Environ Earth Sci (2016) 75:665
DOI 10.1007/s12665-016-5424-9

ORIGINAL ARTICLE

GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran

Mohsen Zabihi[1] · Hamid Reza Pourghasemi[2] · Zohre Sadat Pourtaghi[3] · Morteza Behzadfar[4]

Received: 11 January 2015 / Accepted: 1 February 2016 / Published online: 11 April 2016

Abstract This study evaluated and compared groundwater spring potential maps produced with two different models—namely multivariate adaptive regression spline (MARS) and random forest (RF)—using geographic information system (GIS). In total, 234 spring locations were identified in the Boujnord, North Khorasan, Iran and a GIS spring inventory map was prepared. Of these, 176 (70 %) locations were employed to produce spring potential maps (training), while the remaining 58 (30 %) cases were used to validate the model. The explanatory variables used to predict spring location were altitude, slope aspect, slope degree, slope length, topographic wetness index (TWI), plan curvature, profile curvature, land use, lithology, distance to rivers, drainage density, distance to faults, and fault density. Furthermore, the spatial relationships between spring occurrence and explanatory variables were performed using a Certainty Factor (CF) model. For validation, area under a receiver operating characteristics (ROC) curves (AUC) was used. The validation results showed that the AUC for calibration is almost identical (0.79) in both models, while for prediction, the MARS model (73.26 %) performed better than RF (70.98 %) model. These results indicate that the MARS and RF models are good estimators of groundwater spring potential in the study area. These groundwater spring potential maps can be applied to groundwater management and groundwater resource exploration.

Keywords Groundwater potential mapping · Multivariate adaptive regression spline · Random forest · GIS · Iran

✉ Hamid Reza Pourghasemi
hamidreza.pourghasemi@yahoo.com;
hr.pourghasemi@shirazu.ac.ir

1 Young Researchers and Elite Club, Bojnourd Branch, Islamic Azad University, Bojnourd, Iran

2 Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran

3 Department of Environment Management Engineering, College of Natural Resources, Yazd University, Yazd, Iran

4 Planning and Management Organization, Bojnourd, North-Khorasan Province, Iran

Introduction

Groundwater is one of the most precious natural resources, which supports human civilization (Bera and Bandyopadhyay 2012). Its essential qualities make it an immensely important and dependable source of water supplies in all climatic regions including both urban and rural areas of developed and developing countries (Waikar and Nilawar 2014). Geological strata act both as conduits for transmission of and reservoirs for groundwater. The suitability for exploitation of groundwater in a geological formation primarily depends on storage and transmissivity of the formation. High relief and downhill slopes impart higher runoff, while topographical depressions enhance groundwater recharge (Waikar and Nilawar 2014). Areas of high drainage density also increase surface runoff. Surface water bodies like rivers and ponds can operate as recharge zones (Murugesan et al. 2012; Waikar and Nilawar 2014).

Groundwater is not an unlimited resource so its use should be properly planned based on the understanding of the groundwater systems behavior in order to ensure its sustainable use (Bera and Bandyopadhyay 2012). Assessing the potential zone of groundwater recharge is therefore important to protect water quality and manage groundwater use. Groundwater recharge zones can be demarcated with the help of remote sensing (RS) and GIS techniques (Waikar and Nilawar 2014). One key advantage of RS data for hydrological investigations and monitoring is its capability to generate information in spatial and temporal domains, which is valuable for analysis, prediction, and validation (Waikar and Nilawar 2014). In addition, GIS technology provides suitable alternatives for efficient management of large and complex geospatial databases (Waikar and Nilawar 2014). Several studies have been conducted on groundwater evaluation using GIS and RS techniques (Jaiswal et al. 2003; Solomon and Quiel 2006; Jha et al. 2007; Ganapuram et al. 2009; Saha et al. 2010; Pourtaghi and Pourghasemi 2014; Naghibi et al. 2014; Davoodi Moghaddam et al. 2013; Rahmati et al. 2015). For example, Oh et al. (2011), Ozdemir (2011), Kaliraj et al. (2013) and Pourtaghi and Pourghasemi (2014) published various studies that have applied RS and GIS to groundwater spring potential mapping. Extending these techniques, numerous statistical modeling techniques are able to predict the potential distribution of a phenomenon from a set of independent variables: such as logistic multiple regression (LMR: Mair and El-Kadi 2013), generalized additive model (GAM: Sorichetta et al. 2013), random forest (RF: Rodriguez-Galiano et al. 2014; Naghibi and Pourghasemi 2015; Naghibi et al. 2016), and multivariate adaptive regression splines (MARS: Gutiérrez et al. 2009). In recent years, with the rapid development of information technology and database technology, data mining algorithms have seen applications beyond information technology into other societal applications (Yao et al. 2013). Data mining is a process of extracting potentially helpful information and knowledge, unknown in advance, from a large, incomplete, and noisy, fuzzy and random practical dataset (Yao et al. 2013). Although the MARS (multivariate adaptive regression spline) and RF (random forest) methods have been applied for landslide susceptibility mapping (Youssef et al. 2015), gully erosion modeling (Gutiérrez et al. 2009), and regional or local assessments of nitrate and pesticide contamination (Rodriguez-Galiano et al. 2014); this approach (MARS) and its comparison with a RF has not yet been used for groundwater spring potential mapping.

This study evaluates the GIS-based MARS and RF models for groundwater spring potential mapping at the Bojnurd Township in northern Khorasan Province, Iran (Fig. 1). The main objective of the study is to contribute towards systematic groundwater studies utilizing RF and MARS models to delineate groundwater spring potential areas which could be applied in other similar areas.

Study area

The study area, as shown in Fig. 1, lies in the southern region of Bojnourd Township in North-Khorasan Province, Iran, known as the Bojnourd Plain. This 1243 km^2 area is located between 55°44 and 56°18 longitude, and 38°17′ to 37°13′ latitude. The area’s elevation varies from 887 to 2967 m above mean sea level (m.s.l.) and the annual rainfall in the area is approximately 266.4 mm. The study area slopes gently from south to north and forms foothills for mountains to the north. The Bojnourd Plain is located in Kopet-Dagh geological formation that mainly covered by Quaternary sediments. The groundwater elevation in wells in the study area is between 1025 and 1074 m m.s.l. The groundwater depth varies between 4 and 80 m, and groundwater flows from the south and southwest to the north and northeast.

Methodology

Figure 2 shows a flowchart of the methodology applied in the current study. This figure demonstrates the explanatory variables used in the analysis and the processes applied according to the models. In the first step, the dataset for model development and application were assembled. Next, a certainty factor (CF) model was applied to determine spatial relationships among spring occurrence and explanatory variables. Then, MARS and RF models were applied to map groundwater spring potential. Finally, constructed models were validated and tested using the receiver operating characteristic (ROC) curve (success rate and prediction rate curves).

