sustainability

MDPI

Article

Evaluation of Logistic Regression and Multivariate Adaptive Regression Spline Models for Groundwater Potential Mapping Using R and GIS

Soyoung Park [1], Se-Yeong Hamm [2], Hang-Tak Jeon [2] and Jinsoo Kim [3,*]

[1] BK21 Plus Project of the Graduate School of Earth Environmental Hazard System, Pukyong National University, Busan 48513, Korea; yac100@pknu.ac.kr

[2] Department of Geological Sciences, Pusan National University, Busan 46241, Korea; hsy@pusan.ac.kr (S.-Y.H.); jeonhangtak@gmail.com (H.-T.J.)

[3] Department of Spatial Information Engineering, Pukyong National University, Busan 48513, Korea

* Correspondence: jinsookim@pknu.ac.kr; Tel.: +82-51-629-6658

Received: 7 June 2017; Accepted: 29 June 2017; Published: 2 July 2017

Abstract: This study mapped and analyzed groundwater potential using two different models, logistic regression (LR) and multivariate adaptive regression splines (MARS), and compared the results. A spatial database was constructed for groundwater well data and groundwater influence factors. Groundwater well data with a high potential yield of ≥70 m^3/d were extracted, and 859 locations (70%) were used for model training, whereas the other 365 locations (30%) were used for model validation. We analyzed 16 groundwater influence factors including altitude, slope degree, slope aspect, plan curvature, profile curvature, topographic wetness index, stream power index, sediment transport index, distance from drainage, drainage density, lithology, distance from fault, fault density, distance from lineament, lineament density, and land cover. Groundwater potential maps (GPMs) were constructed using LR and MARS models and tested using a receiver operating characteristics curve. Based on this analysis, the area under the curve (AUC) for the success rate curve of GPMs created using the MARS and LR models was 0.867 and 0.838, and the AUC for the prediction rate curve was 0.836 and 0.801, respectively. This implies that the MARS model is useful and effective for groundwater potential analysis in the study area.

Keywords: groundwater potential; logistic regression; multivariate adaptive regression splines; groundwater potential map; groundwater potential analysis

1. Introduction

Groundwater is defined as water in the saturated zone that fills the pore spaces between mineral grains and the cracks and fractures within a rock mass [1]. It results from the interactions of climatic, geological, hydrological, physiographical, and ecological factors [2]. Globally, groundwater makes up 50% of the present potable water supplies, 40% of the industrial water demand, and 20% of the water used for irrigation [3]. Therefore, it is not only an essential element of life, but also an essential natural resource. Due to the rapid population increase and economic development, the demand for groundwater resources for agricultural, industrial, and potable uses has been increasing [4]. Because groundwater is a limited resource, it is necessary to devise effective and efficient plans to use it based on an understanding of the behavior of groundwater systems and identification of the current status of the local groundwater system through groundwater exploration [5].

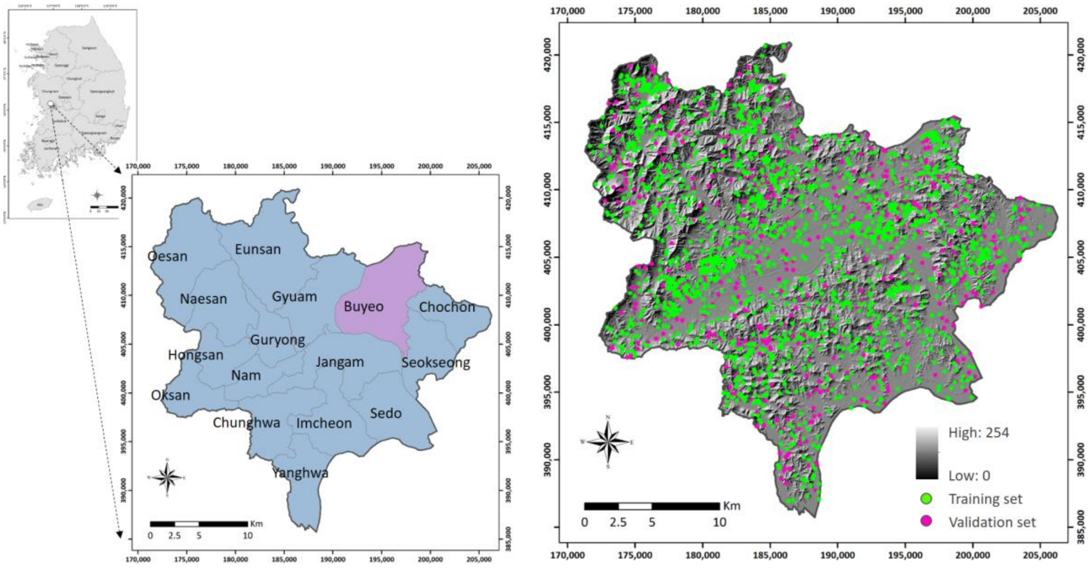
Traditional methods of exploring groundwater, which is a hidden natural resource, include drilling, geophysical, geological, and hydro-geological methods. However, such methods entail large expenses and the use of time and human resources for field surveys [6,7]. Groundwater potential maps (GPMs), based on geographic information system (GIS) and remote sensing (RS) data, have been widely used to solve this problem. GIS offers suitable alternatives for the effective management of large and complex geospatial databases [8]. In addition, it can be useful for groundwater exploitation and groundwater resource conversion, as it provides insights into the future availability of groundwater resources [9,10].

GPM has been applied with various methods, including the frequency ratio (FR) [4,11–13], logistic regression (LR) [14,15], weight of evidence [15,16], multi-criterion decision analysis [17,18], evidential belief function [19–21], index of entropy [12,22], and certainty factor [18]. With the recent rapid development of information technology and database technology, data mining algorithms are now being applied to diverse areas beyond information technology [23]. The fields of geology and hydrogeology have also widely used artificial neural networks, random forests, support vector machines, and decision trees for mapping landslide susceptibility [24–28], gullies [29], mineral potential [30–32], groundwater potential [16,20,33–35], and groundwater levels [36–38]. In recent years, techniques such as K-nearest neighbor (KNN) [39], linear discriminant analysis (LDA) [40], multivariate adaptive regression splines (MARS) [41,42], and quadratic discriminant analysis (QDA) [9,43] have also been used.

As shown in previous studies, diverse data mining techniques can be employed, but few such studies have been performed. Very few studies have attempted to create GPMs using the MARS model. The purpose of this study was to use the MARS model, a data mining technique, with the widely used LR model to create GPMs. The model performance of these GPMs was comparatively analyzed using receiver operating characteristic (ROC) curves. The ultimate aim was to evaluate the efficacy of the MARS model for creating GPMs.

2. Study Area

This study was conducted at a site in Buyeo-gun, Chungcheongnam-do, Korea, with a surface area of 625 km^2, located between 127°03′ and 126°44′ east longitude and 36°04′ and 36°23′ north latitude (Figure 1). The total population of Buyeo-gun was 71,143 in 2015, of which 31.2% or 22,213 individuals were engaged in farming [44]. The elevation of the study area is 0–640 m, and 72.8% of the overall area is formed as lowland with an elevation of 100 m or less. The study area is a basin with a high temperature and large daily temperature range in the summer, as well as large amounts of dew and fog due to the influence of the Geumgang River. Based on 2015, the annual precipitation is 848.8 mm and more than half of the annual precipitation occurs in the summer. The annual mean temperature is 12.9 °C, with a maximum temperature of 35.8 °C in the summer and a minimum winter temperature of −14.2 °C [44]. In terms of ground cover, most of the study area is composed of agricultural (40.1%) and forest (47.0%) areas. The lithology indicates that 52.05% of the study area is covered with metamorphic rock. This study area contains one river and 51 streams. As of 2015, the water and sewer distribution rates in the study area were 73.6% and 50.6%, respectively. In 2015, Buyeo-gun used 35,899,226 m^3 of groundwater annually, which is about 8% of the total groundwater use of Chungcheongnam-do (475,376,469 m^3/year). Most of this water is used for farming (about 71%; [45]). Considering that other cities and districts in Chungcheongnam-do only use about 7% of groundwater annually, Buyeo-gun has a relatively high dependence on groundwater.



(**a**) (**b**)

Figure 1. Location of the study area. (**a**) administrative map showing one town and 15 townships; (**b**) groundwater well locations divided into training and validation datasets.

3. Materials and Methods

In this study, a GPM was created through the three major steps described below. The first step was spatial database construction, in which a spatial database was created containing groundwater well locations and groundwater influence factors. The second step was groundwater potential assessment. LR and MARS models were used to analyze the relationships between well location and groundwater influence factors, and a GPM from each model was created for the overall study area. The third step was the validation process. The performance of the GPM created by each model was evaluated using ROC curves. A flow chart of the methodology used in this study is presented in Figure 2.

