ML approaches are automatic methods that analyze historical data to establish an analytical model for predicting soil

erosion (Aggarwal 2018; Conoscenti et al. 2018; Kavzoglu et al. 2019; Saha et al. 2020; Vu et al. 2020). Recently, some ML methods, such as support vector machine (SVM) (Dinh et al. 2021; Meshram et al. 2020; Vu Dinh et al. 2021), random forest (RF) (Madarász et al. 2021; Paul et al. 2021), artificial neural networks (ANNs) (Gholami et al. 2021), k-nearest neighbor (kNN) (Abu El-Magd et al. 2021; Pacheco et al. 2021; Zhang et al. 2018) and extreme gradient boosting (XGBoost) (Arabameri et al. 2021; Li et al. 2020), became popular tools for interpreting remote sensing images and for analyzing soil erosion susceptibility.

In Morocco, soil erosion is one of the primary natural hazards threatening soil functions and quality. Overall, 90% of the territory is subject to a desertification process that is especially pronounced due to the arid climate and the soils being vulnerable to erosion (Ghanam 2003). Indeed, the increased exposure to water-based erosion is significantly associated with changes in land use because of agricultural practices, deforestation, and overgrazing. Numerous studies have been conducted to measure water-related soil erosion, and these have recorded that mountain regions such as the current study area are the most vulnerable areas. Soil erosion has been investigated using several quantitative models combined with remote sensing and GIS environment (Brahim et al., 2020; El Jazouli et al. 2017, 2019b; El Mouatassime et al. 2019; Elaloui et al. 2017; Meliho et al. 2020). However, qualitative modeling methods are less frequently applied to evaluate soil erosivity by water. Thus, the present study explored the potential of three machine learning models, namely RF, kNN, and XGBoost, integrated with RS and GIS techniques to map the Soil Erosion Susceptibility (SES) in the Oum Er Rbia basin. To date, there are still no research studies of this area in the literature that are based on these ML approaches and their comparison.

2 Methods

2.1 Study Area

The present study was conducted in the Oum Er Rbia River Basin upstream of the dam El Massira, located between 31°00' and 33°00' N latitude and 5°00' to 9°50' W longitude (Fig. 1). It has an approximate surface area of 29,801 km² and has an irregular elevation varying between 257 (around El Massira dam) to 4008 m.a.s.l (top, Central high Atlas Mountain). Due to the topography, the study area is open and exposed to erosion by surface runoff that can affect soil function and cause silting of dams in the Oum Er Rbia river system (Barakat et al. 2016, 2018; El Jazouli et al. 2017, 2019a). The soils are composed of different types, including Lithosols, Xerosols, Kastanozems, Luvisols, Rendzinas, Cambisols, Phaeozems, Regosols, and Gleyic Solonchaks.

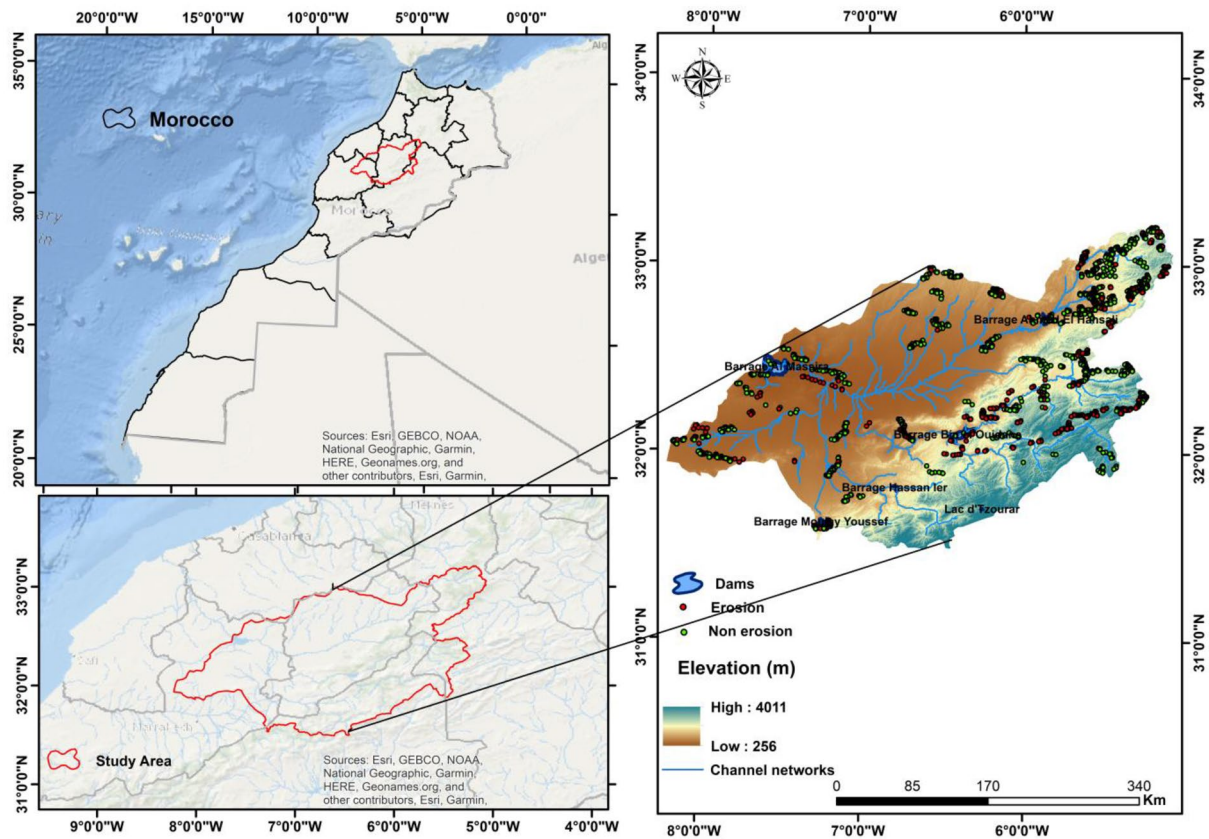


Fig. 1 Location of the study area

The basin experiences an arid or semi-arid climate with annual mean rainfall ranging from 1100 mm in the furthest upstream part of the basin to 300 mm in the downstream part. The average minimum and maximum temperatures varies from 10° to 50 °C. Hydrologically, the basin constitutes an extensive water reservoir due to its complex hydrological system. It is drained by the Oum Er Rbia River and its main tributaries of Oued Srou, Oued El Abid, and Tasaout. Cropland and arboriculture are the dominant land uses in this region. However, the forest canopy covers a large area on the mountain slopes. The primary economic activity in the study area is agriculture. The two common agricultural activities are crop cultivation and grazing. The conversion of land from forest to agriculture in the study area, combined with climate change factors, promotes land degradation processes, including water-based erosion and landslides (El Jazouli et al. 2019a, 2020).

2.2 Materials and Methodology

2.2.1 Methodology

The methodology adopted in the present study for evaluating the soil erosion rate in the Oum Er Rbia basin upstream of

the El Massira Dam is shown in Fig. 2. It included the following steps: (a) extraction of soil erosion and non-erosion sites using different sources, and randomly dividing them into two groups; one for training and the other for validation; (b) preparation using different sources of factors that potentially control soil erosion by water; (c) exploration using the Boruta algorithm to select the most effective factors in the modeling; (d) calibration and validation of the kNN, RF, XGBoost ML models; (e) preparation of the SES maps; (f) validation using the ROC curve by calculating the area under the curve (AUC); (g) comparison of the models and the SES maps based on their validation results.

2.2.2 Data Acquisition and Preparation

To prepare the thematic layers of factors (e.g., geomorphological, hydrological, soil properties, and LULC change) that might possibly control the soil erosion, the following database was collected from various sources, as listed in Table 1.

2.2.3 Erosion Data

The erosion location map is essential to making spatial predictions with various predictive models and has been treated

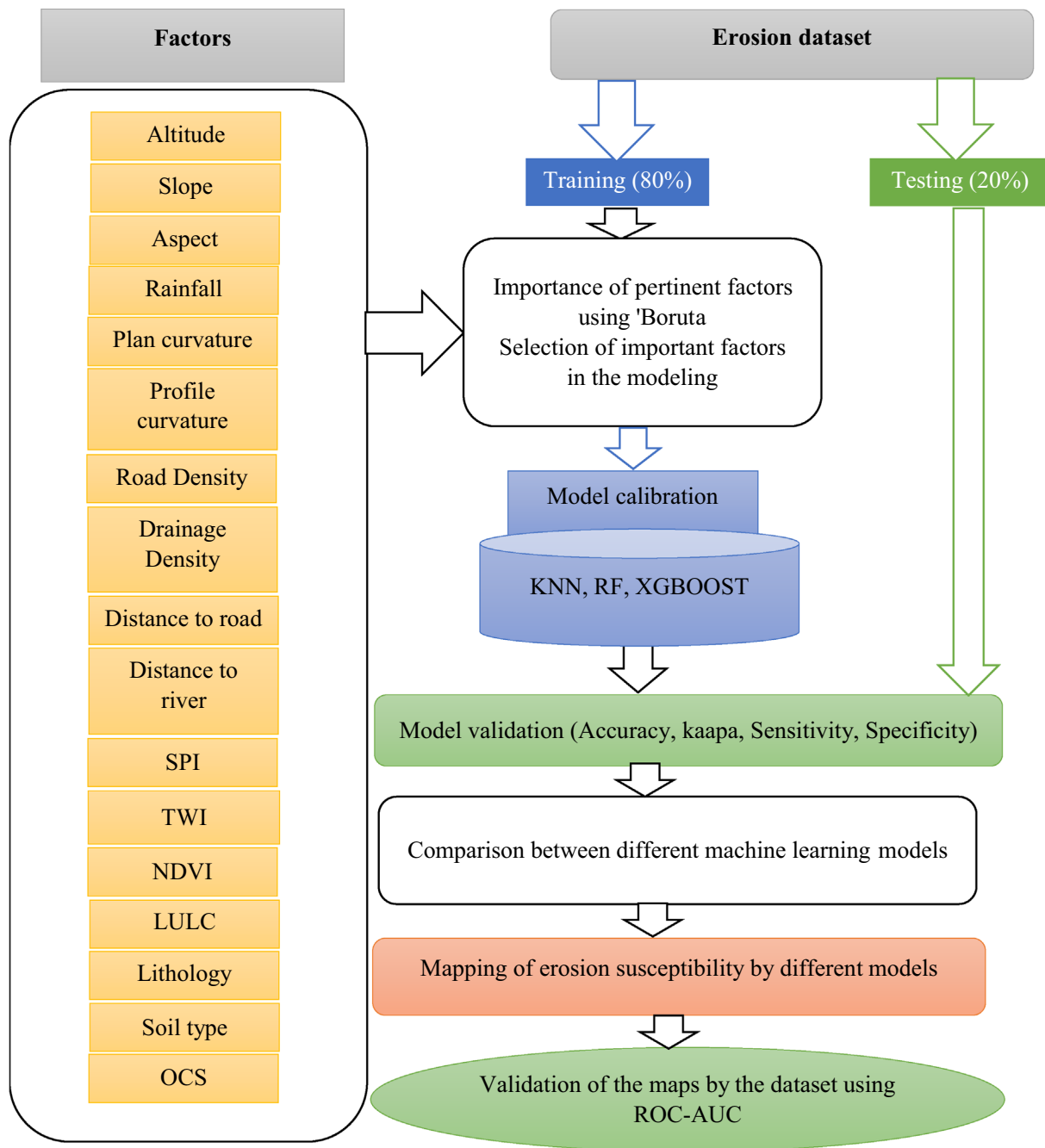


Fig. 2 Flow chart of the methodology used to provide water erosion susceptibility maps

as a dependent variable in this field of study (Saha et al., 2020). The locations of eroded and non-eroded areas are utilized to model the soil erosion susceptibility by considering the occurrence and non-occurrence of erosion (Mosavi et al. 2020). The identified soil erosion occurrences included various forms of water-based erosion, such as rill erosion and gully erosion.

