




Article

Spatial Modelling of Gully Erosion Using GIS and R Programming: A Comparison among Three Data Mining Algorithms

Alireza Arabameri ¹ , Biswajeet Pradhan ^{2,*} , Hamid Reza Pourghasemi ³ , Khalil Rezaei ⁴ and Norman Kerle ⁵

¹ Department of Geomorphology, Tarbiat Modares University, Tehran 36581-17994, Iran; alireza.ameri91@yahoo.com

² Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW 2007, Australia

³ Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz 71441-65186, Iran; hm_porghasemi@yahoo.com

⁴ Faculty of Earth Sciences, Kharazmi University, Tehran 14911-15719, Iran; kh.rezaei@gmail.com

⁵ Department of Earth Systems Analysis (ESA), Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7522 Enschede, The Netherlands; kerle@itc.nl

* Correspondence: biswajeet24@gmail.com or Biswajeet.Pradhan@uts.edu.au

Received: 13 July 2018; Accepted: 10 August 2018; Published: 14 August 2018



Abstract: Gully erosion triggers land degradation and restricts the use of land. This study assesses the spatial relationship between gully erosion (GE) and geo-environmental variables (GEVs) using Weights-of-Evidence (WoE) Bayes theory, and then applies three data mining methods—Random Forest (RF), boosted regression tree (BRT), and multivariate adaptive regression spline (MARS)—for gully erosion susceptibility mapping (GESM) in the Shahroud watershed, Iran. Gully locations were identified by extensive field surveys, and a total of 172 GE locations were mapped. Twelve gully-related GEVs: Elevation, slope degree, slope aspect, plan curvature, convergence index, topographic wetness index (TWI), lithology, land use/land cover (LU/LC), distance from rivers, distance from roads, drainage density, and NDVI were selected to model GE. The results of variables importance by RF and BRT models indicated that distance from road, elevation, and lithology had the highest effect on GE occurrence. The area under the curve (AUC) and seed cell area index (SCAI) methods were used to validate the three GE maps. The results showed that AUC for the three models varies from 0.911 to 0.927, whereas the RF model had a prediction accuracy of 0.927 as per SCAI values, when compared to the other models. The findings will be of help for planning and developing the studied region.

Keywords: gully erosion; environmental variables; data mining techniques; SCAI; GIS

1. Introduction

Today, reducing natural resources, especially soil and water, is one of the major problems and major threats to human life and is one of the most important environmental problems worldwide that has intensified in recent years, with increasing population and the alternation of human activities [1]. According to the data from United Nations research, the world's population is growing at a rate of 1.8% per year and it is expected to rise from 8 billion in 2025 to 9.4 billion in 2050 [2]. This increase in world population would demand the need for food, water, forage, and others, which consequently would add huge pressure on land exploitation, non-standard exploitation, and eventually lead to an increase in erosion rates [1,3]. Soil erosion is one of the factors that endangers water and soil [1]. Soil erosion by water, such as GE, is considered as a major cause of land degradation around the world [4,5]. It leads

to a range of problems, such as desertification, flooding and sediment deposition in reservoirs [6,7], the destructive effects on the ecosystem reducing soil fertility, and imposes huge economic costs [8]. GE is typically defined as a deep channel that has been eroded by concentrated water flow, removing surface soils and materials [9,10]. The amount of moisture and its changes as a result of the dry and wet seasons is a main parameter in creating cracks and grooves in fine-grained clay formations containing clay and silt, and ultimately developing rilled erosion and gullies [10]. The alternation of warm and dry seasons makes it possible to create cracks, in the formation of fine grains, in warm seasons with the drying of the land and the wilting of the vegetation, and these cracks at the time of the first sudden rainfall concentrate the runoff and therefore cause rill and GE to emerge [11]. GE occurs when the erosion of the water flow or the erodibility of the sediments or the formation of the area is higher than the geomorphological threshold of the area [11]. Mapping gully erosion systems is essential for implementing soil conservation measures [6]. GEVs that influence gully occurrence are rainfall, topography-derived factors such as elevation, slope degree, slope aspect, and plan curvature, lithology [12], soil properties [13], and LU/LC [14]. The distribution of precipitation affects the hydraulic flow and moisture content of the soil, and the erosion strength of the flow and soil resistance to erosion is different before and after erosion [11]. Generally, the amount and volume of flow are controlled by the topographic features of the area including slope, aspect, and drainage area of the area. Depth and morphology of the cross section of the gullies are controlled by soil erodibility features of the geological layers of the area. The characteristics of the region's soil affect the subsurface flow and the phenomenon of piping erosion, and the pipes cause a gully when their ceiling collapses [10].

Susceptibility maps of GE are essential for conservation of natural resources, and for evaluating the relationship among gully occurrence and relevant GEVs [12]. Several models have been applied to assess soil erosion and GE rate in a quantitative and qualitative way, such as the Universal Soil Loss Equation (USLE) [1,15], Erosion Potential Method, Modified Pacific Southwest Interagency Committee Model (MPSIAC) [16], Water Erosion Prediction Project (WEPP) [17], European Soil Erosion Model (EUROSEM) [18], Ephemeral Gully Erosion Model (EGEM) [19], and Chemicals, Runoff, and Erosion from Agricultural Management Systems (CREAMS) [20].

Within the soil conservation research field, the distribution of soil erosion is one of the primary sources of information. This is also relevant for GE; however, in the above mentioned methods, spatial distribution of gullies has not been addressed. Remote sensing-based methods to identify GE have been developed [21], including with RF machine learning, though they serve more to validate susceptibility models and to explain the actual erosion presence and distribution. In recent years, scientific research for susceptibility analysis of GE, and work on the statistical relationships between GEVs and the spatial distribution of gullies, have been addressed using various statistical and machine learning methods including bivariate statistics (BS) [1], weights-of-evidence (WoE) [13], index of entropy (IofE) [8], logistic regression (LR) [22–26], information value (IV) [24,25], random forest (RF) [27], bivariate statistical models [28,29], maximum entropy (ME) [30,31], frequency ratio (FR) [28], analytical hierarchy processes (AHP) [29], artificial neural network (ANN) [12,31], support vector machine (SVM) [31], and boosted regression trees (BRT) [12]. For this purpose, various GEVs such as topography (e.g., elevation, slope, aspect, plan curvature, profile curvature, slope length), lithology, land use, soil properties (e.g., soil texture, soil type, erosivity, soil water content), land use, climate (rainfall intensity, rainfall period, and spatial distribution of rainfall), infrastructures (road, bridge) and hydrology (e.g., TWI, SPI, drainage density) were used.

A comprehensive literature review shows that there are still dimensions that require further research, and that a large number of potentially useful methods have not yet been fully implemented to provide GE susceptibility maps. The main objectives of this study are: (i) To determine the relationship between gully occurrence and conditioning factors using Weights-of-Evidence Bayes theory, (ii) assessing the capability of RF, MARS, and BRT data mining/machine learning models to predict GE susceptibility; and (iii) validation of models using the AUC curve and SCAI methods. Study

of the research background showed that using MARS, BRT, and RF data mining models in GE zonation is very new. It will help managers in future planning to prevent human intervention in sensitive areas.

2. Materials and Methods

2.1. Study Area

The Shahroud watershed, with an area of about 848 km² and elevation range from 1084 to 2131 m a.s.l., is located in the northeastern part of Semnan Province, Iran (Figure 1). The study area receives an average rainfall of less than 250 mm has an arid and semi-arid climate [32]. Various types of lithological formations cover this watershed, and the landforms are mainly low level pediment fans and valley terrace deposits. The dominant land use is rangelands, but irrigation farming and bare lands are also present.

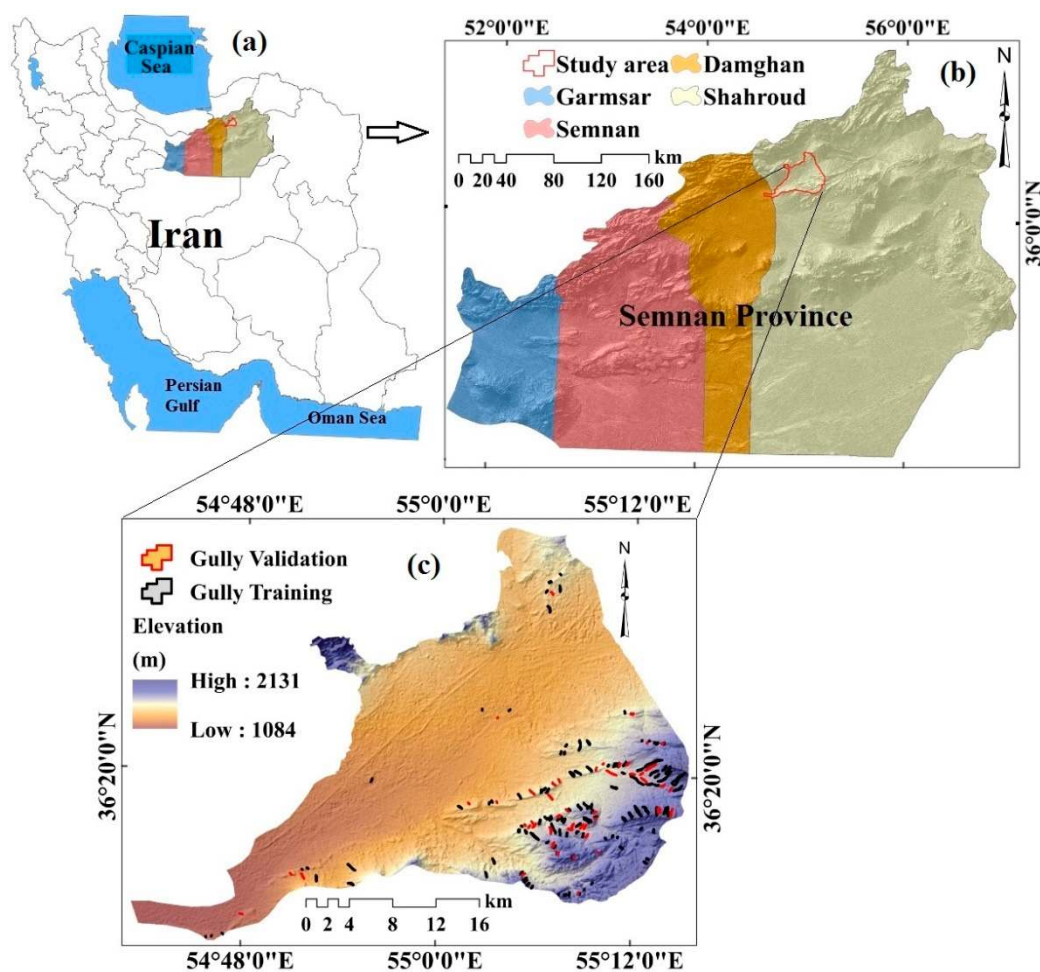


Figure 1. (a) Location of the Semnan provinces in Iran, (b) location of study area, and (c) gully erosion locations with the digital elevation model map of the Shahroud watershed.

