water

MDPI

Article

Groundwater Potential Mapping Using an Integrated Ensemble of Three Bivariate Statistical Models with Random Forest and Logistic Model Tree Models

S. Vahid Razavi-Termeh [1,†], Abolghasem Sadeghi-Niaraki [1,2,*,†] and Soo-Mi Choi [2]

[1] Geoinformation Tech, Center of Excellence, Faculty of Geomatics, K.N. Toosi University of Technology, Tehran 19697, Iran
[2] Department of Computer Science and Engineering, Sejong University, Seoul 143-747, Korea
* Correspondence: a.sadeghi.ni@gmail.com
† These authors contributed equally to this work.

Received: 12 June 2019; Accepted: 26 July 2019; Published: 31 July 2019

Abstract: In the future, groundwater will be the major source of water for agriculture, drinking and food production as a result of global climate change. With increasing population growth, demand for groundwater has increased. Therefore, sustainable groundwater storage management has become a major challenge. This study introduces a new ensemble data mining approach with bivariate statistical models, using FR (frequency ratio), CF (certainty factor), EBF (evidential belief function), RF (random forest) and LMT (logistic model tree) to prepare a groundwater potential map (GPM) for the Booshehr plain. In the first step, 339 wells were chosen and randomly split into two groups with groundwater yields above 11 m^3/h. A total of 238 wells (70%) were used for model training, and 101 wells (30%) were used for model validation. Then, 15 effective factors, including topographic and hydrologic factors, were selected for the modeling. The accuracy of the groundwater potential maps was determined using the ROC (receiver operating characteristic) curve and the AUC (area under the curve). The results show that the AUC obtained using the CF-RF, EBF-RF, FR-RF, CF-LMT, EBF-LMT and FR-LMT methods were 0.927, 0.924, 0.917, 0.906, 0.885 and 0.83, respectively. Therefore, it can be inferred that the ensemble of bivariate statistic and data mining models can improve the effectiveness of the methods in developing a groundwater potential map.

Keywords: groundwater potential mapping (GPM); bivariate statistic models; data mining models; GIS; hybrid model

1. Introduction

In recent decades, in many countries, including Iran, due to population growth and industrialization, groundwater has been identified as one of the greatest natural resources [1,2]. It provides about 50% of the water needed for drinking, 40% of the water needed for industry and 20% of the water for agriculture [3,4]. Among the advantages of groundwater versus surface water is it is naturally stored, does not occupy a large area, is safe from evapotranspiration, is less impacted by pollution and sudden drought and has the ability to be used in all seasons [5]. Considering that the use of groundwater for various purposes is increasing and most of the aquifers and groundwater water reservoirs are over-exploited, identifying areas with varying groundwater potential is important [6]. Groundwater storage potential here relates to the maximum amount of permanent storage in aquifers [7]. This information can play a major role in decision making in the regions. Considering the fact that developing countries, such as Iran, face a lot of restrictions in access to hydrological information, it is essential to identify the current status of the groundwater system [8].

Traditional methods of groundwater exploration such as drilling, geophysical and geological methods, require high costs and a lot of time and human resources [4,9]. GIS and RS (remote sensing) methods are very effective in preparing a groundwater potential map (GPM) and are able to improve the accuracy and speed of groundwater studies [10]. The GIS-based GPM was developed using various methods such as frequency ratio (FR) [5], certainty factor (CF) [11], evidential belief function (EBF) [12,13], logistic regression (LR) [14,15], weight of evidence (WOE) [16] and entropy [17]. Data mining algorithms with the recent advances in IT and big data, such as random forest (RF) [18,19], logistic model tree (LMT) [20], DT (decision trees) [2], CART (classification and regression trees) [17], ANN (artificial neural networks) [21], SVM (support vector machines) [22] and an ensemble of metaheuristic algorithms with an ANFIS (adaptive neuro-fuzzy inference system) [23] are used widely in GPM.

Groundwater data analysis requires strong and flexible analytical methods that can control non-linear relationships, interactions and lost information [24]. Furthermore, understanding and presenting the results by these methods should be simple and easily interpretable [25]. Although DT is easy for classification and interpretation, a major weakness is that the fitted model has a large variance and this makes it difficult to interpret the classification [2]. To overcome this problem, several solutions have been suggested, all of which need fitting of various trees to data and averaging the predictions from trees [26]. RF and LMT models are one of the most powerful methods in this field.

One of the important issues in data mining algorithms is preprocessing and preparing data and selecting the best input data for the algorithms to increase accuracy [27,28]. So far, few studies have used the ensemble of bivariate statistical models and data mining algorithms to provide GPM [29]. This research also aims to prepare a GPM using a set of three bivariate statistical models (FR, EBF and CF) with RF and LMT models and choosing the best hybrid model for the Booshehr plain.

2. Material and Methods

To provide the GPM, this study was conducted in five steps. At first, the characteristics of the studied area were described and the existing wells were identified. In the second step, the required data were collected and the spatial database for effective criteria were created. A well distribution map and fifteen factors, including altitude, slope angle, slope aspect, plan curvature, profile curvature, slope length, topographic wetness index (TWI), rainfall, distance from river, distance from fault, drainage density, fault density, lithology, land use and soil, were chosen and ready for modeling. In the third step, the spatial relationship was calculated using FR, CF and EBF models between existing wells and effective criteria. In the fourth step, the weights from the three statistical models were considered as input values for the RF and LMT models, and the GPM was then prepared. In the fifth step, the GPM was validated using the ROC and AUC and the best model was finally chosen. Figure 1 shows a summary of the research steps.

2.1. Study Area

The Booshehr plain lies between the east longitude of 51°20′ and 52°10′ and between the north latitude of 27°50′ and 28°30′, with an area of 2696 km^2 in south-eastern Iran (Figure 1). A large part of the Booshehr plain has low altitude. The amount of altitude of the Booshehr plain is between 3 and 1490 m above sea level. In this region, the average annual temperature is 24 °C and the highest temperature is 50 °C in the summer season. The rainfall is low in the Booshehr plain, with an average of less than 255 mm. Most rain occurs between November and May. Rainfall in spring and autumn is short but intense, while in winter it is irregular and sparse. The average annual humidity in this region is above 71%. Figure 2 shows the study area. Table 1 presents the parameters for the regional lithology unit.

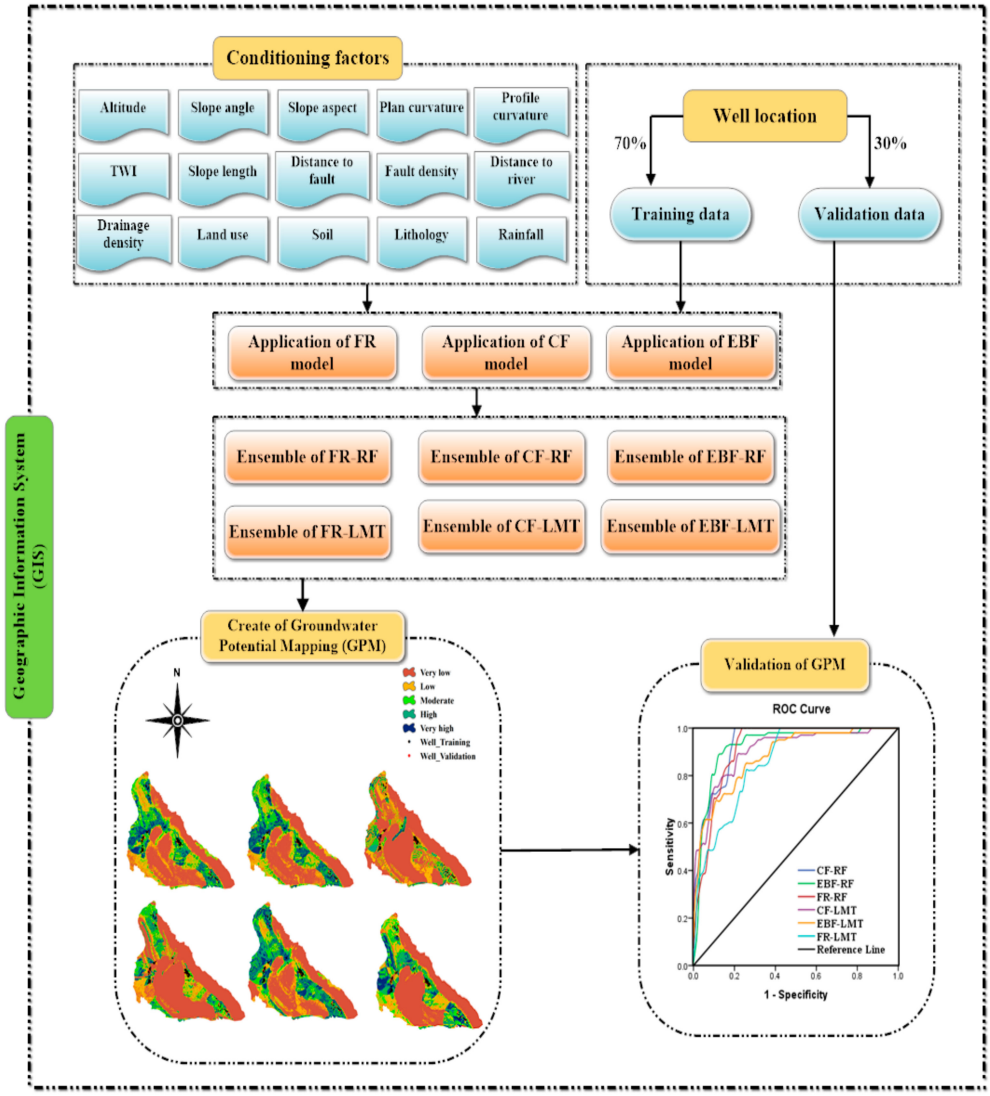

Figure 1. Flowchart of studies.

2.2. Well Inventory

Information about the groundwater wells was provided by the water resource administration of Booshehr Province. Based on previous studies and water management reports, only data of high potential groundwater wells (≥11 m^3/h) [9,30] and an average pH and electric conductivity (EC) of 6.9 and 495 µmhos/cm, respectively, were used. These wells were divided randomly into two training and testing datasets in this study. A total of 70% of the places (238 wells) were regarded as training and the rest of the wells (30% (101 wells)) were regarded as validation. The well places are shown in Figure 2.