They were identified from field observations, Google Earth images, and the literature (El Jazouli et al. 2017,

2019a; Elaloui et al. 2017). The recorded soil erosion sites showed various types of water-based erosion, including rill and gully erosion (Fig. 1). 1517 soil erosion inventoried sites and 1517 soil non-erosion sites inventoried were used to formulate the models and then to validate the soil erosion susceptibility maps obtained by all models.

Table 1 Data sources

Data types	Sources	Scale	Year
ASTER DEM	Earthexplor.usgs.gov	30×30	2014
Landsat 8OLI/TIRS	Earthexplor.usgs.gov	30×30	2018
Rainfall	Oum Er Rbia Hydraulic Basin Agency, PERSIANN-Cloud Classification System (PERSIANN-CCS)	4 km×4 km	2018
Geological map	Geological Survey, Ministry of Moroccan Ministry of Energy, Mines, Water and the Environment	1/100000	1985
Soil type map	http://www.fao.org/	1/5000000	2007
Organic carbon stock	https://soilgrids.org/	250×250	2020

2.2.4 Soil Erosion Influence Factors

They knew that selecting effective erosion conditioning parameters is essential to identify areas prone to soil erosion by water. A literature review (Arabameri et al. 2019a; Garosi et al. 2019a; Rahmati et al. 2017; Sajedi-Hosseini et al. 2018) was conducted to select independent (predictor) factors that were considered as candidates to predict the spatial distribution of water-based erosion. In addition, the components of the ML models, as reported in some published works (El Jazouli et al. 2019b; Rahmati et al. 2017; Sajedi-Hosseini et al. 2018), and any existing data for this study area were utilized. A total of 17 effective factors, classified as topography, hydro-climate, geology, land cover, and soil properties, were finally considered to model soil erosion and generate the erosion susceptibility map in the Oum Er Rbia basin. The selected factors are briefly described hereafter.

2.2.5 Topographic Factors

Regarded as variables that mainly influence the soil erosion rate (Garosi et al. 2019a; Gómez-Gutiérrez et al. 2015; Sajedi-Hosseini et al. 2018), topographic parameters including elevation, slope, aspect, plan curvature, and profile curvature were used in the current research on soil erosion modeling. A digital elevation model (DEM) with a cell size of 30×30 m was employed to prepare this set of topographic variables.

Elevation that significantly affects precipitation, and the associated runoff, is widely used in geohazard modeling such as water-based erosion (Conoscenti et al. 2013). For example, Zabihi et al. (2018) reported that elevation is one of the significant explanatory variables in gully erosion susceptibility assessment. The elevations in the watershed ranged from 257 to 4008 m (Fig. 3a).

Slope has always been used as one of the main factors for soil erosion mapping since it largely determines surface runoff, infiltration, drainage density pattern, and soil erosion (Arabameri et al. 2020a; Chakraborty et al. 2020; Conforti et al. 2011). In the study area, slopes vary between 0 and 75.49° (Fig. 3b). Aspect is also considered in this

study because it plays an essential role in controlling some climatic parameters such as sun and wind exposure (dry or wet), precipitation intensity, and soil moisture (Conforti et al. 2014; Lucà et al. 2011). The generated aspect map of the study area is presented in Fig. 3c.

The map of the plan curvature produced is shown in Fig. 3e. The positive values on the plan curvature map indicate that the surface is convex laterally to this cell, while the negative values mean that the surface is concave laterally. A value of zero indicates a flat surface. The profile curvature map of the study area was also generated, as presented in Fig. 3f. A positive value means that the surface is concave upwards within this cell. A negative value means that the surface is convex upwards within this cell, so that the flow will be decelerated. A value of zero indicates a flat surface.

2.2.6 Hydroclimatic Factors

Seven hydro-climatic factors controlling soil erosion were chosen: rainfall, drainage density, stream power index (SPI), topographic wetness index (TWI), distance from rivers, road density, and distance to roads.

Rainfall is one of the major factors in water-based erosion. Data from meteorological stations and satellite-based PERSIANN-CCS were used to prepare the rainfall map of the study area by applying inverse distance weighting (IDW) interpolation in a GIS environment. The annual rainfall in the study area varied between 186 and 784 mm.

The drainage density representing the length of all the rivers in the watershed is one of the more important factors and is widely used to model soil erosion by water. The drainage density in watersheds indicates the resistance to the surface and deep soil erosion. It is low because the deep soil layers have high permeability and the soil surface is well vegetated. Conversely, in areas where the deep soil layers are impermeable, and the soil surface is bare, the drainage density is high, indicating that runoff discharges rapidly, resulting in a higher erosion risk (Arabameri et al. 2020a; Conoscenti et al. 2014; Lucà et al. 2011; Mosavi et al. 2020). The drainage density map showed density values ranging from 0 to more than 0.48 km⁻¹ (Fig. 3h).

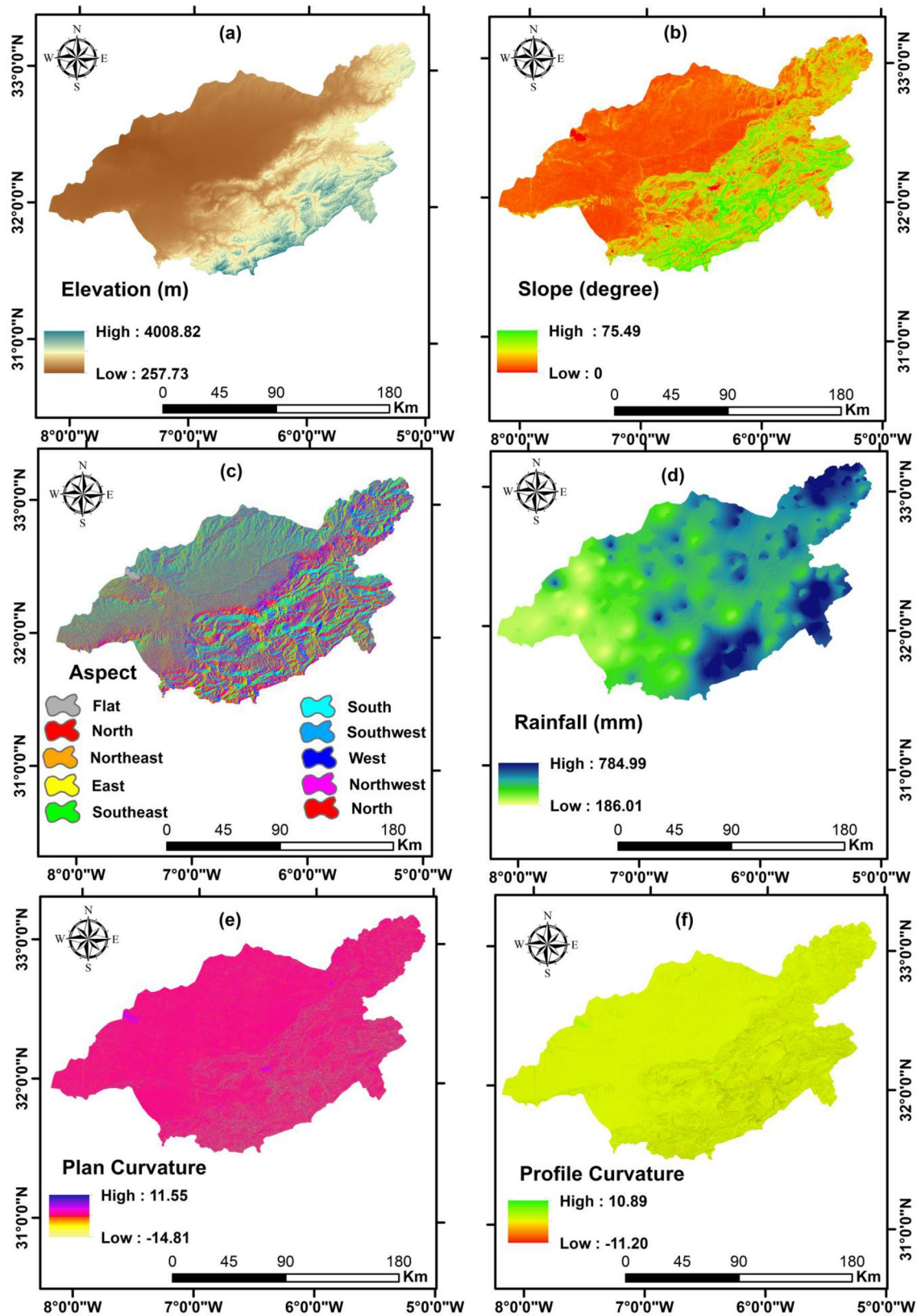


Fig. 3 Controlling factors: (a) DEM, (b) slope, (c) aspect, (d) rainfall, (e) plan curvature, (f) profile curvature, (g) road density, (h) drainage density, (i) distance from road, (j) distance from river, (k) SPI, (l) TWI, (m) NDVI, (n) LILC, (o) lithology, (p) soil type, and (q) OCS