2.2. Data and Method

Figure 2 shows the methodological approach applied to map GE susceptibility in the Shahroud watershed using BRT, MARS, and RF models. For preparing an accurate and reliable gully inventory map, extensive field surveys with a DGPS device were performed in the study area to determine the location of the Gullies [27,28]. Then, among 172 detected gully locations, randomly (70/30 ratio), 121 gully locations (70%) and 51 gully locations (30%) in the polygon format were used for training the

testing models [28]. The locations of training and testing gullies are shown in Figure 1. Interventionary studies involving animals or humans, and other studies require ethical approval must list the authority that provided approval and the corresponding ethical approval code.

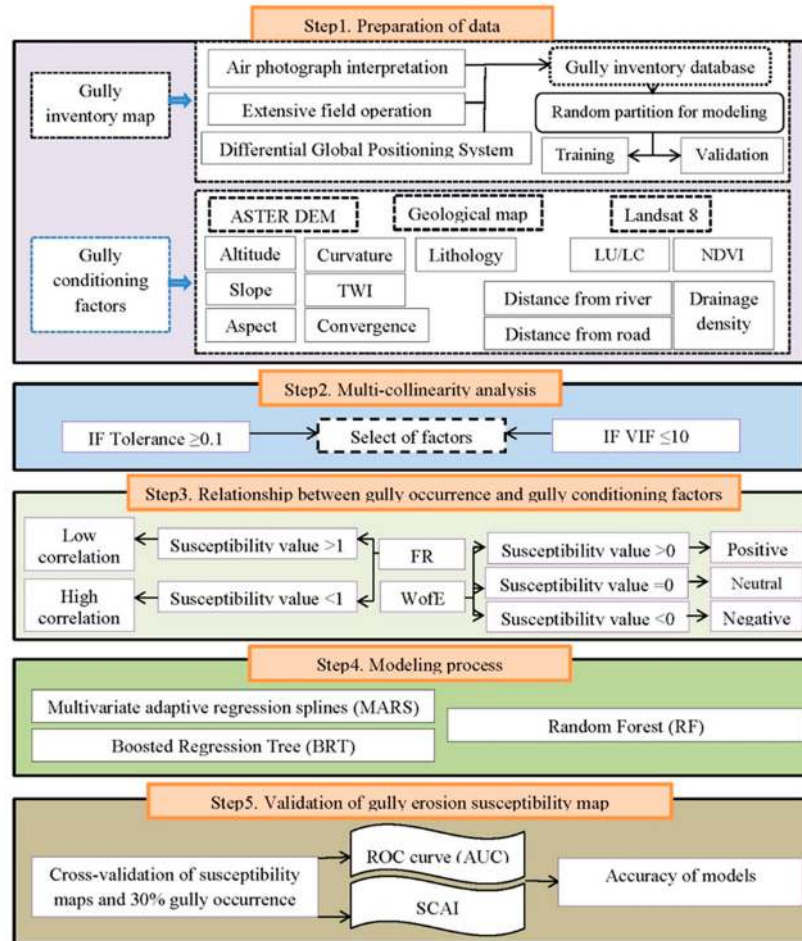


Figure 2. Flowchart of research methodology.

The tools used in present study are ArcGIS10.5, ENVI 4.8, SAGA-GIS 2.1.1, and a DGPS. The basic maps used were geological maps [33], at a scale of 1:100,000, topographic maps, at a scale of 1:50,000, satellite images acquired by Landsat8, and ASTER GDEM with spatial resolution of 30 m [34]. In this study, based on literature review [24,26,31] and local conditions of the study area, twelve factors were selected. Elevation map was divided into six classes: <1200 m, 1200–<1350 m, 1350–<1450 m, 1450–1600 m, and >1600 m (Figure 3a). Slope degree affects surface runoff [35], soil erosion, and pattern of drainage density. Slope degree map was classified into six classes [24,26]: <5°, 5–<10°, 10–<15°, 15–<20°, 20–<25°, 25–30°, and >30° (Figure 3b).

The aspect map was classified into nine classes (Figure 3c). Positive and negative values of plan curvatures define convexity and concavity of slope curvature, whereas zero is flat surface. The plan curvature map was divided into 3 categories: Concave, Flat, and Convex. The TWI indicator is important for identifying prone areas to GE [36]. TWI is calculated by Equation (1):

$$TWI = \ln\left(\frac{S}{\tan \alpha}\right) \tag{1}$$

TWI map of study area is divided into four classes [24,26,37] including <5, 5–<7.5, 7.5–11, and >11 (Figure 3e). The convergence index (CI) gives a measure of how flow in a cell diverges (convergence

index in negative and positive values) [38]. The CI map was prepared in SAGA-GIS 2.1.1 and divided into 3 classes: <0, 0–10, and >10 (Figure 3f). In this research, for the computation of the effect of drainage network and infrastructures on GE, the distance from rivers and roads was considered [14] and divided into four classes: <170 m, 170–<370 m, 370–650 m, and >650 m for rivers (Figure 3g) and <500 m, 500–<1500 m, 1500–3000 m, and >3000 m for roads (Figure 3h). The line density tool in ArcGIS 10.5 was used for calculating drainage density and then its map was divided into four categories: <1.4, 1.4–<2.4, 2.4–3.7, and >3.7 km/km² (Figure 3i). A geological map at a 1:100,000 scale was used to prepare the lithological unit layer. The lithological units were classified into ten categories based on their sensitivity to gully occurrence using expert knowledge method (Figure 3j and Table 1). The advantage of this method is it is easy to use, however this method has certain disadvantages, such as the possibility of a mistake by the expert.

Table 1. Lithology of the study area.

Code	Lithology	Geological Age
Murmg	Gypsiferous marl	Miocene
Qft2	Low level piedment fan and vally terrace deposits	Quaternary
Ku	Upper cretaceous, undifferentiated rocks	Cretaceous
Jd	Well—bedded to thin—bedded, greenish—grey argillaceous limestone with intercalations of calcareous shale (DALICHAJ FM)	Jurassic
PeEz	Reef-type limestone and gypsiferous marl (ZIARAT FM)	Paleocene-Eocene
PIQc	Fluvial conglomerate, Piedmont conglomerate and sandstone.	Pliocene-Quaternary
Jl	Light grey, thin—bedded to massive limestone (LAR FM)	Jurassic-Cretaceous
E2c	Conglomerate and sandstone	Eocene
PIQc	Fluvial conglomerate, Piedmont conglomerate and sandstone.	Pliocene-Quaternary
E1c	Pale-red, polygenic conglomerate and sandstone	Paleocene-Eocene

The LU/LC map was obtained using Landsat 8 images [39–41]. The main LU/LC types identified in the study area were range, irrigation farming, and bare lands (Figure 3k). The NDVI map was also produced using Landsat 8 images and classified into 3 categories: <0.11, 0.11–0.25, and >0.25 (Figure 3l).

For multi-collinearity checking, the tolerance (TOL) and variance inflation factor (VIF) were used. If during modeling there is collinearity among the variables, the accuracy of the model’s prediction decreases. Values of TOL and VIF were ≤0.1 and ≥10, respectively, indicating that multi-collinearity among parameters [28].

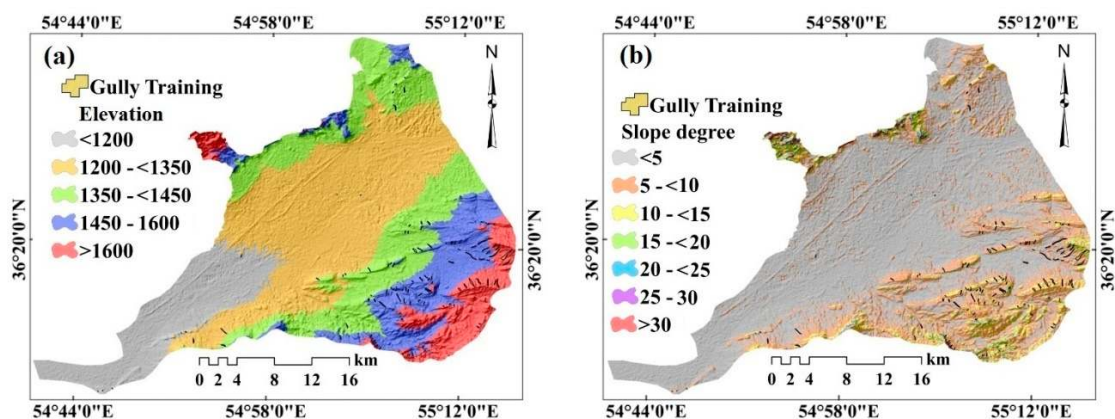


Figure 3. Cont.