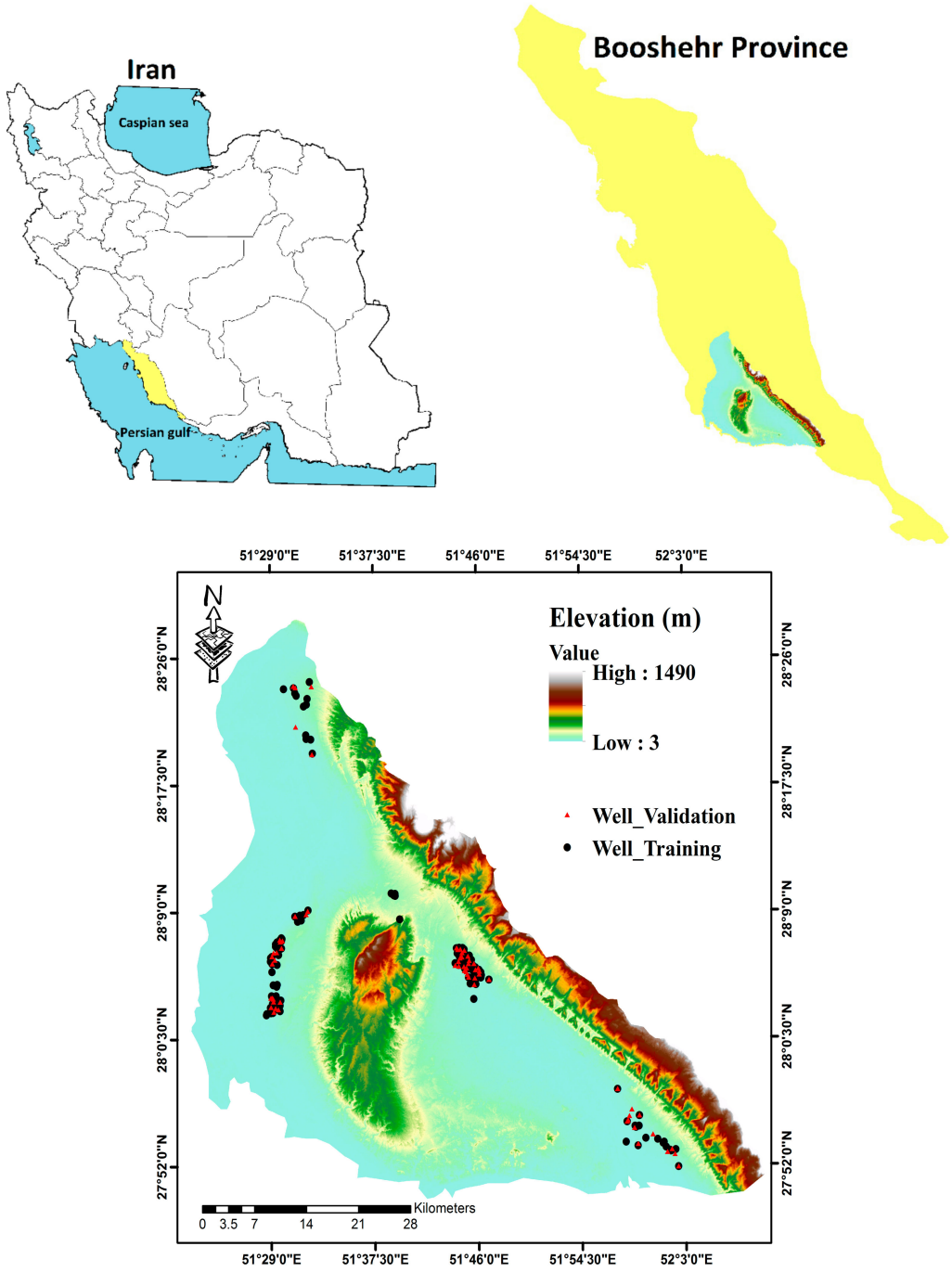

Figure 2. Location of the study area.

2.3. Conditioning Factors

2.3.1. Topographic Parameters

The ASTER digital elevation model (DEM) was downloaded in a spatial resolution of 30 m × 30 m (https://gdex.cr.usgs.gov/gdex/). Topographic parameters of the study area, consisting of altitude, profile curvature, slope length, slope angle, plan curvature and slope aspect, were obtained from the DEM. These layers were created in ArcGIS®10.3 and SAGA GIS®2.1.2 software. The classification of all maps was based on the natural break technique as well as the characteristics of the region and previous research. Various altitudes create different climatic conditions, creating different types of vegetation and soil [1].

The altitude map was divided into five classes (<108, 108–287, 287–535, 553–851 and >851 m) (Figure 3a). The slope angle mainly controls the process of feeding groundwater, infiltration and runoff, as well as the speed of groundwater movement [13]. The slope angle layer is categorized into five classes and includes 0°–6°, 6°–14°, 14°–24°, 24°–39° and >39° (Figure 3b).

The slope aspect is influenced by rainfall and the physiographical process and influences the amount of precipitation and vegetation type [29]. This criterion was grouped into nine categories (Figure 3c). Slope length, which is a function of catchment area and slope steepness, was measured by Equation (1) [31]:

$$LS = \left(\frac{B_s}{22.13}\right)^{0.6}\left(\frac{\sin\alpha}{0.0896}\right)^{1.3} \quad (1)$$

where LS is slope length, B_s is the catchment area (m^2) and α is the slope angle. Slope length was classified into five classes (0–10, 10–20, 20–30, 30–40 and >40 m) (Figure 3d).

The plan curvature influences flow convergence and divergence, and the profile curvature is consistent with the maximum slope aspect direction and mainly affects the surface flow velocity [32]. The plan curvature and profile curvature layers were classified into five classes respectively (<−2.22, −2.22–0.8, 0.8–0.4, 0.4–2.2 and >2.2 100/m) and (<−3.4, −3.4–1, −1–0.4, −0.4–2.8 and >2.8 100/m) (Figure 3e,f).

2.3.2. Hydrological Parameters

In hydrogeological systems, hydrological parameters such as distance to river, TWI and drainage density play an important role. In soil moisture, slope stability, groundwater flow and the TWI plays an important role [33]. TWI represents topographic control over hydrological processes. TWI was calculated according to Equation (2):

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

where A_s is the area of the cumulative upslope and α is the slope angle in radians. This factor was classified into five classes: <2.92, 2.92–3.84, 3.84–4.69, 4.69–6.57 and >6.57 (Figure 3g).

An area's drainage system relies on the nature and composition of geological formations, ability to absorb soil, permeability and slope [34]. High drainage density increases surface runoff and decreases infiltration. The high drainage density areas are not suitable for groundwater resource production [35]. Distance to river is divided to seven classes: <100, 100–200, 200–500, 500–1000, 1000–1500, 1500–2000 and >2000 m (Figure 3h). Drainage density was classified into five classes: <0.13, 0.13–0.27, 0.27–0.4, 0.4–0.58 and >0.58 (km/km^2) (Figure 3i).

2.3.3. Geological Parameters

Lithology influences groundwater potential through hydraulic conductivity [35]. The lithology layer was prepared from the Iran Geology Survey (1997) (scale 1:100,000) and was used to construct a lithological map in ArcGIS10.3. The lithology map in the Booshehr plain has 11 types, including Qft2, MuPlaj, Plbk, Mmn, Mgs, Eoas-ja, KEpd-gu, Kbgp, JKkgp, OMr and Pc-ch (Figure 3j). The fault mainly controls the distribution of spatial networks and the accumulation of groundwater [36]. The fault layer in the plain of Booshehr was determined at a scale of 1:100,000 from the geological map of the province of Booshehr and distance from fault was created and classified into seven classes including <100, 100–200, 200–500, 500–1000, 1000–2000, 2000–5000 and >5000 m (Figure 3k). The fault density was also calculated and divided into five classes: <0.03, 0.03–0.09, 0.09–0.13, 0.13–0.19 and >0.19 km/km^2 (Figure 3l).

2.3.4. Climate Parameters

Rain is the climate parameter that most influences groundwater recharge. Rainfall is very important for assessing water flow into the basin area and for understanding the nutritional status of the basin [18]. The rainfall map was prepared from statistics over a 30-year period collected from four

stations in the Booshehr plain as well as the neighboring basin using the Kriging interpolation method in ArcGIS®10.3 software. The rainfall layer was divided into five classes: <247, 247–264, 264–281, 281–297 and >297 mm (Figure 3m).

2.3.5. Ecological Parameters

The most significant ecological parameters are land use and soil. Land use directly and indirectly effects on permeability, runoff and evapotranspiration [37]. The land-use map was prepared on a scale of 1:100,000 from the Natural Resources Organization of Booshehr Province. The land-use map was divided into eleven different classes: Mangrove forest, forest, urban, agriculture, salt land, very low forest, poor range land, sand dune, moderate range land, afforestation, rock and water body (Figure 3n). One of the major variables for the generation and accumulation of surface runoff and subsurface runoff is the soil type [38]. A soil map was prepared by the Regional Water Organization of the Province of Booshehr and categorized into three groups (Entisols/Aridisoils, rock outcrops/Entisols and badlands) (Figure 3o).