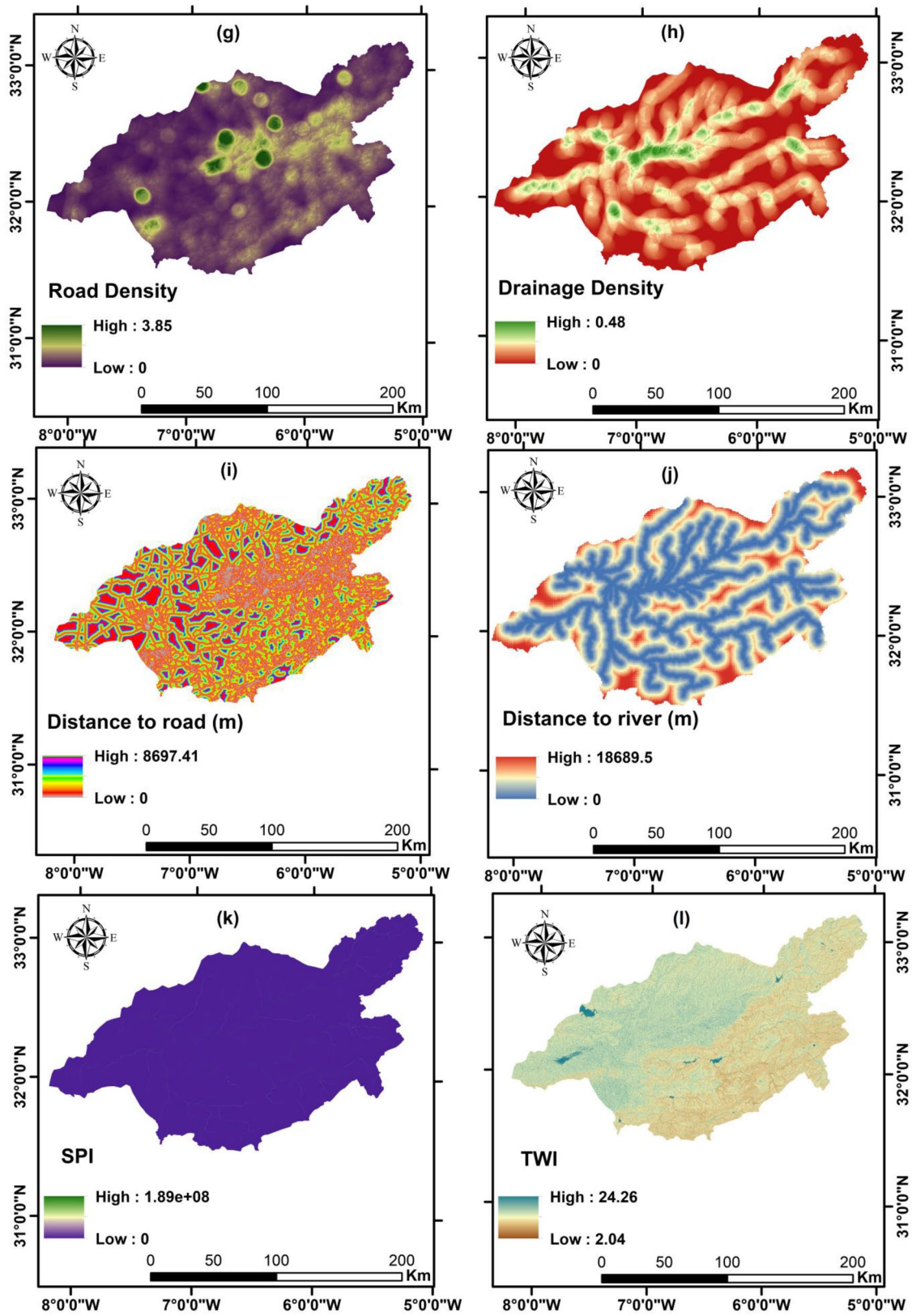


Fig. 3 (continued)

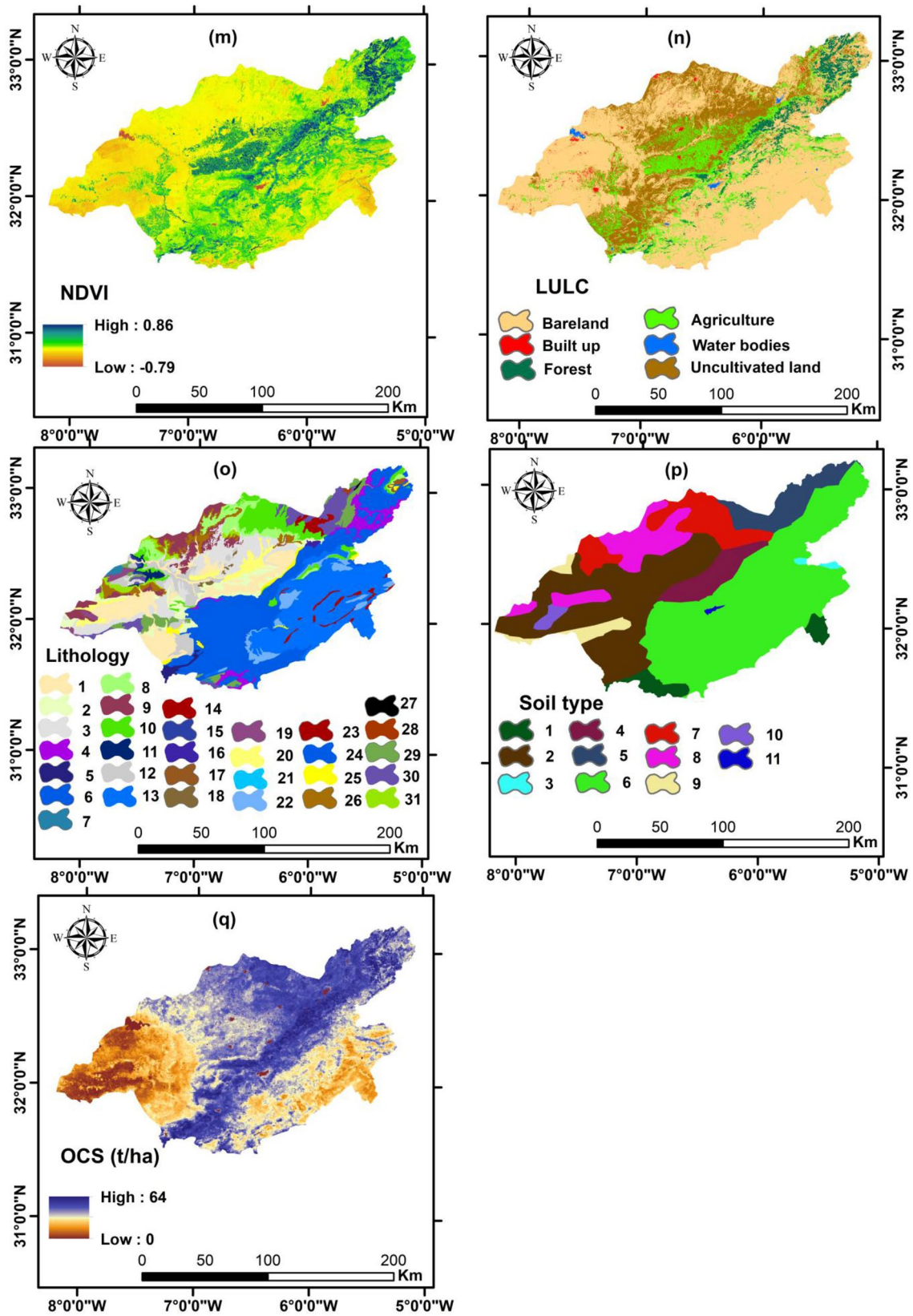


Fig. 3 (continued)

The SPI index, extensively used in assessments of soil loss, measures the degree of soil erosion by water flow and surface runoff (Moore et al. 1991). SPI was calculated in GIS environment using the following equation:

$$SPI = A \tan \beta \quad (1)$$

where A_s and β are upstream contributing area (m^2) and slope angle (in degrees), respectively. The spatial distribution map of SPI was prepared and revealed SPI values varying between 0 and 18,690 (Fig. 3k).

The TWI index is commonly employed with the SPI to predict the erosion-prone area. TWI indicates the topographic influence on the spatial distribution of saturated runoff source areas. It is calculated according to the following equation (Moore et al. 1991):

$$TWI = \ln \left(\frac{A_s}{\tan \beta} \right) \quad (2)$$

The TWI map prepared showed a TWI value range of 2–24 (Fig. 3l).

The distance from rivers represents an important parameter in predicting erosion susceptibility, because the drainage system decreases slope stability by erosion. The map shown in Fig. 3j indicates that the distance from rivers varies between 0 and 18,690 m. Generally, areas close to the river or stream (0–50 m) will be the most sensitive to erosion and most prone to flooding (Nekhay et al. 2009).

The density and distance from roads also contribute to soil erosion on the road because they can cause changes in the slope and in hydrology and drainage, leading to the slope equilibrium, and consequently promote erosion (Ayalew and Yamagishi 2005; Du et al. 2017; El Jazouli et al. 2019a;

Table 3 Soil types in the study area

Code	Description	Code	Description
1	Lithosols	7	Calcic Cambisols
2	Calcic Xerosols	8	Luvic Phaeozems
3	Kastanozems	9	Eutric Regosols
4	Calcic Kastanozems	10	Gleyic Solonchaks
5	Chromic Luvisols	11	Water
6	Rendzinas		

Nyssen et al. 2002). Figure 3h, i represents the maps generated for road density and distance from roads.

2.2.7 Geological Factors

Two important geological factors affect erosion: lithology and soil type (Mosavi et al. 2020).

The lithological units in the study area were digitized based on an available 1:1,000,000-scale geological map of Morocco. The lithological units used in this study are provided in Fig. 3o and Table 2.

Soil type and surface are also primary factors controlling soil erosion (Mosavi et al. 2020). Figure 3p presents the soil types in the study area (Table 3).

Soil properties are widely useful to map erosion. Garosi et al. (2019b) used soil properties and mentioned that soil organic carbon (SOC) greatly influences the soil erosion. The SOC map of the study area was constructed based on Soil Grids 250 m data downloaded via <http://soilgrids.org> (Fig. 3q). The SOC values varied from 0 to 64 T/ha at a depth of 0–30 cm.

Table 2 Lithology of the study area

Code	Description	Code	Description
1	Alluvium	17	Lacustrine facies
2	Current alluvium	18	Lacustrine or palustrine facies
3	Modern alluvium and silty colluvium	19	Schisto-volcanic facies
4	Clay, red marl, sandstone	20	Lake or marsh formations
5	Altered Basalt	21	Sandstone
7	Calcareous	22	Sandstone, sandstone pelites, siltstones, conglomerates
8	Beige limestone with flint, brown sand, phosphate	23	Sandstone with pelitic intercalations
9	White saccharoid limestone	24	Marls, massive Dolomites, in platelets
10	White to beige limestone, Marls	25	Miocene or Mio-Pliocene continental
11	Meso-dolomitic limestones, marls	26	Phosphate: sandstone sands, and white
12	Conglomerates	27	Quartzites, quartzite schists, metaconglomerates
13	Limestone ledges	28	Rhyolites, dacites, latites and trachy-andesites
14	Diorites and gabbros of the central high atlas	29	Shale
15	Red detrital facies	30	Olistolite zone
16	Red detritic facies, phosphate, Conglomerates	31	Phosphate facies

2.2.8 Land Use/Land Cover Factors

Land cover factors such as normalized difference vegetation index (NDVI), and land use, are also important for stimulating soil erosion processes.

Vegetation cover plays a positive role in protecting landscapes from erosion (Sajedi-Hosseini et al. 2018). Therefore, NDVI is commonly used to describe the vegetation characteristics and consequently the resistance force to topsoil erosion. It is calculated as expressed in the following equation:

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \quad (3)$$

NIR and Red are the near-infrared and red regions of spectral reflectance, respectively. In this study, Landsat 8 OLI/TIRS (Operational Land Imager/Thermal Infrared Sensor) imagery from July 2019 (2019–07-18, 2019–07-18, 2019–07-19, 2019–07-19) was used to calculate NDVI. The NDVI map generated was divided into three classes, namely bare soil ($\text{NDVI} < 0.2$), poor vegetation ($0.2 < \text{NDVI} < 0.5$), and dense vegetation ($\text{NDVI} > 0.5$), according to (Choubin et al. 2017; Julien et al. 2011) (Fig. 3m). NDVI values ranging from -0.79 and 0.86 in the study area reveal that the existing classes in the Oum Er Rbia basin are water, soil, and vegetation.

Land use categories strongly influence soil erosion processes. In general, bare and sparsely vegetated areas experience faster erosion than forests, where vegetation cover greatly reduces the erosive action of surface runoff (Chen et al. 2018; Liu et al. 2018; Yang and Lu 2018). The land use of the study area was mapped from Landsat 8 OLI images by a maximum likelihood algorithm using “ENVI” software. The land use map was created with six categories including bare soil, habitat, agriculture, forest, uncultivated land, and water (Fig. 3n). Most of the Oum Er Rbia watershed is bare soil.