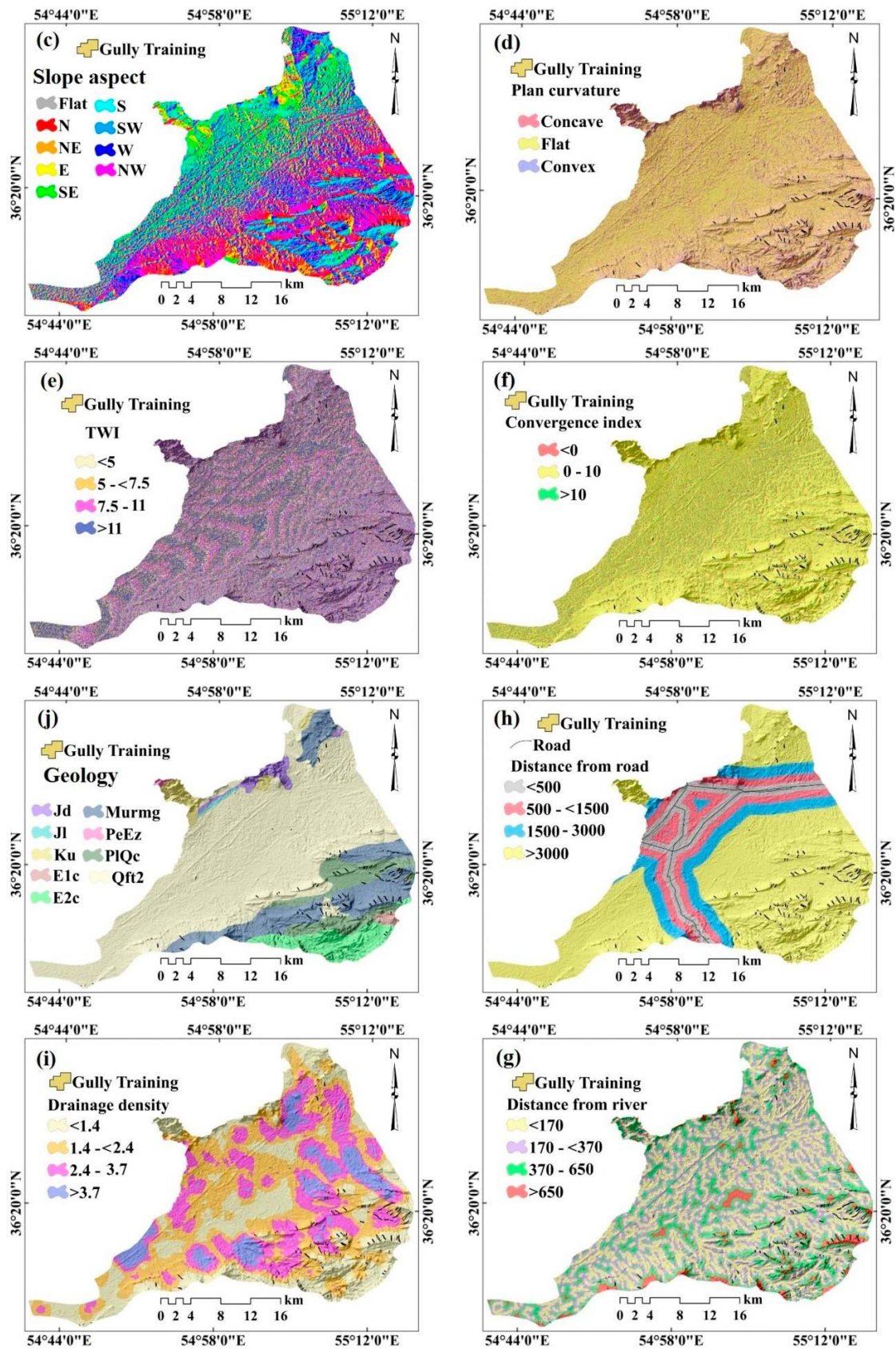


Figure 3. Cont.

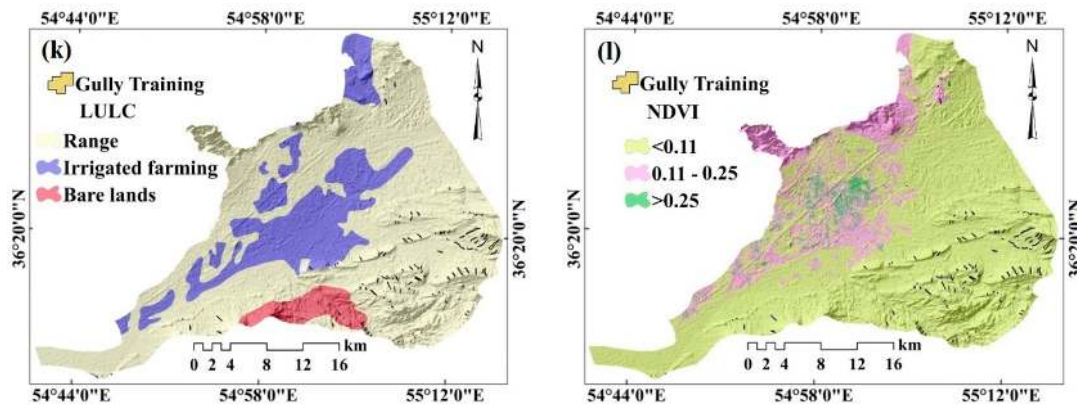


Figure 3. Gully erosion conditioning factors: (a) Elevation, (b) slope, (c) aspect, (d) plan curvature, (e) TWI, (f) convergence index, (j) geology, (h) distance from road, (i) drainage density, (g) distance from road, (k) land use/land cover (LU/LC), and (l) NDVI.

2.3. Gully Erosion Modelling

2.3.1. WoE Model

WoE is according to the Bayesian probability framework, to predict the significance of effective factor classes through a statistical approach [42–51]. In this method, the spatial relationship between GE areas and GEVs are identified. The WoE model is based on the calculation of positive (W^+) and negative (W^-) weights. This model computes the weight of each GEVs according to the existence or absence of the gully inventory [52] as follows:

$$W_i^+ = \ln \left(\frac{P\{B|L\}}{P\{B|\bar{L}\}} \right) \tag{2}$$

$$W_i^- = \ln \left(\frac{P\{\bar{B}|L\}}{P\{\bar{B}|\bar{L}\}} \right) \tag{3}$$

$$C = W^+ + W^- \tag{4}$$

$$S(C) = \sqrt{S^2(W^+) + S^2(W^-)} \tag{5}$$

$$S^2(W^+) = \frac{1}{N\{B \cap L\}} + \frac{1}{\{B \cap \bar{L}\}} \tag{6}$$

$$S^2(W^-) = \frac{1}{\{\bar{B} \cap L\}} + \frac{1}{\{\bar{B} \cap \bar{L}\}} \tag{7}$$

$$W = \left(\frac{C}{S(C)} \right) \tag{8}$$

where \ln is the natural log function and P is the probability, B and \bar{B} indicate the presence and absence of the gully geo-environmental factor, respectively, L is the presence of gully, and \bar{L} is the absence of a gully. W^+ and W^- are positive and negative weights, with W^+ indicating that a geo-environmental factor is present in the gully inventory. $S^2(W^+)$ is the variance of the W^+ and $S^2(W^-)$ is the variance of W^- . C indicates the overall association between GEVs and gully occurrence. $S(C)$ is the standard deviation of the contrast and W is final weight of each class factor.

2.3.2. RF Model

RF is a controlled learning method that uses multiple trees in the classification [21]. The RF algorithm, by replacing and continuously changing the factors that affect the target, leads to the creation

of a large number of decision trees, then all trees are combined to make decisions [21]. The RF consists of 3 user-defined parameters, which include: (1) The number of variables used in the construction of each tree, which expresses the power of each independent tree; (2) number of trees in RF; and (3) minimum number of nodes [43]. RF prediction power increases with the increasing strength of independent trees and reducing the correlation between them [44]. This algorithm uses 66% of the data to grow a tree called Bootstrap, and then a predictor variable is introduced randomly during the growing process to split a node in the tree construction. The remaining 33% of the data is also used to evaluate the fitted tree [45]. This process is repeated several times and the average of all predicted values is used as the final prediction of the algorithm. In this model, two factors, including the mean decrease accuracy and mean decrease Gini, are used to prioritize of each of the effective factors. The use of the mean decrease accuracy in comparison to mean decrease Gini index is more effective in determining the priority of effective factors, especially in the context of the relationship between environmental factors [46]. The RF analyses were carried out in R 3.3.1, using the “Randomforest” package [21].

2.3.3. BRT Model

BRT is one of several techniques that can help improve the performance of a single model by combining multiple models [47]. BRT uses two algorithms for modeling: Boosting and regression [48]. Boosting is a way to increase the accuracy of the model, and based on this, the construction, combination, and averaging of a large number of models are better and more accurate than an individual model on its own [49]. BRT overcomes the greatest weakness of the single decision tree, which is relatively weak in data processing. In BRT, only the first tree of all the training data is constructed, the next trees are grown on the remaining data from the tree before it; trees are not built on all data and only use a number of data [50]. The main idea in this method is to combine a set of weak predictor models (high predictive error) to arrive at strong prediction (low predictive error) [51]. Thus, in this study, BRT was used for GE spatial modeling using GMB (Generalized Boosted Models) and dismo (Species Distribution Modeling) packages in R 3.3.1.3.

2.3.4. MARS Model

The MARS model is a form of regression algorithm that was introduced by Friedman in 1991 to predict continuous numerical outputs [52]. This technique generates flexible regression models for predicting the target variable by means of dividing the problem space into intervals of input variables and processing a basic function in each interval.

The base function represents information in relation to one or more independent variables. A base function is defined in a given interval, in which the primary and end points are called knot. The knot is the key concept in this method and represents the point at which the behavior of the function changes at that point. The base function expresses the relationship between the input variables and the target variable and is in the form of $Max(0, X - c)$ or $Max(0, c - X)$, in which c is threshold value and X is the impute variable. The general form of the MARS model is as follows:

$$f(x) = \beta_0 + \sum_{j=1}^P \sum_{b=1}^B \left[\beta_{jb(+)} Max(0, x_i - H_{bj}) + \beta_{jb(-)} Max(0, H_{bj} - x_j) \right] \quad (9)$$

where x = input, $f(x)$ = output, P = predictor variables, and B = basis function. $Max(0, x - H)$ and $Max(0, H - x)$ are basis function and do not have to be present if their coefficients are 0. β_0 is constant, β_{jb} is the coefficient of the j th base function (BF), and the H values are called knots. The MARS model includes three main steps: (1) A forward stepwise algorithm to select certain spline basis functions, (2) a backward stepwise algorithm to delete base functions (BFs) until the best set is found, and (3) a smoothing method which gives the final MARS approximation a certain degree of continuity [52]. First, the MARS model estimates the value of the target function with a constant value, and then generates the best processing in the forward direction by searching among the variables.