Dataset for models development and application

Dataset and construction of a spatial database of explanatory variables are important parts of any research (Pourghasemi et al. 2013; Davoodi Moghaddam et al. 2013). At first, the spring locations were compiled from Iranian Department Water Resources Management (http://www.wrm.ir/index.php?l=EN) and extensive field surveys. In total, 234 springs were detected in Boujnurd watershed, North Khorasan, Iran (Fig. 1). 176 (70 %) of the spring locations were used for groundwater spring potential mapping and 58 (30 %) were set aside for validation. For conducting a spring potential map (SPM), it is necessary to evaluate mappable explanatory variables with the spring inventory map (Davoodi Moghaddam et al. 2013). In this study, 13 such explanatory variables were considered. These were; altitude, slope aspect, slope degree, slope-length (LS), topographic wetness index (TWI),

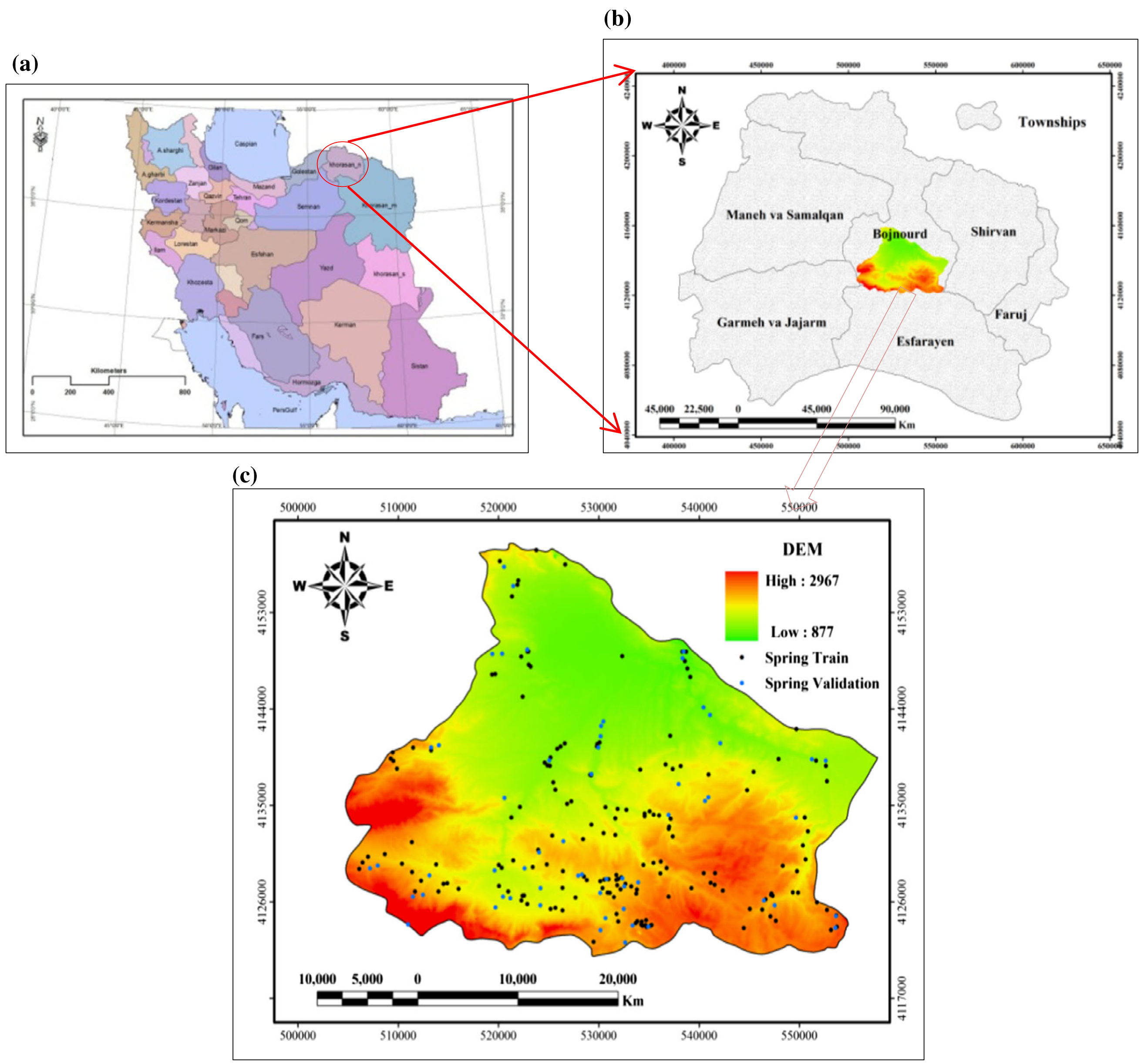


Fig. 1 Location of the study area; **a** Iran map, **b** North Khorasan Province map, **c** spring location map of study area

plan curvature, profile curvature, land use, lithology, distance to rivers, drainage density, distance to faults, and fault density. A digital elevation model (DEM) was created using topographical maps at 1:50,000 scale. The DEM has a cell size of 30 m with 1284 rows and 1768 columns. The DEM was used to derive the altitude, slope aspect, slope degree, LS, TWI, plan curvature, and profile curvature values. The altitude map for the study area with cell size 30 m $\times$ 30 m was produced from the DEM and classified into five classes (Fig. 3a). Slope aspect strongly affects hydrologic processes via evapotranspiration (Sidle and Ochiai 2006) and has been categorized into nine classes (Fig. 3b). The slope map of the study area is obtained from the DEM using the slope function in ILWIS-GIS (http://www.ilwis.org/). These slope values (in degree) are divided into four classes (Fig. 3c). Slope-length (LS) is the combination of slope steepness (S) and slope length (L) which is implemented to represent soil loss potential from the combined slope properties (Fig. 3d). The LS factor was calculated according to Eq. 1 (Moore and Burch 1986) and classified into four categories.

$$\mathrm{LS} = \left(\frac{B_s}{22.13}\right)^{0.6} \times \left(\frac{\sin_\beta}{0.0896}\right)^{1.3}, \tag{1}$$

where B_s = specific catchment's area, β = slope angle.