2.3 Soil Erosion Modeling

2.3.1 Importance Assessment of Controlling Factors

After preparing the conditioning factors and before using the susceptibility models, it is necessary to identify a minimum optimal set of factors that might be more useful as significant indicators of erosion. However, utilizing multiple conditioning factors has some disadvantages, such as slowing down the algorithms, and decreasing the accuracy when the number of variables is much higher than the optimal number (Kohavi and John 1997; Kursa and Rudnicki 2010). Therefore, selecting a small (possibly minimal)

set of factors (Nilsson et al. 2007), and giving the best possible classification results, became a desirable step in applying machine learning methods. Many algorithms have been developed for this purpose, particularly that of Boruta proposed by Kursa and Rudnicki (2010). Executed by the Boruta package in R, the Boruta algorithm uses a forest of random trees to measure the importance of the controlling variables and then determine variables statistically to keep.

In this study, a minimal-optimal set of controlling factors was selected using the Boruta variable selection method due to its unbiased selection ability, numerical stability, ability to account for interactions between variables, and ability to handle fluctuations related to importance measures (Mercier 2017).

2.3.2 Erosion Susceptibility Approaches

After preparation of the dependent variable (i.e., the inventory map of erosion) and selection of the independent variables (controlling factors), three supervised ML algorithms: kNN (k-nearest neighbor), RF (random forest), and XGBoost (extreme gradient boosting), were used to map water-related soil erosion susceptibility in the Oum Er Rbia basin. The models were calibrated and tested on the data collected from different sources. To create training and testing subsets, different ratios are applied in the literature to split the dataset, and the most common followed is the Pareto ratio of 80%:20% or sometimes 70%:30% or 90%:10% (Bui et al. 2012; Vasu and Lee 2016). In this study, we opted for the common splitting ratio of 80:20, and two sets are randomly partitioned for the dataset: 80% of data is used in the training set, and the remainder of 20% is used for the testing set. Finally, the different models were implemented in the Rstudio environment using the R programming language.

2.3.2.1 RF RF is a supervised ML, using a multivariate and nonparametric algorithm introduced by Breiman (2001). It is widely used for randomly generating a forest combining a group of decision trees, and often results in a more accurate and more reliable prediction (Goldblatt et al. 2016). The RF method is proven for soil erosion delineation (Arabameri et al. 2019a; Lei et al. 2020; Phinzi et al. 2021). For the reasons stated above and for its efficiency when the number of observations relative to the predictor is small (Rodriguez-Galiano et al. 2012), we employed it to map the soil erosion in the study area.

2.3.2.2 kNN The kNN algorithm, developed in 1951 by Fix and Hodges (1989), is a supervised learning method that can be used for regression and classification. It is a useful non-parametric statistical tool in data analysis (He and Wang, 2007), as it identifies the class of each test data-point by

voting for its neighbor classes of training data. It belongs to the class of algorithms that can classify an unknown entity if we have data with specific properties (dependent variable) and the value of the relation (independent variable) (Mitchell 1997).

XGBoost.

XGBoost, introduced by Chen and Guestrin (2016), is applied in this study because it represents state of the art within the machine learning community (Arabameri et al. 2021). This algorithm is based on classification trees (Breiman et al., 1984) and the gradient boosting framework (Friedman 2001). XGBoost is a popular and scalable machine learning system used to boost the performance of classification trees. A classification tree typically establishes a set of rules to categorize each erosion case based on a set of predisposing factors in a graphical structure. The main explanation for the success of XGBoost is its scalability in all situations and its capability to handle sparse data (Chen et al. 2015; Gumus and Kiran 2017).

2.3.3 Model Assessment

All ML models applied in this study were calibrated and evaluated based on training/calibration using 80% and validation using 20% of the available data to predict erosion-prone occurrences and generate the soil erosion susceptibility maps. The performance of the models was measured by employing sensitivity, specificity, the kappa coefficient, accuracy, and the receiver-operating characteristics (ROC) curve.

Sensitivity includes all pixels with erosion, correctly recognized as sensitive, while specificity includes all pixels without erosion, correctly recognized as non-sensitive (Garosi et al. 2019b). Precision represents the proportion of erosion occurrence and non-occurrence pixels correctly classified. These statistical indices were calculated as follows:

$$\text{Sensitivity} = \frac{VP}{VP + FN} \quad (4)$$

$$\text{Specificity} = \frac{VN}{VN + FP} \quad (5)$$

$$\text{Precision} = \frac{VP + VN}{VP + VN + FP + FN} \quad (6)$$

where VP represents true positives, VN true negatives, FP false positives, and FN false negatives.

The kappa coefficient represents the difference between actual observations and final model predictions. It was calculated using the likelihood of the model classification according to the following equations:

$$\text{Kappa} = \frac{P_{obs} - P_e}{1 - P_e} \quad (7)$$

$$P_{obs} = \frac{VP + VN}{N} \quad (8)$$

$$P_e = \frac{((VP + FN) * (VP + FP)) + ((FP + VN) * (FN + VN))}{N^2} \quad (9)$$

where N is the total number of pixels in the map.

The performance of all models was represented as accuracy percentages.

The effectiveness of each ML model used to produce a SES map in the study area was determined by drawing the ROC curve and by evaluating their respective average area under curve (AUC) values (Swets 1988). The ROC curve represents the true positive rate or TPR (the sensitivity) versus the false-positive rate or FPR (1- specificity) for all possible thresholds. The shape of the ROC curves suggests that a model predictive performance is higher when the ROC curve is closer to the upper left corner (Garosi et al. 2019b). The AUC indicates the overall performance of the models used in the prediction process (Pereira et al. 2012). The AUC is interpreted as the mean sensitivity value for all possible specificity values. According to (Yesilnacar and Topal 2005), the estimated AUC values vary between 0.50 and 1.00. In terms of the computed AUC value, Garosi et al. (2019a) classified the predictive performance as acceptable for $AUC \geq 0.7$, excellent for $AUC \geq 0.8$, or outstanding for $AUC \geq 0.9$. The AUC is estimated from the following equation:

$$\text{AUC} = \frac{\sum VP + \sum VN}{P + N} \quad (10)$$

where VP is the true positive, VN is the true negative, and P and N represent the total number of pixels with and without gully erosion, respectively. Values closer to 1 indicate better performance in producing the SES prediction map.

3 Results

3.1 Factor Importance

The Boruta algorithm was used to select the effective factors controlling water-based soil erosion and determine each factor's importance in the study area. As presented in Fig. 4, all 17 factors were confirmed as effective based on the Boruta algorithm; consequently, spatial modeling of the soil loss in the study area used all chosen factors. According to Fig. 5, showing the average importance analysis, SPI, NDVI, LULC, elevation, and slope were the most significant

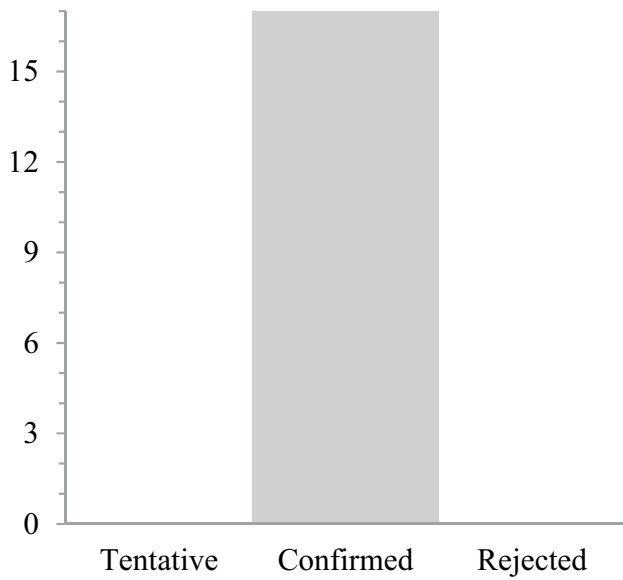


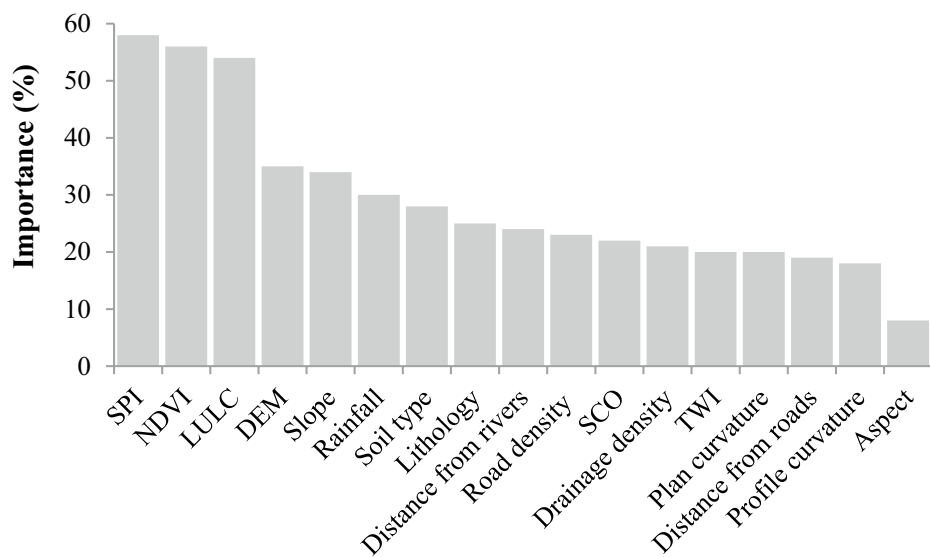
Fig. 4 Selection using Boruta Algorithm of the effective factors in modeling erosion susceptibility

factors, followed in descending order of importance by precipitation, soil types, lithology, distance from rivers, road density, SOC, and drainage density. In contrast, TWI, plan curvature, distance from roads, profile curvature, and aspect had the least impact on erosion.

3.2 Model Calibration and Validation

After confirming the influence of all factors in SES and understanding the importance and relationship between the different erosion conditioning factors, the ML models employed to predict the spatial distribution of soil loss were

Fig. 5 Importance of selected factors using Boruta algorithm.



built using the calibration dataset (80% of samples) and evaluated using the remaining 20%, as determined by accuracy measures (sensitivity, specificity, precision, Kappa) representing discrimination and reliability as different aspects of performance (Table 4).

As summarized in Table 4, the RF, kNN and XGBoost models showed high values of sensitivity (0.90, 0.87, and 0.91, respectively) and specificity (0.90, 0.86, and 0.89, respectively). Based on ML model performance, many previous studies related to water-based soil erosion assessment (Arabameri et al. 2021; Avand et al. 2019; Garosi et al. 2019b; Gayen et al. 2019; Lei et al. 2020; Mosavi et al. 2020; Pourghasemi et al. 2020; Saha et al. 2020) reported a classification error rate of 0.15 to 0.35. The Kappa index of the ML models has values of 0.80, 0.73, and 0.80 for the RF, kNN, and XGBoost models, respectively, as illustrated in Table 4. Accuracy values of 0.87, 0.90, and 0.90 were obtained for the RF, XGBoost, and kNN models, respectively. Based on the Kappa index and accuracy values, we may conclude that kNN, RF and XGBoost are very good to excellent models for water-based soil erosion prediction in the study area.