The search process continues as long as all possible (BFs) are added to the model. At this stage, a very complex model with a large number of knots is obtained. In the next step, through the process of pruning backward, BFs that are less important are identified and deleted by using the generalized cross-validation (GCV) criterion [27]. GCV is a criterion for data fitting and eliminates a large number of BFs and reduces the probability of overfitting. This indicator is obtained by using Equation (10):

$$GCV = \frac{1}{2} \sum_{i=1}^N [y_i - f(x_i)]^2 / \left[1 - \frac{C(B)}{N} \right]^2 \quad (10)$$

where N is the number of data and $C(B)$ is a complexity penalty that increases with the number of BF in the model and which is defined as:

$$C(B) = (B - 1) + dB \quad (11)$$

where d is a penalty for each BF included into the model. This process continues until a complete review of all the basic functions, and at the end of the optimal model is obtained by applying base functions [52]. MARS model is an adaptive approach, since the selection of BFs and node locations is based on the data and type of purpose. After determining the optimal MARS model, the analysis of variance (ANOVA) method can be used to estimate the participation rate of each of the input variables and BFs. A detailed description of the MARS model can be found in [45]. MARS was run with R 3.3.1 and the “Earth” package [53].

2.4. Validation of GESMs Using Three Data Mining Models

A single criterion is not enough to select the best model among a large number of models, and judging about choosing a superior model by one criteria. It is not a suitable approach and it raises the chance of mistake in choosing the suitable model [27,37,54]. In this study, to compare the performance between data mining models and select the appropriate model, AUC and SCAI were used [28,36,55]. For calculating AUC, different thresholds were considered from 0 to 1, and for each threshold, the number of cells detected by the model as gully erosion was compared with observed gully erosion cells and positive and negative ratio indicators was calculated. After calculating these two indicators, we arranged them in ascending order, then they were plotted to calculate AUC. The AUC values range from 0.5 to 1. If a model cannot estimate the occurrence of an event better than a probable or random viewpoint, its AUC is 0.5 and therefore it will have the least accuracy, while if the AUC is equal to one, the model will have the highest accuracy [56,57]. The quantitative–qualitative relationship between AUC value and prediction accuracy can be classified as follows: 0.5–0.6, poor; 0.6–0.7, average; 0.7–0.8, good; 0.8–0.9, very good; and 0.9–1, excellent. SCAI is the ratio of the percentage area of each of the zoning classes to the percentage of gullies occurring on each class. Based on the SCAI indicator, the values of SCAI in very high sensitivity class are lower than very low sensitivity class.

3. Results

3.1. Multi-Collinearity Analysis

Multicollinearity is a condition of very high inter-correlations or inter-associations among the independent variables. Therefore, it is a type of disturbance in the data, and if present in the data, the statistical conclusions of the data may not be reliable [27]. A TOL value less than 0.1 or a VIF value larger than 10 indicates a high multicollinearity [56]. The outcomes of the coherent analysis among the 12 GEVs are shown in Table 2. The outcomes showed that the TOL and VIF of all GEVs were ≥ 0.1 and ≤ 5 , respectively. As a result, no multi-collinearity is seen among the GEVs.

Table 2. Multi-collinearity of effective factors using tolerance (TOL) and variance inflation factor (VIF).

Conditioning Factors	Collinearity Statistics	
	Tolerance	VIF
Constant Coefficient	-	-
Slope degree	0.998	1.002
Distance from road	0.672	1.489
Distance from river	0.323	3.094
Plan curvature	0.674	1.483
Lithology	0.945	1.058
LU	0.864	1.158
Drainage density	0.826	1.211
Elevation	0.920	1.087
Convergence index	0.666	1.503
Aspect	0.299	3.343
TWI	0.942	1.062
NDVI	0.941	1.063

3.2. Spatial Relationship Using WoE Model

The outcomes of WoE model are shown in Table 3. In elevation, the results of WoE indicate that there is a direct correlation between classes of elevation and GE, and with an increase in elevation, GE also increases. Therefore, the class of >1600 m with WoE 47.95 had the greatest impact on gully occurrence. The result of slope degree indicate that classes 5–<10 with WoE 34.96 had a strong relation with GEIM. For slope aspect, NE-facing slopes with a value of 19.46 show high probability of gully occurrence. In the case of plan curvature, among the three classes of concave, flat, and convex, the concave class had the highest value (78.04), and thus a positive correlation with GE. This result is in line with [11,50]. In TWI, the class of >11 has the strongest relationship with GE with the highest value (78.04). In the case of the convergence index, the class of 0–10 with values of 13.18 has a positive relation with gully occurrence. With respect to distance from river, class of >650 with value of 25.86 and regarding distance from road the class of >3000 m with values of 16.25 had the greatest effect on gully occurrence. For the drainage density factor, the class of <1.4 km/km² showed the highest value (14.23) and thus high correlation with gully occurrence. According to the lithology factor, Gypsiferous marl with greatest value (51.23) is more prone to GE than other lithology units. Concerning LU/LC, most gullies are located in the range land use type and this class with the highest value (21.02) has the strongest relationship with gully occurrence. In NDVI, results indicated that all gullies are located in the class of <0.11, showing that very low vegetation density renders slopes susceptible to GE.

Table 3. Relationship between conditioning factors and gully erosion using weights-of-evidence (WoE) model.

Factor	Class	Number of Pixels in Domain	Pixels of Gullies	Weights-of-Evidence (WoE)				
				C	S2 (w ⁺)	S2 (w ⁻)	S	W
1	<1200	144,200	21	-3.16	0.05	0.00	0.22	-14.41
	1200-<1350	348,463	89	-2.87	0.01	0.00	0.11	-26.60
	1350-<1450	230,735	502	-0.37	0.00	0.00	0.05	-7.52
	1450-1600	133,305	1057	0.33	0.00	0.00	0.00	0.00
	>1600	85,376	1074	1.88	0.00	0.00	0.04	47.95
2	<5	705,163	896	-1.83	0.00	0.00	0.04	-44.90
	5-<10	171,923	1259	1.34	0.00	0.00	0.04	34.96
	10-<15	38,854	397	1.38	0.00	0.00	0.05	25.36
	15-<20	13,936	121	1.13	0.01	0.00	0.09	12.13
	20-<25	6223	50	1.03	0.02	0.00	0.14	7.22
	25-30	3396	15	0.42	0.07	0.00	0.26	1.62
>30	2584	5	-0.41	0.20	0.00	0.45	-0.92	

Table 3. Cont.

Factor	Class	Number of Pixels in Domain	Pixels of Gullies	Weights-of-Evidence (WoE)				
				C	S2 (w ⁺)	S2 (w ⁻)	S	W
3	Flat	16,770	2	-3.22	0.50	0.00	0.71	-4.55
	N	72,345	208	-0.01	0.00	0.00	0.07	-0.19
	NE	79,383	209	-0.11	0.00	0.00	0.07	-1.52
	E	72,794	43	-1.66	0.02	0.00	0.15	-10.81
	SE	91,567	54	-1.68	0.02	0.00	0.14	-12.24
	S	114,731	246	-0.34	0.00	0.00	0.07	-5.13
	SW	119,263	396	0.15	0.00	0.00	0.05	2.81
	W	142,533	459	0.12	0.00	0.00	0.05	2.35
4	NW	232,693	1126	0.76	0.00	0.00	0.04	19.46
	Concave	54,613	1493	2.99	0.00	0.00	0.04	78.04
	Flat	574,180	749	-1.43	0.00	0.00	0.04	-33.33
5	Convex	313,286	501	-0.80	0.00	0.00	0.05	-16.27
	<7	24,272	63	0.00	0.00	0.00	0.02	-0.15
	5-<7.5	42,453	91	-0.32	0.01	0.00	0.11	-3.00
	7.5-11	89,328	225	-0.16	0.00	0.00	0.07	-2.29
6	>11	786,026	2364	0.21	0.00	0.00	0.06	3.87
	<0	75,370	26	-2.21	0.04	0.00	0.20	-11.21
	0-10	776,920	2534	0.95	0.00	0.00	0.07	13.18
7	>10	89,789	183	-0.39	0.01	0.00	0.08	-5.08
	<170	382,383	522	-1.07	0.00	0.00	0.05	-21.99
	170-<370	329,586	835	-0.21	0.00	0.00	0.04	-4.99
	370-650	179,671	914	0.75	0.00	0.00	0.04	18.62
8	>650	50,444	472	1.31	0.00	0.00	0.05	25.86
	<500	90,285	0	-0.10	0.00	0.00	0.02	-5.29
	500-<1500	102,453	0	-0.12	0.00	0.00	0.02	-6.05
	1500-3000	113,685	12	-3.44	0.08	0.00	0.29	-11.91
9	>3000	635,661	2731	4.70	0.00	0.08	0.29	16.25
	<1.4	277,251	1150	0.55	0.00	0.00	0.04	14.23
	1.4-<2.4	353,215	1000	-0.04	0.00	0.00	0.04	-1.12
	2.4-3.7	231,503	573	-0.21	0.00	0.00	0.05	-4.49
10	>3.7	80,115	20	-2.54	0.05	0.00	0.22	-11.32
	Murmg	144,412	1544	1.97	0.00	0.00	0.04	51.23
	Qft2	617,176	417	-2.36	0.00	0.00	0.05	-44.47
	Ku	23,972	0	-0.03	0.00	0.00	0.02	-1.35
	Jd	18,232	0	-0.02	0.00	0.00	0.02	-1.03
	PeEz	1449	0	0.00	0.00	0.00	0.02	-0.08
	PIQc	71,058	600	1.24	0.00	0.00	0.05	26.85
	Jl	3,274	0	0.00	0.00	0.00	0.02	-0.18
11	E2c	58,380	174	0.03	0.01	0.00	0.08	0.33
	E1c	4,820	8	-0.56	0.13	0.00	0.35	-1.60
	Range	708,879	2669	2.48	0.00	0.01	0.12	21.02
12	Farming	193,682	33	-3.06	0.03	0.00	0.18	-17.47
	Bare land	39,523	41	-1.06	0.02	0.00	0.16	-6.75
	<0.11	863,198	2743	0.09	0.00	0.00	0.02	4.59
12	0.11-0.25	56,745	0	-0.06	0.00	0.00	0.02	-3.26
	>0.25	22,140	0	-0.02	0.00	0.00	0.02	-1.25

1. Elevation, 2. Slope degree, 3. Slope aspect, 4. Plan curvature, 5. topographic wetness index (TWI), 6. Convergence index, 7. Distance from river, 8. Distance from road, 9. Drainage density, 10. Lithology, 11. land use (LU), 12. NDVI.