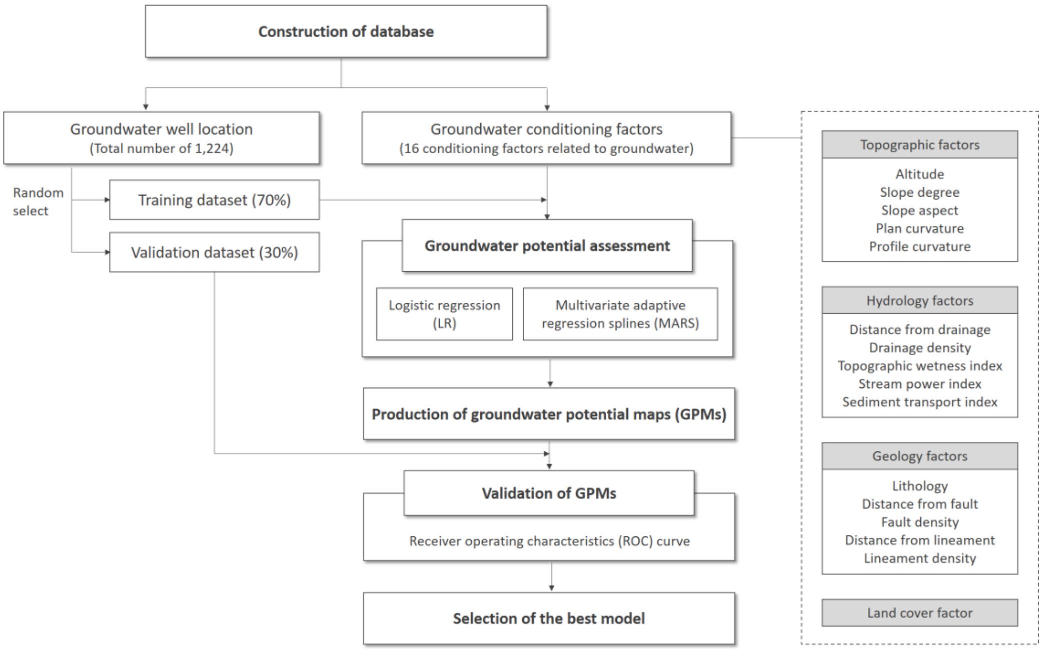


Figure 2. Flow chart of the methodology used in this study.

3.1. Data Preparation

3.1.1. Well Data

The groundwater well data used in this study were collected from extensive field surveys and governmental reports. Well water was used for a variety of purposes including livestock, farming, and human drinking water. The groundwater yield was calculated from the results of a pumping test of a groundwater well. The groundwater potential was based on the prediction of the best potential for groundwater extraction in the study area [9]. Based on previous studies and groundwater productivity reports, an actual pumping test was conducted using the groundwater well data for the study area. The high productivity value was based on a yield value $\geq$70 m^3/d. The groundwater productivity data from 1224 wells were selected and randomly divided into a training dataset containing 70% or 859 wells and a validation dataset with 30% or 365 wells. In addition, it was necessary to obtain sampling data from areas without groundwater wells. The data for the same number of groundwater wells (859) were selected as non-well occurrence data and allocated a value of 0 for application to the LR and MARS models. Figure 1 shows the locations of the groundwater well data used in this study.

3.1.2. Groundwater Influencing Factors

As presented in Table 1 and Figure 3, 16 groundwater influence factors were used in this study. These factors were largely divided into topographical, hydrological, geological, and land cover factors. Groundwater influence factors were created using ArcGIS 10.2 software (ESRI, Redlands, CA, USA), and were converted into a raster file with a spatial resolution of 10 × 10 m prior to use for groundwater potential assessment.

Topographic factors included altitude, slope degree, slope aspect, plan curvature, and profile curvature. Areas with different elevations have notable differences in weather and climatic conditions, which lead to differences in soil conditions and vegetation [46]. The slope degree is a factor that is mainly used to determine groundwater recharge processes, as gentle slope areas have a low surface runoff and high rates of percolation, while the opposite is true for high slope areas [19]. The slope aspect is a factor related to precipitation direction and physiographic trends, and it affects the soil water content [12]. Curvature represents the morphology of the topography, and is composed of three aspects: profile, plan, and total, the latter of which combines profile and plan. Profile curvature and plan curvature mainly affect the acceleration and deceleration of flow, as well as flow convergence and divergence, on the ground surface [22]. These factors were extracted from the digital elevation model (DEM) using a spatial analyst tool of ArcGIS 10.2 software. The DEM was created using contour lines and points extracted from a 1:5000 digital map provided by the Korean National Geographic Information Institute. The ArcGIS triangular irregular network (TIN) module was used for this process, and the generated TIN was converted into a raster file with a pixel size of 10 m.

Hydrologic factors such as the topographic wetness index (TWI), stream power index (SPI), sediment transport index (STI), distance from drainage, and drainage density were considered when estimating the flow of surface water and groundwater according to topographical factors. TWI is a secondary topographic index that has been used to describe spatial moisture patterns and explain the effects of topographic conditions on these patterns [47]. It plays an important role in influencing the movement and accumulation of runoff on the ground surface [11]. SPI is a factor that estimates the degree of slope erosion due to flowing water. TWI and SPI are calculated using the following equations [47]:

$$\mathrm{TWI} = \ln\left(\frac{A_s}{tan\beta}\right) \tag{1}$$

$$\mathrm{SPI} = A_s tan\beta \tag{2}$$

where A_s is the cumulative upslope area and β is the slope gradient. STI combines slope steepness and slope length, and is used to measure the sediment transport capacity of overland flow within the universal soil loss equation [48]. STI is calculated using the following equation [49]:

$$\mathrm{STI} = (\frac{A_s}{22.13})^{0.6}(\frac{sin\beta}{0.0896})^{0.6} \tag{3}$$

where A_s is the cumulative upslope area and β is the slope gradient. Drainage lines were used to create a drainage density map and distance map of the study area. Drainage density has an inverse relationship with permeability. A high drainage density decreases infiltration and increases surface runoff, and is therefore not appropriate for groundwater development [50]. The drainage density is calculated by dividing the surface area (km^2) by the sum of drainage lengths (km) for the corresponding cell. The drainage density and distance from drainage were determined using the line density tool and Euclidean distance tool in ArcGIS 10.2 software, respectively.

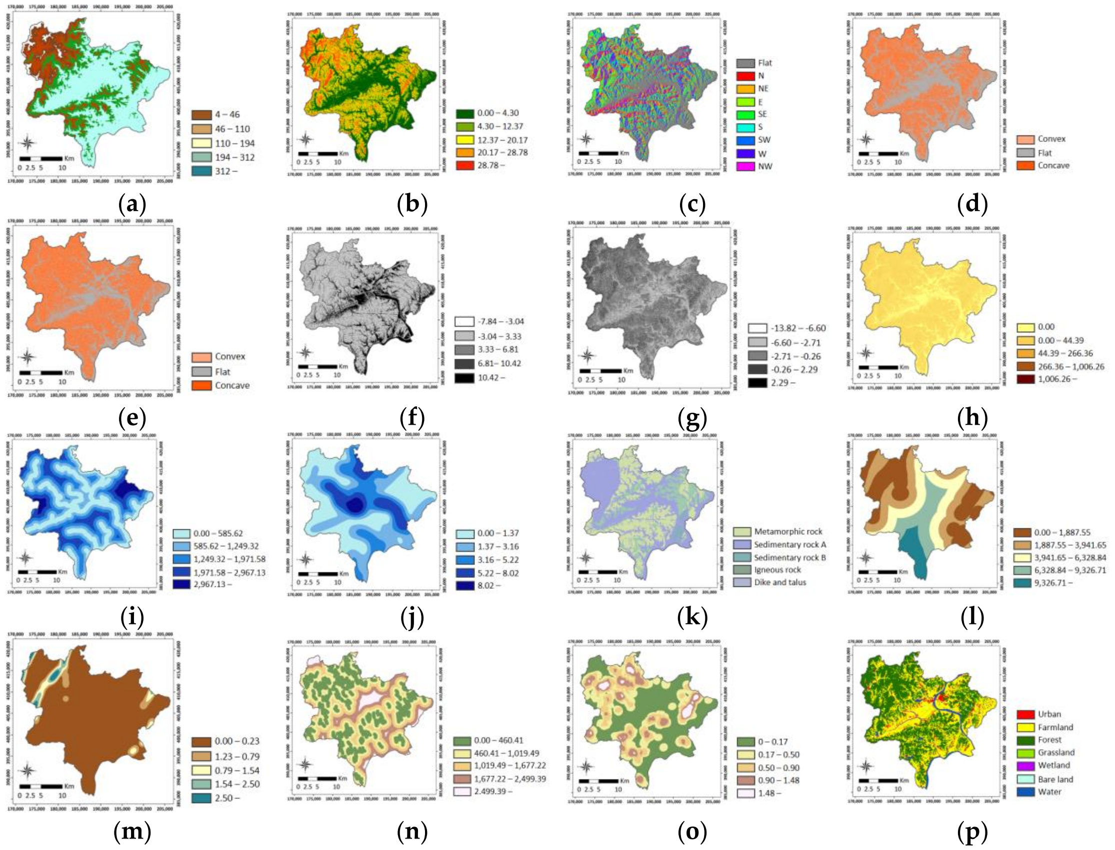
Geological factors affect the porosity and permeability of aquifer materials, and are thus considered indicators of hydrological features. Geological factors are composed of lithology, distance from fault, fault density, distance from lineament, and lineament density. These factors were determined using a digital geological map at a 1:50,000 scale, obtained from the Korea Institute of Geoscience and Mineral Resources.