Another topographic factor is TWI which is defined in Eq. 2 (Beven and Kirkby 1979; Moore et al. 1991):

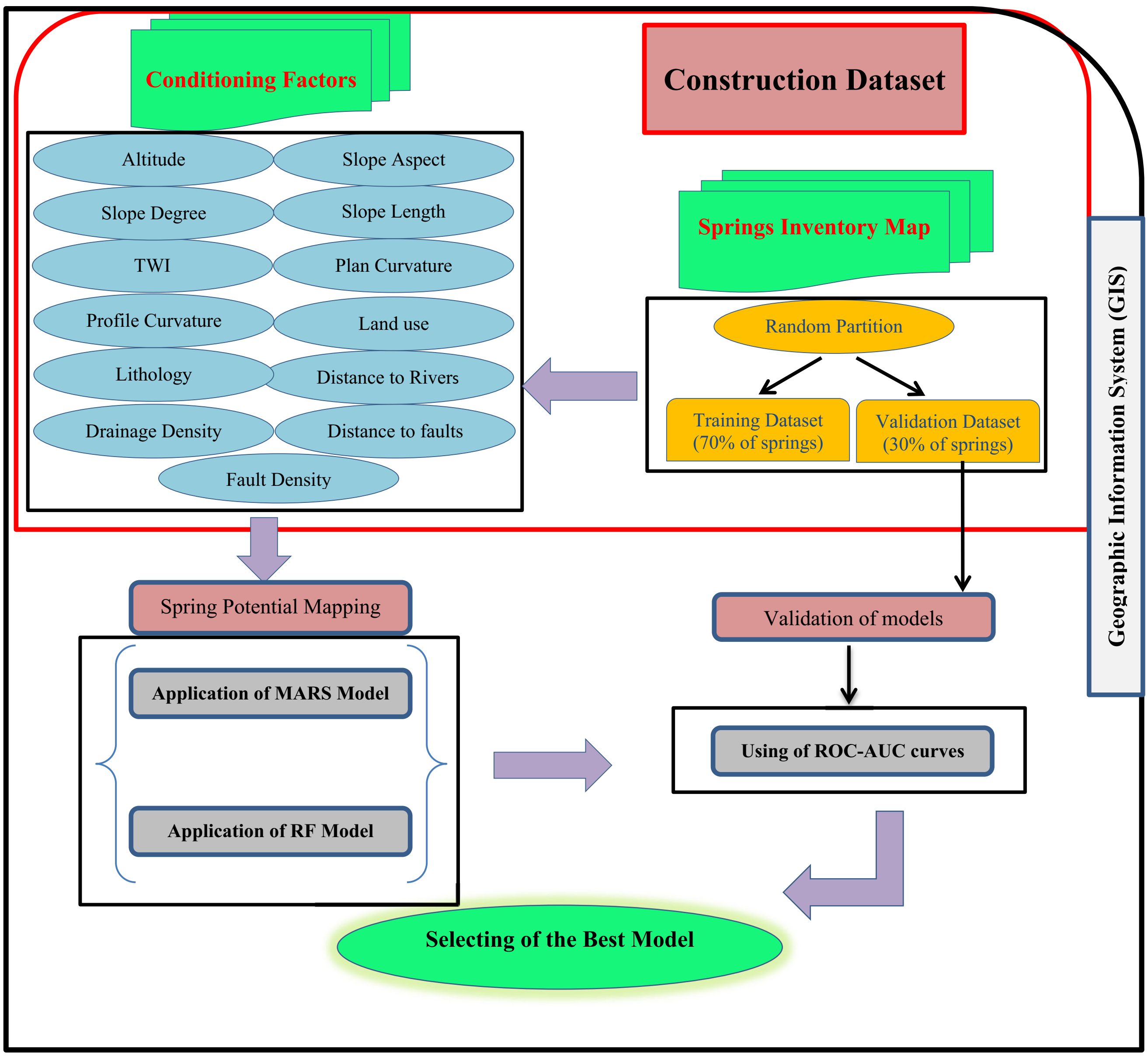


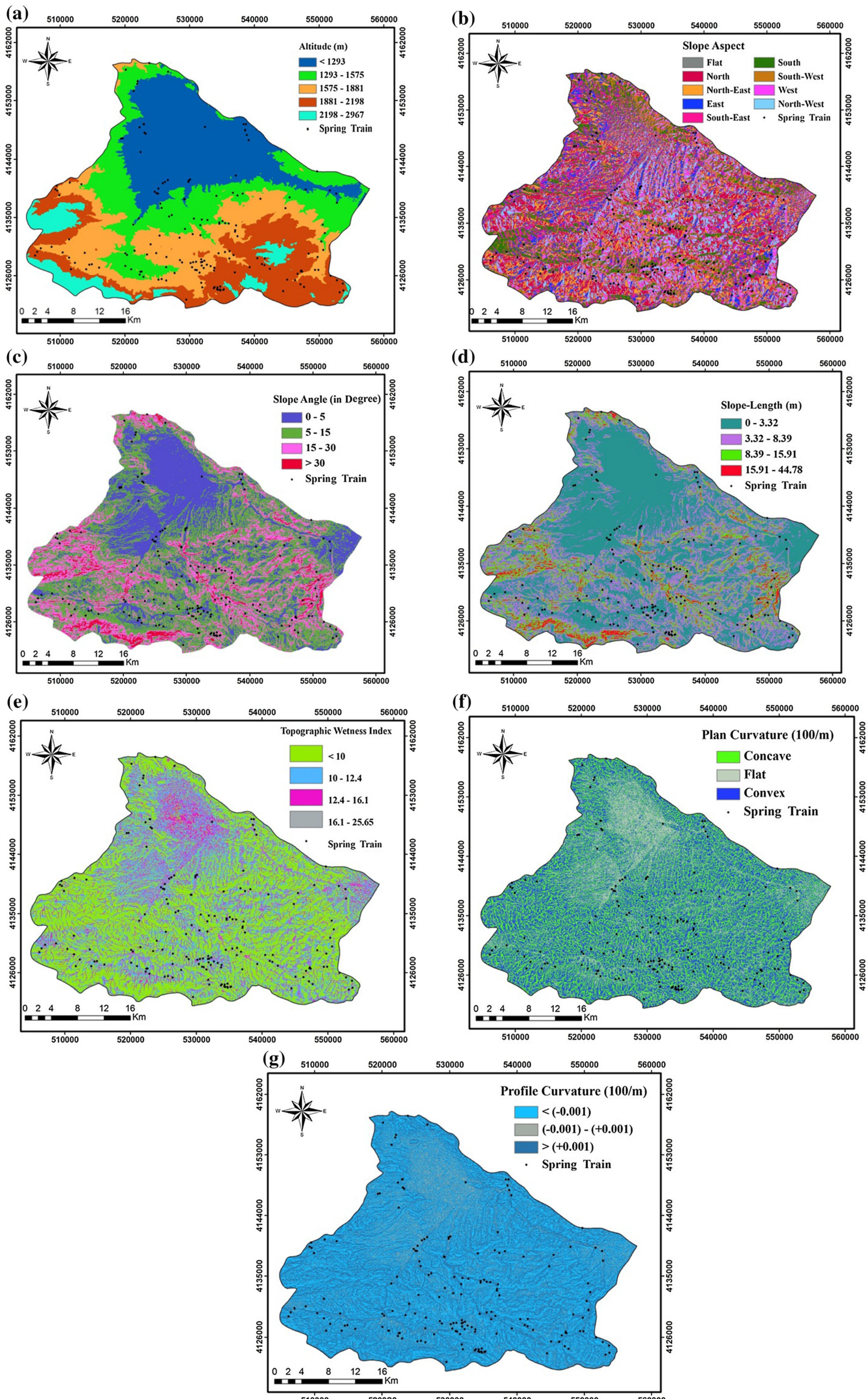
Fig. 2 Flow chart of methodology used in spring potential mapping