3.3. Applying RF Model

The outcomes of the confusion matrix for RF model are shown in Table 4. The result shows that the model predicted 2487 non-gully pixels as non-gullies and 256 non-gullies as gully. On the other hand, the RF model predicted 2677 gullies as gullies and 66 gullies as non-gullies. Moreover, the out-of-bag error (OOB) for RF was 5.82%. This means that the model has a precision of 94.18%, which expresses the excellent accuracy of the model in predicting gully erosion.

Table 4. Confusion matrix from the random forest (RF) model (0 = no gully, 1 = gully).

	0	1	Class Error
0	2487	256	0.0933
1	66	2677	0.0240

Prioritization results of RF are shown in Table 5 and Figure 4. The results show that the distance from roads (381.67, 22%), elevation (335.06, 19%), and lithology (234.21, 14%) had the highest values, followed by slope degree, drainage density, distance from river, NDVI, convergence index, slope aspect, TWI, plan curvature, and LU/LC.

Table 5. Relative influence of effective conditioning factors in the RF model.

Conditioning Factors	Weight
Distance from road	381.67
Elevation	335.06
Lithology	234.21
Slope degree	153.85
Drinage density	126.72
Distance from river	106.84
NDVI	105.26
Convergence index	73.97
Slope aspect	72.41
TWI	71.3
Plan curvature	42.43
LU	25.38

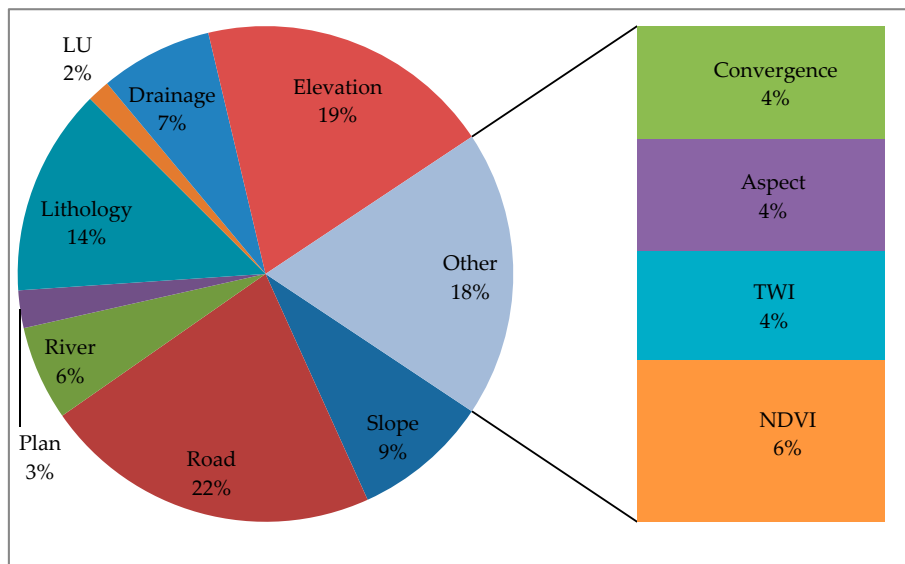


Figure 4. Relative influence of effective conditioning factors in the random forest (RF) model.

Finally, the GESM by the RF model was prepared in ArcGIS 10.5 and divided into five classes from very low to very high (Figure 5a), using a natural break classification [8]. According to the results, of the entire study area (847.87 km²), 525.97 km² (62.03%) are located in the very low susceptibility class, 148.28 km² (17.49%) in the low susceptibility, 79.42 km² (9.37%) in the moderate class, 56.34 km² (6.64%) in the high class, and 37.88 km² (4.47%) are located in the very high susceptibility class. Of the total area of GE (0.729 km²) in the study area, 0.86% (0.01 km²) are located in the very low susceptibility class,

5.67% (0.04 km²) in the low susceptibility, 14.80% (0.11 km²) in the moderate susceptibility, 21.95% (0.16 km²) in the high susceptibility, and 56.72% (0.41 km²) in the very high susceptibility classes.

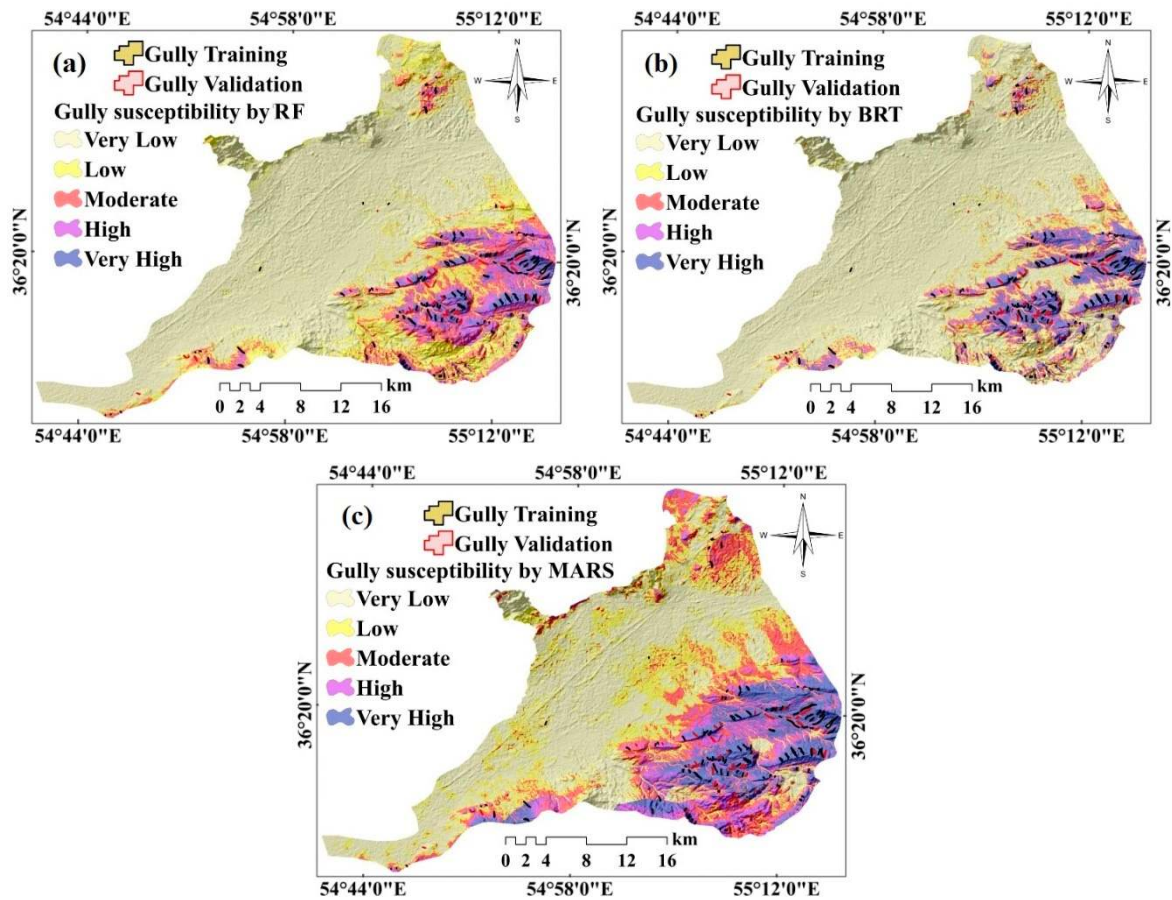


Figure 5. Gully erosion susceptibility maps using: (a) RF model, (b) BRT model, and (c) multivariate adaptive regression spline (MARS) model.

3.4. Applying BRT Model

The BRT model was used to reveal the spatial correlation between the existing GE and the GEVs in the study area. The results of the model are shown in Figure 6. They indicate that the factors distance from roads (31.1%), elevation (27.2%), and lithology (11%) had the highest importance on GE, mirroring the outcomes of the RF model, followed by slope degree (7%), drainage density (6.7%), distance from river (5.1%), slope aspect (3.8%), convergence index (2.4%), NDVI (2.2%), plan curvature (1.6%), TWI (1.6%), and LU/LC (0.3%). The gully susceptibility map by the BRT model was also prepared in ArcGIS 10.5 and divided into five classes of very low to very high (Figure 6c). The results of the GE susceptibility class by the BRT model covered 847.87 km² of the study area an area distribution in the very low, low, moderate, high, and very high susceptibility classes are 605.37 km², 88.38 km², 52.01 km², 34.13 km², and 67.98 km², and percentage distribution in the susceptibility classes of are 71.40, 10.42, 6.13, 4.03, and 8.02, respectively. Of the actual GE area of 0.729 km², 0.04 (5.55%), 0.03 (4.56%), 0.06 (8.26%), 0.08 (11.34%), and 0.51 km² (70.28%) are located in the very low to very high susceptibility classes, respectively.

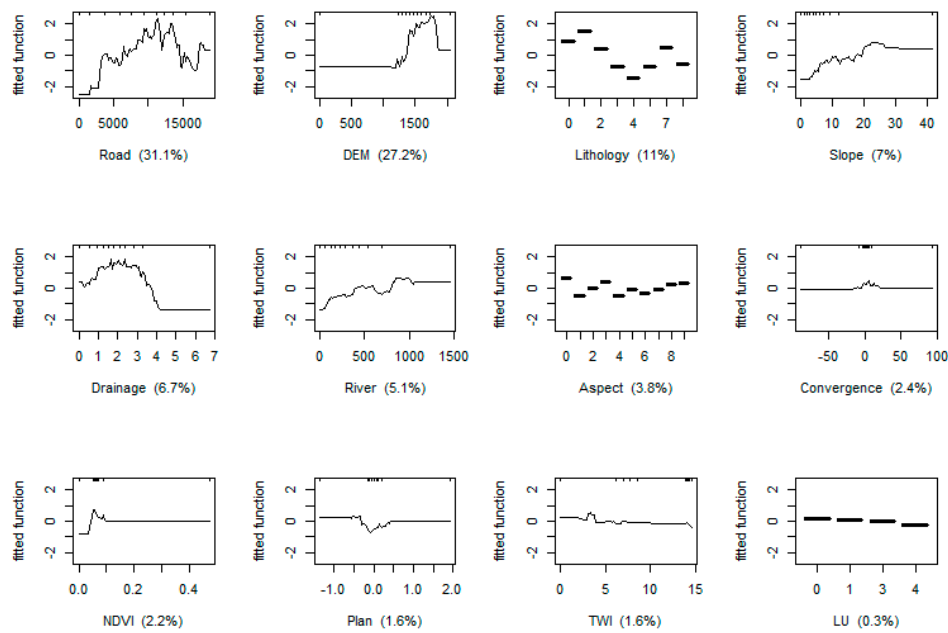


Figure 6. Relative influence of effective conditioning factors in boosted regression tree (BRT) model.