Table 1. Data sources used in this study.

Category	Factor	Source	Scale (Resolution)	GIS and Data Type
Well location		National research paper Local research paper Field survey		Point
Topographical factors	Altitude	Topographic digital map [1]	1:5000	Polyline, point
	Slope degree Slope aspect Plan curvature Profile curvature	Digital elevation map	10 × 10 m	Raster
Hydrological factors	Topographic wetness index Stream power index Sediment transport index Distance from drainage Drainage density	Digital elevation map	10 × 10 m	Raster
Geological factors	Lithology	Geology map [2]	1:50,000	Polygon
	Distance from fault Fault density	Geology map	1:50,000	Polyline
	Distance from lineament Lineament density	Hill-shaded map	10 × 10 m	Raster
Land cover	Land cover	Land cover map [3]	1:25,000	Polygon

[1] Topographic digital maps were obtained from the National Geographic Information Institute, Korea; [2] Geology maps were obtained from the Korea Institute of Geoscience and Mineral Resources; [3] Land cover maps were obtained from the Ministry of Environment, Korea.

The study area was divided into 37 lithology units according to the type of lithology and geological age. In this study, these lithology units were classified according to their characteristics into metamorphic rock, sedimentary rock A, sedimentary rock B, igneous rock, and dike and talus. Here, sedimentary rocks were classified based on permeability. Sedimentary rock A is permeable rock made of sandstone or gravel, whereas sedimentary rock B is impermeable rock made of shale or clay. Faults extracted from the digital geology map were used to determine the distance from the fault and fault density. A lineament is defined as a straight or slightly curved surface feature of natural origin directly observed from the image [51,52]. Because lineaments are related to discontinuities such as joints, faults, and folds, they have been used for structural analysis, lithological relationship analysis, and groundwater productivity assessment [13]. In this study, lineaments were extracted from a hill-shaded map created from the DEM using the geological and geophysical analysis tool in Geomatica 2016 software (PCI Geomatics, Markham, ON, Canada). The hill-shaded map was created by combining images in three directions where the sun altitude was 45° and the sun azimuth was 45°, 90°, or 135°. The extracted lineament lines were used to determine the distance from the lineament and lineament density using the Euclidean distance tool and line density tool in ArcGIS 10.2 software.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**) (**j**) (**k**) (**l**) (**m**) (**n**) (**o**) (**p**)

Figure 3. Factors influencing groundwater. (**a**) altitude; (**b**) slope degree; (**c**) slope aspect; (**d**) plan curvature; (**e**) profile curvature; (**f**) topographic wetness index; (**g**) stream power index; (**h**) sediment transport index; (**i**) distance from drainage; (**j**) drainage density; (**k**) lithology; (**l**) distance from fault; (**m**) fault density; (**n**) distance from lineament; (**o**) lineament density; and (**p**) land cover.

Land cover represents the biological state of a geographic feature on the surface of the earth, and has been used to demarcate groundwater availability [22]. The type of land cover contributes to variation in the soil condition, and discontinuities resulting from this variation affect the occurrence, storage, and movement of groundwater. In this study, land cover factors were constructed using a digital land cover map provided by the Ministry of Environment. The digital land cover map was prepared on a 1:25,000 scale using Korea Multi-Purpose SATellite-2, and this map included 22 land cover categories that were reclassified into seven groups including urban, farmland, forest, grassland, wetland, bare land, and water.

3.2. Groundwater Potential Mapping

In this study, groundwater potential assessment was analyzed using the LR and MARS models. Analysis of the LR and MARS models was performed using the "glm" and "earth" packages in R 3.3.0 software (R Foundation for Statistical Computing, Vienna, Austria), respectively.

3.2.1. Logistic Regression

The LR model is a type of multivariate regression used to explain the relationships among a dichotomous dependent variable coded into 0 and 1, and one or more categorical or numerical independent variables [53]. In this study, the dependent variable was a binary variable indicating the presence or absence of groundwater wells, with a value of 1 or 0, respectively. Independent variables were the groundwater influencing factors that affect the groundwater wells. In general, the LR model can be expressed as follows [14,21]:

$$P = \frac{e^z}{(1 + e^z)} \tag{4}$$

where P is the probability of an occurrence and Z is the linear combination function of the independent variables showing a linear relationship. Z can be expressed as follows:

$$\mathrm{Z} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{5}$$

where α is the intercept, n is the number of independent variables, β_n represents the regression coefficients that represent the contribution of each independent variable to the probability value (P), and x_n represents the independent variables. In Equation (5), the value of Z ranges from $-\infty$ to $+\infty$. Positive regression coefficients indicate that the dependent variable has a positive correlation with the independent variables, whereas negative regression coefficients indicate that the independent variables have a negative effect on the dependent variable. However, if the dependent variable is a binary variable divided into presence and absence, as in this study, the value of the dependent variable is coded as 1 and 0. Thus, the predicted value of the dependent variable is a probability estimate. The probability value has an upper limit of 1 and a lower limit of 0, and the relationship between the dependent variable and this probability cannot be expressed as a linear function. Therefore, the upper and lower limits of the probability are removed by converting the probability into a logit function. The relationship between the dependent variable and logit can be expressed as a linear function. Probability can be converted into a logit function using the following equation:

$$Logit(P) = ln\frac{P}{1 - P} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{6}$$

where $\frac{P}{1-P}$ is the odds or likelihood ratio representing the ratio between the probability P that the dependent variable is present and the probability 1 – P that the dependent variable is absent. The natural logarithm of odds is called logit(P) and is a linear function of the independent variables ranging from $-\infty$ to $+\infty$. To be more precise, if the value of the probability P increases, the value of logit(P) also increases [14].

3.2.2. Multivariate Adaptive Regression Splines Model

The MARS model is a statistical method introduced by Friedman (1991) that is used to fit the relationship between the dependent and independent variables. It is a nonlinear and nonparametric regression method that combines classic linear regression, the mathematical construction of splines, binary recursive partitioning, and brute and intelligent algorithms [54,55]. A benefit of this method is that specific assumptions of the underlying functional relationship between the independent and dependent variables are unnecessary [56,57]. The MARS model predicts a function using linear combinations and interactions of the adaptive piecewise linear regression known as the "basis function (BF)". Accordingly, f(X) of the MARS model can be expressed by the following equation [42,57]:

$$f(\mathrm{X}) = \beta_0 + \sum_{i=1}^{n} \beta_i \lambda_i(X) \tag{7}$$

where β_0 is a constant, β_i is the coefficient of the ith BF, $\lambda_i(X)$ is a BF, and n is the number of BFs in the model. All of the coefficients were estimated using the least-squares method. The BFs are functions that take the following form [58]:

$$\max(0, x - \alpha)\; or \max(0, \alpha - x) \tag{8}$$

where x is an independent variable and α is a constant corresponding to a knot (or hinge). Two adjacent splines intersect at the knot to maintain the presence of the BF [33]. The MARS model was developed in two steps: the forward stepwise algorithm and the backward stepwise algorithm. The first step, the forward stepwise algorithm, adds BFs to Equation (1) and finds potential knots to obtain a better model performance. However, obtaining too many BFs in this process can result in overfitting the MARS model. The second step, the backward stepwise algorithm, is used to lessen this problem. In this step, redundant BFs that have the smallest contributions to the model are removed from the BFs used in the forward stepwise algorithm to find the best sub-model. Generalized cross-validation (GCV) is used to remove redundant BFs from the MARS model, and is calculated as follows [23,37]:

$$\mathrm{GCV} = \frac{\frac{1}{N}\sum_{i=1}^{N}[y_i - f(X_i)]^2}{\left[1 - \frac{M + d \times \frac{M-1}{2}}{N}\right]^2} \tag{9}$$

where N is the amount of data, $f(X_i)$ is the predicted value of the MARS model, M is the number of BFs, and d is the penalizing parameter. If the value of d is large, it can result in a small number of knots being used. The optimum value of d is considered to be in the range of $2 \leq d \leq 4$ [56]. In this study, a default value of 3 was used for d.

4. Results

4.1. Preliminary Analysis

In this study, multicollinearity and FRs were analyzed as preliminary analyses prior to groundwater potential assessment.

4.1.1. Multicollinearity Analysis

Multicollinearity refers to a linear relationship that exists among two or more variables. If multicollinearity exists among the independent variables during regression analyses, the variance of the regression coefficient increases. Error also increases with multicollinearity, reducing the accuracy of the model's prediction. Multicollinearity can be assessed by various means, and tolerance (TOL) and variance inflation factor (VIF) assessments were used in this study. TOL and VIF indicate multicollinearity among independent variables when their values are ≤0.1 and ≥10, respectively [59]. The results of the multicollinearity analysis on the 16 independent variables used in this study are presented in Table 2. This analysis showed that the TOL and VIF of all variables used in this study were ≥0.1 and ≤10, respectively. This suggests that there was no problem of multicollinearity among the independent variables used in this study, so LR analyses were performed with all of the variables.

Table 2. Multicollinearity diagnostic indices for independent variables.