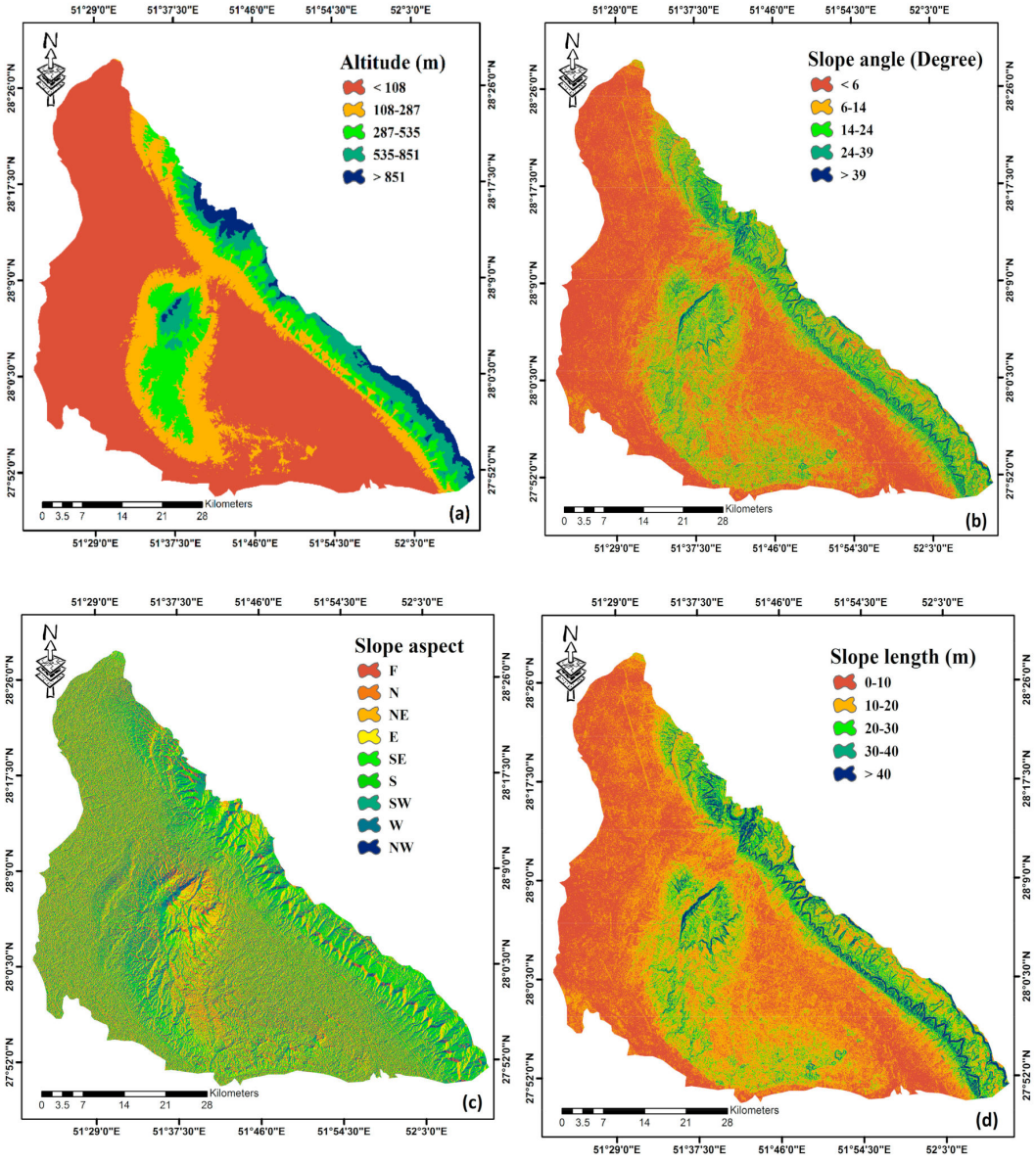

Figure 3. *Cont.*

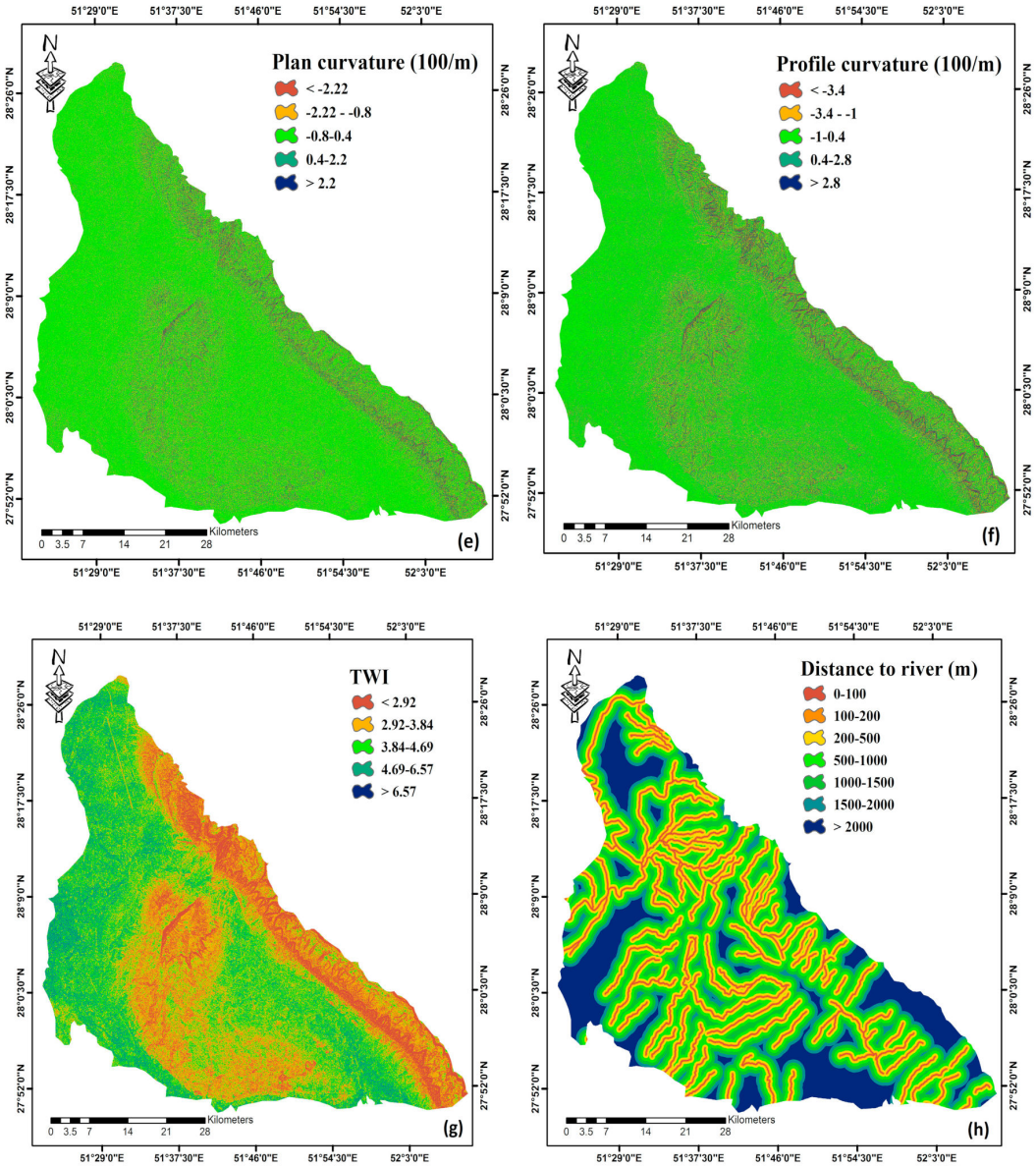

Figure 3. *Cont.*

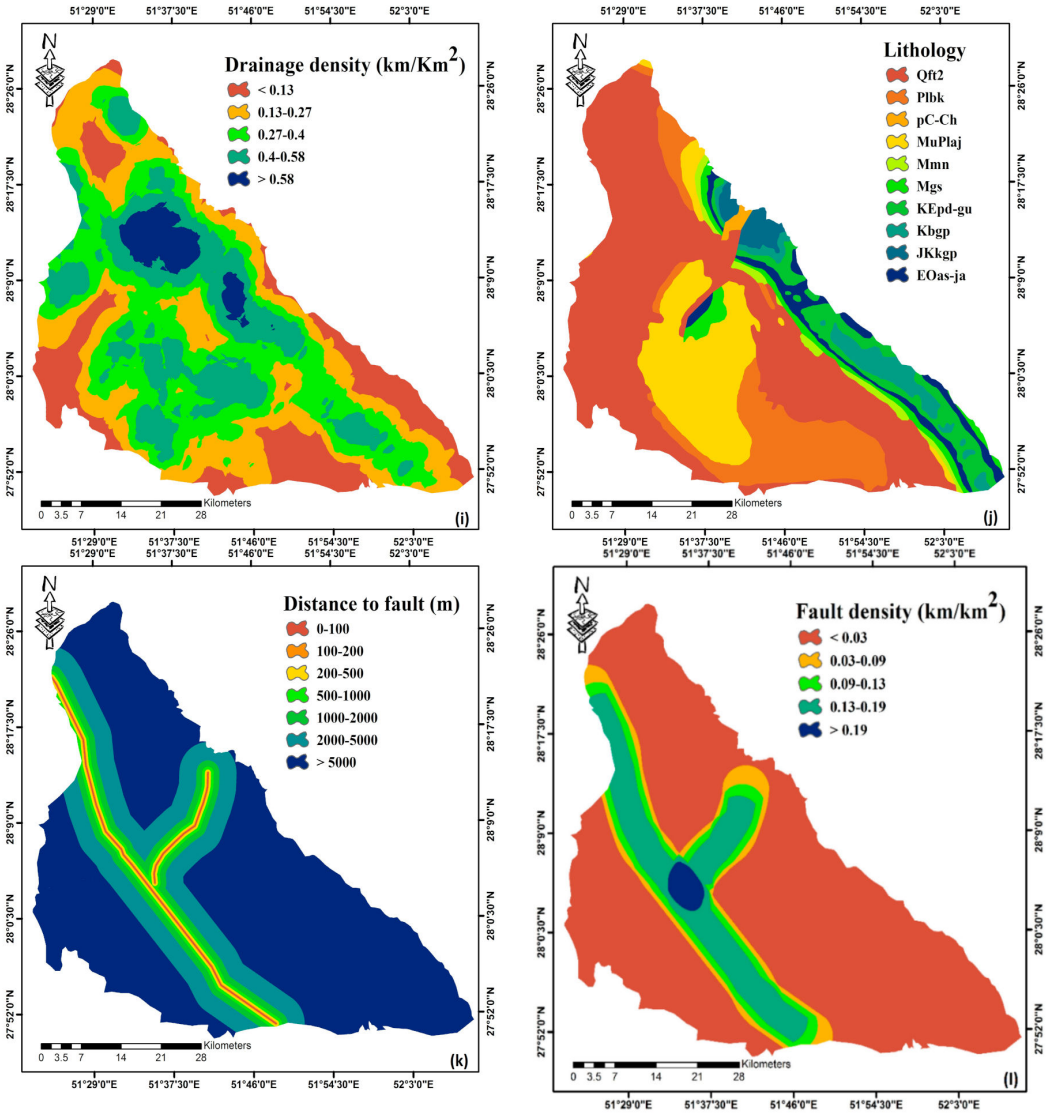

Figure 3. *Cont.*

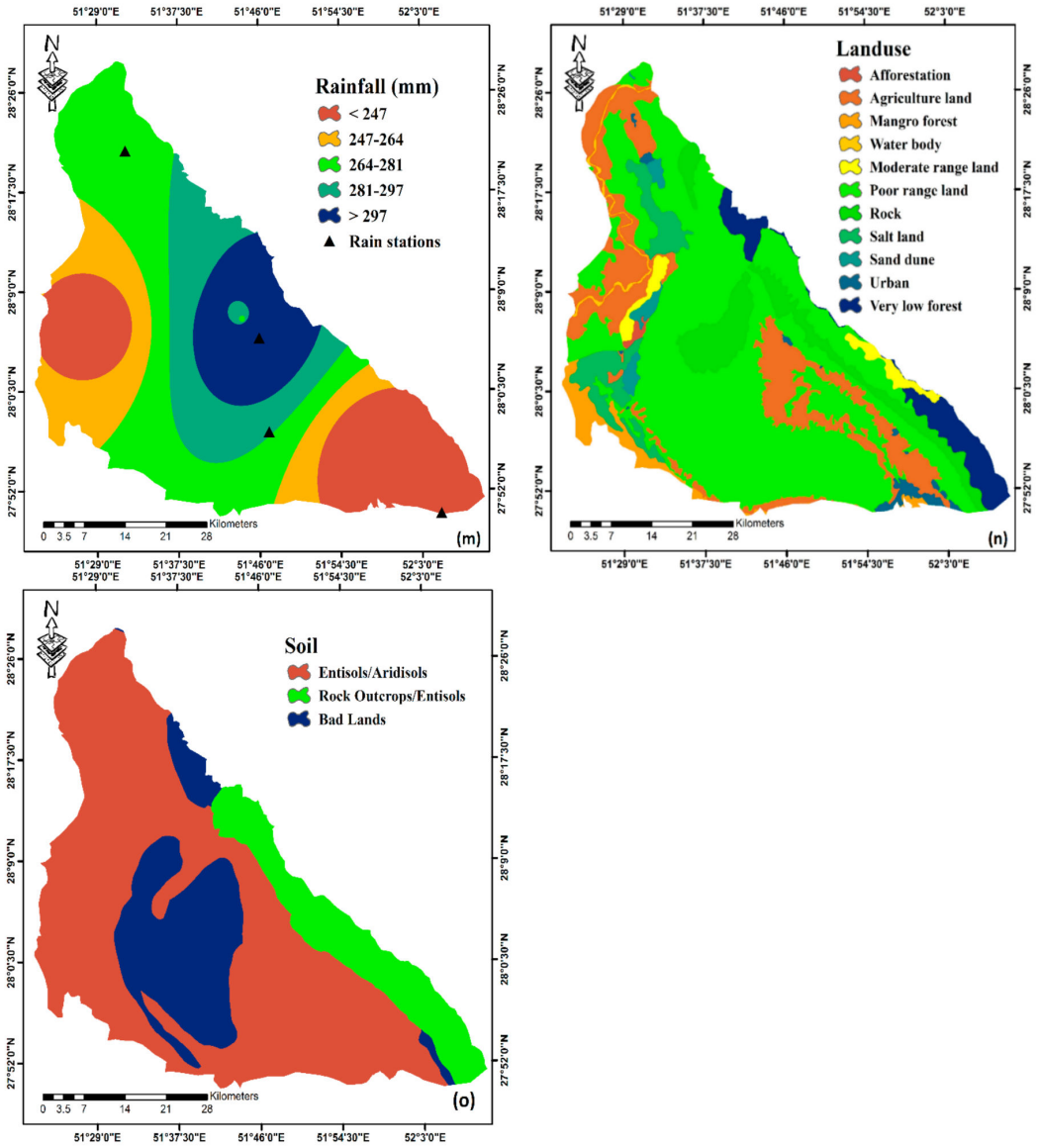

Figure 3. Groundwater effective factors: (**a**) altitude; (**b**) slope angle; (**c**) slope aspect; (**d**) slope length plan curvature; (**e**) plan curvature; (**f**) profile curvature; (**g**) topographic wetness index (TWI); (**h**) distance to river; (**i**) drainage density; (**j**) lithology; (**k**) distance to fault; (**l**) fault density; (**m**) rainfall; (**n**) land use; and (**o**) soil.

Table 1. Lithology characteristics of the Booshehr plain.

Unit	Lithology	Unit	Lithology
Qft2	Low-level piedmont fan	KEpd-gu	Massive fossiliferous limestone
MuPlaj	Siltstone, sandstone, red marl (Aghajari formation)	Kbgp	Mostly limestone and shale.
Plbk	Conglomerate locally with sandstone (Bakhtyari formation)	pC-Ch	Rock salt, rhyolite basalt, and trachyte

Table 1. *Cont.*

Unit	Lithology	Unit	Lithology
Mmn	Gray marls with low weather (Mishan formation)	OMr	Silty red, gray and green marls, little ribs of sandstone (RAZAK FM)
Mgs	Red marl, anhydrite, salt locally with argillaceous limestone (Gachsaran formation)	JKkgp	Undivided group of Khami, made up of huge thin limestone bedded
EOas-ja	Undivided formation of Asmari and Jahrum		

2.4. Models

2.4.1. FR Model

The FR illustrates the relationship between wells and groundwater's effective factors [6]. The FR is the ratio of the area where the wells are located to the total area of study. The ratio of well occurrence to nonoccurrence is obtained in order to calculate the FR value for each class or factor. The FR model is calculated by Equation (3) [5]:

$$FR = \frac{\frac{Npix\ (SX_i)}{\sum_{i=1}^{m} Npix\ (SX_i)}}{\frac{Npix\left(X_j\right)}{\sum_{j=1}^{n} Npix\left(X_j\right)}} \quad (3)$$

where $Npix(SX_i)$ is the total value of pixels in each class of each criterion with well locations, $Npix\left(X_j\right)$ is the number of pixels in each class of each criterion j, and m and n are, respectively, the number of classes per criterion and the total number of criteria. In this research, 15 parameters affecting groundwater have been used, each of these parameters is called a criterion, and each criterion is divided into different categories, each of which is called a class. Although the FR model uses simple and understandable concepts, it can analyze bivariate statistical analyzes and classifications of each factor as well. The disadvantage of FR is that the relationship between variables is ignored in this method [39].

2.4.2. CF Model

The CF model was first presented by Shortliffe and Buchanan in 1975 as a bivariate statistical model and modified by Longman in 1986 [40,41]. This model integrates the results in a spatial database using GIS. The CF model is calculated by frequency of well events in each class of layer via Equation (4) [42]:

$$\begin{cases} cf = \frac{pp_a - pp_s}{pp_a\{1 - pp_s} if \ \ pp_a \geq pp_s \\ cf = \frac{pp_a - pp_s}{pp_s\{1 - pp_a} if \ \ pp_a < pp_s \end{cases} \quad (4)$$

where pp_a is the conditional likelihood of wells in a class and pp_s is the previous likelihood of wells in the region. pp_a is the ratio of the number of well-pixels in a class to the total pixels of that class, and pp_s is the ratio of the total number of pixels with wells in the study area to the total pixels of the map. The CF model varies from −1 to +1, with positive values indicating an increase in certainty and negative values indicating a decline in certainty [43].

2.4.3. EBF Model

The EBF model is a statistical bivariate technique based on the theory of Dempster Shafer [44]. In the EBF model, Bel, Unc, Dis and Pls parameters are the rank of belief, the rank of uncertainty, the rank of disbelief and the rank of plausibility, respectively [45]. In the EBF model, the Bel parameter takes into account the pessimistic mode and low probability and the Pls parameter considers the optimistic mode and high probability state, so the value of the Bel parameter is smaller or equal than the Pls parameter, and the difference between these two parameters is called Unc. The data

extracted from this model not only estimates the spatial correlation between the effective factors and the occurrence of wells, but also the spatial correlation between each class factor. This method is calculated via Equations (5) and (6) [13]:

$$Bel_{C_{ij}} = \frac{W_{C_{ij}D}}{\sum_{j=0}^{m} W_{C_{ij}D}}, \quad (5)$$

$$W_{C_{ij}D} = \frac{N\left(C_{ij} \cap D\right)/N\left(C_{ij}\right)}{N(D) - N\left(C_{ij} \cap D\right)/N(T) - N\left(C_{ij}\right)}. \quad (6)$$

The conditional probability that shows the probability of an existing well (i.e., groundwater occurrences) in the absence of C_{ij} (each class of each factor) is shown in Equation (6). $W_{C_{ij}D}$ is the weight of C_{ij} which represents the belief that there is a well instead of it being lacking. In these relations, m represents the number of criteria considered for modeling, i represents each class of each criterion and j represents each criterion. In these relations, N (T) and N (D) represent the total number of pixels in the study area and the total number of well-pixels in the study area, respectively.

2.4.4. RF Model