3.5. Applying MARS Model

The optimal MARS model included 28 terms, and the GCV was 0.157. MARS model provides the optimal model only by selecting the necessary parameters. In this research, nine GEVs including lithology, distance from road, distance from river, drainage density, elevation, aspect, convergence index, slope, and NDVI were used to construct the optimal model from the 12 GEVs. The GESM by the MARS model was implemented in ArcGIS 10.5 using Equation (12). According to Equation (12), distance from roads, elevation, and lithology were the most important variables. Values of GESM by MARS model varies from -9.8 to 7.3 . At first, GESM classified using quantile, equal interval, natural break, and geometrical interval classification techniques, then, by comparatively analyses of the distribution of training and validation gullies in high and very high classes, the natural break classification technique was most accurate. As a result, GESM by MARS were classified into very low (-9.86 – -6.24), low (-6.24 – -2.31), moderate (-2.3 – 0.04), high (0.04 – 0.38), and very high (0.38 – 7.32) gully erosion susceptibility zones by natural break classification technique (Figure 5c). The results indicate that 0.02 km^2 (2.10%) of GE in the study area are located in the very low susceptibility class, with 339.01 km^2 (39.98% of total study area) and 0.58 km^2 (79.16%) located in the very high susceptibility class with 105.50 km^2 (0.58%) (Table 6). In general, the results indicate that for all three models with increasing susceptibility (from very low to very high), the area of the respective classes decreased, while in contrast the areas of GE increased. These results is in line with Youssef et al. (2015).

Table 6. Area under the curve (AUC) values of RF, MARS, and BRT data mining models.

Models	AUC	Standard Error	Asymptotic Significant	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
RF	0.927	0.007	0.000	0.914	0.941
MARS	0.911	0.008	0.000	0.896	0.926
BRT	0.919	0.007	0.000	0.905	0.933

3.6. Validation of Models

The results of the validation of the models using the AUC curve and SCAI indicator are shown in Figure 7, and in Tables 6 and 7. The results show that the values of the AUC for the three models vary from 0.911 to 0.927, indicating very good prediction accuracy for all models, with RF resulting in the highest value. In addition, the SCAI values for the three models, RF (61.08–0.00), MARS (10.45–0.03), BRT (12.59–0.01), show that the RF model has higher SCAI values compared to the other models in the very low, low, and very high susceptibility classes (Figure 7). In spite of the high efficiency and accuracy of the RF model for GE sensitivity mapping, so far this model has not been used by the research community.

$$\begin{aligned}
 GESP_{MARS} = & 0.74 + (0.659 \times Lithology1) + (0.656 \times Lithology7) - 0.0001 \\
 & \times \max(0, 13445 - Distance\ from\ road) + 0.0001 \\
 & \times \max(0, Distance\ from\ road - 13445) - 0.0002 \\
 & \times \max(0, 2907.97 - Distance\ from\ River) - 0.087 \\
 & \times \max(0, 2.377 - Drainage\ density) - 0.106 \\
 & \times \max(0, Drainage\ density - 2.377) + 0.001 \times \max(0, 1793 \\
 & - Elevation) - 0.002 \times \max(0, Elevation - 1793) - 0.605 \\
 & \times Lithology7 \times Aspect4 - 0.0001 \times \max(0, 7355.32 \\
 & - Distance\ from\ road) \times Lithology1 - 0.0001 \\
 & \times \max(0, 11249.2 - Distance\ from\ road) \times Lithology7 \\
 & - 0.00002 \times \max(0, Distance\ from\ road - 11249.2) \\
 & \times Lithology7 + 0.0001 \times \max(0, 13445 \\
 & - Distance\ from\ road) \times Lithology10 - 0.005 \\
 & \times \max(0, 84.853 - Distance\ from\ River) \times Lithology1 \\
 & - 0.0003 \times \max(0, Distance\ from\ River - 84.853) \\
 & \times Lithology1 + 0.001 \times Lithology2 \times \max(0, Elevation \\
 & - 1249) - 0.001 \times Lithology2 \times \max(0, 1249 - Elevation) \\
 & - 0.019 \times Lithology7 \times \max(0, 0.772 - Convergence) - 23.54 \\
 & \times Lithology7 \times \max(0, NDVI - 0.055) - 22.23 \times Lithology7 \\
 & \times \max(0, 0.055 - NDVI) - 0.00001 \times \max(0, 7.65 - Slope) \\
 & \times \max(0, Distance\ from\ road - 13445) + 0.00001 \\
 & \times \max(0, Slope - 7.65) \times \max(0, Distance\ from\ road - 13445) \\
 & - 0.0001 \times \max(0, 8.186 - Slope) \times \max(0, 1793 - Elevation) \\
 & + 0.00004 \times \max(0, Slope - 8.19) \times \max(0, 1793 - Elevation) \\
 & - 0.0000001 \times \max(0, Distance\ from\ road - 3877.78) \\
 & \times \max(0, 907.97 - Distance\ from\ River) + 0.00000004 \\
 & \times \max(0, 8861.03 - Distance\ from\ road) \times \max(0, 907.97 \\
 & - Distance\ from\ River) + 0.0000001 \times \max(0, Road \\
 & - 8861.03) \times \max(0, 907.97 - Distance\ from\ River) - 0.001 \\
 & \times \max(0, Distance\ from\ road - 13445) \\
 & \times \max(0, Drainage\ density - 1.821) - 0.000001 \\
 & \times \max(0, Distance\ from\ road - 11435.4) \times \max(0, 1793 \\
 & - Elevation) - 0.000001 \times \max(0, Distance\ from\ road \\
 & - 13238.3) \times \max(0, 1793 - Elevation)
 \end{aligned} \tag{12}$$

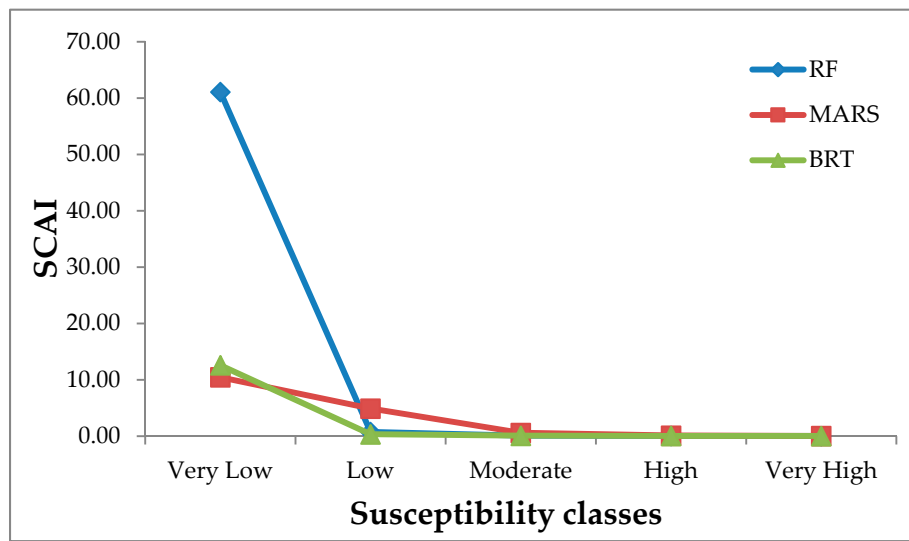


Figure 7. Seed cell area index (SCAI) values for different susceptibility classes in RF, MARS, and BRT data mining models.

Table 7. Seed cell area index (SCAI) values in RF, multivariate adaptive regression spline (MARS), and boosted regression tree (BRT) data mining models.

Model	Susceptibility Classes	Total Area of Classes		Gully in Classes		No Gully Area (km)	Seed Cell (%)	SCAI
		Area (km)	%	Area (km)	%			
RF	Very Low	525.97	62.03	0.01	0.86	525.96	0.01	61.08
	Low	148.28	17.49	0.04	5.67	148.24	0.24	0.74
	Moderate	79.42	9.37	0.11	14.80	79.31	1.15	0.08
	High	56.34	6.64	0.16	21.95	56.18	2.41	0.03
	Very High	37.88	4.47	0.41	56.72	37.46	9.27	0.00
MARS	Very Low	339.01	39.98	0.02	2.10	339.00	0.04	10.45
	Low	194.83	22.98	0.01	1.48	194.82	0.05	4.89
	Moderate	131.17	15.47	0.04	5.67	131.13	0.27	0.58
	High	77.35	9.12	0.08	11.59	77.26	0.93	0.10
	Very High	105.50	12.44	0.58	79.16	104.92	4.64	0.03
BRT	Very Low	605.37	71.40	0.04	5.55	605.33	0.06	12.59
	Low	88.38	10.42	0.03	4.56	88.34	0.32	0.33
	Moderate	52.01	6.13	0.06	8.26	51.95	0.98	0.06
	High	34.13	4.03	0.08	11.34	34.05	2.06	0.02
	Very High	67.98	8.02	0.51	70.28	67.46	6.40	0.01