Factor	Tolerance	VIF
Altitude	0.447	2.235
Slope degree	0.223	4.483
Slope aspect	0.522	1.916
Plan curvature	0.589	1.698
Profile curvature	0.747	1.338

Table 2. *Cont.*

Factor	Tolerance	VIF
Topographic wetness index	0.166	6.010
Stream power index	0.180	5.543
Sediment transport index	0.380	2.629
Distance from drainage	0.604	1.655
Drainage density	0.609	1.641
Lithology	0.943	1.061
Distance from fault	0.833	1.201
Fault density	0.762	1.312
Distance from lineament	0.492	2.031
Lineament density	0.497	2.012
Land cover	0.845	1.183

4.1.2. Spatial Relationship between Groundwater Wells and Influence Factors

The FR was used to analyze the probabilistic relationship between groundwater wells and groundwater influence factors in this study. FR is defined as the ratio between areas in which groundwater wells occur to the total study area, and is calculated using the following equation [18]:

$$\text{FR} = \frac{a/b}{c/d} \tag{10}$$

where a is the number of pixels with groundwater wells for each groundwater influence factor, b is the total number of groundwater wells in the study area, c is the number of pixels in the factor's class, and d is the total number of pixels in the study area. FR is considered to show an average relationship if its value is 1, high correlation if larger than 1, and low correlation if lower than 1 [60]. Among groundwater influence factors, FR values for continuous factors (e.g., altitude) were calculated after dividing the values into nine interclasses by quantile classification.

The results of the FR analysis are presented in Table 3. The value of FR for altitude was larger than 1 in classes 7–59. The value of FR was larger than 1 when the slope was 10.76 degrees or below, and was highest at 2.92, in the 2.96–6.72 class. Regarding the slope, the values of FR were relatively high for flat, northeast-facing, southeast-facing, and southwest-facing slopes. The value of FR for plan curvature and profile curvature was highest in the flat class and lowest in the convex class. Regarding the TWI, the value of FR was high at 2.42 and 2.09 in the −5.20–0.57 and 3.45–5.01 classes, respectively, and low at 0.19 in the −7.84–−5.2 class. SPI had an FR value larger than 1 when the value of SPI was lower than −1.60, except in the −8.60–−6.26 class. The FR value for STI was high at 1.16 and 1.14 in the 0 and 1–14.80 classes, respectively. Regarding the distance from drainage, the FR value was highest at 1.17 in the 917.47–1190.76 and 1483.56–1834.94 classes and lowest at 0.83 in the 0–195.206 class. For drainage density, the highest FR value was found in the 0.18–0.73 class. For lithology, FR was larger than 1 in the igneous rock and dike and talus classes, indicating a high probability of well occurrence. In the case of distance from a fault, the value of FR was highest at 2.07 in the 2997.87–3997.16 class and lowest at 0.41 in 0.00–777.23 class. The value of FR for fault density was highest in the 0 class. For distance from a lineament, the highest FR value of 1.29 was found in the 443.97–624.85 class and the lowest FR value in the 0.00–131.55 class (0.58). For lineament density, the highest FR value was found in the 0.46–0.60 class (1.31) and lowest in the >1.09 class (0.58). For land cover, the values of FR were high at 2.48, 1.95, and 1.79 in the urban, farmland, and bare land classes, respectively, indicating a high probability of well occurrence.

Table 3. Spatial relationships between groundwater wells and groundwater influencing factors determined using the frequency ratio model.

Factor	Class	No. of Pixels for Domain	% of Domain	No. of Wells	% of Wells	Frequency Ratio
Altitude	4–7	739,906	11.83	31	3.61	0.30
	7–13	758,490	12.13	121	14.09	1.16
	13–23	694,227	11.10	192	22.35	2.01
	23–39	706,011	11.29	186	21.65	1.92
	39–59	685,604	10.96	118	13.74	1.25
Altitude	59–85	671,769	10.74	76	8.85	0.82
	85–124	673,522	10.77	53	6.17	0.57
	124–189	662,166	10.59	68	7.92	0.75
	>189	661,265	10.58	14	1.63	0.15
Slope degree	0.00	1,346,466	21.53	219	25.49	1.18
	0.00–2.96	540,629	8.65	143	16.65	1.93
	2.96–6.72	848,902	13.58	341	39.70	2.92
	6.72–10.76	578,596	9.25	110	12.81	1.38
	10.76–14.52	584,424	9.35	26	3.03	0.32
	14.52–18.29	595,343	9.52	7	0.81	0.09
	18.29–22.05	570,843	9.13	7	0.81	0.09
	22.05–27.16	592,491	9.48	4	0.47	0.05
	>27.16	595,266	9.52	2	0.23	0.02
Slope aspect	Flat	1,346,466	21.53	219	25.49	1.18
	N	550,564	8.80	47	5.47	0.62
	NE	666,731	10.66	96	11.18	1.05
	E	656,324	10.50	83	9.66	0.92
	SE	682,654	10.92	109	12.69	1.16
	S	635,319	10.16	82	9.55	0.94
	SW	664,071	10.62	107	12.46	1.17
	W	530,399	8.48	60	6.98	0.82
	NW	520,432	8.32	56	6.52	0.78
Plan curvature	Convex	1,912,371	30.58	166	19.32	0.63
	Flat	2,574,062	41.17	505	58.79	1.43
	Concave	1,766,527	28.25	188	21.89	0.77
Profile curvature	Convex	2,035,086	32.55	212	24.68	0.76
	Flat	1,992,425	31.86	381	44.35	1.39
	Concave	2,225,449	35.59	266	30.97	0.87
Topographic wetness index	−7.84–5.2	728,730	11.65	19	2.21	0.19
	−5.20–0.57	489,555	7.83	163	18.98	2.42
	0.57–1.65	1,308,964	20.93	57	6.64	0.32
	1.65–2.49	1,336,212	21.37	125	14.55	0.68
	2.49–3.45	717,823	11.48	176	20.49	1.78
	3.45–5.01	525,554	8.40	151	17.58	2.09
	5.01–7.89	649,627	10.39	104	12.11	1.17
	7.89–10.30	436,178	6.98	51	5.94	0.85
	>10.30	60,317	0.96	13	1.51	1.57
Stream power index	−3.82–8.60	668,024	10.68	197	22.93	2.15
	−8.60–6.26	720,080	11.52	36	4.19	0.36
	−6.26–3.26	713,700	11.41	125	14.55	1.27
	−3.26–1.60	740,146	11.84	131	15.25	1.29
	−1.60–0.71	713,224	11.41	92	10.71	0.94
	−0.71–0.07	750,100	12.00	67	7.80	0.65
	0.07–0.73	674,154	10.78	66	7.68	0.71
	0.73–1.62	656,256	10.50	64	7.45	0.71
	>1.62	617,276	9.87	81	9.43	0.96
Sediment transport index	0.00	2,608,602	41.72	415	48.31	1.16
	0.00–14.80	2,731,789	43.69	426	49.59	1.14
	14.80–29.60	731,160	11.69	15	1.75	0.15
	29.60–44.39	123,380	1.97	2	0.23	0.12
	44.39–59.19	32,564	0.52	1	0.12	0.22
	59.19–73.99	12,138	0.19	0	0.00	0.00
	73.99–103.59	7,863	0.13	0	0.00	0.00
	103.59–162.78	3,257	0.05	0	0.00	0.00
	>162.78	2,207	0.04	0	0.00	0.00