3.3 Spatial Prediction of SES

After calibration, validation, and accuracy verification of all the models used (kNN, RF and XGBoost), soil erosion probability maps were produced and classified into very low, low, medium, and high levels according to the natural break classification method (Arabameri et al. 2021, 2019b, 2018; Garosi et al., 2019b; Gayen et al. 2019; Gianinetto et al. 2020; Lei et al. 2020; Mosavi et al. 2020; Pourghasemi et al. 2020; Rahmati et al. 2017). All model soil erosion susceptibility maps reflected the probability of soil erosion by water

Table 4 Accuracy statistics for the test data set, for all used ML models

Statistical measures	Models		
	RF	kNN	XGBoost
VP	273	261	271
VN	273	264	275
FP	30	42	33
FN	30	39	27
Sensitivity	0.90	0.87	0.91
Specificity	0.90	0.86	0.89
Accuracy	0.90	0.87	0.90
Kappa	0.80	0.73	0.80

occurrence in each pixel of the whole studied basin. These maps identify the spatial distribution of ephemeral and permanent eroded areas and indicate the portions that do not show erosion evidence at present but are more susceptible to soil erosion in the future (Conoscenti et al. 2013; Garosi et al. 2019b).

The soil erosion susceptibility maps generated by the three models (Fig. 6a, b, c) show that the areas with high susceptibility to water-based erosion are located mainly in the southeast part, which contains a lot of broken and mountainous terrain. In addition, the susceptibility maps show that the high and very high susceptibility classes are close to the drainage network. The areas with low to moderate sensitivity to soil erosion have a moderate elevation and are located along the northwestern and southwestern banks of the Oum Er Rbia River. The areas with very low erosion risk are located in the center parts of the watershed, characterized by a smoother topography.

Visual interpretation of the maps suggests that the results obtained for the three ML models are very close to each other, with clear differences in class superficialities, particularly compared to the XGBoost model distribution. It can also be observed that the field erosion locations frequently coincided with the high susceptibility area for the present erosion modeling results.

The map derived from the RF model shows that 50% (14,950.63 km²), 15% (4499.32 km²) and 10% (2917.63 km²) of the total area (29,802 km²) has very low, low, and moderate erosion potential, respectively (Fig. 6a). The class with high risk occupies about 25% (7433.77 km²) of the total area (Table 5), and consists mostly of rugged mountainous terrain.

The map generated using kNN (Fig. 6b) shows that 15% (4412.86 km²) of the total studied basin is classified as having high erosion risk. The very low, low, and moderate risk areas represented 51% (15,216.32 km²), 14% (4230.99 km²), and 20% (5941.18 km²) of the total watershed area, respectively (Table 5).

The map developed using the XGBoost model shown in Fig. 6c shows that 11% (3398.22 km²) of the whole basin is susceptible to severe soil erosion and is located mainly on steep slopes or/and on the upland areas, near the drainage system. The areas classified as having very low, low, and moderate soil erosion susceptibility occupy, respectively, about 68% (20,392.28 km²), 10% (3100.48 km²), and 11% (3398.22 km²) of the whole basin, as summarized in Table 5. The closeness of these percentages does not necessarily mean similar spatial distributions in the basin.

ROC and AUC tools were applied to validate the soil erosion susceptibility maps derived from all prediction models (Fig. 7). According to Fig. 7, the AUC curve values for the RF, kNN, XGBoost are 0.92, 0.91, and 0.91, with prediction accuracies of 92%, 90%, and 91% and standard errors of 0.08, 0.1, and 0.09, respectively. It can be observed that all models performed comparably well, with RF providing a better predictive performance than the others.

3.4 Relevancy Analysis of Input Factors

Water-based soil erosion is a threshold-dependent process under the influence of a wide range of effective factors. Each factor has its importance in the models used, and identifying the relevance and importance of the erosion conditioning factors is a pivotal part of all spatial erosion susceptibility modeling. Therefore, most ML algorithms employed to model SES may be used to better understand the relative importance of each factor in the erosion formation. In this study, the relevancy analysis result is shown in Fig. 8, also confirms and enhances the result carried out using Boruta algorithm shown in Fig. 5.

The relative importance of the effective controlling factors on soil erosion calculated by the RF model followed a decreasing trend ordered as follows: NDVI, SPI, LULC, elevation, slope, rainfall, road density, lithology, organic carbon stock, soil types, drainage density, distance to river, distance to road, TWI, plan curvature, profile curvature and aspect (Fig. 8a). Based on the kNN model, the mean importance of erosion conditioning factors show the following order: SPI > LULC > slope > elevation > plan curvature > profile curvature > NDVI > soil types > rainfall > TWI > lithology > drainage density > organic carbon stock > road density > distance from road > distance from river > aspect (Fig. 8a). Figure 8c shows that in the XGBoost model, the importance of the input factors follow the order: LULC > SPI > elevation > NDVI > rainfall > slope > lithology > road density > distance from road > distance from river > organic carbon stock > drainage density > TWI > profile curvature > plan curvature > soil types > aspect. Overall, there is not a perfect match in the factor relevancy analysis between all the ML models and Boruta algorithm; however, Fig. 5 carried out by Boruta and Fig. 8a by RF are very

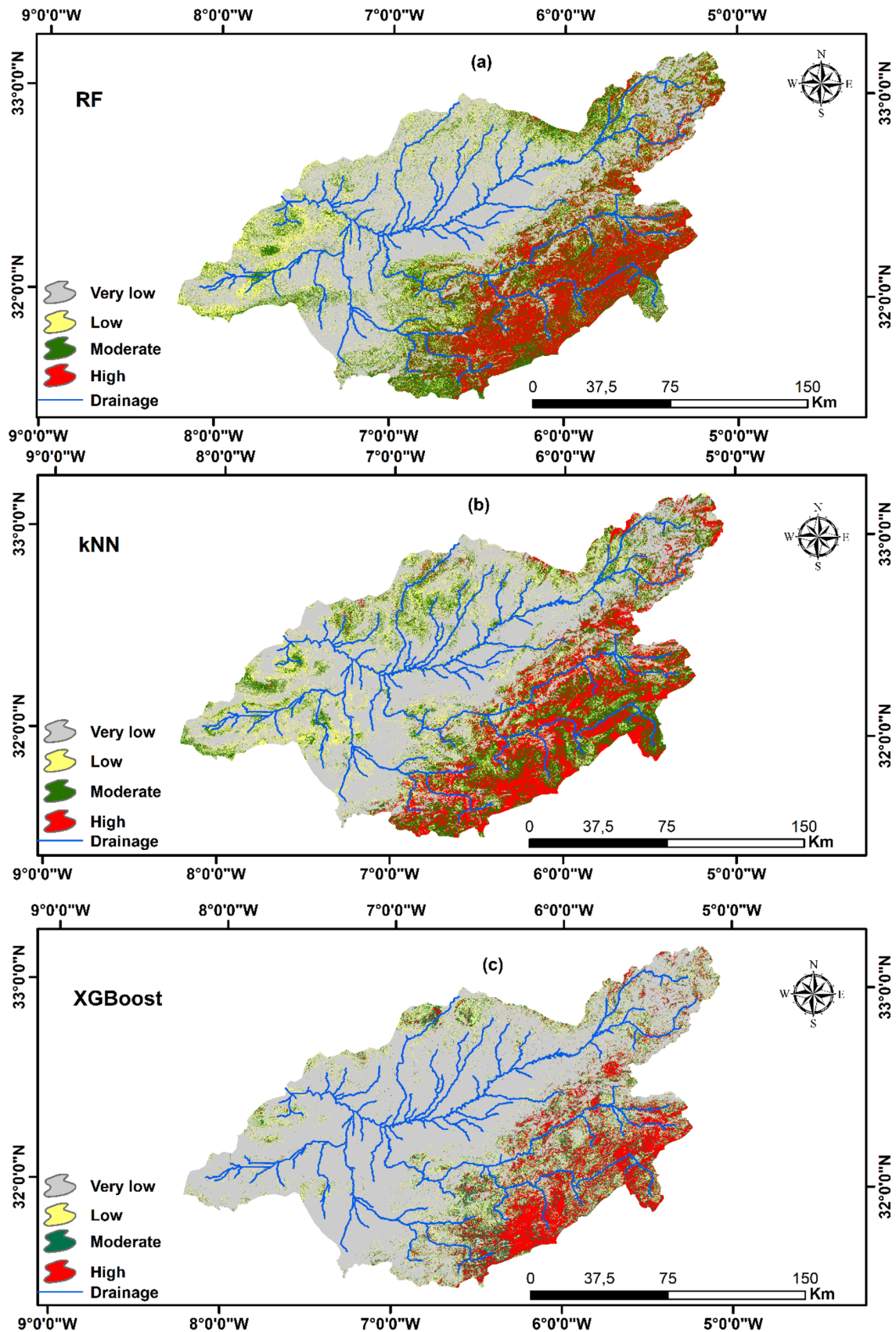


Fig. 6 Spatial prediction of soil water erosion using RF (a), kNN (b), and XGBoost (c) models

Table 5 Surfaces of the erosion susceptibility classes according to RF, kNN and XGBoost models

Susceptibility class	RF		kNN		XGBoost	
	Area (km ²)	Area (%)	Area (km ²)	Area (%)	Area (km ²)	Area (%)
Very low	14,950.63	50%	15,216.32	51%	20,392.28	68%
Low	4499.32	15%	4230.99	14%	3100.48	10%
Medium	2917.63	10%	5941.18	20%	2910.39	10%
High	7433.77	25%	4412.86	15%	3398.22	11%
Total	29,802	100%	29,802	100%	29,802	100%

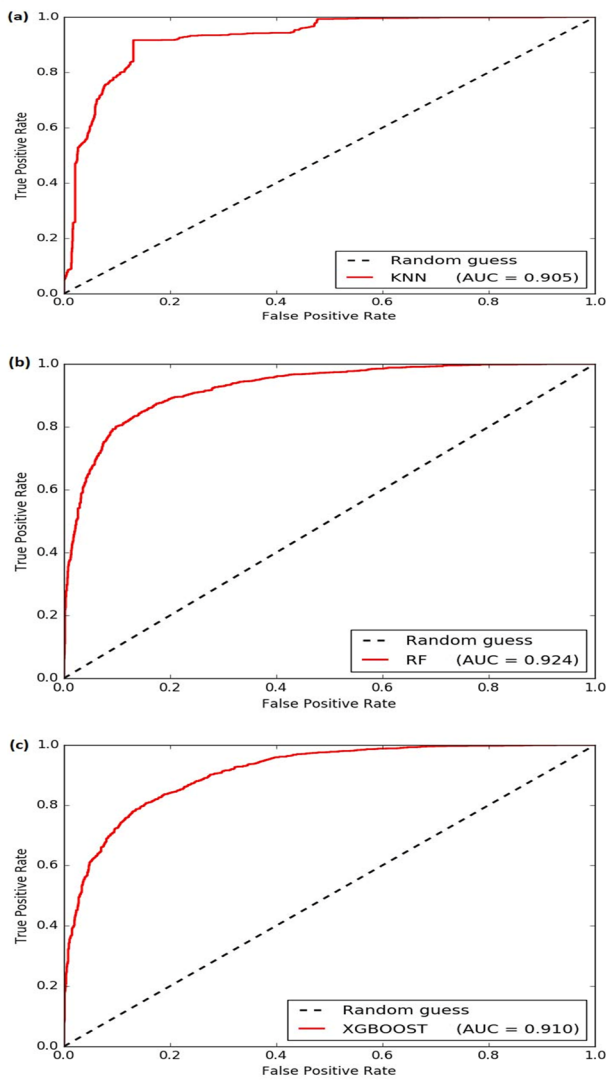


Fig. 7 Performance evaluation of the kNN (a), RF (b) and XGBoost (c) models by ROC and AUC curves

similar due to the principle of two algorithm that is based on a forest of random trees. Under the hood, Boruta applies ML methods; in particular, RF for feature importance estimation (Breiman 2001). Overall, however, SPI, LULC, NDVI, elevation, slope, rainfall were identified as the most important factors controlling water-based soil erosion.