The RF model is one of the data mining algorithms used by various trees in the classification [46]. By replacing and altering the variables influencing the target, the RF model generates a big amount of decision trees. Then in a prediction, the algorithm integrates all trees [47]. In the training process, the original data of each tree is selected randomly [48]. The RF includes three user-defined parameters, including the number of factors used in the construction of each tree, the number of trees and the minimum number of tree nodes. By enhancing the strength of autonomous trees and reducing the correlation between them, the power of the RF model forecast improves [49]. The RF system does not use all accessible information to grow the tree, but utilizes 66% of the Bootstrap information. Then during the growing phase, a predictor variable is implemented randomly and to generate a node in a tree, this variable is used. The decision tree is thus produced in its maximum size [50]. For the evaluation of the fitted tree, 33% of the remaining information are also used. This process is repeated several times, and the algorithm's final forecast is used as the average of all expected values [49].

2.4.5. LMT Model

This model is a nonparametric method that predicts quantitative variables or classified variables based on a collection of quantitative and qualitative predictor variables. Indeed, a hierarchical model's decision tree comprises of decision tools that return to the decomposition into homogeneous areas of independent variables [51]. Decision trees are a way of defining a set of laws leading to a category or value. One of the distinctions between the building techniques of the decision tree is how this distance is measured. Decision trees used to forecast discrete variables are called classification trees because they categorize samples. Decision trees are called regression trees, which are used to forecast continuous variables [52]. The decision tree's objective is to discover an approach in the form of a sequence of rules to present the outcomes of the predictions extracted from the set of input factors. This model of classification is a combination of the method of logistic regression and decision tree learning. To isolate an increase in logistic type data, the LogitBoost algorithm is used to produce the LR model in each tree leaf and the tree is cut using the CART algorithm. For each class C (well or non-well), the LogitBoost algorithm uses logistic regression of an additive with minimum squares as the Equation (7) [53]:

$$L_c(X) = \beta_0 + \sum_{i=1}^{D} \beta_i, \quad (7)$$

$$(C|X) = \frac{\exp(L_c(X))}{\sum_{c'=1}^{C} \exp(L_{c'}(X))} \quad (8)$$

where c is the class number.

2.5. Validation

One of the appropriate methods for evaluating the results of classification and assessing its capability to identify a specified class is to use the ROC curve and the AUC to validation the sensitivity of the method [54]. The sensitivity means the relationship between the classified values and unclassified values. The higher the deviation from the baseline for a particular class in the ROC curve, the more efficient the classifier is in identifying the class. In addition to considering the trend diagram of a specific class, the AUC is also calculated [55]. This area indicates the probability that a randomly selected value will be classified correctly. Higher values show the reliability of the method [56]. This index assesses the values properly assigned to the target class (True Positive), the values assigned to the incorrect class (False Positive), the values not assigned to the defined class (True Negative) and the values not assigned to the incorrect class (False Negative). This curve consists of a horizontal axis (X axis) and a vertical axis (Y axis) which are calculated in Equations (9) and (10) [57]:

$$X = 1 - \frac{TN}{TN + FP}, \quad (9)$$

$$Y = \frac{TP}{TP + FN}. \quad (10)$$

When the actual output is positive and the prediction value is positive, this state is called TP (True positive), whereas FN (False negative) represents the state where the actual output is negative and the prediction value is also negative. TN (True negative) represents the state where the actual output is positive and the prediction value is negative, and FP (False positive) is the state where the actual output is negative and the prediction value is positive. These indices are derived from the confusion matrix and the ROC curve is calculated on this basis.

3. Results

3.1. Result of Bivariate Statistical Models

Table A1 presents the outcomes of bivariate statistical models. Based on the results of the FR model, in the altitude factor, the class with less than 108 m (1.56) had the highest correlation with groundwater, while the altitude class of more than 287 m showed no correlation with groundwater. The results for altitude with less than 108 m for EBF and CF are 0.983 and 0.361, respectively. Areas with low altitude have more permeable against runoff [6]. In the slope aspect factor, the northwest direction has the highest weight in FR and CF with weights of 1.485 and 0.326, respectively. This is due to the fact that the northwest receives more rainfall and moisture content than other directions in creating groundwater. In the EBF model, the highest weight (0.505) is dedicated to the southeast. Based on the slope angle factor, the slope class with less than 6° has the maximum weight. These weights for FR, EBF and CF are 1.526, 0.649 and 0.345, respectively. The results show that the lower slope is strongly correlated with groundwater, while the higher slope has less effect on the occurrence of groundwater due to increased runoff in high slopes [31].