Table 3. *Cont.*

Factor	Class	No. of Pixels for Domain	% of Domain	No. of Wells	% of Wells	Frequency Ratio
Distance from drainage	0–195.206	674,229	10.78	77	8.96	0.83
	195.206–429.45	740,195	11.84	107	12.46	1.05
	429.45–663.70	705,573	11.28	90	10.48	0.93
	663.70–917.47	710,245	11.36	104	12.11	1.07
	917.47–1190.76	706,354	11.30	114	13.27	1.17
	1190.76–1483.56	674,175	10.78	89	10.36	0.96
	1483.56–1834.94	684,110	10.94	110	12.81	1.17
	1834.94–2342.47	694,454	11.11	105	12.22	1.10
	>2342.47	663,625	10.61	63	7.33	0.69
Drainage density	0.00–0.18	647,899	10.36	67	7.80	0.75
	0.18–0.73	764,117	12.22	145	16.88	1.38
	0.73–1.24	696,058	11.13	121	14.09	1.27
	1.24–1.74	735,696	11.77	109	12.69	1.08
	1.74–2.47	692,591	11.08	81	9.43	0.85
	2.47–3.21	692,813	11.08	69	8.03	0.72
	3.21–4.31	687,447	10.99	92	10.71	0.97
	4.31–5.73	669,019	10.70	93	10.83	1.01
	>5.73	667,320	10.67	82	9.55	0.89
Lithology	Metamorphic rock	1,993,700	31.88	235	27.36	0.86
	Sedimentary rock A	3,251,216	51.99	440	51.22	0.99
	Sedimentary rock B	86,184	1.38	1	0.12	0.08
	Igneous rock	892,839	14.28	174	20.26	1.42
	Dike and talus	29,019	0.46	9	1.05	2.26
Distance from fault	0.00–777.23	680,735	10.89	38	4.42	0.41
	777.23–1498.94	725,020	11.59	73	8.50	0.73
	1498.94–2220.65	738,889	11.82	78	9.08	0.77
	2220.65–2997.87	689,019	11.02	122	14.20	1.29
	2997.87–3997.16	684,571	10.95	195	22.70	2.07
	3997.16–5163.00	700,837	11.21	120	13.97	1.25
	5163.00–6550.90	675,093	10.80	84	9.78	0.91
	6550.90–8493.97	685,678	10.97	74	8.61	0.79
	>8493.97	673,118	10.76	75	8.73	0.81
Fault density	0.00	4,958,065	79.29	764	88.94	1.12
	0.00–0.05	411,100	6.57	43	5.01	0.76
	0.05–0.13	145,678	2.33	12	1.40	0.60
Fault density	0.13–0.28	137,777	2.20	13	1.51	0.69
	0.28–0.47	128,683	2.06	5	0.58	0.28
	0.47–0.74	118,625	1.90	8	0.93	0.49
	0.74–1.13	118,172	1.89	5	0.58	0.31
	1.13–1.83	117,638	1.88	8	0.93	0.50
	>1.83	117,222	1.87	1	0.12	0.06
Distance from lineament	0.00–131.55	673,029	10.76	54	6.29	0.58
	131.55–279.54	714,676	11.43	95	11.06	0.97
	279.54–443.97	750,304	12.00	120	13.97	1.16
	443.97–624.85	714,445	11.43	127	14.78	1.29
	624.85–855.05	722,406	11.55	92	10.71	0.93
	855.05–1151.03	698,358	11.17	100	11.64	1.04
	1151.03–1529.23	663,276	10.61	100	11.64	1.10
	1529.23–2022.53	658,337	10.53	92	10.71	1.02
	>2022.53	658,129	10.53	79	9.20	0.87
Lineament density	0.00	1,943,652	31.08	268	31.20	1.00
	0.00–0.13	583,577	9.33	76	8.85	0.95
	0.13–0.23	578,400	9.25	85	9.90	1.07
	0.23–0.35	593,925	9.50	104	12.11	1.27
	0.35–0.46	540,809	8.65	48	5.59	0.65
	0.46–0.60	518,321	8.29	93	10.83	1.31
	0.60–0.80	517,855	8.28	83	9.66	1.17
	0.80–1.09	495,919	7.93	64	7.45	0.94
	>1.09	480,502	7.68	38	4.42	0.58
Land cover	Urban	322,571	5.16	110	12.81	2.48
	Farmland	2,504,403	40.05	672	78.23	1.95
	Forest	2,941,173	47.04	53	6.17	0.13
	Grassland	134,727	2.15	6	0.70	0.32
	Wetland	67,362	1.08	1	0.12	0.11
	Bare land	52,777	0.84	13	1.51	1.79
	Water	229,947	3.68	4	0.47	0.13

Based on the results of FR analysis, the classes of groundwater influence factors used in this study had different FR values, and some factors showed a broad range of values. For example, the slope degree had an FR range of 0.02–2.92. In addition, each groundwater influence factor had at least one class with an FR value larger than 1, showing a high correlation with well occurrence. Therefore, the 16 factors used in this study are appropriate for use as groundwater influence factors.

4.2. Groundwater Potential Assessment and Mapping

4.2.1. Application of the Logistic Regression Model

The results of analysis using the LR model are presented in Table 4. Among the independent variables used in this study, factors such as the slope degree, profile curvature 2 (flat), TWI, SPI, distance from drainage, lithology 3 (sedimentary rock B), lithology 4 (igneous rock), fault density, distance from lineament, land cover 3 (forest), land cover 5 (wetland), and land cover 7 (water) had significant effects on groundwater well occurrence at the 5% significance level. The results of the β coefficient, altitude, SPI, distance from drainage, and lineament density also had positive effects on groundwater well occurrence. However, the slope degree, TWI, STI, drainage density, distance from fault, and distance from lineament had negative β coefficient values, indicating negative effects on groundwater well occurrence. For categorical variables such as the slope aspect, plan curvature, profile curvature, lithology, and land cover, the plan curvature 3 (concave), profile curvature 2 (flat), lithology 2 (sedimentary rock A), lithology 4 (igneous rock), lithology 5 (dike and talus), and land cover 6 (bare land) classes had positive effects on groundwater well occurrence. On the other hand, some classes of slope aspect, plan curvature, profile curvature, lithology, and land cover had negative effects on groundwater well occurrence. In addition, the results of the LR model showed a null deviance of 2381.7 with 1717 degrees of freedom. The residual deviance was 1722.9 with 1684 degrees of freedom and an Akaike Information Criterion of 1790.9.

Table 4. β coefficients of groundwater influence factors used in the logistic regression model.

Factors	β	Std. Error	z Value	Pr (>|z|)
(Intercept)	5.016	1.816	2.762	0.006 *
Altitude	0.002	1.5×10^{-3}	1.419	0.156
Slope degree	−0.231	3.9×10^{-2}	−5.937	2.9×10^{-9} *
Slope aspect2 (N)	−1.412	1.015	−1.391	0.164
Slope aspect3 (NE)	−0.981	9.7×10^{-1}	−1.016	0.309
Slope aspect4 (E)	−1.250	1.013	−1.234	0.217
Slope aspect5 (SE)	−1.286	9.5×10^{-1}	−1.351	0.177
Slope aspect6 (S)	−1.126	1.013	−1.112	0.266
Slope aspect7 (SW)	−1.365	9.4×10^{-1}	−1.455	0.146
Slope aspect8 (W)	−1.384	1.026	−1.349	0.177
Slope aspect9 (NW)	−1.498	9.8×10^{-1}	−1.532	0.126
Plan curvature2 (Flat)	−0.036	2.7×10^{-1}	−0.135	0.893
Plan curvature3 (Concave)	0.278	2.5×10^{-1}	1.132	0.258
Profile curvatire2 (Flat)	0.629	2.4×10^{-1}	2.662	0.008 *
Profile curvature3 (Concave)	−0.051	1.9×10^{-1}	−0.272	0.785
Topographic wetness index	−0.367	1.4×10^{-1}	−2.682	0.007 *
Stream power index	0.351	1.4×10^{-1}	2.557	0.011 *
Sediment transport index	-5.8×10^{-4}	6.6×10^{-3}	−0.087	0.931
Distance from drainage	2.3×10^{-4}	9.5×10^{-5}	2.392	0.017 *
Drainage density	−0.020	3.3×10^{-2}	−0.609	0.543
Lithology2 (Sedimentary rock A)	0.172	1.5×10^{-1}	1.165	0.244
Lithology3 (Sedimentary rock B)	−2.873	1.077	−2.668	0.008 *
Lithology4 (Igneous rock)	0.581	1.9×10^{-1}	3.059	0.002 *
Lithology5 (Dike and talus)	0.187	6.5×10^{-1}	0.287	0.774
Distance from fault	-1.0×10^{-5}	2.3×10^{-5}	−0.444	0.657

Table 4. *Cont.*

Factors	β	Std. Error	z Value	Pr (>|z|)
Fault density	−0.717	3.0×10^{-1}	−2.43	0.015 *
Distance from lineament	-2.3×10^{-4}	1.2×10^{-4}	−1.971	0.049 *
Lineament density	0.227	2.4×10^{-1}	0.938	0.348
Land cover2 (Farmland)	−0.332	2.1×10^{-1}	−1.613	0.107
Land cover3 (Forest)	−1.992	2.8×10^{-1}	−7.064	1.6×10^{-12} *
Land cover4 (Grassland)	−0.962	5.9×10^{-1}	−1.618	0.106
Land cover5 (Wetland)	−3.055	1.068	−2.86	0.004 *
Land cover6 (Bare land)	0.980	7.4×10^{-1}	1.321	0.187
Land cover7 (Water)	−2.717	5.8×10^{-1}	−4.659	3.2×10^{-6} *

* $p < 0.05$.

4.2.2. Application of Multivariate Adaptive Regression Splines Model

The optimal MARS model included 25 terms, and the GCV was 0.165. The MARS model generates the optimal model by only selecting the necessary independent variables [55]. Of the 16 independent variables included in this study, only 10 variables (altitude, slope degree, distance from drainage, drainage density, lithology, distance from fault, fault density, distance from lineament, lineament density, and land cover) were used to construct the optimal model. Categorical variables, such as lithology and land cover, only included classes, for example, sedimentary rock A, sedimentary rock B, igneous rock, forest, wetland, and water. Based on an analysis of the MARS model, a BF was created for each independent variable and each BF had a different β coefficient. Continuous variables have one or more constants corresponding to a knot within the variable, which lead to different effects on groundwater well occurrence.

In the MARS model, it is possible to estimate the relative importance of variables. The results of the selections and contributions of various independent variables are shown in Table 5. Here, nsubset is a criterion of the number of model subsets that include each variable. Variables that are included in more subsets are considered more important. GCV provides a generalized cross-validation of the model. The GCV criterion first calculates the decrease in the GCV for each subset relative to the previous subset. Then, for each variable, it sums these decreases over all subsets that include that variable. Finally, for ease of interpretation, the summed decreases are scaled so the largest summed decrease is 100. In addition, RSS is the residual sum-of-squares of the mode. In the case of RSS and GCV, variables that cause larger net decreases are considered more important [58,61].