4. Discussion

Determining effective parameters in GE and providing a GESM are the first steps in risk management. In regards to this, prediction of areas susceptible to erosion is associated with uncertainty, various models can be used to predict it accurately. Over the past decades, numerous statistical and empirical models have been developed to predict environmental hazards, such as GE, by various researchers around the world [12,14,28,30,31,45]. Due to some of the limitations of the aforementioned models such as time consuming, complexity, costly, and need a lot of data, in recent years data mining methods have been presented. Data mining is a process of discovery of relationships, patterns, and trends that consider the vast amount of information stored in databases with template recognition technology [51,58,59]. The most important applications of data mining are categorization, estimation, forecasting, group dependency, clustering, and descriptions. The results of data mining models show that in RF, BRT, and MARS mode, distance from roads had the highest impact in the occurrence of gully erosion in the study area. This result is in line with [10,49]. If the engineering measures

are not considered in site selection and construction of roads as anthropogenic structures in nature, they can act as a causative factor in environmental hazards such as landslide and gully erosion. The construction of roads in bare lands with erosion-sensitive formations has led to the expansion of gully erosion in the study area, so that the construction of a road without proper culverts causes disrupted of natural drainage and runoff concentrations, thus eroding the bare lands and resulting in the formation of a gully. The results of the validation of data mining models showed that the RF model more accurately predicted areas that are sensitive to gully erosion. These results are consistent with the results of [36,43,46,59], which introduced the RF model as a strong and high-performance model. One of the most widely used data mining methods is the RF model. The advantages of the RF method over other models is that this model can apply several input factors without eliminating any factors, and return a very small set of categories that support high prediction accuracy [6]. The classification accuracy of this model is affected by many factors such as the number, scale, type, and precision of input data. Thus, in the processing, the use of all suitable factors causes the accuracy of the model to increase. Compared with other models, RF has higher sufficiency to apply a very high number of datasets [6]. The RF model has the potential as a tool of spatial model for assessing environmental issues and environmental hazards. The RF model combines several tree algorithms to generate a repeated prediction of each phenomenon. This method can learn complicated patterns and consider the nonlinear relationship between explanatory variables and dependent variables. It can also incorporate and combine different types of data in the analysis, due to the lack of distribution of assumptions about the data used. This model can use and apply thousands of input variables without deleting one of them. This method is less sensitive to artificial neural networks, in case of noise data, and can better estimate the parameters [60]. The greatest advantages of RF model are high predictive accuracy, the ability to learn nonlinear relationships, the ability to determine the important variables in prediction, its nonparametric nature, and in dealing with distorted data, it works better than other algorithms for categorization. The main disadvantages of this algorithm include high memory occupation, hard and time-consuming in implementation for large datasets, high cost of pruning, high number of end nodes in case of overlap, and the accumulation of layers of errors in the case of the tree growing. [15,61] stated that the CART, BRT, and RF models showed better accuracy compared to bivariate and multivariate methods. Pourghasemi et al. concluded that the RF and maximum entropy (ME), models have high performance and precision in modeling [31]. Mojaddadi et al. showed that BRT, CART, and RF methods are suitable for modelling [55]. Chen et al. indicates that the MARS and RF models are good estimators for mapping [36]. Lai et al. indicated that the RF model has significant potential for weight determination on landslide modelling [62]. Kuhnert et al stated that RF with AUC = 97.0 is suitable for landslide susceptibility [27]. Lee et al stated that the prediction accuracy of RF model is high (90.8) and that this model had a high capability for landslide prediction [43]. They applied RF and boosted-tree models for spatial prediction of flood susceptibility in Seoul metropolitan city, Korea [43]. They stated that the RF model has better performance compared to boosted-tree. As a scientific achievement, the methodology framework used in this research has shown that the proper selection of effective variables in gully erosion, along with the use of modern data mining models and Geography Information System (GIS) technique, are able to successfully identify areas susceptible to gully erosion. The susceptibility map prepared using this methodology is a suitable tool for sustainable planning to protect the land against gully erosion processes. Therefore, this methodology can be used to assess gully erosion in other similar areas, especially in arid and semi-arid regions.

5. Conclusions

GE is one of the main processes causing soil degradation and there is a need to improve methods to predict susceptible areas and responsible environmental factors, to allow early intervention to prevent, limit, or reverse gully formation. The utility of three data mining models, RF, BRT, and MARS, to predict GE in the Shahroud watershed, Iran, was assessed. For this purpose, twelve causative

factors and 121 gully locations (70%) are used for applying the models. In addition, 51 gully locations (30%) are used for validation of models. The correlation between GE and conditioning factor classes was researched with a WoE Bayes theory. Distance to roads, elevation, and lithology were the key factors. Validation of the models showed that all three models have high accuracy for GE mapping. Data mining/machine learning methods have a unique ability and accuracy for GESM. The results also showed that the southwestern part of the study region has a high susceptibility to GE.

Therefore, it is recommended that the following suggestions should be made to prevent and reduce soil erosion and its subsequent risks in the Sharoud watershed: (1) Control of gullies by restoration of vegetation adaptable with the natural conditions of the area; (2) gully controlling by building dams that could prevent soil erosion by slowing down the flow of water and aggravation of sedimentation; (3) awareness of farmers by environmental officials of the region, in terms of the type and principles of proper cultivation and prevention of overgrazing and destruction of vegetation; (4) correction of land use based on natural ability and restrictions related to geomorphologic and physiographic soil characteristics of the area.

Author Contributions: Conceptualization, A.A., B.P., and H.R.P.; Data curation, A.A.; Formal analysis, A.A., and H.R.P.; Investigation, A.A., B.P., and H.R.P.; Methodology, B.P., A.A., H.R.P., and K.R.; Resources, B.P. and A.A.; Software, H.R.P., and A.A.; Supervision, B.P., N.K. and H.R.P.; Validation, H.R.P., and A.A.; Writing—original draft, A.A.; Writing—review and editing, B.P., A.A., H.R.P., N.K. and K.R.

Funding: This research was supported by the UTS under grant number 321740.2232335 and 321740.2232357.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Magliulo, P. Assessing the susceptibility to water-induced soil erosion using a geomorphological, bivariate statistics-based approach. *Environ. Earth Sci.* **2012**, *67*, 1801–1820. [CrossRef]
- UNEP. The Emissions Gap Report. United Nations Environment Programme (UNEP). Nairobi, 2017. Available online: www.unenvironment.org/resources/emissions-gap-report (accessed on 13 January 2018).
- Haregeweyn, N.; Tsunekawa, A.; Poesen, J.; Tsubo, M.; Meshesha, D.T.; Fenta, A.A.; Nyssen, J.; Adgo, E. Comprehensive assessment of soil erosion risk for better land use planning in river basins: Case study of the Upper Blue Nile River. *Sci. Total Environ.* **2017**, *574*, 95–108. [CrossRef] [PubMed]
- Nampak, H.; Pradhan, B.; Mojaddadi Rizeei, H.; Park, H.-J. Assessment of Land Cover and Land Use Change Impact on Soil Loss in a Tropical Catchment by Using Multi-Temporal SPOT-5 Satellite Images and RUSLE model. *Land Degrad. Dev.* **2018**. [CrossRef]
- Rizeei, H.M.; Saharkhiz, M.A.; Pradhan, B.; Ahmad, N. Soil erosion prediction based on land cover dynamics at the Semenyih watershed in Malaysia using LTM and USLE models. *Geocarto Int.* **2016**, *31*, 1158–1177. [CrossRef]
- Zhang, X.; Fan, J.; Liu, Q.; Xiong, D. The contribution of gully erosion to total sediment production in a small watershed in Southwest China. *Phys. Geogr.* **2018**, *39*, 246–263. [CrossRef]
- Mojaddadi, H.; Habibnejad, M.; Solaimani, K.; Ahmadi, M.; Hadian-Amri, M. An Investigation of Efficiency of Outlet Runoff Assessment. *J. Appl. Sci.* **2009**, *9*, 105–112.
- Zabihi, M.; Mirchooli, F.; Motevalli, A.; Darvishan, A.K.; Pourghasemi, H.R.; Zakeri, M.A.; Sadighi, F. Spatial modelling of gully erosion in Mazandaran Province, northern Iran. *Catena* **2018**, *161*, 1–13. [CrossRef]
- Kirkby, M.; Bracken, L. Gully processes and gully dynamics. *Earth Surf. Process. Landf. J. Br. Geomorphol. Res. Group* **2009**, *34*, 1841–1851. [CrossRef]
- Torri, D.; Poesen, J.; Borselli, L.; Bryan, R.; Rossi, M. Spatial variation of bed roughness in eroding rills and gullies. *Catena* **2012**, *90*, 76–86. [CrossRef]
- McCloskey, G.; Wasson, R.; Boggs, G.; Douglas, M. Timing and causes of gully erosion in the riparian zone of the semi-arid tropical Victoria River, Australia: Management implications. *Geomorphology* **2016**, *266*, 96–104. [CrossRef]
- Rahmati, O.; Tahmasebipour, N.; Haghizadeh, A.; Pourghasemi, H.R.; Feizizadeh, B. Evaluating the influence of geo-environmental factors on gully erosion in a semi-arid region of Iran: An integrated framework. *Sci. Total Environ.* **2017**, *579*, 913–927. [CrossRef] [PubMed]