$$\mathrm{TWI} = \ln\left(\frac{\alpha}{\tan\beta}\right) \tag{2}$$

where $\alpha =$ is the cumulative up slope area from a point (per unit contour length) and $\beta =$ is the slope angle at the point (Fig. 3e). The plan curvature demonstrates the morphology of the topography. A positive curvature represents that the surface is upwards convex at that cell, and a negative curvature shows that the surface is upwards concave at that cell. A value of zero indicates a flat surface (Oh and Lee 2010) (Fig. 3f). The profile curvature shows the flow acceleration, erosion (negative values)/deposition (positive values) rate and it controls the change of speed of mass flowing down the slope (Yesilnacar 2005; Talebi et al. 2007). In this study, the profile curvature was prepared and classified into three groups based on common standard classification scheme (Pourghasemi et al. 2013) (Fig. 3g).

A land use layer was produced from Landsat-7/ETM^{+} satellite images using a supervised classification and maximum likelihood algorithm (Rahmati et al. 2016). The area is covered by six land use types; forest, rangeland,

Fig. 3 Maps of explanatory variables in the study area; **a** altitude (m), **b** slope aspect, **c** slope degree, **d** slope length (LS), **e** topographic wetness index (TWI), **f** plan curvature (100/m), **g** profile curvature (100/m)


(a)
Altitude (m)
< 1293
1293 - 1575
1575 - 1881
1881 - 2198
2198 - 2967
Spring Train
(b)
Slope Aspect
Flat
North
North-East
East
South-East
South
South-West
West
North-West
Spring Train
(c)
Slope Angle (in Degree)
0 - 5
5 - 15
15 - 30
> 30
Spring Train
(d)
Slope-Length (m)
0 - 3.32
3.32 - 8.39
8.39 - 15.91
15.91 - 44.78
Spring Train
(e)
Topographic Wetness Index
< 10
10 - 12.4
12.4 - 16.1
16.1 - 25.65
Spring Train
(f)
Plan Curvature (100/m)
Concave
Flat
Convex
Spring Train
(g)
Profile Curvature (100/m)
< (-0.001)
(-0.001) - (+0.001)
> (+0.001)
Spring Train
0 2 4 8 12 16 Km

dray farming, irrigation farming, residential area, and bare land. The details of land use type are shown in Fig. 4 and summarized in Table 1. Lithological features of study region are represented in the geologic map (Fig. 5), which is derived from the geologic map at 1:100,000-scale prepared by Geological Survey of Iran (Geology Survey of Iran (GSI) 1997), digitized in ILWIS-GIS (version 3.8), and divided into 12 classes (Table 1).

The distance to rivers was calculated using the vector river lines by manually applying the distance function in ArcGIS (version 9.3). Five classes corresponding to distance to rivers were calculated at 200-m intervals (Fig. 6a). The drainage density exhibits the flow of water through the study area and is defined as the ratio of sum of the drainage lengths in the cell and the area of the corresponding cell (Sarkar and Kanungo 2004; Pourghasemi et al. 2013). The drainage density was computed for each 30 × 30 m grid cell which ranges from 1.81 to 7.99 km/km^2 and is classified into four classes (Fig. 6b). The distance to faults map was extracted from geologic maps at 1:100,000 scale, and then the buffer categories were defined (Fig. 7a). Finally, the fault density map was produced. The length of the faults from geological maps at 1:100,000 scale of the study area were extracted and divided by area for each 30 × 30 m grid cell with results ranging from 1.81 to 15.94 km/km^2. The results were classified into four classes (Fig. 7b).

Models

Certainty factor (CF) model

In this study, the CF model was implemented to demonstrate the spatial link joining spring occurrence and explanatory variables. The CF (an approach that has seen widespread use in rule-based expert systems), is based on probabilistic reasoning (Chung and Leclerc 1994). This is one strategy to handle the problem of blending of different data layers and the heterogeneity and unreliability of the input data. The CF, defined as a function of probability, was originally suggested by Shortliffe and Buchanan (1975) and later modified by Heckerman (1986) (Kanungo et al. 2011):

$$\mathrm{CF} = \begin{cases} \dfrac{pp_m - pp_n}{pp_m(1 - pp_n)} & \text{if } pp_m \geq pp_n \\ \dfrac{pp_m - pp_n}{pp_n(1 - pp_m)} & \text{if } pp_m < pp_n \end{cases} \quad (3)$$

where, pp_m is the conditional probability of having a number of spring events occurring in category m and pp_n is the prior probability of having the total number of spring events occurring in the study area. The range of variation of the CF is [−1.0 to 1.0], where a positive value means an increasing certainty in spring occurrence, while a negative value corresponds to a decreasing certainty in spring occurrence. A value close to 0 means that the prior