Based on Table 5, land cover 3 (forest) was the most important variable explaining the spatial distribution of groundwater wells in the study area, followed by altitude, slope degree, land cover 7 (water), and distance from the fault. These independent variables had lower values (<0.15) for the frequency ratio compared with other variables. In addition, altitude had no significant effect on groundwater well occurrence at the 5% significance level in the LR model. The influence of the independent variables differed depending on the result of FR, LR, and the MARS model.

Table 5. The contributions of various independent variables in the MARS model.

Factors	Nsubset	GCV	RSS
Land cover3 (Forest)	24	100	100
Altitude	23	60.8	65.3
Slope degree	22	49.2	55.4
Land cover7 (Water)	18	22.6	34.3
Distance from fault	17	16.9	30.6
Land cover5 (wetland)	15	15.8	28.7
Distance from lineament	13	13.3	26
Lineament density	13	13.3	26

Table 5. *Cont.*

Factors	Nsubset	GCV	RSS
Lithology3 (Sedimentary rock B)	13	12.7	25.8
Distance from drainage	11	10.5	23.2
Lithology2 (Sedimentary rock A)	9	9.4	20.9
Drainage density	7	7.7	18.2
Lithology4 (Igneous rock)	3	5	11.9
Fault density	2	3.2	9.4

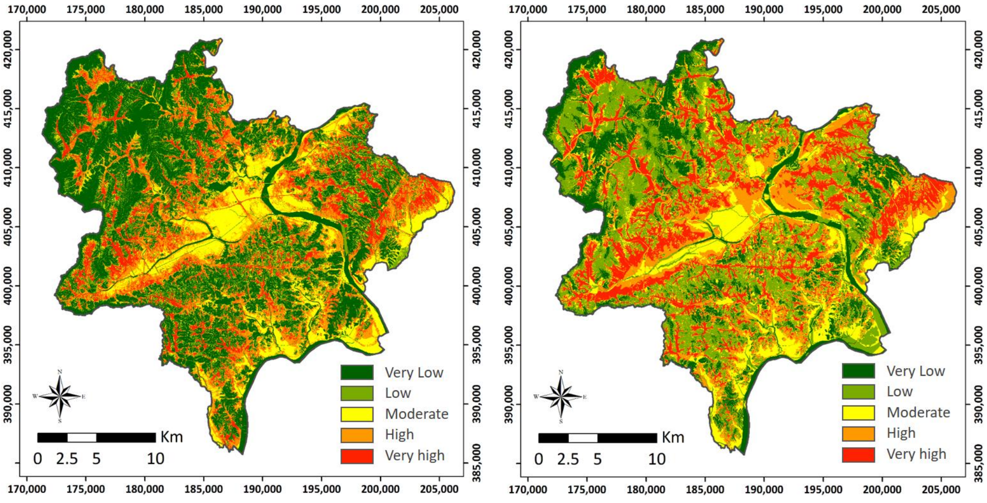
4.2.3. Groundwater Potential Mapping

The following equations were used to apply the analysis results of the LR and MARS models to the creation of a GPM for the overall study area:

$$\begin{aligned} GPM_{LR} = 5.016 &+ (0.002 \times Altitude) - (0.213 \times Slope\ degree) \\ &- (1.412 \times Slope\ aspect2\ [N]) - (0.981 \times Slope\ aspect3\ [NE]) \\ &- (1.250 \times Slope\ aspect4\ [E]) - (1.286 \times Slope\ aspect5\ [SE]) \\ &- (1.126 \times Slope\ aspect6\ [S]) - (1.365 \times Slope\ aspect7\ [SW]) \\ &- (1.384 \times Slope\ aspect8\ [W]) - (1.498 \times Slope\ aspect9\ [NW]) \\ &- (0.036 \times Plan\ curvature2\ [Flat]) \\ &+ (0.278 \times Plan\ curvature3\ [Concave]) \\ &+ (0.629 \times Profile\ curvature2\ [Flat]) \\ &- (0.051 \times Profile\ curvature3\ [Concave]) - (0.367 \times TWI) \\ &+ (0.251 \times SPI) - (5.79E - 04 \times STI) \\ &+ (2.28E - 04 \times Distance\ from\ drainage) \\ &- (0.020 \times Drainage\ density) \\ &+ (0.172 \times Lithology2\ [Sedimentary\ rock\ A]) \\ &- (2.873 \times Lithology3\ [Sedimentary\ rock\ B]) \\ &+ (0.581 \times Lithology4\ [Igneous\ rock]) \\ &+ (0.187 \times Lithology5\ [Dike\ and\ talus]) \\ &- (1.01E - 05 \times Distance\ from\ lineament) - (0.717 \times Fault\ density) \\ &- (2.28E - 04 \times Distance\ from\ lineament) \\ &+ (0.227 \times Lineament\ density) \\ &- (0.332 \times Land\ cover\ 2\ [Farmland]) \\ &- (1.992 \times Land\ cover\ 3\ [Forest]) \\ &- (0.962 \times Land\ cover\ 4\ [Grassland]) \\ &- (3.055 \times Land\ cover\ 5\ [Wetland]) \\ &+ (0.980 \times Land\ cover\ 6\ [Bareland]) \\ &- (2.717 \times Land\ cover\ 7\ [Water]) \end{aligned} \tag{11}$$

$$
\begin{aligned}
GPM_{MARS} = 3.241 & -(0.041\times\max(0,\ Altitude-13))-(0.046\times\max(0,\ 28\\
& -Altitude)+(0.303\times\max(Slope\ degree-5.15))-(0.370\\
& \times\max(Slope\ degree-5.44))+(0.061\times\max(Slope\ degree\\
& -12.18))-(0.001\times\max(Distance\ from\ drainage-150.33))\\
& -(0.001\times\max(2020-Distance\ from\ drainage))+(0.001\\
& \times\max(Distance\ from\ drainage-2020))-(0.115\\
& \times\max(Distance\ density-0.89))+(0.180\\
& \times\max(Distance\ density-1.85))-(0.070\\
& \times\max(Distance\ density-3.72))+(0.090\\
& \times Lithology2\ [Sedimentary\ rock\ A])-(0.232\\
& \times Lithology3\ [Sedimentary\ rock\ B])+(0.078\\
& \times Lithology4\ [Igneous\ rock])+(2.32E-04\\
& \times\max(Distance\ from\ fault-2,801.79)-(4.82E-04\\
& \times\max(Distance\ from\ fault-3,548.03)+(2.47E-04\\
& \times\max(Distance\ from\ fault-4,105.85)+(0.298\times\max(0.33\\
& -Fault\ density)-(0.01\times\max(226.72\\
& -Distance\ from\ lineament)-(0.143\times\max(0.677\\
& -Lineament\ density)-(0.280\times Land\ cover\ [Forest])-(0.292\\
& \times Land\ cover\ [Wetland])-(0.342\times Land\ cover\ [Water])
\end{aligned}
\tag{12}
$$

When the GPM was classified by the groundwater potential zone using four classification techniques including a natural break, quantile, equal interval, and geometrical interval, and when the distribution of training and validation groundwater wells in high and very high zones was comparatively analyzed, the quantile classification technique was most accurate [15]. Based on this finding, the GPMs created using the LR and MARS models were classified into very low, low, moderate, high, and very high groundwater potential zones using the quantile classification technique. Figure 4 shows the GPMs created by the LR and MARS models.



(**a**) (**b**)

Figure 4. Groundwater potential maps produced by the (**a**) LR and (**b**) MARS model.

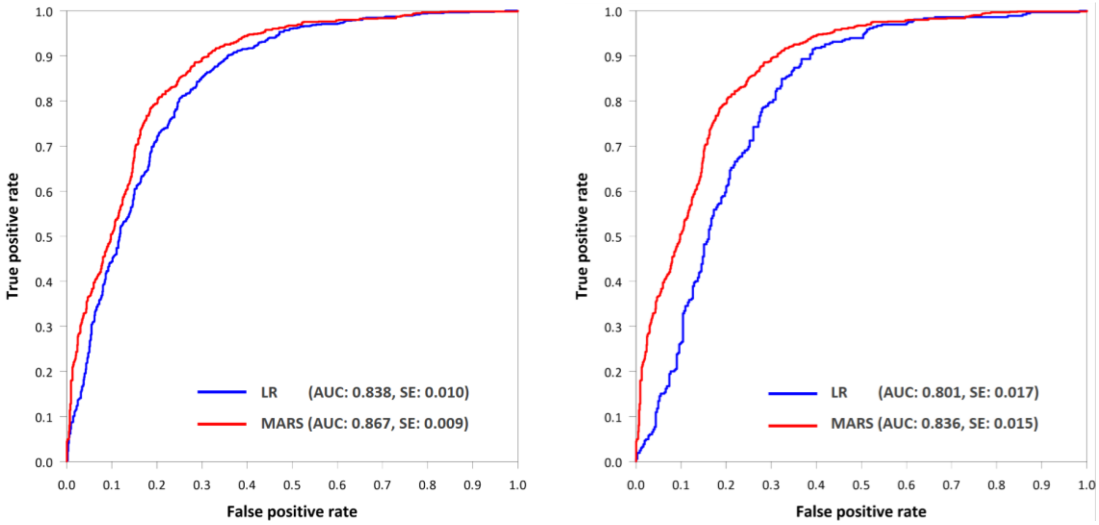
The surface area of the GPM created using the LR model in high and very high zones was 248 km^2, which is 39.7% of the overall surface area. In addition, the surface area of the GPM created with the

MARS model in high and very high zones was 251 km^2 (40.1%), which is slightly larger than that of the GPM created using the LR model. Comparing the surface areas of the GPM in each zone, the difference in the surface area of the very high zone (0.62 km^2) was not large. However, the surface area of the GPM created with the MARS model in the high zone was 1.78 km^2 larger than that in the GPM created using the LR model, a relatively large difference in surface area.