The class of (−0.4–0.8) had the maximum weight (FR = 1.23, Bel = 0.611 and CF = 0.189) based on the plan curvature factor and had the greatest effect on groundwater incidence. This category of the plan curvature holds more water over a long period of time [6]. The class of 0.4–1 has the highest weight, equivalent to 1.39, 0.596 and 0.283 for FR, EBF and CF, respectively, according to the profile curvature. In the slope length parameter, the class of 0–10 m has the most effect on groundwater (FR = 1.5, Bel = 0.762 and CF = 0.336). As the slope length increases, the weight of its classes decreases,

which indicates that the lower values of this parameter have a greater effect on groundwater. According to the results of TWI, the class of 4.69–6.57 has the highest weight (FR = 1.64, Bel = 0.44 and CF = 0.39). The TWI shows the impact of topography on the location and size of runoff saturation areas [31]. The class of 0–100 m has the highest weight based on the outcome of the distance to river factor. With regard to the results, less distances to rivers have had a greater impact on groundwater [6]. This weight is equal to 1.75 for FR and 0.429 for CF, while in the EBF, the class greater than 2000 m has the greatest effect on groundwater occurrence (0.343). In the distance from fault factor, the class of 1000–2000 m has the highest weight in FR and CF (FR = 1.26 and CF = 0.206). While in the EBF model, the class of 2000–5000 m has the most effect on the occurrence of groundwater with the weight of 0.482. The distance from fault and fault density can control the water exchange between the ground and the basement, so the distance closer to the fault can have a positive effect on the occurrence of groundwater [6].

In the drainage density factor, for the FR and CF, the maximum weight is dedicated to the class greater than 0.58 with 4.19 and 0.761 for FR and CF, respectively. While in the EBF model, the highest weight is assigned to the class 0.4–0.58 (0.383). Drainage density represents the lithology structure of an area and has a significant impact on the identification of groundwater resources [31]. According to the results of the fault density factor, the class of 0.9–0.13 has the highest weight in FR and CF (FR = 1.69 and CF = 0.409), and in the EBF (0.808), the highest weight is related to the class with less than 0.03. In the rainfall factor, the class of 297 mm has the highest weight in the FR (2.72) and CF (0.632), while in the EBF model, the highest weight is related to the class 0–247 with 0.737. Considering the fact that the study area has a small annual precipitation, more rainfall classes indicate more groundwater occurrence. According to the lithology factor, the Qft2 class has the highest weight (FR = 1.95, Bel = 0.857 and CF = 0.489). This is because most of the area is comprised of the Qft2 unit. According to the results of the land use factor, the class of moderate range land has the highest weight in the FR (10.42) and CF (0.904), while in the EBF model (0.49) the class of agriculture has the greatest impact on the occurrence of groundwater, which can be due to water penetration from irrigation of agricultural areas to the groundwater system and charging the aquifer. In the soil factor, the Entisols/Aridosols class has the highest weight (FR = 1.535, Bel = 1 and CF = 0.348).

3.2. Application of Ensemble Models

To prepare the GPM with the ensemble model, the weights obtained from bivariate statistic models (Table A1) are considered as inputs to the RF and LMT models. In this research, the RF and LMT model were implemented in the Waikato Environment for Knowledge Analysis (WEKA) data mining software [58,59]. In modeling the RF, the values and importance of each parameter in modeling can be determined. The results show that in the FR-RF model, the greatest importance is related to the slope aspect (0.41) following distance to river (0.38), rainfall (0.33) and TWI (0.32). The least importance is related to soil (0.25), slope length (0.24) and lithology (0.21), respectively. In the CF-RF model, the greatest importance is related to the slope angle (0.39), then distance to river, drainage density, profile curve, rainfall and slope length with the value of 0.32. The least importance is related to lithology (0.22), soil (0.22) and altitude (0.19), respectively.

According to the EBF-RF model, distance from river and slope aspect have the most importance (0.360) followed by rainfall and profile curvature with 0.34. The least importance is related to slope length (0.25), lithology (0.22) and soil (0.19). The results of importance in the three models are shown in Figure 4. Slope aspect, distance to river and rainfall in all three models shows high importance. Figure 5 shows the results of tree building in the LMT model.

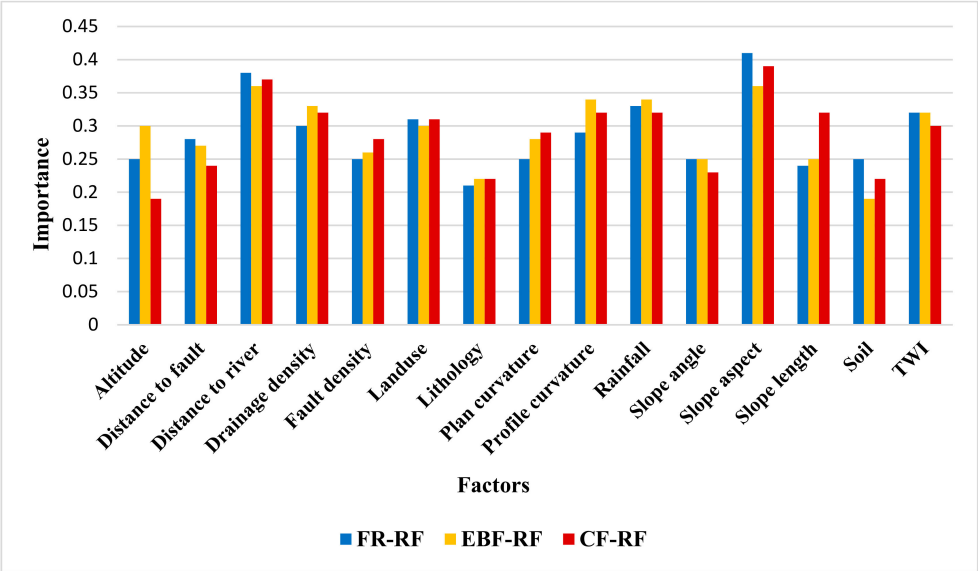

Figure 4. The relative importance of the factors using the random forest (RF) model.

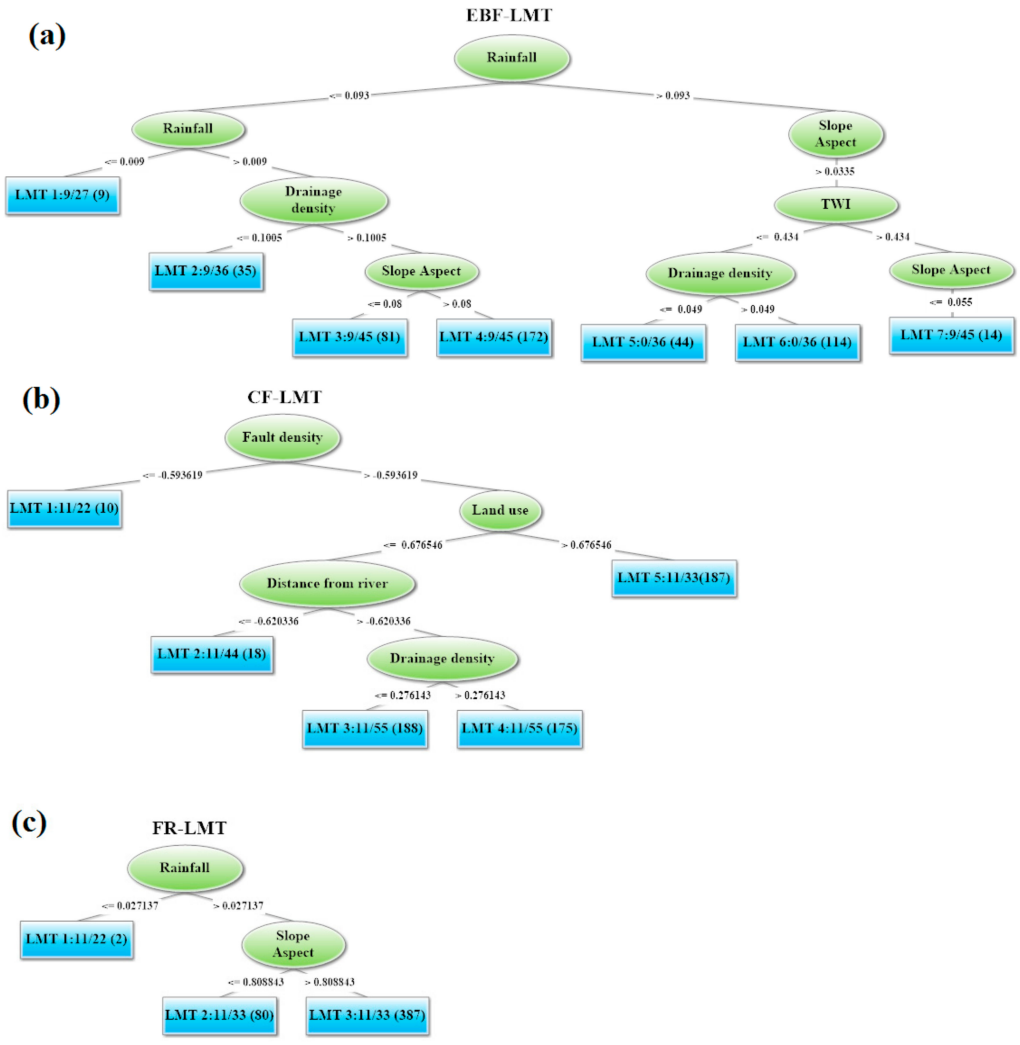

Figure 5. Result of the logistic model tree (LMT) model.

After modeling the RF and LMT in hybrid with the other bivariate statistical models in the WEKA software, the model was generalized to the total pixels of the study area and was calculated in ArcGIS 10.3 software for each pixel, which represents the effect of that pixel on groundwater. After generalizing the model to the total pixels in the area, each pixel has a weight that indicates the occurrence of groundwater. In order to describe the numerical data, the natural breaks classifier was used. In this technique, the results were classified as rising in very low potential, low potential, moderate potential, high potential and very high potential. Very low and low potential classes show a low probability of occurrence of groundwater in these areas, and very high and high potential classes also indicate the probable occurrence of groundwater in these areas. A very high potential class that indicates the likelihood that more groundwater will occur in these areas (Figure 6a–f).