Fig. 4 Land use map of study area

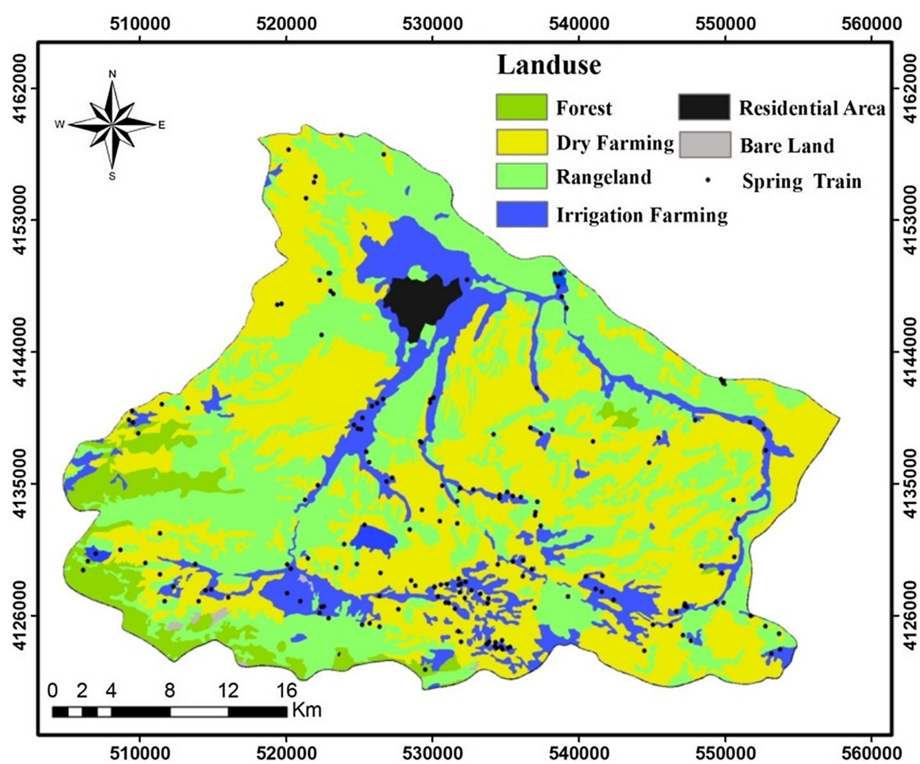


Table 1 Lithology of the study area

Code	Lithology	Age	Era
Qft1	High level piedmont fan and valley terrace deposits	Quaternary	Cenozoic
Qft2	Low level piedmont fan and valley terrace deposits	Quaternary	Cenozoic
Ksr	Ammonite Bearing shall and thin layers of siltstone and sandstone	Early Cretaceous	Mesozoic
Ksn	Grey to black shale and thin layers of siltstone and sandstone	Early Cretaceous	Mesozoic
Ktr	Grey oolitic and bioclastic orbitolina limestone	Early Cretaceous	Mesozoic
JKsj	Pale red argillaceous limestone, marl, Gypsiferous marl, sandstone and conglomerate	Jurassic-Cretaceous	Mesozoic
Kat	Olive green glauconitic sandstone and shale	Cretaceous	Mesozoic
Mur	Red marl, Gypsiferous marl, sandstone and conglomerate	Miocene	Cenozoic
Jl	Light grey, thin-bedded to massive limestone	Jurassic-Cretaceous	Mesozoic
TRJs	Dark grey shale and sandstone	Triassic-Jurassic	Mesozoic
DCkh	Yellowish, thin to thick-bedded, Fossileferous argillaceous limestone, dark grey limestone, greenish marl and shale, locally including gypsum	Devonian	Paleozoic
Jd	Well-bedded to thin-bedded, greenish-grey argillaceous limestone with intercalations of calcareous shale	Jurassic	Mesozoic
Sn	Greenish grey, shale, sandstone, sandy-lime, coral limestone and dolomite (Niur Formation)	Silurian	Paleozoic

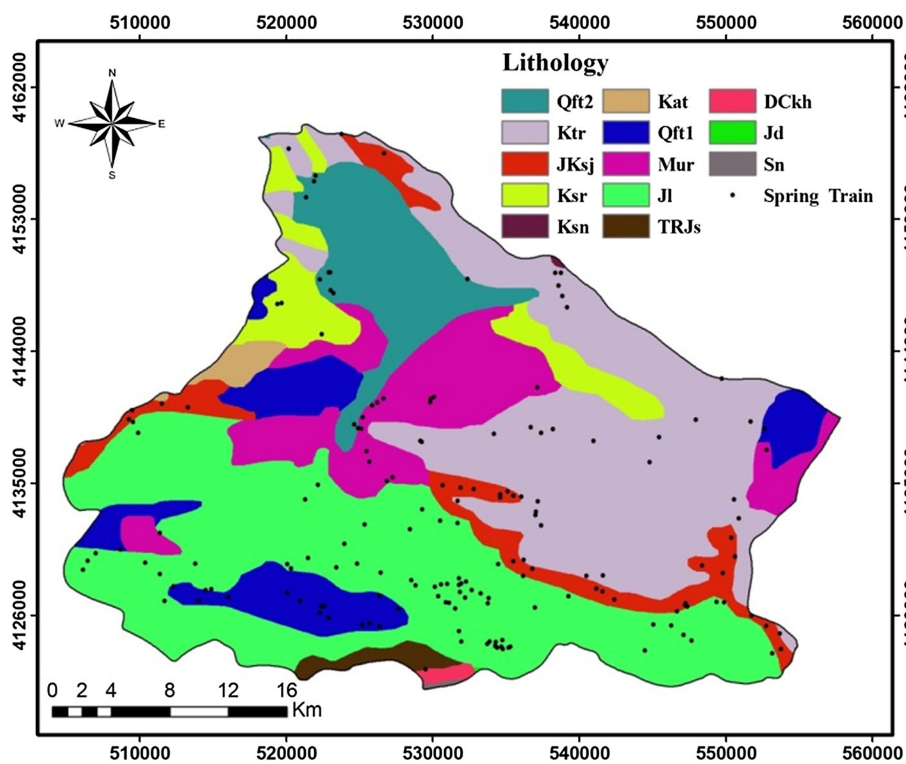


Fig. 5 Lithology map of study area

probability is very similar to the conditional one, so it is difficult to give any indication about the certainty of the spring occurrence. The favorability values (pp_m, pp_n) are derived from overlaying each data layer with the existing spring distribution layer in GIS environment and calculating the spring occurrence frequency. CF values are then calculated for each layer and their sub-classes in Microsoft Excel 2010 (Kanungo et al. 2011).