4.3. Validation and Comparison

ROC curves were used to evaluate the performance of the GPMs created in this study. An ROC curve is a scientific technique used to describe the efficiency of probabilistic and deterministic detection and prediction systems [62], and is formed by plotting the trade-off between the true positive rate (sensitivity) on the X-axis and the false positive rate (1-specificity) on the Y-axis. ROC curves can be divided into success rate curves and prediction rate curves according to the dataset used. A success rate curve is formed using a training dataset, and represents how well the model fits the groundwater wells observed. A prediction rate curve is formed using a validation dataset, and represents how well the model predicts groundwater wells [63]. The ROC curve can be used as a quantitative measure through the calculation of the area under the curve (AUC). The value of AUC is between 0.5 and 1.0, and a value closer to 1 indicates a model with a better predictive capability. AUC values can be evaluated as follows. Poor: 0.5–0.6, Average: 0.6–0.7, Good: 0.7–0.8, Very good: 0.8–0.9, and Excellent: 0.9–1.0 [64]. The ROC curves and AUC of GPMs created using the LR and MARS models are shown in Figure 5.

In our success rate curve, the AUC value of the GPMs created using the LR and MARS models were 0.838 and 0.867, respectively. Thus, the AUC value was 0.029 higher for the GPM created using the MARS model compared to the GPM created with the LR model. The AUCs of the prediction rate curves were similar, with the AUC (0.836) of the GPM created with the MARS model being slightly higher than the AUC (0.801) of the GPM created with the LR model. These results, showing that the GPMs created in this study have AUC values of 0.8 or above, indicate an excellent predictive capability for both models, with the MARS model performing better than the LR model.



(**a**) (**b**)

Figure 5. Results of model validation for each GPM. (**a**) success rate and (**b**) prediction rate curves.

5. Discussion and Conclusions

Groundwater is an important natural resource, and the spatial distribution of groundwater can be detected and predicted by creating GPMs using various factors to ensure its continued availability. In this study, the LR and MARS models were used to evaluate groundwater potential and create GPMs.

Groundwater well data (with high potential yields of ≥70 m^3/d) were classified into a training dataset (70%, 859 groundwater well locations) and validation dataset (30%, 365 groundwater well locations). This study used 16 groundwater influence factors for groundwater potential assessment, including topographic factors (altitude, slope degree, slope aspect, plan curvature, profile curvature), hydrologic factors (TWI, SPI, STI, distance from drainage, drainage density), geological factors (lithology, distance from fault, fault density, distance from lineament, lineament density), and land cover. Groundwater well locations and groundwater influence factors were applied to the LR and MARS models to analyze groundwater potential, and GPMs were created based on the results of this analysis. The accuracy of the models was tested using an ROC curve. The GPMs created in this study all exhibited AUC values of 0.8 or above, indicating an excellent model performance. The AUCs for the success rate and prediction rate curves of the GPM created with the MARS model were, respectively, 0.029 and 0.035 higher than those of the GPM created using the LR model.

It was also possible to estimate the groundwater potential in the MARS model. The results showed that land cover, altitude, slope degree, and distance from a fault made large contributions to groundwater occurrence. According to another study, altitude, TWI, distance from rivers, land cover, fault density, slope, and lithology were important factors [9,20,39]. These results were similar to our results. However, in our study, lithology, distance from rivers, and fault density were less important factors. This could be due to the study area conditions and method used.

Based on the results, the MARS model is more robust and has a better predictive capability than the LR model for the evaluation and mapping of groundwater potential in this study area. The MARS model has the following advantages compared to traditional regression-based analysis. The MARS model creates its final results by only selecting the important variables from the multiple variables used in the results [65]. This can effectively reduce the time needed for researchers to select the groundwater influence factors during GPM analysis. Even if unnecessary variables are used during this review process, the optimal variables can be selected using the MARS model. In addition, the MARS model is an easy-to-interpret model that can extract complex data in a computationally efficient manner for multivariate problems involving large volumes of data [57].

The GPMs created in this study suggest the possibility of groundwater occurrence in the study area. Through a comprehensive understanding of groundwater potential, GPMs can be used to drive the exploration of groundwater resources effectively and economically, and prevent undesirable effects due to water resource development. Therefore, GPMs can be useful for decision-makers and planners to devise plans for sustainable water resource management, ecologically friendly land use, and environmental preservation of the study area. However, it is still necessary to apply the MARS model in more diverse areas and to compare it with other models to make a reliable judgment of its efficacy. Additional detailed spatial data reflecting geological and hydrogeological conditions should also be used to analyze groundwater potential in the future.

Acknowledgments: This research was financially supported by the BK21 plus Project of the Graduate School of Earth Environmental Hazard System. In addition, this work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2017R1A2B2009033).

Author Contributions: Soyoung Park wrote the paper and analyzed the data; Se-Yeong Hamm managed the paperwork, Hang-Tak Jeon collected and prepared the input data; Jinsoo Kim suggested the idea for the study.

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

References

1. Fitts, C.R. *Groundwater Science*; Academic Press: San Diego, CA, USA, 2002.
2. Shahid, S.; Nath, S.K.; Roy, J. Groundwater potential modeling in a soft rock area using a GIS. *Int. J. Remote Sens.* **2000**, *21*, 1919–1924. [CrossRef]
3. Qadir, M.; Wichelns, D.; Raschid-Sally, L.; Minhas, P.S.; Drechsel, P.; Bahri, A.; McCornick, P.G.; Abaidoo, R.; Attia, F.; El-Guindy, S. *Water for Food, Water for Life: A Comprehensive Assessment of Water Management in Agriculture*; Molden, D., Ed.; IWMI & Earthscan: London, UK, 2007.

4. Mannap, M.A.; Nampak, H.; Pradhan, B.; Lee, S.; Soleiman, 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]
5. Bera, K.; Bandyopadhyay, J. Ground water potential mapping in Dulung watershed using remote sensing & GIS techniques, West Bengal, India. *Int. J. Sci. Res. Publ.* **2012**, *2*, 1–7.
6. Sander, P.; Chesley, M.M.; Minor, T.B. Groundwater assessment using remote sensing and GIS in a rural groundwater project in Ghana: lessons learned. *Hydrogeol. J.* **1996**, *4*, 40–49. [CrossRef]
7. Singh, A.K.; Prakash, S.R. An integrated approach of remote sensing, geophysics and GIS to evaluation of groundwater potentiality of Ojhala sub-watershed, Mirjapur district, UP, India. In Proceedings of the First Asian Conference on GIS, GPS, Aerial Photography and Remote Sensing, Bangkok, Thailand, 7–9 August 2002.
8. Waikar, M.L.; Nilawar, A.P. Identification of groundwater potential zone using remote sensing and GIS technique. *Int. J. Innov. Res. Sci. Eng. Technol.* **2014**, *3*, 12163–12174.
9. 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.* **2016**, 1–21. [CrossRef]
10. Reilly, T.E.; Dennehy, K.F.; Alley, W.M.; Cunningham, W.L. *Ground-Water Availability in the United States*; Circular 1323; U.S. Geological Survey: Reston, VA, USA, 2008.
11. Elmahdy, S.I.; Mohamed, M.M. Probabilistic frequency ratio model for groundwater potential mapping in Al Jaww plain, UAE. *Arab. J. Geosci.* **2015**, *8*, 2405–2416. [CrossRef]
12. Naghibi, S.A.; Pourghasemi, H.R.; Pourtaghi, Z.S.; Rezaei, A. Groundwater qanat potential mapping using frequency ratio and Shannon's entropy models in the Moghan watershed, Iran. *Earth Sci. Inform.* **2015**, *8*, 171–186. [CrossRef]
13. Oh, H.J.; Kim, Y.S.; Choi, J.K.; Park, E.; Lee, S. GIS mapping of regional probabilistic groundwater potential in the area of Pohang City, Korea. *J. Hydrol.* **2011**, *399*, 158–172. [CrossRef]
14. Ozdemir, A. Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). *J. Hydrol.* **2011**, *405*, 123–136. [CrossRef]
15. 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]
16. Corsini, A.; Cervi, F.; Ronchetti, F. Weight of evidence and artificial neural networks for potential groundwater spring mapping: An application to the Mt. Modino area (Northern Apennines, Italy). *Geomorphology* **2009**, *111*, 79–87. [CrossRef]
17. Adiat, K.A.N.; Nawawi, M.N.M.; Abdullah, K. Assessing the accuracy of GIS-based elementary multi criteria decision analysis as a spatial prediction tool—A case of predicting potential zones of sustainable groundwater resources. *J. Hydrol.* **2012**, *440*, 75–89. [CrossRef]
18. 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]
19. Mogaji, K.A.; Lim, H.S.; Abdullah, K. Regional prediction of groundwater potential mapping in a multifaceted geology terrain using GIS-based Dempster-Shafer model. *Arab. J. Geosci.* **2015**, *8*, 3235–3258. [CrossRef]
20. Naghibi, S.A.; Pourghasemi, H.R. A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping. *Water Resour. Manag.* **2015**, *29*, 5217–5236. [CrossRef]
21. 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]
22. 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]
23. Yao, D.; Yang, J.; Zhan, X. A novel method for disease prediction: hybrid of random forest and multivariate adaptive regression splines. *J. Comput.* **2013**, *8*, 170–177. [CrossRef]
24. Hong, H.; Pourghasemi, H.R.; Pourtaghi, Z.S. Landslide susceptibility assessment in Lianhua County (China): A comparison between a random forest data mining technique and bivariate and multivariate statistical models. *Geomorphology* **2016**, *259*, 105–118. [CrossRef]