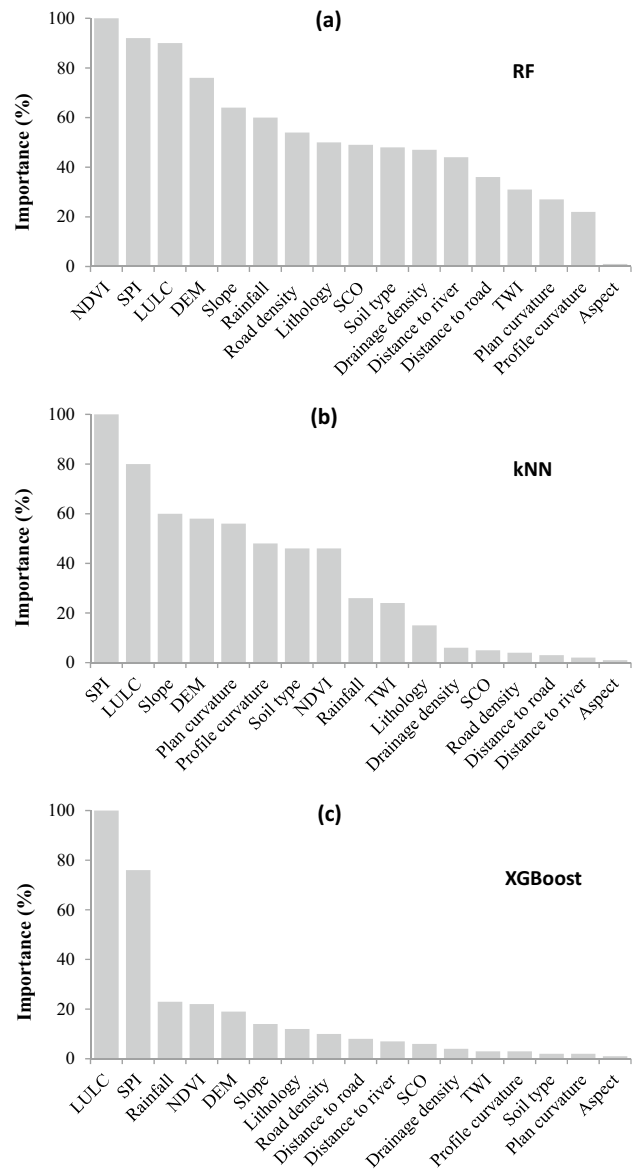


Fig. 8 Importance of factors influencing susceptibility to water erosion by RF (a), kNN (b) and XGBoost (c) models

4 Discussion

Several ML models were employed to assess the SES maps in the last decades, and have usually outperformed the traditional statistical methods (Reichenbach et al. 2018; Sahin 2020). In the present study, the tree methods RF, kNN, and XGBoost were tested to assess water-based soil erosion and model the spatial distribution of its risks in the Oum Er Rbia basin.

Our results clearly show that the RF model was the most accurate in the present research, followed by XGBoost and kNN, according to the accuracy evaluation and the kappa coefficient. The particular success of the RF model may be due to its ability to handle large databases without variable deletion and its ability to handle data assimilation, and nonlinearities between dominant factors (Catani et al. 2013; Gayen et al. 2019; Naghibi et al. 2016). This model has proved to be useful in different research areas as well, such as potential groundwater mapping, wildfire prediction, sediment yield modeling, soil erosion, and landslide susceptibility mapping (Amiri et al. 2019; Arabameri et al. 2020a, 2021; Avand et al. 2019; Chakraborty et al. 2020; Cheng et al. 2018; Pourghasemi et al. 2020; Saha et al. 2020). A study by Avand et al. (2019) on the comparison between RF and kNN for gully erosion mapping, mentions that RF had better performance than kNN. In addition, based on the ROC and AUC values that are widely used for evaluating the performance differences of ML models (Arabameri et al. 2020a, 2021; Avand et al. 2019; Lei et al. 2020; Pourghasemi et al. 2020; Rahmati et al. 2017), RF achieved a better prediction than the other two algorithms. Chen et al. (2021) confirmed that XGBoost making spatial predictions with $AUC = 0.92$ represents excellent performance. Pourghasemi et al. (2020) also compared different ML models, and concluded that RF was the best performing model with $AUC = 0.985$.

The soil erosion susceptibility classes classified according to the natural break and the susceptibility maps produced by all ML models reveal that less than 20% of the Oum Er Rbia basin is designated as having high susceptibility to water-based soil erosion, while most of the studied area was designated as having low to very low susceptibility. These results are compatible with those obtained by El Jazouli et al. (2017), El Jazouli et al. (2019b), and Elaloui et al. (2017). Comparing the maps of predicted erosion susceptibility (Fig. 6a, b, c) and NDVI (Fig. 3m), the majority of low erosion risk is distributed in areas having higher vegetation values, while the high category of erosion risk is located in areas with lower NDVI values. This indicates that the NDVI is the most important driving factor of soil erosion in the study area. Most of the eroded areas are found in areas that have low vegetation

and high altitudes or high precipitation, which results in water runoff and sediment transport as well as the removal of nutrients from the soil along with the drainage networks (Garosi et al. 2019b; Gómez-Gutiérrez et al. 2015; Poesen et al. 2003; Rahmati et al. 2017). In addition, the present study shows that LULC, SPI, precipitation, slope, lithology are of more importance in producing water-based soil erosion. These results align with those obtained by Amiri et al. (2019).

Several studies quantifying soil loss have been done at the level of different sub-basins of the Oum Er Rbia basin (Barakat 2020; El Jazouli et al. 2017, 2019b; Elaloui et al. 2017). These showed that the areas that lose a lot of soil correspond to the high sensitivity category of our results. This confirms the high performance of the models used to create the SES maps that identify areas at risk of soil erosion by water despite, even though the models cannot easily differentiate map errors from sensitive areas where erosion has not yet occurred. Nevertheless, the errors do not necessarily represent high or medium susceptibility of actual eroded areas. They showed that these areas present favorable conditions for the development and occurrence of the soil erosion but that it has not yet occurred, indicating that these areas are possibly at risk of soil erosion (Garosi et al. 2019b; Gutiérrez et al. 2009). Finally, the soil erosion susceptibility maps produced for the study area identify locations where soil erosion by water is present and those susceptible to erosion. Thus, these maps could be an important tool for regional managers or planners (Garosi et al. 2019b; Zabihi et al. 2018).

5 Conclusion

This study aimed to assess the susceptibility to water-based soil erosion in the Oum Er Rbia Basin upstream of the El Massira dam (Morocco) using RF, kNN, and XGBoost ML models. The Boruta algorithm, used to select the effective factors controlling the erosion, confirmed that all 17 factors chosen (SPI, NDVI, LULC, elevation, slope, rainfall, soil types, lithology, distance from rivers, road density, SOC, drainage density, TWI, plan curvature, distance from roads, profile curvature, aspect) had important roles in water-based soil erosion in the Oum Er Rbia Basin. The accuracy measures (sensitivity, specificity, precision, Kappa) reported that kNN, RF and XGBoost are good to excellent models for analyzing soil water erosion prediction in the study area.

The maps generated by all ML models categorized into four different classes display a very low (50–68% of the entire basin), low (10–15%), moderate (10–20%), high and very high (11–25%) severity of erosion potential. The areas with high susceptibility to water-based erosion are mainly located in the southeast and strongly related to broken and

mountainous terrain. The areas with low to moderate sensitivity to water-based erosion are connected to sites with a moderate elevation located on the northwestern and southwestern banks of the Oum Er Rbia River. The areas with very low erosion risk are located in the center parts of the watershed, characterized by a smoother topography. The relevancy analysis of the chosen factors indicated that LULC, SPI, elevation NDVI, precipitation, slope, and lithology are the most important ones contributing to soil erosion. In contrast, TWI, soil types, and aspect have a relatively low effect. Validation of the erosion susceptibility maps using ROC and AUC revealed that all models performed comparably well and notably high accuracies. The RF model had a slightly better predictive performance (AUC = 92%) than the others (kNN AUC = 90%, XGBoost AUC = 91%).

Finally, the methodological framework used in this study demonstrated that the appropriate choice of effective water-based erosion factors and the use of data-driven techniques could allow for the correct assessment of erosion effects. Its major limitation was that the input factor data collection and sampling were not done at the same scale. Despite these limitations, the methodology adopted in this study, based on ML algorithms, can be a helpful tool for soil erosion modeling and mapping in similar settings elsewhere. Moreover, the results generated in this study are important for decision-makers responsible for proposing appropriate measures to avoid soil loss in the Oum Er Rbia Basin.

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Declarations

Conflict of Interest No potential conflict of interest was reported by the authors.