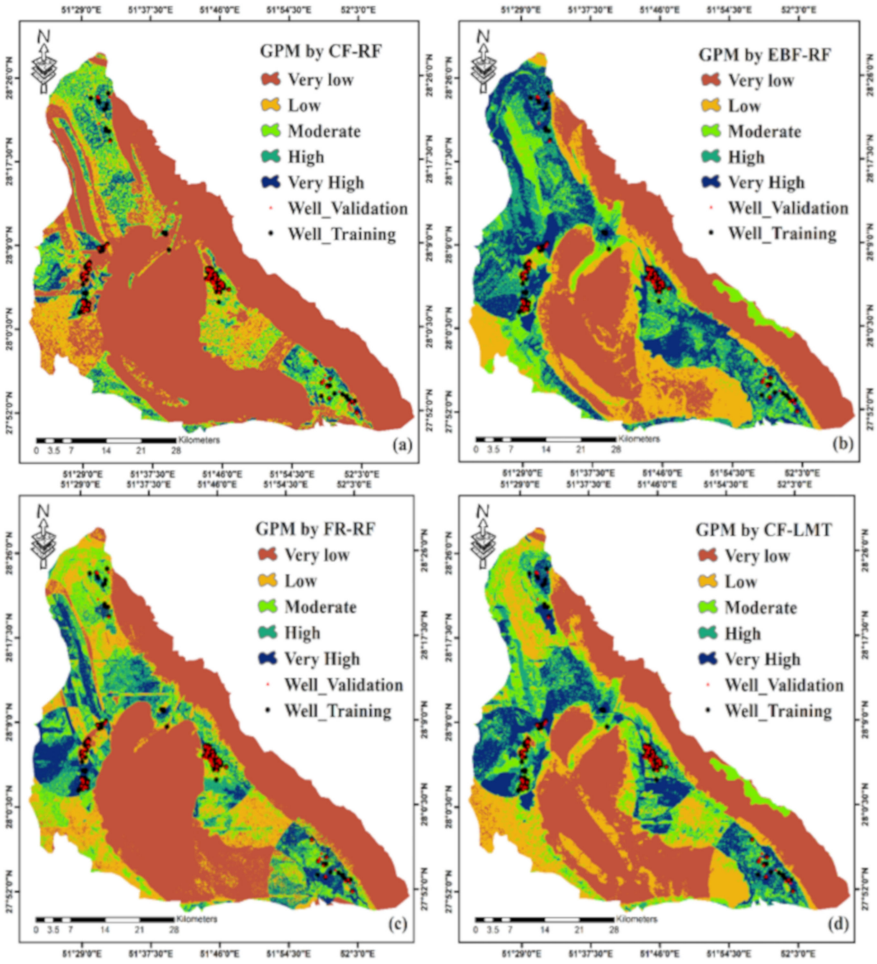

Figure 6. *Cont*.

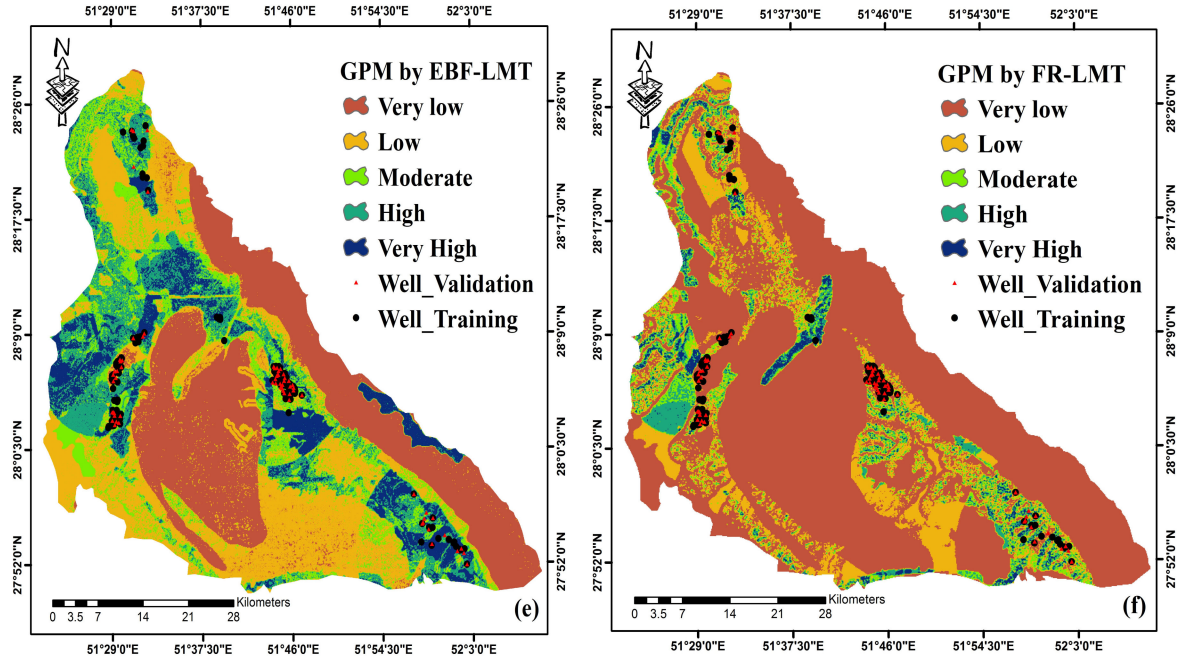

Figure 6. Groundwater potential map (GPM) of the Booshehr plain made from (**a**) certainty factor (CF)-RF, (**b**) evidential belief function (EBF)-RF, (**c**) frequency ratio (FR)-RF, (**d**) CF-LMT, (**e**) EBF-LMT and (**f**) FR-LMT models.

3.3. Validation of Models

Figure 7 presents the ROC curve for the six ensemble models. The results show that the highest accuracy is related to the CF-RF model (0.927), EBF-RF (0.924), FR-RF (0.917), CF-LMT (0.906), EBF-LMT (0.885) and FR-LMT (0.830) models, respectively (Table 2). The higher accuracy of EBF and CF models than the FR model is due to the uncertainty regarding the occurrence of groundwater in the results. In Figure 7, the X axis represents a sensitivity that expresses the prediction value correct in front of all positive outputs, and also the Y axis represents the specificity that represents the predicted negative value correct in front of all negative outputs. The AUC is between zero and one; values less than 0.5 represent model integrity and for models larger than 0.5 has a higher accuracy. The results show high accuracy in combining statistical models with the RF model in preparing a groundwater potential map.

Table 2. The results of the receiver operating characteristic (ROC) area under the curve (AUC) of three GPM models.

Test Result Variable(s)	AUC	Std. Error [a]	Asymptotic Sig. [b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
CF-RF	**0.927**	0.018	0.000	0.892	0.963
EBF-RF	**0.924**	0.021	0.000	0.884	0.965
FR-RF	**0.917**	0.020	0.000	0.877	0.957
CF-LMT	**0.906**	0.021	0.000	0.865	0.947
EBF-LMT	**0.885**	0.023	0.000	0.841	0.929
FR-LMT	**0.830**	0.029	0.000	0.773	0.886

[a] Under the nonparametric assumption; [b] Null hypothesis: True area = 0.5.

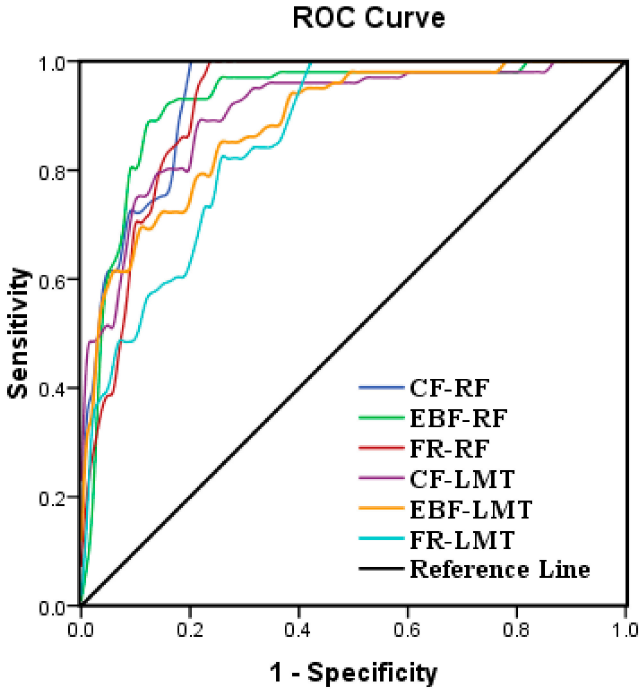

Figure 7. Assessment of model performance using the ROC curve.

4. Discussion

Groundwater potential mapping can be explored using GIS-based models, with less time and cost. So far, a number of statistical bivariate techniques were used to prepare GPMs. However, due to the need for more precision in determining the potential areas, data mining methods and their combination with bivariate statistical models are used. Decision tree is one of the non-parametric supervised learning algorithms that is widely used in classification. One of the concepts that should be taken into account in deciding trees is the issue of overfitting. Another disadvantage of this method is the imbalance in the two concepts of bias and variance.

The concept of bias refers to how much the system outputs are far from the actual expected value, and the concept of variance refers to how much the system accuracy varies by changing the sample or using out-data samples [2]. If we select a tree with low depth, it suffers from bias errors and reduces variance, and if we use a tree with a high depth, it will have a higher variance, while reducing lower bias. Therefore, the decision tree must meet the balance between bias and variance. The bagging method is used to reduce variance and combines the results of prediction from several predictive systems based on some factors, such as mean, median, and so on [49]. One of the implementation methods in bagging for decision trees is the RF model. On the other hand, the main part in modeling the data mining method is preparing and processing data. Therefore, in this research, a combination of bivariate statistical models, such as FR, CF and EBF with RF and LMT models, were used to prepare groundwater potential mapping for the Booshehr plain.

The fundamental difference between bivariate statistical models and data mining methods is based on the hypothesis or nature of the data being processed. As a general rule, the statistical technique hypothesis is based on the reality that information distribution is clear and normal and the accuracy or inaccuracy of the final outcomes depends on the validity of the initial assumption [60]. In contrast, data mining techniques do not use any hypotheses. Another advantage of data mining techniques over bivariate statistical models is that data mining models perform much better where data is incomplete or contradictory. Because in this case, lost data is retrieved based on the pattern in the data [53]. However, in bivariate statistical models a lost parameter leads to uselessness of the data. By increasing the number of parameters, bivariate statistical models lose the ability to find patterns, and due to their

linear nature, it is impossible for them to discover the nonlinear and complex relationships between variables. But data mining techniques are designed to find complex relationships between several parameters in the database.

Based on the ROC curve outcomes, the CF-RF model has the highest accuracy with the AUC (0.927), followed by EBF-RF, FR-RF, CF-LMT, EBF-LMT and FR-LMT models (0.924, 0.917, 0.906, 0.885 and 0.83). The input and output of the FR model are simple and easy to calculate [57], while the EBF model has more computational complexity. EBF and CF models are able to combine the confidence of different sources and are flexible against uncertainty [61]. In CF and BF models, areas where there is no wells are considered in modeling, while in the FR model, only the areas in which the well is located is used in modeling. According to the results, CF and EBF are more accurate than FR by including uncertainty in their results. The weights obtained from the FR model are greater than zero, while the weights obtained from the CF model have values between −1 and 1. The reason for the different results of these two models in the hybrid model is to consider the negative values in the CF model, so that in this model negative values indicate the negative effect of that parameter on the potential of groundwater, but in the FR model this negative effect is not specified precisely. According to previous research, the combination of bivariate statistical models with data mining algorithms has increased the accuracy of groundwater potential mapping [5,17]. The results show that the RF model has a higher accuracy than the LMT model, which is because the RF model requires no assumptions on the distribution of factors and can also calculate the interaction between the factors. Another advantage of the RF model over the LMT model is overcoming overfitting and coping with the bias in the data. One of the advantages of the RF model is its usability, both for classification and regression issues, which is dominated by current machine learning systems. RF accepts and executes thousands of input variables without deleting one of them, and it can also determine which variables are important in predicting the model. The RF algorithm is very useful and easy to use, because its default hyper parameters often produce good predictive results. The number of hyper parameters is also not high and easy to understand.