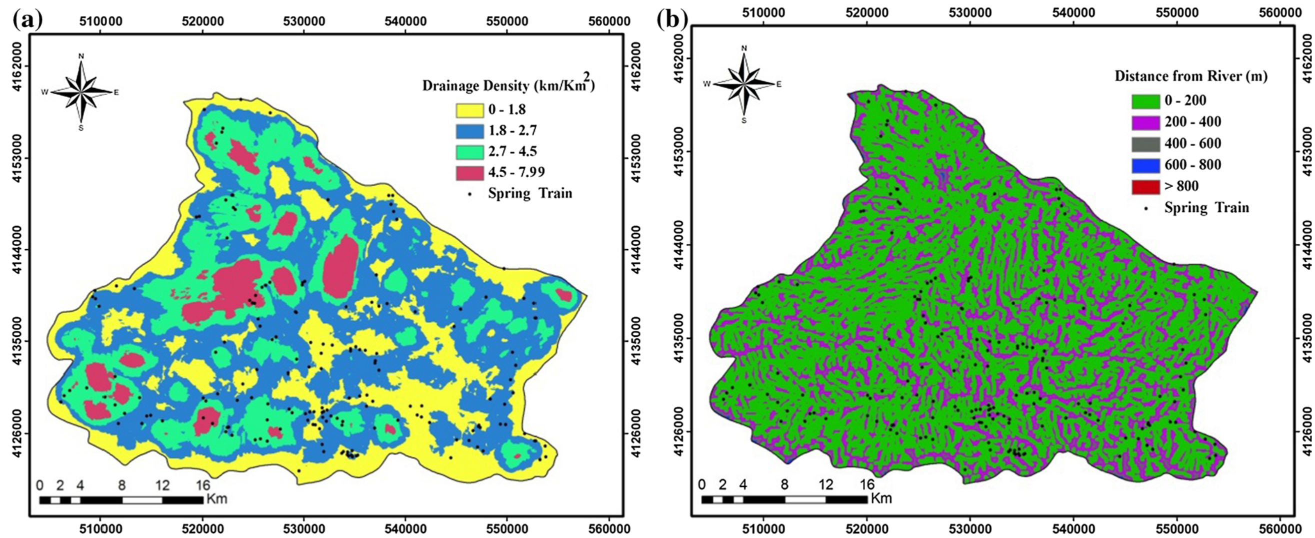


Fig. 6 **a** Distance to rivers (m), **b** drainage density (Km/Km^2) in the study area

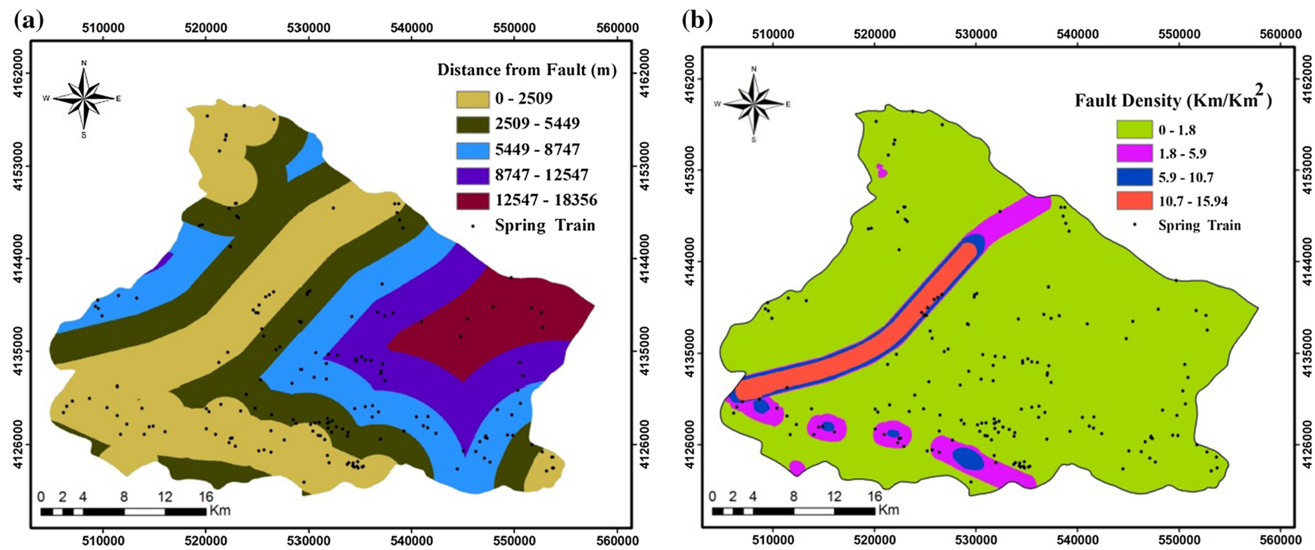


Fig. 7 **a** Distance to faults (m), **b** fault density (Km/Km^2) in the study area

Random forest (RF) model

"Random forest is an ensemble method which compounds multiple decision tree algorithms to produce repeated predictions of the same phenomenon. Random forests (RF) are very flexible ensemble classifiers based on decision trees, first developed by Breiman (2001)" (Breiman 2001; Catani et al. 2013; Micheletti et al. 2014). Decision trees can be separated to classification trees and regression trees (Rodriguez-Galiano et al. 2014). A regression tree (RT) indicates a set of restrictions or conditions which are hierarchically structured, and which are successively applied from a root to a terminal node or leaf of the tree (Breiman et al. 1984; Quinlan 1993). In order to derive the RT, recursive partitioning and multiple regressions are carried out from the dataset. From the root node, the data splitting process in each internal node of a rule of the tree is consecutive until a stop condition previously specified is reached. Each of the terminal nodes, or leaves, has joined to it a simple regression model which applies in that node only. Once the tree's exaction process is finished, pruning can be applied with the aim of improving the tree's generalization capacity by reducing its structural complexity. The number of cases in nodes can be derived as pruning criteria (Rodriguez-Galiano et al. 2014).

The RF algorithm handles random binary trees which use a subset of the observations through bootstrapping techniques: from the original data set a random choice of the training data is sampled and used to build the model, the data not included are referred to as "out-of-bag" (OOB) (Breiman 2001; Catani et al. 2013). Furthermore, a random selection of predictor variables is applied to split each node of the trees. Each tree is expanded to minimize classification errors, but the random selection influences the results, thus making a single-tree classification very unstable. The RF algorithm estimates the importance of a variable by looking for how much the prediction error increases when OOB data for that variable is permuted while all others are left unchanged (Liaw and Wiener 2002; Catani et al. 2013). This capability can be profitably applied to study the relative importance of the different explanatory variables, a critically important but often neglected aspect of SPM (spring potential mapping). In the R statistical package application of RF used in this work (the "randomForest" package in R 2.0.3 (Breiman and Cutler 2006)), the model output is a membership probability to one of the two possible classes "Spring" and "No spring". Random forests need two parameters to be tuned by the user: (1) the number of trees T, (2) the number of variables m to be stochastically chosen from the available set of features. It is suggested (Breiman 2001; Micheletti et al. 2014) to pick a large number of trees and the square root of the dimensionality of the input space for m (Micheletti et al. 2014). Based on two parameters, the number of trees in RF has been fixed to 1000 after an introductory analysis and the number m of variables sampled at each node has been selected to be three to analyze the conjunct contribution of subsets of features while maintaining fast convergence during iterations. Moreover, two types of error were calculated: mean decrease in accuracy and mean decrease in node impurity (mean decrease Gini). These importance measures can be used for ranking variables and for variable selection (Calle and Urrea 2010).