References

- Abu El-Magd SA, Ali SA, Pham QB (2021) Spatial modeling and susceptibility zonation of landslides using random forest, naïve bayes and K-nearest neighbor in a complicated terrain. *Earth Sci Inf.* <https://doi.org/10.1007/s12145-021-00653-y>
- Aggarwal CC (2018) *Neural networks and deep learning*. Springer 10:978-973
- Amiri M, Pourghasemi HR, Ghanbarian GA, Afzali SF (2019) Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. *Geoderma* 340:55–69. <https://doi.org/10.1016/j.geoderma.2018.12.042>
- Arabameri A, Asadi Nalivan O, Chandra Pal S, Chakraborty R, Saha A, Lee S, Pradhan B, Tien Bui D (2020a) Novel machine learning approaches for modelling the gully erosion susceptibility. *Remote Sens* 12:2833
- Arabameri A, Chandra Pal S, Costache R, Saha A, Rezaie F, Seyed Danesh A, Pradhan B, Lee S, Hoang N-D (2021) Prediction of gully erosion susceptibility mapping using novel ensemble machine learning algorithms. *Geomat Nat Haz Risk* 12:469–498
- Arabameri A, Chen W, Loche M, Zhao X, Li Y, Lombardo L, Cerda A, Pradhan B, Bui DT (2020b) Comparison of machine learning models for gully erosion susceptibility mapping. *Geosci Front* 11:1609–1620. <https://doi.org/10.1016/j.gsf.2019.11.009>
- Arabameri A, Pradhan B, Rezaei K (2019a) Gully erosion zonation mapping using integrated geographically weighted regression with certainty factor and random forest models in GIS. *J Environ Manage* 232:928–942. <https://doi.org/10.1016/j.jenvman.2018.11.110>
- Arabameri A, Pradhan B, Rezaei K (2019b) Spatial prediction of gully erosion using ALOS PALSAR 23a data and ensemble bivariate and data mining models. *Geosci J* 23:669–686
- Arabameri A, Pradhan B, Rezaei K, Yamani M, Pourghasemi HR, Lombardo L (2018) Spatial modelling of gully erosion using evidential belief function, logistic regression, and a new ensemble of evidential belief function–logistic regression algorithm. *Land Degrad Dev* 29:4035–4049. <https://doi.org/10.1002/ldr.3151>
- Avand M, Janizadeh S, Naghibi SA, Pourghasemi HR, Khosrobeigi Bozchaloei S, Blaschke T (2019) A comparative assessment of random forest and k-nearest neighbor classifiers for gully erosion susceptibility mapping. *Water* 11:2076
- Ayalew L, Yamagishi H (2005) The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology* 65:15–31. <https://doi.org/10.1016/j.geomorph.2004.06.010>
- Barakat A, El Baghdadi M, Rais J, Aghezzaf B, Slassi M (2016) Assessment of spatial and seasonal water quality variation of Oum Er Rbia River (Morocco) using multivariate statistical techniques. *International Soil and Water Conservation Research* 4:284–292
- Barakat A, Meddah R, Afdali M, Touhami F (2018) Physicochemical and microbial assessment of spring water quality for drinking supply in Piedmont of Béni-Mellal Atlas (Morocco). *Phys Chem Earth Parts a/b/c* 104:39–46
- Barakat A, Ouargaf Z, Khellouk R, El Jazouli A, Touhami F (2019) Land Use/land cover change and environmental impact assessment in Béni-Mellal District (Morocco) using remote sensing and GIS. *Earth Syst Environ* 3:113–125. <https://doi.org/10.1007/s41748-019-00088-y>
- Barakat HmaA (2020) Cartographie et estimation de l'érosion dans le bassin versant de l'Oued Srou (Moyen Atlas, Maroc) par le modèle empirique « EPM » et la susceptibilité magnétique. In: mellal Fdsetb (ed). p 73
- Brahim B, Meshram SG, Abdallah D, Larbi B, Driss S, Khalid M, Khedher KM (2020) Mapping of soil sensitivity to water erosion by RUSLE model: case of the Inaouene watershed (Northeast Morocco). *Arab J Geosci* 13:1–15
- Breiman L (2001) *Random Forests*. *Mach Learn* 45:5–32
- Bui DT, Pradhan B, Lofman O, Revhaug I, Dick OB (2012) Landslide susceptibility mapping at Hoa Binh province (Vietnam) using an adaptive neuro-fuzzy inference system and GIS. *Comput Geosci* 45:199–211
- Cabral AIR, Silva S, Silva PC, Vanneschi L, Vasconcelos MJ (2018) Burned area estimations derived from landsat ETM+ and OLI data: comparing genetic programming with maximum likelihood and classification and regression trees. *ISPRS J Photogramm Remote Sens* 142:94–105. <https://doi.org/10.1016/j.isprsjprs.2018.05.007>
- Catani F, Lagomarsino D, Segoni S, Tofani V (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Nat Hazard* 13:2815–2831
- Chakraborty R, Pal SC, Sahana M, Mondal A, Dou J, Pham BT, Yunus AP (2020) Soil erosion potential hotspot zone identification using

- machine learning and statistical approaches in eastern India. *Nat Hazards* 104:1259–1294
- Chen H, Zhang X, Abla M, Lü D, Yan R, Ren Q, Ren Z, Yang Y, Zhao W, Lin P, Liu B, Yang X (2018) Effects of vegetation and rainfall types on surface runoff and soil erosion on steep slopes on the Loess Plateau, China. *CATENA* 170:141–149. <https://doi.org/10.1016/j.catena.2018.06.006>
- Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp 785–794
- Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H (2015) Xgboost: extreme gradient boosting. R Package Version 04–2:1
- Chen W, Lei X, Chakraborty R, Pal SC, Sahana M, Janizadeh S (2021) Evaluation of different boosting ensemble machine learning models and novel deep learning and boosting framework for head-cut gully erosion susceptibility. *J Environ Manage* 284:112015
- Cheng Z, Lu D, Li G, Huang J, Sinha N, Zhi J, Li S (2018) A random forest-based approach to map soil erosion risk distribution in Hickory Plantations in western Zhejiang Province. *China Remote Sens* 10:1899
- Choubin B, Solaimani K, Roshan MH, Malekian A (2017) Watershed classification by remote sensing indices: a fuzzy c-means clustering approach. *J Mt Sci* 14:2053–2063
- Conforti M, Aucelli PP, Robustelli G, Scarciglia F (2011) Geomorphology and GIS analysis for mapping gully erosion susceptibility in the Turbolo stream catchment (Northern Calabria, Italy). *Nat Hazards* 56:881–898
- Conforti M, Robustelli G, Scarciglia F, Aucelli PPC (2014) Studio geomorfopedologico dei fenomeni erosivi nel bacino del T. Turbolo (Calabria settentrionale) con il contributo della spettrometria della riflettanza.
- Conoscenti C, Agnesi V, Angileri S, Cappadonia C, Rotigliano E, Märker M (2013) A GIS-based approach for gully erosion susceptibility modelling: a test in Sicily, Italy. *Environ Earth Sci* 70:1179–1195
- Conoscenti C, Agnesi V, Cama M, Caraballo-Arias NA, Rotigliano E (2018) Assessment of gully erosion susceptibility using multivariate adaptive regression splines and accounting for terrain connectivity. *Land Degrad Dev* 29:724–736. <https://doi.org/10.1002/ldr.2772>
- Conoscenti C, Angileri S, Cappadonia C, Rotigliano E, Agnesi V, Märker M (2014) Gully erosion susceptibility assessment by means of GIS-based logistic regression: a case of Sicily (Italy). *Geomorphology* 204:399–411
- Croke J, Mockler S (2001) Gully initiation and road-to-stream linkage in a forested catchment, southeastern Australia. *Earth Surface Process Landforms* 26:205–217
- De Jong S, Paracchini M, Bertolo F, Folving S, Megier J, De Roo A (1999) Regional assessment of soil erosion using the distributed model SEMMED and remotely sensed data. *CATENA* 37:291–308
- Dinh TV, Nguyen H, Tran X-L, Hoang N-D (2021) Predicting rainfall-induced soil erosion based on a hybridization of adaptive differential evolution and support vector machine classification. *Math Probl Eng* 2021:6647829. <https://doi.org/10.1155/2021/6647829>
- Du G-l, Zhang Y-s, Iqbal J, Yang Z-h, Yao X (2017) Landslide susceptibility mapping using an integrated model of information value method and logistic regression in the Bailongjiang watershed, Gansu Province, China. *J Mt Sci* 14:249–268
- El Jazouli A, Barakat A, Ghafri A, El Moutaki S, Ettaqy A, Khellouk R (2017) Soil erosion modeled with USLE, GIS, and remote sensing: a case study of Ikkour watershed in Middle Atlas (Morocco). *Geoscience Letters* 4:1–12
- El Jazouli A, Barakat A, Khellouk R (2019a) GIS-multicriteria evaluation using AHP for landslide susceptibility mapping in Oum Er Rbia high basin (Morocco). *Geoenvironmental Disasters* 6:1–12
- El Jazouli A, Barakat A, Khellouk R (2020) Geotechnical studies for Landslide susceptibility in the high basin of the Oum Er Rbia river (Morocco). *Geology, Ecology, and Landscapes*:1–8.
- El Jazouli A, Barakat A, Khellouk R, Rais J, El Baghdadi M (2019b) Remote sensing and GIS techniques for prediction of land use land cover change effects on soil erosion in the high basin of the Oum Er Rbia River (Morocco). *Remote Sens Appl* 13:361–374
- El Moutassime S, Boukdir A, Karaoui I, Skataric G, Nacka M, Darvishan AK, Sestras P, Spalevic V (2019) Modelling of soil erosion processes and runoff for sustainable watershed management: Case study Oued el Abid Watershed, Morocco. *Poljoprivreda i Sumarstvo* 65:241–250
- Elaloui A, Marrakchi C, Fekri A, Maimouni S, Aradi M (2017) USLE-based assessment of soil erosion by water in the watershed upstream Tessaoute (Central High Atlas, Morocco). *Modeling Earth Syst Environ* 3:873–885
- Esa E, Assen M, Legass A (2018) Implications of land use/cover dynamics on soil erosion potential of agricultural watershed, northwestern highlands of Ethiopia. *Environ Syst Res* 7:21. <https://doi.org/10.1186/s40068-018-0122-0>
- Fix E, Hodges JL (1989) Discriminatory analysis. Nonparametric discrimination: consistency properties. *Int Stat Rev/revue Internationale De Statistique* 57:238–247
- Flanagan D, Nearing M (1995) USDA-water erosion prediction project: hillslope profile and watershed model documentation. *Nserl Rep* 10:1–123
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat*:1189–1232.
- Garosi Y, Shekhabadi M, Conoscenti C, Pourghasemi HR, Van Oost K (2019a) Assessing the performance of GIS-based machine learning models with different accuracy measures for determining susceptibility to gully erosion. *Sci Total Environ* 664:1117–1132. <https://doi.org/10.1016/j.scitotenv.2019.02.093>
- Garosi Y, Shekhabadi M, Conoscenti C, Pourghasemi HR, Van Oost K (2019b) Assessing the performance of GIS-based machine learning models with different accuracy measures for determining susceptibility to gully erosion. *Sci Total Environ* 664:1117–1132
- Gayen A, Pourghasemi HR, Saha S, Keesstra S, Bai S (2019) Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. *Sci Total Environ* 668:124–138
- Ghanam M (2003) La désertification au Maroc-Quelle stratégie de lutte. 2nd FIG Regional Conference Marrakech. pp 2–5
- Gholami V, Sahour H, Hadian Amri MA (2021) Soil erosion modeling using erosion pins and artificial neural networks. *CATENA* 196:104902. <https://doi.org/10.1016/j.catena.2020.104902>
- Gianinetto M, Aiello M, Vezzoli R, Polinelli FN, Rulli MC, Chiarelli DD, Bocchiola D, Ravazzani G, Soncini A (2020) Future scenarios of soil erosion in the Alps under climate change and land cover transformations simulated with automatic machine learning. *Climate* 8:28
- Goldblatt R, You W, Hanson G, Khandelwal AK (2016) Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine. *Remote Sensing* 8:634
- Gómez-Gutiérrez Á, Conoscenti C, Angileri SE, Rotigliano E, Schnabel S (2015) Using topographical attributes to evaluate gully erosion proneness (susceptibility) in two mediterranean basins: Advantages and limitations. *Nat Hazards* 79:291–314
- Gumus M, Kiran MS (2017) Crude oil price forecasting using XGBoost. 2017 International conference on computer science and engineering (UBMK). IEEE. pp 1100–1103