13. Dube, F.; Nhapi, I.; Murwira, A.; Gumindoga, W.; Goldin, J.; Mashauri, D. Potential of weight of evidence modelling for gully erosion hazard assessment in Mbire District–Zimbabwe. *Phys. Chem. Earth Part A/B/C* **2014**, *67*, 145–152. [[CrossRef](#)]
14. Zakerinejad, R.; Maerker, M. An integrated assessment of soil erosion dynamics with special emphasis on gully erosion in the Mazayjan basin, southwestern Iran. *Nat. Hazards* **2015**, *79*, 25–50. [[CrossRef](#)]
15. Pham, T.G.; Degener, J.; Kappas, M. Integrated universal soil loss equation (USLE) and Geographical Information System (GIS) for soil erosion estimation in A Sap basin: Central Vietnam. *Int. Soil Water Conserv. Res.* **2018**, *6*, 99–110. [[CrossRef](#)]
16. Pournader, M.; Ahmadi, H.; Feiznia, S.; Karimi, H.; Peirovan, H.R. Spatial prediction of soil erosion susceptibility: An evaluation of the maximum entropy model. *Earth Sci. Inform.* **2018**, *11*, 389–401. [[CrossRef](#)]
17. Althuwaynee, O.F.; Pradhan, B.; Park, H.-J.; Lee, J.H. A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping. *Catena* **2014**, *114*, 21–36. [[CrossRef](#)]
18. Morgan, R.; Quinton, J.; Smith, R.; Govers, G.; Poesen, J.; Auerswald, K.; Chisci, G.; Torri, D.; Styczen, M. The European Soil Erosion Model (EUROSEM): A dynamic approach for predicting sediment transport from fields and small catchments. *Earth Surf. Process. Landf. J. Br. Geomorphol. Res. Group* **1998**, *23*, 527–544. [[CrossRef](#)]
19. Barber, M.; Mahler, R. Ephemeral gully erosion from agricultural regions in the Pacific Northwest, USA. *Ann. Wars. Univ. Life Sci.-SGGW. Land Reclam.* **2010**, *42*, 23–29. [[CrossRef](#)]
20. Leonard, R.; Knisel, W.; Still, D. GLEAMS: Groundwater loading effects of agricultural management systems. *Trans. ASAE* **1987**, *30*, 1403–1418. [[CrossRef](#)]
21. Liaw, A.; Breiman, W.M. Cutler’s Random Forests for Classification and Regression. Available online: <https://www.rdocumentation.org/packages/randomForest> (accessed on 1 April 2018).
22. Akgün, A.; Türk, N. Mapping erosion susceptibility by a multivariate statistical method: A case study from the Ayvalık region, NW Turkey. *Comput. Geosci.* **2011**, *37*, 1515–1524. [[CrossRef](#)]
23. Conoscenti, C.; Angileri, S.; Cappadonia, C.; Rotigliano, E.; Agnesi, V.; Märker, M. Gully erosion susceptibility assessment by means of GIS-based logistic regression: A case of Sicily (Italy). *Geomorphology* **2014**, *204*, 399–411. [[CrossRef](#)]
24. Conforti, M.; Aucelli, P.P.; Robustelli, G.; Scarciglia, F. Geomorphology and GIS analysis for mapping gully erosion susceptibility in the Turbolo stream catchment (Northern Calabria, Italy). *Nat. Hazards* **2011**, *56*, 881–898. [[CrossRef](#)]
25. Lucà, F.; Conforti, M.; Robustelli, G. Comparison of GIS-based gully susceptibility mapping using bivariate and multivariate statistics: Northern Calabria, South Italy. *Geomorphology* **2011**, *134*, 297–308. [[CrossRef](#)]
26. Meyer, A.; Martinez-Casasnovas, J. Prediction of existing gully erosion in vineyard parcels of the NE Spain: A logistic modelling approach. *Soil Tillage Res.* **1999**, *50*, 319–331. [[CrossRef](#)]
27. Kuhnert, P.M.; Henderson, A.K.; Bartley, R.; Herr, A. Incorporating uncertainty in gully erosion calculations using the random forests modelling approach. *Environmetrics* **2010**, *21*, 493–509. [[CrossRef](#)]
28. Rahmati, O.; Haghizadeh, A.; Pourghasemi, H.R.; Noormohamadi, F. Gully erosion susceptibility mapping: The role of GIS-based bivariate statistical models and their comparison. *Nat. Hazards* **2016**, *82*, 1231–1258. [[CrossRef](#)]
29. Svoray, T.; Michailov, E.; Cohen, A.; Rokah, L.; Sturm, A. Predicting gully initiation: Comparing data mining techniques, analytical hierarchy processes and the topographic threshold. *Earth Surf. Proc. Land.* **2012**, *37*, 607–619. [[CrossRef](#)]
30. Zakerinejad, R.; Märker, M. Prediction of Gully erosion susceptibilities using detailed terrain analysis and maximum entropy modeling: A case study in the Mazayejan Plain, Southwest Iran. *Geogr. Fis. Din. Quat.* **2014**, *37*, 67–76.
31. Pourghasemi, H.R.; Yousefi, S.; Kornejady, A.; Cerdà, A. Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. *Sci. Total Environ.* **2017**, *609*, 764–775. [[CrossRef](#)] [[PubMed](#)]
32. I.R. of Iran Meteorological Organization. 2012. Available online: <http://www.mazandaranmet.ir/> (accessed on 12 October 2017).

33. Geological Survey Department of Iran (GSDI). 2012. Available online: <http://www.mazandaranmet.ir/> (accessed on 12 October 2017).
34. Althuwaynee, O.F.; Pradhan, B.; Lee, S. Application of an evidential belief function model in landslide susceptibility mapping. *Comput. Geosci.* **2012**, *44*, 120–135. [[CrossRef](#)]
35. Rizeei, H.M.; Pradhan, B.; Saharkhiz, M.A. Surface runoff prediction regarding LULC and climate dynamics using coupled LTM, optimized ARIMA, and GIS-based SCS-CN models in tropical region. *Arab. J. Geosci.* **2018**, *11*, 53. [[CrossRef](#)]
36. Chen, W.; Xie, X.; Wang, J.; Pradhan, B.; Hong, H.; Bui, D.T.; Duan, Z.; Ma, J. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* **2017**, *151*, 147–160. [[CrossRef](#)]
37. Tehrany, M.S.; Pradhan, B.; Mansor, S.; Ahmad, N. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena* **2015**, *125*, 91–101. [[CrossRef](#)]
38. Claps, P.; Fiorentino, M.; Oliveto, G. Informational entropy of fractal river networks. *J. Hydrol.* **1996**, *187*, 145–156. [[CrossRef](#)]
39. Aal-Shamkhi, A.D.S.; Mojaddadi, H.; Pradhan, B.; Abdullahi, S. Extraction and modeling of urban sprawl development in Karbala City using VHR satellite imagery. In *Spatial Modeling and Assessment of Urban Form*; Springer: Berlin, Germany, 2017; pp. 281–296.
40. Abdullahi, S.; Pradhan, B.; Mojaddadi, H. City compactness: Assessing the influence of the growth of residential land use. *J. Urban Technol.* **2018**, *25*, 21–46. [[CrossRef](#)]
41. Rizeei, H.M.; Shafri, H.Z.; Mohamoud, M.A.; Pradhan, B.; Kalantar, B. Oil palm counting and age estimation from WorldView-3 imagery and LiDAR data using an integrated OBIA height model and regression analysis. *J. Sensors* **2018**, *2018*, 2536327. [[CrossRef](#)]
42. Xie, Z.; Chen, G.; Meng, X.; Zhang, Y.; Qiao, L.; Tan, L. A comparative study of landslide susceptibility mapping using weight of evidence, logistic regression and support vector machine and evaluated by SBAS-InSAR monitoring: Zhouqu to Wudu segment in Bailong River Basin, China. *Environ. Earth Sci.* **2017**, *76*, 313. [[CrossRef](#)]
43. Lee, S.; Kim, J.-C.; Jung, H.-S.; Lee, M.J.; Lee, S. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1185–1203. [[CrossRef](#)]
44. Cutler, D.R.; Edwards Jr, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. *Ecology* **2007**, *88*, 2783–2792. [[CrossRef](#)] [[PubMed](#)]
45. Simpson, G.L.; Birks, H.J.B. Statistical learning in palaeolimnology. In *Tracking Environmental Change Using Lake Sediments*; Springer: Berlin, Germany, 2012; pp. 249–327.
46. Nicodemus, K.K. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Brief. Bioinform.* **2011**, *12*, 369–373. [[CrossRef](#)] [[PubMed](#)]
47. Bui, D.T.; Bui, Q.-T.; Nguyen, Q.-P.; Pradhan, B.; Nampak, H.; Trinh, P.T. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. *Agr. For. Meteorol.* **2017**, *233*, 32–44.
48. Elith, J.; Leathwick, J.R.; Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **2008**, *77*, 802–813. [[CrossRef](#)] [[PubMed](#)]
49. Regmi, A.D.; Devkota, K.C.; Yoshida, K.; Pradhan, B.; Pourghasemi, H.R.; Kumamoto, T.; Akgun, A. Application of frequency ratio, statistical index, and weights-of-evidence models and their comparison in landslide susceptibility mapping in Central Nepal Himalaya. *Arab. J. Geosci.* **2014**, *7*, 725–742. [[CrossRef](#)]
50. Aertsen, W.; Kint, V.; Van Orshoven, J.; Özkan, K.; Muys, B. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecol. Model.* **2010**, *221*, 1119–1130. [[CrossRef](#)]
51. Krishnaiah, V.; Narsimha, G.; Chandra, N.S. Heart disease prediction system using data mining techniques and intelligent fuzzy approach: A review. *Heart Dis.* **2016**, *136*, 43–51. [[CrossRef](#)]
52. Oh, H.-J.; Pradhan, B. Application of a neuro-fuzzy model to landslide-susceptibility mapping for shallow landslides in a tropical hilly area. *Comput. Geosci.* **2011**, *37*, 1264–1276. [[CrossRef](#)]
53. Torgo, L. *Data Mining with R: Learning with Case Studies*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2016.

54. Umar, Z.; Pradhan, B.; Ahmad, A.; Jebur, M.N.; Tehrany, M.S. Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio and logistic regression models in West Sumatera Province, Indonesia. *Catena* **2014**, *118*, 124–135. [[CrossRef](#)]
55. Mojaddadi, H.; Pradhan, B.; Nampak, H.; Ahmad, N.; Ghazali, A.H.B. Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1080–1102. [[CrossRef](#)]
56. Pourghasemi, H.R.; Beheshtirad, M.; Pradhan, B. A comparative assessment of prediction capabilities of modified analytical hierarchy process (M-AHP) and Mamdani fuzzy logic models using Netcad-GIS for forest fire susceptibility mapping. *Geomat. Nat. Hazards Risk* **2016**, *7*, 861–885. [[CrossRef](#)]
57. Hong, H.; Naghibi, S.A.; Dashtpajardi, M.M.; Pourghasemi, H.R.; Chen, W. A comparative assessment between linear and quadratic discriminant analyses (LDA-QDA) with frequency ratio and weights-of-evidence models for forest fire susceptibility mapping in China. *Arab. J. Geosci.* **2017**, *10*, 167. [[CrossRef](#)]
58. Mezaal, M.R.; Pradhan, B.; Shafri, H.; Mojaddadi, H.; Yusoff, Z. Optimized Hierarchical Rule-Based Classification for Differentiating Shallow and Deep-Seated Landslide Using High-Resolution LiDAR Data. In *Global Civil Engineering Conference*; Springer: Berlin, Germany, 2017.
59. Rizeei, H.M.; Pradhan, B.; Saharkhiz, M.A. An integrated fluvial and flash pluvial model using 2D high-resolution sub-grid and particle swarm optimization-based random forest approaches in GIS. *Complex Intell. Syst.* **2018**, 1–20. [[CrossRef](#)]
60. Kantardzic, M. *Data mining: Concepts, Models, Methods, and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
61. Pham, B.T.; Pradhan, B.; Bui, D.T.; Prakash, I.; Dholakia, M. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environ. Model. Softw.* **2016**, *84*, 240–250. [[CrossRef](#)]
62. Lai, C.; Chen, X.; Wang, Z.; Xu, C.-Y.; Yang, B. Rainfall-induced landslide susceptibility assessment using random forest weight at basin scale. *Hydrol. Res.* **2017**, nh2017044. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).