25. Pham, B.T.; Pradhan, B.; Bui, D.T.; Prakash, I.; Dholakia, M.B. 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]
26. Saito, H.; Nakayama, D.; Matsuyama, H. Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan. *Geomorphology* **2009**, *109*, 108–121. [CrossRef]
27. Trigila, A.; Iadanza, C.; Esposito, C.; Scarascia-Mugnozza, G. Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). *Geomorphology* **2015**, *249*, 119–136. [CrossRef]
28. Wu, X.; Ren, F.; Niu, R. Landslide susceptibility assessment using object mapping units, decision tree, and support vector machine models in the Three Gorges of China. *Environ. Earth Sci.* **2014**, *71*, 4725–4738. [CrossRef]
29. Shruthi, R.B.; Kerle, N.; Jetten, V.; Stein, A. Object-based gully system prediction from medium resolution imagery using random forests. *Geomorphology* **2014**, *216*, 283–294. [CrossRef]
30. Carranza, E.J.M.; Laborte, A.G. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). *Comput. Geosci.* **2015**, *74*, 60–70. [CrossRef]
31. Leite, E.P.; de Souza Filho, C.R. Probabilistic neural networks applied to mineral potential mapping for platinum group elements in the Serra Leste region, Carajás Mineral Province, Brazil. *Comput. Geosci.* **2009**, *35*, 675–687. [CrossRef]
32. Rigol-Sanchez, J.P.; Chica-Olmo, M.; Abarca-Hernandez, F. Artificial neural networks as a tool for mineral potential mapping with GIS. *Int. J. Remote Sens.* **2003**, *24*, 1151–1156. [CrossRef]
33. Kisi, O.; Parmar, K.S. Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *J. Hydrol.* **2016**, *534*, 104–112. [CrossRef]
34. Lee, S.; Lee, C.W. Application of decision-tree model to groundwater productivity-potential mapping. *Sustainability* **2015**, *7*, 13416–13432. [CrossRef]
35. Rahmati, O.; Pourghasemi, H.R.; Melesse, A.M. Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. *Catena* **2016**, *137*, 360–372. [CrossRef]
36. Gusyev, M.A.; Haitjema, H.M.; Carlson, C.P.; Gonzalez, M.A. Use of nested flow models and interpolation techniques for science-based management of the sheyenne national grassland, North Dakota, USA. *Groundwater* **2013**, *51*, 414–420. [CrossRef] [PubMed]
37. Xu, T.; Valocchi, A.J.; Choi, J.; Amir, E. Use of machine learning methods to reduce predictive error of groundwater models. *Groundwater* **2014**, *52*, 448–460. [CrossRef] [PubMed]
38. Yoon, H.; Jun, S.; Hyun, Y.; Bae, G.; Lee, K. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *J. Hydrol.* **2011**, *396*, 128–138. [CrossRef]
39. Naghibi, S.A.; Pourghasemi, H.R.; Abbaspour, K. A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. *Theor. Appl. Climatol.* **2017**, 1–18. [CrossRef]
40. Ramos-Cañón, A.M.; Prada-Sarmiento, L.F.; Trujillo-Vela, M.G.; Macías, J.P.; Santos-R, A.C. Linear discriminant analysis to describe the relationship between rainfall and landslides in Bogotá, Colombia. *Landslides* **2016**, *13*, 671–681. [CrossRef]
41. Conoscenti, C.; Ciaccio, M.; Caraballo-Arias, N.A.; Gómez-Gutiérrez, Á.; Rotigliano, E.; Agnesi, V. Assessment of susceptibility to earth-flow landslide using logistic regression and multivariate adaptive regression splines: A case of the Belice River basin (western Sicily, Italy). *Geomorphology* **2015**, *242*, 49–64. [CrossRef]
42. Wang, L.J.; Guo, M.; Sawada, K.; Lin, J.; Zhang, J. Landslide susceptibility mapping in Mizunami City, Japan: A comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models. *Catena* **2015**, *135*, 271–282. [CrossRef]
43. Eker, A.M.; Dikmen, M.; Cambazoğlu, S.; Düzgün, Ş.H.; Akgün, H. Evaluation and comparison of landslide susceptibility mapping methods: A case study for the Ulus district, Bartın, northern Turkey. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 132–158. [CrossRef]

44. Buyeo-gun office. *Statistical Yearbook of Buyeo-Gun*; Buyeo-gun: Chungcheongnam-do, Korea, 2016.
45. Ministry of Environment. *Groundwater Annual Report*; Ministry of Environment: Sejong-si, Korea, 2016.
46. Aniya, M. Landslide-susceptibility mapping in the Amahata River basin, Japan. *Ann. Assoc. Am. Geogr.* **1985**, *75*, 102–114. [CrossRef]
47. Moore, I.D.; Grayson, R.B.; Ladson, A.R. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrol. Process.* **1991**, *5*, 3–30. [CrossRef]
48. Wischmeier, W.H.; Smith, D.D. *Predicting Rainfall Erosion Losses: A Guide to Conservation Planning*; United States Department of Agriculture: Washington, DC, USA, 1978.
49. Moore, I.D.; Burch, G.J. Sediment transport capacity of sheet and rill flow: application of unit stream power theory. *Water Res.* **1986**, *22*, 1350–1360. [CrossRef]
50. Dinesh Kumar, P.K.; 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]
51. Koike, K.; Nagano, S.; Kawaba, K. Construction and analysis of interpreted fracture planes through combination of satellite-image derived lineaments and digital elevation model data. *Comput. Geosci.* **1998**, *24*, 573–583. [CrossRef]
52. O'Leary, D.W.; Friedman, J.D.; Pohn, H.A. Lineament, linear, lineation: Some proposed new standards for old terms. *Geol. Soc. Am. Bull.* **1976**, *87*, 1463–1469. [CrossRef]
53. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*, 2nd ed.; John Wiley and Sons Inc.: New York, NY, USA, 2000.
54. Felicísimo, Á.M.; Cuartero, A.; Remondo, J.; Quirós, E. Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: A comparative study. *Landslides* **2013**, *10*, 175–189. [CrossRef]
55. Gutiérrez, Á.G.; Schnabel, S.; Contador, J.F.L. Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. *Ecol. Model.* **2009**, *220*, 3630–3637. [CrossRef]
56. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–141. [CrossRef]
57. Zhang, W.; Goh, A.T.; Zhang, Y. Multivariate adaptive regression splines application for multivariate geotechnical problems with big data. *Geotech. Geol. Eng.* **2016**, *34*, 193–204. [CrossRef]
58. 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*, 1–19. [CrossRef]
59. Menard, S. *Applied Logistic Regression Analysis*, 2nd ed.; SAGE University Series on Quantitative Applications in the Social Sciences; SAGE: Thousand Oaks, CA, USA, 1995.
60. Ozdemir, A.; Altural, T. A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan Mountains, SW Turkey. *J. Asian Earth Sci.* **2013**, *64*, 180–197. [CrossRef]
61. Milborrow, S. Notes on the Earth Package. Available online: https://www.milbo.org/doc/earth-varmod.pdf (accessed on 23 June 2017).
62. Swets, J.A. Measuring the accuracy of diagnostic systems. *Science* **1973**, *240*, 1285–1293. [CrossRef]
63. Bui, D.T.; Pradhan, B.; Lofman, O.; Revhaug, I.; Dick, O.B. Landslide susceptibility mapping at Hoa Binh Province (Vietnam) using an adaptive neuro-fuzzy inference system and GIS. *Comput. Geosci.* **2012**, *45*, 199–211. [CrossRef]
64. Yesilnacar, E.; Topal, T. Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Eng. Geol.* **2005**, *79*, 251–266. [CrossRef]
65. Kennison, R.F.; Cox, J. Health and functional limitations predict depression scores in the health and retirement study: Results straight from MARS. *Calif. J. Health Promot.* **2013**, *11*, 97–108.