5. Conclusions

The purpose of this research was to make an ensemble of bivariate statistical models including FR, EBF and CF with RF and LMT models in order to prepare a GPM for the Booshehr Plain. The results obtained from this research are as follows:

1. Based on the results from the ROC curve and AUC, the CF-RF model is more accurate in providing GPM, followed by the EBF-RF and FR-RF models.
2. The results show that CF and EBF are more accurate than FR in combining with the random forest model via considering the uncertainty in the results.
3. In combined models, slope aspect, distance from waterway, rainfall, and topography curve parameters have the most importance, and lithology and soil parameters have the least importance.
4. According to the results from FR and CF, the maximum weight is dedicated to an elevation class of less than 108 m, the slope angle class of less than 6°, northwest slope aspect, the topographic curve, the slope length class of less than 10 m, the topographic humidity index between 4.69 and 6.57, the distance from waterway class of less than 100 m, the distance from fault between 1000 and 2000 m, the water density of greater than 0.58, the density of fault between 0.09 and 0.19, rainfall greater than 297 mm, the lithology class of Qft2 unit, the moderate rangeland in land-use class and the Entisols class in the soil parameter. The results of the evidential belief model are largely similar to the other two models, but for some parameters the results are different. According to the results of the evidential belief model, the highest weight is dedicated to the southeast slope aspect, distance from waterway in the 200 to 500 m class, distance from fault in the 2000 to 5000 m class, the water density in the 0.4–0.58 class, the fault density in the class of less than 0.03, rainfall in the class of 0 to 274 mm, and the agricultural class in the land-use parameter.

Author Contributions: Conceptualization, S.V.R.-T. and A.S.-N.; Data curation, S.V.R.-T.; Formal analysis, S.V.R.-T.; Funding acquisition, S.-M.C.; Investigation, S.V.R.-T.; Methodology, A.S.-N.; Project administration, S.-M.C.; Resources, A.S.-N.; Software, S.V.R.-T. and A.S.-N.; Supervision, A.S.-N.; Validation, S.V. R.-T. and A.S.-N.; Visualization, S.V.R.-T.; Writing—original draft, S.V.R.-T. and A.S.-N.; Writing—review & editing, A.S.-N. and S.-M.C.

Funding: This research was supported by MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2016-0-00312) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1. Result of three bivariate statistic models.

Class	No. Pixels in Domain	No. of Wells	FR	Bel	CF
Altitude (m)					
<108	427576	237	1.56	0.983	0.361
108–287	108357	1	0.026	0.0164	0.0164
287–535	736786	0	0	0	0
535–851	421447	0	0	0	0
>851	202301	0	0		0
Slope Angle (degree)					
<6	340276	184	1.526	0.649	0.345
6–14	183016	53	0.817	0.347	−0.182
14–24	856988	0	0	0	−1
24–39	475130	1	0.06	0.0025	−0.94
>39	154827	0	0	0	−1
Slope Aspect					
F	470400	13	0.78	0.047	−0.219
N	656895	8	0.343	0.02	−0.656
NE	730046	33	1.27	0.077	0.216
E	791429	25	0.891	0.053	−0.108
SE	101153	30	0.837	0.505	−0.162
S	957842	47	1.385	0.083	0.278
SW	941836	36	1.079	0.065	0.073
W	798851	27	0.954	0.057	−0.045
NW	361045	19	1.485	0.089	0.326
Plan Curvature (100/m)					
<−2.22	146963	0	0	0	−1
−1.42	695619	15	0.6	0.03	−0.391
−0.8–0.4	423317	185	1.23	0.611	0.189
0.4–2.2	148465	38	0.722	0.358	−0.277
>2.2	159465	0	0	0	−1
Soil					
Entisols/Aridosols	436865	238	1.535	1	0.348
Bad lands	140085	0	0	0	−1
Rock outcrops/Entisols	939041	0	0	0	−1
Profile Curvature (100/m)					
<−3.4	137472	1	0.2	0.008	−0.794
−2.4	924569	3	0.091	0.004	−0.908
−1–0.4	337923	167	1.39	0.596	0.283
0.4–2.8	206784	67	0.914	0.39	−0.085
>2.8	210755	0	0	0	−1
Slope Length (m)					
0–10	404924	216	1.5	0.762	0.336

Table A1. *Cont.*

Class	No. Pixels in Domain	No. of Wells	FR	Bel	CF
10–20	134087	21	0.442	0.223	−0.557
20–30	740597	0	0	0	−1
30–40	453420	1	0.062	0.003	−0.937
>40	135739	0	0	0.01	−1
TWI					
<2.92	115281	1	0.024	0.006	−0.975
2.92–3.84	187497	31	0.466	0.125	−0.533
3.84–4.69	212059	120	1.59	0.428	0.374
4.69–6.57	147887	86	1.64	0.44	0.39
>6.57	92617	0	0	0	−1
Distance to river (m)					
0–100	418540	26	1.75	0.039	0.429
100–200	418903	18	1.21	0.027	0.175
200–500	122099	60	1.38	0.309	0.279
500–1000	170413	44	0.72	0.162	−0.271
1000–1500	112869	20	0.5	0.111	−0.499
1500–2000	563509	6	0.259	0.006	−0.74
>2000	117338	64	1.539	0.343	0.35
Distance to fault (m)					
0–100	52047	1	0.542	0.023	−0.457
100–200	46406	0	0	0	−1
200–500	140802	0	0	0	−1
500–1000	226477	0	0	0	−1
1000–2000	425525	19	1.26	0.053	0.206
2000–5000	124594	50	1.13	0.482	0.117
>5000	458318	168	1.03	0.44	0.033
Drainage Density (km/km^2)					
<0.13	118694	52	1.24	0.3067	0.195
0.13–0.27	174085	27	0.44	0.108	−0.559
0.27–0.4	199824	28	0.39	0.098	−0.602
0.4–0.58	149777	82	1.55	0.383	0.356
>0.58	331397	49	4.19	0.103	0.761
Fault Density (km/km^2)					
<0.03	506907	190	1.058	0.808	0.054
0.03–0.09	331072	0	0	0	−1
0.09–0.13	399735	24	1.69	0.129	0.409
0.13–0.19	833523	24	0.812	0.062	−0.187
>0.19	84751	0	0	0	−1
Rainfall (mm)					
0–247	153067	119	2.2	0.737	0.545
247–264	104382	2	0.054	0.018	0.00003
264–281	209726	21	0.283	0.095	0.000025
281–297	114294	7	0.173	0.058	0.000029
>297	926820	89	2.72	0.091	0.632
Lithology					
Qft2	320270	222	1.95	0.853	0.489
MuPlaj	112135	13	0.327	0.142	−0.672
Plbk	959539	3	0.088	0.003	−0.911
Mmn	123084	0	0	0	−1
Mgs	221836	0	0	0	−1
Eoas-ja	288275	0	0	0	−1
KEpd-gu	389238	0	0	0	−1
Kbgp	258465	0	0	0	−1
JKkgp	125464	0	0	0	−1
Pc-ch	30462	0	0	0	−1

References

1. Jothibasu, A.; Anbazhagan, S. Modeling groundwater probability index in Ponnaiyar river basin of South India using analytic hierarchy process. *Model. Earth Syst. Environ.* **2016**, *2*, 109. [CrossRef]
2. Lee, M.-J.; Park, I.; Lee, S. Forecasting and validation of landslide susceptibility using an integration of frequency ratio and neuro-fuzzy models: A case study of Seorak mountain area in Korea. *Environ. Earth Sci.* **2015**, *74*, 413–429. [CrossRef]
3. Molden, D. *Water for Food Water for Life: A Comprehensive Assessment of Water Management in Agriculture*; Routledge: London, UK, 2013.
4. Park, S.; Hamm, S.-Y.; Jeon, H.-T.; Kim, J. Evaluation of logistic regression and multivariate adaptive regression spline models for groundwater potential mapping using R and GIS. *Sustainability* **2017**, *9*, 1157. [CrossRef]
5. Manap, M.A.; Nampak, H.; Pradhan, B.; Lee, S.; Sulaiman, W.N.A.; Ramli, M.F. Application of probabilistic-based frequency ratio model in groundwater potential mapping using remote sensing data and GIS. *Arab. J. Geosci.* **2014**, *7*, 711–724. [CrossRef]
6. Moghaddam, D.D.; Rezaei, M.; Pourghasemi, H.; Pourtaghie, Z.; Pradhan, B. Groundwater spring potential mapping using bivariate statistical model and GIS in the Taleghan watershed, Iran. *Arab. J. Geosci.* **2015**, *8*, 913–929. [CrossRef]
7. Kebede, S. Groundwater potential, recharge, water balance: Vital numbers. In *Groundwater in Ethiopia*; Springer: Berlin, Germany, 2013; pp. 221–236.
8. Arabgol, R.; Sartaj, M.; Asghari, K. Predicting nitrate concentration and its spatial distribution in groundwater resources using support vector machines (SVMs) model. *Environ. Model. Assess.* **2016**, *21*, 71–82. [CrossRef]
9. Arabameri, A.; Rezaei, K.; Cerda, A.; Lombardo, L.; Rodrigo-Comino, J. GIS-based groundwater potential mapping in Shahroud plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. *Sci. Total Environ.* **2019**, *658*, 160–177. [CrossRef]
10. Naghibi, S.A.; Dashtpagerdi, M.M. Evaluation of four supervised learning methods for groundwater spring potential mapping in Khalkhal region (Iran) using GIS-based features. *Hydrogeol. J.* **2017**, *25*, 169–189. [CrossRef]
11. Razandi, Y.; Pourghasemi, H.R.; Neisani, N.S.; Rahmati, O. Application of analytical hierarchy process, frequency ratio, and certainty factor models for groundwater potential mapping using GIS. *Earth Sci. Inform.* **2015**, *8*, 867–883. [CrossRef]
12. Pourghasemi, H.R.; Beheshtirad, M. Assessment of a data-driven evidential belief function model and GIS for groundwater potential mapping in the Koohrang watershed, Iran. *Geocarto Int.* **2015**, *30*, 662–685. [CrossRef]
13. Mogaji, K.; Omosuyi, G.; Adelusi, A.; Lim, H. Application of GIS-based evidential belief function model to regional groundwater recharge potential zones mapping in hardrock geologic terrain. *Environ. Process.* **2016**, *3*, 93–123. [CrossRef]
14. Pourtaghi, Z.S.; Pourghasemi, H.R. GIS-based groundwater spring potential assessment and mapping in the Birjand Township, southern Khorasan Province, Iran. *Hydrogeol. J.* **2014**, *22*, 643–662. [CrossRef]
15. Chen, W.; Li, H.; Hou, E.; Wang, S.; Wang, G.; Panahi, M.; Li, T.; Peng, T.; Guo, C.; Niu, C. GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. *Sci. Total Environ.* **2018**, *634*, 853–867. [CrossRef] [PubMed]
16. Lee, S.; Kim, Y.-S.; Oh, H.-J. Application of a weights-of-evidence method and GIS to regional groundwater productivity potential mapping. *J. Environ. Manag.* **2012**, *96*, 91–105. [CrossRef] [PubMed]
17. Naghibi, S.A.; Moghaddam, D.D.; Kalantar, B.; Pradhan, B.; Kisi, O. A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. *J. Hydrol.* **2017**, *548*, 471–483. [CrossRef]
18. Zabihi, M.; Pourghasemi, H.R.; Pourtaghi, Z.S.; Behzadfar, M. GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. *Environ. Earth Sci.* **2016**, *75*, 665. [CrossRef]
19. Golkarian, A.; Naghibi, S.A.; Kalantar, B.; Pradhan, B. Groundwater potential mapping using C5. 0, random forest, and multivariate adaptive regression spline models in GIS. *Environ. Monit. Assess.* **2018**, *190*, 149. [CrossRef] [PubMed]