- Gutiérrez ÁG, Schnabel S, Contador JFL (2009) Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. *Ecol Model* 220:3630–3637
- He QP, Wang J (2007) Fault Detection Using the k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes. *IEEE Trans Semicond Manuf* 20:345–354. <https://doi.org/10.1109/TSM.2007.907607>
- Ionita I, Fullen MA, Zgłobicki W, Poesen J (2015) Gully erosion as a natural and human-induced hazard. *Nat Hazards* 79:1–5. <https://doi.org/10.1007/s11069-015-1935-z>
- Jarrah M, Mayel S, Tatarko J, Funk R, Kuka K (2020) A review of wind erosion models: data requirements, processes, and validity. *CATENA* 187:104388. <https://doi.org/10.1016/j.catena.2019.104388>
- Julien Y, Sobrino JA, Mattar C, Ruescas AB, Jimenez-Munoz JC, Soria G, Hidalgo V, Atitar M, Franch B, Cuenca J (2011) Temporal analysis of normalized difference vegetation index (NDVI) and land surface temperature (LST) parameters to detect changes in the Iberian land cover between 1981 and 2001. *Int J Remote Sens* 32:2057–2068
- Kavzoglu T, Colkesen I, Sahin EK (2019) Machine learning techniques in landslide susceptibility mapping: a survey and a case study. *Landslides: Theory, practice and modelling*:283–301.
- Kohavi R, John GH (1997) Wrappers for feature subset selection. *Artif Intell* 97:273–324
- Kursa MB, Rudnicki WR (2010) Feature selection with the Boruta package. *J Stat Softw* 36:1–13
- Lafren JM, Elliot W, Flanagan D, Meyer C, Nearing M (1997) WEPP-predicting water erosion using a process-based model. *J Soil Water Conserv* 52:96–102
- Lei X, Chen W, Avand M, Janizadeh S, Kariminejad N, Shahabi H, Costache R, Shahabi H, Shirzadi A, Mosavi A (2020) GIS-based machine learning algorithms for gully erosion susceptibility mapping in a semi-arid region of Iran. *Remote Sensing* 12:2478
- Li X, Luo J, Jin X, He Q, Niu Y (2020) Improving soil thickness estimations based on multiple environmental variables with stacking ensemble methods. *Remote Sens* 12:3609
- Liu J, Gao G, Wang S, Jiao L, Wu X, Fu B (2018) The effects of vegetation on runoff and soil loss: Multidimensional structure analysis and scale characteristics. *J Geog Sci* 28:59–78. <https://doi.org/10.1007/s11442-018-1459-z>
- Lucà F, Conforti M, Robustelli G (2011) Comparison of GIS-based gully susceptibility mapping using bivariate and multivariate statistics: Northern Calabria, South Italy. *Geomorphology* 134:297–308
- Madarász B, Jakab G, Szalai Z, Juhas K, Kotrocó Z, Tóth A, Ladányi M (2021) Long-term effects of conservation tillage on soil erosion in Central Europe: a random forest-based approach. *Soil and Tillage Research* 209:104959. <https://doi.org/10.1016/j.still.2021.104959>
- Meliho M, Khattabi A, Mhammdi N (2020) Spatial assessment of soil erosion risk by integrating remote sensing and GIS techniques: a case of Tensift watershed in Morocco. *Environ Earth Sci* 79:1–19
- Mercier L-M (2017) Modifications de l’algorithme SELECT () appliquées à une problématique du cancer du sein. *École Polytechnique de Montréal*
- Merritt WS, Letcher RA, Jakeman AJ (2003) A review of erosion and sediment transport models. *Environ Model Softw* 18:761–799. [https://doi.org/10.1016/S1364-8152\(03\)00078-1](https://doi.org/10.1016/S1364-8152(03)00078-1)
- Meshram SG, Singh VP, Kisi O, Karimi V, Meshram C (2020) Application of artificial neural networks, support vector machine and multiple model-ANN to sediment yield prediction. *Water Resour Manage* 34:4561–4575. <https://doi.org/10.1007/s11269-020-02672-8>
- Mitchell TM (1997) Machine learning.
- Mohan A, Singh AK, Kumar B, Dwivedi R (2021) Review on remote sensing methods for landslide detection using machine and deep learning. *Trans Emerging Telecommun Technol* 32:e3998
- Moore ID, Grayson R, Ladson A (1991) Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrol Process* 5:3–30
- Mosavi A, Sajedi-Hosseini F, Choubin B, Taramideh F, Rahi G, Dineva AA (2020) Susceptibility mapping of soil water erosion using machine learning models. *Water* 12:1995
- Naghibi SA, Pourghasemi HR, Dixon B (2016) GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ Monit Assess* 188:1–27
- Nekhay O, Arriaza M, Boerboom L (2009) Evaluation of soil erosion risk using analytic network process and GIS: a case study from Spanish mountain olive plantations. *J Environ Manage* 90:3091–3104
- Nilsson R, Pena JM, Björkegren J, Tegnér J (2007) Consistent feature selection for pattern recognition in polynomial time. *J Mach Learn Res* 8:589–612
- Nyssen J, Poesen J, Moeyersons J, Luyten E, Veyret-Picot M, Deckers J, Haile M, Govers G (2002) Impact of road building on gully erosion risk: a case study from the Northern Ethiopian Highlands. *Earth Surf Proc Land* 27:1267–1283. <https://doi.org/10.1002/esp.404>
- Pacheco AdP, Junior JAdS, Ruiz-Armenteros AM, Henriques RFF (2021) Assessment of k-nearest neighbor and random forest classifiers for mapping forest fire areas in central Portugal using Landsat-8, Sentinel-2, and Terra imagery. *Remote Sens* 13:1345
- Pal SC, Arabameri A, Blaschke T, Chowdhuri I, Saha A, Chakraborty R, Lee S, Band SS (2020) Ensemble of machine-learning methods for predicting gully erosion susceptibility. *Remote Sens* 12:3675
- Parajuli A, Gautam AP, Sharma SP, Bhujel KB, Sharma G, Thapa PB, Bist BS, Poudel S (2020) Forest fire risk mapping using GIS and remote sensing in two major landscapes of Nepal. *Geomat Nat Haz Risk* 11:2569–2586. <https://doi.org/10.1080/19475705.2020.1853251>
- Paul SS, Li J, Li Y, Shen L (2021) Assessing land use–land cover change and soil erosion potential using a combined approach through remote sensing, RUSLE and random forest algorithm. *Geocarto Int* 36:361–375. <https://doi.org/10.1080/10106049.2019.1614099>
- Pereira S, Zêzere JL, Bateira C (2012) Assessing predictive capacity and conditional independence of landslide predisposing factors for shallow landslide susceptibility models. *Nat Hazard* 12:979–988
- Phinzi K, Ngetar NS, Ebhuoma O (2021) Soil erosion risk assessment in the Umzintlawa catchment (T32E), Eastern Cape, South Africa, using RUSLE and random forest algorithm. *S Afr Geogr J* 103:139–162. <https://doi.org/10.1080/03736245.2020.1716838>
- Poesen J, Nachtergaele J, Verstraeten G, Valentin C (2003) Gully erosion and environmental change: importance and research needs. *CATENA* 50:91–133
- Poesen J, Vandaele K, Van Wesemael B (1996) Contribution of gully erosion to sediment production on cultivated lands and rangelands. *IAHS Publ-Ser Proc Reports-Intern Assoc Hydrol Sci* 236:251–266
- Pourghasemi H, Yousefi S, Kornejady A, Cerda A (2017) Applying different new ensemble data mining techniques for Gully erosion mapping with Geographical Information Systems. *Sci Total Environ* 609:764–775.
- Pourghasemi HR, Sadhasivam N, Kariminejad N, Collins AL (2020) Gully erosion spatial modelling: role of machine learning algorithms in selection of the best controlling factors and modelling process. *Geosci Front* 11:2207–2219

- Puente C, Olague G, Trabucchi M, Arjona-Villicaña PD, Soubervielle-Montalvo C (2019) Synthesis of vegetation indices using genetic programming for soil erosion estimation. *Remote Sens* 11:156
- Rahmati O, Haghizadeh A, Pourghasemi HR, Noormohamadi F (2016) Gully erosion susceptibility mapping: the role of GIS-based bivariate statistical models and their comparison. *Nat Hazards* 82:1231–1258
- Rahmati O, Tahmasebipour N, Haghizadeh A, Pourghasemi HR, Feizizadeh B (2017) Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. *Geomorphology* 298:118–137. <https://doi.org/10.1016/j.geomorph.2017.09.006>
- Reichenbach P, Rossi M, Malamud BD, Mihir M, Guzzetti F (2018) A review of statistically-based landslide susceptibility models. *Earth-Sci Rev* 180:60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- Rodrigo Comino J, Quiquerez A, Follain S, Raclot D, Le Bissonnais Y, Casali J, Giménez R, Cerdà A, Keesstra SD, Brevik EC, Pereira P, Senciales JM, Seeger M, Ruiz Sinoga JD, Ries JB (2016) Soil erosion in sloping vineyards assessed by using botanical indicators and sediment collectors in the Ruwer-Mosel valley. *Agric Ecosyst Environ* 233:158–170. <https://doi.org/10.1016/j.agee.2016.09.009>
- Rodríguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sánchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J Photogramm Remote Sens* 67:93–104. <https://doi.org/10.1016/j.isprsjprs.2011.11.002>
- Saha S, Roy J, Arabameri A, Blaschke T, Tien Bui D (2020) Machine learning-based gully erosion susceptibility mapping: a case study of Eastern India. *Sensors* 20:1313
- Sahin EK (2020) Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences* 2:1308. <https://doi.org/10.1007/s42452-020-3060-1>
- 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 Degrad Dev* 29:3092–3103. <https://doi.org/10.1002/ldr.3058>
- Sansare DA, Mhaske S (2020) Natural hazard assessment and mapping using remote sensing and QGIS tools for Mumbai city, India. *Nat Hazards* 100:1117–1136
- Sarkar H, Soni S, Ahmad I, Verma MK (2020) Assessment of agricultural drought in Upper Seonath Sub-Basin of Chhattisgarh (India) using remote sensing and GIS-based indices. *J Indian Soc Remote Sens* 48:921–933. <https://doi.org/10.1007/s12524-020-01124-5>
- Senanayake S, Pradhan B, Huete A, Brennan J (2020) Assessing soil erosion hazards using land-use change and landslide frequency ratio method: a case study of Sabaragamuwa Province. *Sri Lanka Remote Sensing* 12:1483
- Swets JA (1988) Measuring the accuracy of diagnostic systems. *Science* 240:1285–1293
- Vasu NN, Lee S-R (2016) A hybrid feature selection algorithm integrating an extreme learning machine for landslide susceptibility modeling of Mt. Woomyeon. *South Korea Geomorphology* 263:50–70
- Vu Dinh T, Hoang N-D, Tran X-L (2021) Evaluation of different machine learning models for predicting soil erosion in tropical sloping lands of Northeast Vietnam. *Appl Environ Soil Sci* 2021:6665485. <https://doi.org/10.1155/2021/6665485>
- Vu DT, Tran X-L, Cao M-T, Tran TC, Hoang N-D (2020) Machine learning based soil erosion susceptibility prediction using social spider algorithm optimized multivariate adaptive regression spline. *Measurement* 164:108066. <https://doi.org/10.1016/j.measurement.2020.108066>
- Wischmeier WH, Smith DD (1978) Predicting rainfall erosion losses: a guide to conservation planning. Department of Agriculture, Science and Education Administration
- Yang K, Lu C (2018) Evaluation of land-use change effects on runoff and soil erosion of a hilly basin—the Yanhe River in the Chinese Loess Plateau. *Land Degrad Dev* 29:1211–1221. <https://doi.org/10.1002/ldr.2873>
- Yesilnacar E, Topal T (2005) Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Eng Geol* 79:251–266. <https://doi.org/10.1016/j.enggeo.2005.02.002>
- Zabihi M, Mirchooli F, Motevalli A, Darvishan AK, Pourghasemi HR, Zakeri MA, Sadighi F (2018) Spatial modelling of gully erosion in Mazandaran Province, northern Iran. *CATENA* 161:1–13
- Zhang Z, Jiang T, Li S, Yang Y (2018) Automated feature learning for nonlinear process monitoring—an approach using stacked denoising autoencoder and k-nearest neighbor rule. *J Process Control* 64:49–61. <https://doi.org/10.1016/j.jprocont.2018.02.004>