Multivariate adaptive regression spline (MARS) model

The MARS models (implemented in this work using the "earth" package in R 3.0.2 (Milborrow 2012). Use a nonparametric modeling approach that does not require assumptions about the form of the relationship between the independent and dependent variables (Friedman 1991; Balashi et al. 2009). The MARS algorithm works by division the ranges of the explanatory variables into regions and by producing, for each of these regions, a linear regression equation. Breaks values between regions are called "knots", while the term "basis function" (BF) is used to demonstrate each distinct interval of the predictors. BFs are functions of the following form (Eq. 4):

$$\max(0, x-k) \text{ or } \max\ (0, k-x) \tag{4}$$

where x is an independent variable and k is a constant corresponding to a knot. The general formulation of MARS is:

$$\hat{y} = \hat{f}(x) = \beta + \sum_{m=1}^{M} \alpha_m H_m(x) \tag{5}$$

where, y is the dependent variable predicted by the function $f(x)$, β is a constant, and M is the number of terms, each of them formed by a coefficient α_m and $H_m(x)$ is an individual basis function or a product of two or more BFs (Conoscenti et al. 2014). The MARS models were developed in two steps. In the first step—the forward algorith—basis functions are presented to define Eq. 5. Many basis functions are added in Eq. 5 to get better performance. The developed MARS can experience overfitting due to large a number of basic functions. To mitigate this problem, the second step—the backward algorithm—prevents over fitting by removing redundant basis functions from Eq. 5. MARS adopts Generalized Cross-Validation (GCV) to delete the redundant basis functions (Craven and Wahba 1979; Samui and Kothari 2012). The expression of GCV is written as follows (Eq. 6):

$$\text{GCV} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}(x_i)\right]^2}{\left[1 - \frac{C(H)}{N}\right]^2} \tag{6}$$

where N is the number of data and $C(K)$ is a complexity penalty that increases with the number of basis function (BF) in the model and which is defined as (Eq. 7):

$$C(H) = (\mathrm{H} + 1) + dH \tag{7}$$

where d is a penalty for each BF included into the model and H is number of basic functions in Eq. 5 (Friedman 1991; Samui and Kothari 2012).

Results

Application of certainty factor model

The results of spatial relationship between spring occurrence and explanatory variables using the CF technique are shown in Table 2. Based on Table 2, for altitude, for example, the 1575–1881 m class has the highest CF value (0.33). CF values generally increased with increasing altitude in the study area, and then spring occurrence probability decreases at altitudes above 2189 m. For slope aspect, most of the springs occurred in south and north-west facing slopes, while north east-facing slopes have the lowest abundance. In the study area, the drainage density, rivers, and faults are in the south and north-west facing parts of the study area, so these sites are considered as

Table 2 Spatial relationship between each spring explanatory variable and springs by certainty factor (CF) model

Factors	Class	No. of pixel	% pixel	No. of spring	% spring	FR	Ppa	PPs	Final weight (CF)
Altitude (m)	<1293	375,275	27.16	34	19.32	0.71	9.06E-05	0.0001	−0.29
	1293–1575	332,365	24.05	39	22.16	0.92	0.0001	0.0001	−0.08
	1575–1881	310,603	22.48	59	33.52	1.49	0.0002	0.0001	0.33
	1881–2189	272,704	19.74	44	25	1.27	0.0002	0.0001	0.21
	2189–2967	90,789	6.57	0	0	0	0	0.0001	−1
Slope aspect	Flat	5329	0.39	0	0	0	0	0.0001	−1
	North	268,923	19.46	22	12.5	0.64	8.18E-05	0.0001	−0.36
	North East	214,071	15.49	16	9.09	0.59	7.47E-05	0.0001	−0.41
	East	171,483	12.41	14	7.95	0.64	8.16E-05	0.0001	−0.36
	South East	149,669	10.83	23	13.07	1.21	0.0002	0.0001	0.17
	South	166,474	12.05	34	19.32	1.60	0.0002	0.0001	0.38
	South West	119,584	8.65	20	11.36	1.31	0.0002	0.0001	0.24
	West	122,874	8.89	19	10.8	1.21	0.0002	0.0001	0.18
	North West	163,329	11.82	28	15.91	1.35	0.0002	0.0001	0.26
Slope degree	<5	399,460	28.91	47	19.42	0.67	0.0001	0.0001	−0.08
	5–15	620,236	44.89	99	40.91	0.91	0.0002	0.0001	0.20
	15–30	305,833	22.13	90	37.19	1.68	0.0003	0.0001	0.57
	>30	56,207	4.07	6	2.48	0.61	0.0001	0.0001	−0.16
Slope-length (m)	0–3.32	731,542	52.94	88	50	0.94	0.0001	0.0001	−0.06
	3.32–8.39	415,216	30.05	60	34.09	1.13	0.0001	0.0001	0.12
	8.39–15.91	182,699	13.22	22	12.5	0.95	0.0001	0.0001	−0.05
	>15.91	52,279	3.78	6	3.41	0.90	0.0001	0.0001	−0.10
TWI	<10.03	682,022	49.36	65	36.93	0.75	9.53E-05	0.0001	−0.25
	10.03–12.40	471,608	34.13	68	38.64	1.13	0.0001	0.0001	0.12
	12.40–16.19	174,426	12.62	24	13.64	1.08	0.0001	0.0001	0.07
	16.13–25.65	53,680	3.88	19	10.8	2.78	0.0004	0.0001	0.64
Plan curvature (100/m)	Concave	590,728	42.75	96	54.55	1.28	0.0002	0.0001	0.22
	Flat	168,527	12.2	20	11.36	0.93	0.0001	0.0001	−0.07
	Convex	622,481	45.05	60	34.09	0.76	9.64E-05	0.0001	−0.24
Profile curvature (100/m)	<−0.001	623,840	45.15	64	36.36	0.81	0.0001	0.0001	−0.19
	−0.001 to +0.001	87,448	6.33	5	2.84	0.45	5.72E-05	0.0001	−0.55
	>+0.001	670,448	48.52	107	60.8	1.25	0.0002	0.0001	0.20
Land use	Forest	7550	5.18	4	2.27	0.44	0.0005	0.0001	0.76
	Rangeland	515,867	37.33	33	18.75	0.50	6.40E-05	0.0001	−0.50
	Dry farming	602,337	43.59	71	40.34	0.93	0.0001	0.0001	−0.07
	Irrigation farming	173,374	12.55	68	38.64	3.08	0.0004	0.0001	0.68
	Residential area	15,973	1.16	0	0	0	0	0.0001	−1
	Bare land	2635	0.19	0	0	0	0	0.0001	−1
Lithology	Qft2	118,279	8.56	11	6.25	0.73	9.30E-05	0.0001	−0.27
	Ktr	375,592	27.18	30	17.05	0.63	7.99E-05	0.0001	−0.37
	JKsj	79,875	5.78	29	16.48	2.85	0.0004	0.0001	0.65
	Ksr	71,741	5.19	3	1.7	0.33	4.18E-05	0.0001	−0.67
	Ksn	526	0.04	0	0	0	0	0.0001	−1
	Kat	12,982	0.94	0	0	0	0	0.0001	−1
	Qft1	111,619	8.08	17	9.66	1.20	0.0002	0.0001	0.161
	Mur	160,095	11.59	14	7.95	0.69	8.74E-05	0.0001	−0.31