20. Rahmati, O.; Naghibi, S.A.; Shahabi, H.; Bui, D.T.; Pradhan, B.; Azareh, A.; Rafiei-Sardooi, E.; Samani, A.N.; Melesse, A.M. Groundwater spring potential modelling: Comprising the capability and robustness of three different modeling approaches. *J. Hydrol.* **2018**, *565*, 248–261. [CrossRef]
21. Lee, S.; Hong, S.-M.; Jung, H.-S. GIS-based groundwater potential mapping using artificial neural network and support vector machine models: The case of Boryeong city in Korea. *Geocarto Int.* **2018**, *33*, 847–861. [CrossRef]
22. Naghibi, S.A.; Ahmadi, K.; Daneshi, A. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resour. Manag.* **2017**, *31*, 2761–2775. [CrossRef]
23. Khosravi, K.; Panahi, M.; Tien Bui, D. Spatial prediction of groundwater spring potential mapping based on an adaptive neuro-fuzzy inference system and metaheuristic optimization. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 4771–4792. [CrossRef]
24. Rahmati, O.; Samani, A.N.; Mahdavi, M.; Pourghasemi, H.R.; Zeinivand, H. Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS. *Arab. J. Geosci.* **2015**, *8*, 7059–7071. [CrossRef]
25. Olden, J.D.; Lawler, J.J.; Poff, N.L. Machine learning methods without tears: A primer for ecologists. *Q. Rev. Biol.* **2008**, *83*, 171–193. [CrossRef] [PubMed]
26. Oliveira, R.B.; Papa, J.P.; Pereira, A.S.; Tavares, J.M.R. Computational methods for pigmented skin lesion classification in images: Review and future trends. *Neural Comput. Appl.* **2018**, *29*, 613–636. [CrossRef]
27. Zhang, S.; Zhang, C.; Yang, Q. Data preparation for data mining. *Appl. Artif. Intell.* **2003**, *17*, 375–381. [CrossRef]
28. Kordestani, M.D.; Naghibi, S.A.; Hashemi, H.; Ahmadi, K.; Kalantar, B.; Pradhan, B. Groundwater potential mapping using a novel data-mining ensemble model. *Hydrogeol. J.* **2019**, *27*, 211–224. [CrossRef]
29. Ercanoglu, M.; Gokceoglu, C. Assessment of landslide susceptibility for a landslide-prone area (north of Yenice, NW Turkey) by fuzzy approach. *Environ. Geol.* **2002**, *41*, 720–730.
30. Nampak, H.; Pradhan, B.; Manap, M.A. Application of gis based data driven evidential belief function model to predict groundwater potential zonation. *J. Hydrol.* **2014**, *513*, 283–300. [CrossRef]
31. Moore, I.; Burch, G. Sediment transport capacity of sheet and rill flow: Application of unit stream power theory. *Water Resour. Res.* **1986**, *22*, 1350–1360. [CrossRef]
32. Al-Abadi, A.M.; Al-Temmeme, A.A.; Al-Ghanimy, M.A. A GIS-based combining of frequency ratio and index of entropy approaches for mapping groundwater availability zones at Badra–Al Al-Gharbi–Teeb areas, Iraq. *Sustain. Water Resour. Manag.* **2016**, *2*, 265–283. [CrossRef]
33. Moore, I.D.; Grayson, R.; Ladson, A. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrol. Process.* **1991**, *5*, 3–30. [CrossRef]
34. Dinesh Kumar, P.; Gopinath, G.; Seralathan, P. Application of remote sensing and GIS for the demarcation of groundwater potential zones of a river basin in Kerala, southwest coast of India. *Int. J. Remote Sens.* **2007**, *28*, 5583–5601. [CrossRef]
35. Ayazi, M.H.; Pirasteh, S.; Arvin, A.; Pradhan, B.; Nikouravan, B.; Mansor, S. Disasters and risk reduction in groundwater: Zagros mountain southwest Iran using geoinformatics techniques. *Disaster Adv.* **2010**, *3*, 51–57.
36. Devkota, K.C.; Regmi, A.D.; Pourghasemi, H.R.; Yoshida, K.; Pradhan, B.; Ryu, I.C.; Dhital, M.R.; Althuwaynee, O.F. Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling–Narayanghat road section in Nepal Himalaya. *Nat. Hazards* **2013**, *65*, 135–165. [CrossRef]
37. Tehrany, M.S.; Pradhan, B.; Jebur, M.N. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. *J. Hydrol.* **2013**, *504*, 69–79. [CrossRef]
38. Jamieson, R.; Gordon, R.; Sharples, K.; Stratton, G.; Madani, A. Movement and persistence of fecal bacteria in agricultural soils and subsurface drainage water: A review. *Can. Biosyst. Eng.* **2002**, *44*, 1–9.
39. Abdalla, F. Mapping of groundwater prospective zones using remote sensing and GIS techniques: A case study from the central eastern desert, Egypt. *J. Afr. Earth Sci.* **2012**, *70*, 8–17. [CrossRef]
40. Binaghi, E.; Luzi, L.; Madella, P.; Pergalani, F.; Rampini, A. Slope instability zonation: A comparison between certainty factor and fuzzy Dempster–Shafer approaches. *Nat. Hazards* **1998**, *17*, 77–97. [CrossRef]
41. Komac, B.; Zorn, M. Statistical landslide susceptibility modeling on a national scale: The example of Slovenia. *Rev. Roum Géogr.* **2009**, *53*, 179–195.

42. Sujatha, E.R.; Rajamanickam, G.V.; Kumaravel, P. Landslide susceptibility analysis using Probabilistic Certainty Factor Approach: A case study on Tevankarai stream watershed, India. *J. Earth Syst. Sci.* **2012**, *121*, 1337–1350. [CrossRef]
43. Mohammady, M.; Pourghasemi, H.R.; Pradhan, B. Landslide susceptibility mapping at Golestan Province, Iran: A comparison between frequency ratio, Dempster–Shafer, and weights-of-evidence models. *J. Asian Earth Sci.* **2012**, *61*, 221–236. [CrossRef]
44. Dempster, A.P. A generalization of Bayesian inference. *J. R. Stat. Soc. Ser. B (Methodol.)* **1968**, *30*, 205–232. [CrossRef]
45. Lee, S.; Hwang, J.; Park, I. Application of data-driven evidential belief functions to landslide susceptibility mapping in Jinbu, Korea. *Catena* **2013**, *100*, 15–30. [CrossRef]
46. Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Chapman & Hall: New York, NY, USA, 1984.
47. Vorpahl, P.; Elsenbeer, H.; Märker, M.; Schröder, B. How can statistical models help to determine driving factors of landslides? *Ecol. Model.* **2012**, *239*, 27–39. [CrossRef]
48. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
49. Peters, J.; Verhoest, N.; Samson, R.; Boeckx, P.; De Baets, B. Wetland vegetation distribution modelling for the identification of constraining environmental variables. *Landsc. Ecol.* **2008**, *23*, 1049–1065. [CrossRef]
50. Ließ, M.; Glaser, B.; Huwe, B. Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and random forest models. *Geoderma* **2012**, *170*, 70–79. [CrossRef]
51. Pradhan, B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput. Geosci.* **2013**, *51*, 350–365. [CrossRef]
52. Bui, D.T.; Tuan, T.A.; Klempe, H.; Pradhan, B.; Revhaug, I. Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides* **2016**, *13*, 361–378.
53. Chen, W.; Shahabi, H.; Shirzadi, A.; Li, T.; Guo, C.; Hong, H.; Li, W.; Pan, D.; Hui, J.; Ma, M. A novel ensemble approach of bivariate statistical-based logistic model tree classifier for landslide susceptibility assessment. *Geocarto Int.* **2018**, *33*, 1398–1420. [CrossRef]
54. Pourghasemi, H.R.; Moradi, H.R.; Aghda, S.F.; Gokceoglu, C.; Pradhan, B. GIS-based landslide susceptibility mapping with probabilistic likelihood ratio and spatial multi-criteria evaluation models (north of Tehran, Iran). *Arab. J. Geosci.* **2014**, *7*, 1857–1878. [CrossRef]
55. Tien Bui, D.; Khosravi, K.; Li, S.; Shahabi, H.; Panahi, M.; Singh, V.; Chapi, K.; Shirzadi, A.; Panahi, S.; Chen, W. New hybrids of ANFIS with several optimization algorithms for flood susceptibility modeling. *Water* **2018**, *10*, 1210. [CrossRef]
56. Alatorre, L.C.; Sánchez-Andrés, R.; Cirujano, S.; Beguería, S.; Sánchez-Carrillo, S. Identification of mangrove areas by remote sensing: The ROC curve technique applied to the northwestern Mexico coastal zone using Landsat imagery. *Remote Sens.* **2011**, *3*, 1568–1583. [CrossRef]
57. Termeh, S.V.R.; Kornejady, A.; Pourghasemi, H.R.; Keesstra, S. Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. *Sci. Total Environ.* **2018**, *615*, 438–451. [CrossRef] [PubMed]
58. Aburub, F.; Hadi, W. Predicting groundwater areas using data mining techniques: Groundwater in Jordan as case study. *Int. J. Comput. Electr. Autom. Control Inf. Eng.* **2016**, *10*, 1475–1478.
59. Faridi, M.; Verma, S.; Mukherjee, S. Integration of GIS, Spatial Data Mining, and Fuzzy Logic for Agricultural Intelligence. In *Soft computing: Theories and Applications*; Springer: Singapore, Singapore, 2018; pp. 171–183.
60. Clapcott, J.; Goodwin, E.; Snelder, T. *Predictive Models of Benthic Macroinvertebrate Metrics*; Prepared for Ministry for the Environment; Cawthron Institute: Nelson, New Zealand, 2013.
61. Tehrany, M.S.; Shabani, F.; Javier, D.N.; Kumar, L. Soil erosion susceptibility mapping for current and 2100 climate conditions using evidential belief function and frequency ratio. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1695–1714. [CrossRef]