Table 2 continued

Factors	Class	No. of pixel	% pixel	No. of spring	% spring	FR	Ppa	PPs	Final weight (CF)
	Jl	432,568	31.31	71	40.34	1.29	0.0002	0.0001	0.22
	TRJs	13,419	0.97	0	0	0	0	0.0001	−1
	DCkh	3995	0.29	1	0.57	1.97	0.0003	0.0001	0.49
	Jd	188	0.01	0	0	0	0	0.0001	−1
	Sn	857	0.06	0	0	0	0	0.0001	−1
Distance to rivers (m)	0–200	1,015,757	73.51	164	93.18	1.27	0.0002	0.0001	0.21
	200–400	342,138	24.76	11	6.25	0.25	3.21E-05	0.0001	−0.75
	400–600	22,171	1.6	1	0.57	0.36	4.51E-05	0.0001	−0.65
	600–800	1577	0.11	0	0	0	0	0.0001	−1
	>800	93	0.01	0	0	0	0	0.0001	−1
Drainage density (km/km^2)	<1.81	365,793	26.47	52	29.55	1.12	0.0001	0.0001	0.10
	1.81–2.77	594,115	43	82	46.59	1.08	0.0001	0.0001	0.08
	2.77–4.55	325,750	23.58	36	20.45	0.87	0.0001	0.0001	−0.13
	4.55–7.99	96,078	6.95	6	3.41	0.49	6.24E-05	0.0001	−0.51
Distance to faults (m)	<2509	465,360	33.68	76	43.18	1.28	0.0002	0.0001	0.22
	2509–5449	364,994	26.42	37	21.02	0.80	0.0001	0.0001	−0.20
	5449–8747	259,796	18.8	37	21.02	1.12	0.0001	0.0001	0.11
	8747–12,547	166,330	12.04	18	10.23	0.85	0.0001	0.0001	−0.15
	12,547–18,355	125,256	9.07	8	4.55	0.50	6.39E-05	0.0001	−0.50
Fault density (km/km^2)	<1.87	1,220,870	88.36	149	84.66	0.96	0.0001	0.0001	−0.04
	1.87–5.91	74,325	5.38	18	10.23	1.90	0.0002	0.0001	0.47
	5.91–10.71	31,498	2.28	5	2.84	1.25	0.0002	0.0001	0.20
	10.71–15.94	55,043	3.98	4	2.27	0.57	7.27E-05	0.0001	−0.43

likely zones of groundwater recharge. The north and west regions have lower drainage density and fault values. These areas fall under low suitability for infiltration (Davoodi Moghaddam et al. 2013). For slope angles, the 5°–15° and 15°–30° classes have the highest CF values (0.57 and 0.20, respectively). Slope always plays a very important role in groundwater potential mapping and at the same time the slope increases, the runoff increases as well (Israil et al. 2006) leading to less infiltration (Jaiswal et al. 2003; Davoodi Moghaddam et al. 2013). For slope length index, the 3.32–8.39 m class has the highest CF value (0.12). The CF value for TWI clearly showed that class of 16.13–25.65 has the most effect on spring locations. The TWI factor illustrates the effect of topography on the location and size of saturated source areas of runoff generation under the steady-state assumption and uniform soil properties (i.e., transmissivity is constant throughout the catchments and equal to unity) (Pourghasemi et al. 2013; Davoodi Moghaddam et al. 2013). The relation between plan curvature and spring locations showed that concave class has the highest value of CF (0.22), and for profile curvature, the >0.001 class shows a high CF value (0.20). A concave slope contains more water and holds this water for a longer period especially during heavy rainfall (Lee and Pradhan 2006; Davoodi Moghaddam et al. 2013). In the case of land use, the highest CF value was for the forest land use type (0.76). When comparing the relationship between spring location and lithology, the CF values were positive for the classes of JKsi (Pale red argillaceous limestone, marl, Gypsiferous marl, sandstone and conglomerate) and Jl (Light grey, thin-bedded to massive limestone). In the case of distance to rivers, the 0 and 200 m class has a CF score of 0.21. The drainage density class <1.81 km/km^2 has a CF value of 0.10. In general, we observed that as the drainage density increases, the spring frequency decreases. The drainage density depends on the slope, nature, and attitude of bedrock and the existing regional and local fracture templates. It is a reflection of the lithology and structure of a given area and can be of great value for groundwater resources evaluation (Godebo 2005; Davoodi Moghaddam et al. 2013). Assessment of distance to faults showed that the <2509 m class has high correlation with spring occurrence. Lineaments are linearly fractured zones in the geological structure of an area, such as faults and dykes, and they can control the exchange of water between surface and subsurface (Davoodi Moghaddam et al. 2013). Finally, for fault density, the 1.87–5.91 km/km^2 class has a CF value of 0.47.

Application of random forest model

The results of spatial relationship between springs and explanatory variables using the RF model are shown in Tables 3, 4 and Figs. 8, 9. The aggregate OOB predictions are presented in Fig. 8 and Table 3 (confusion matrix). The OOB results indicate a prediction error rate of about 30.11 %. In other words, the model can be considered 69.89 % accurate. Overall measure of accuracy is then followed by a confusion matrix that records the conflict between that final model's predictions and the present outcomes of the training observations. The present observations are the rows (Table 3), whilst the columns correspond to the model predictions collocated with the observations: the number reflect the counts in each box (Williams 2011). The model incorrectly predicted springs where they were actually absent in 56 cases (type I error) and the absence of springs when they were actually present in 50 cases (Type II error). The model correctly predicted the absence of springs for 120 cases and the presence of springs for 126 cases. Results from variable selection for the RF model are presnted in Fig. 9. This shows the 13 variables ordered by two specific importance measures (mean decrease accuracy and mean decrease Gini). Based on Fig. 9 and Table 4, the higher values indicate that the variable is relatively more important (Williams 2011). The accuracy measure (mean decrease) lists distance to rivers (32.75), altitude (22.29), and land use (18.49) as the most important. The distance to rivers (14.78), altitude (14.13) and distance to faults (10.89) have the highest importance according to the Gini measure. Aside from the first two most important measures, the rankings are different according to the Gini measure relative to the Mean Decrease Accuracy measure.

A full Spring Potential Map for the area was created using the RF model in ArcGIS 9.3 and categorize based on a natural break classification plan (Ozdemir 2011;

Table 3 Confusion matrix for the RF model (0 = spring absent, 1 = spring present)

	Predicted		
	0	1	Class error
Actual			
0	120	56	0.32
1	50	126	0.28

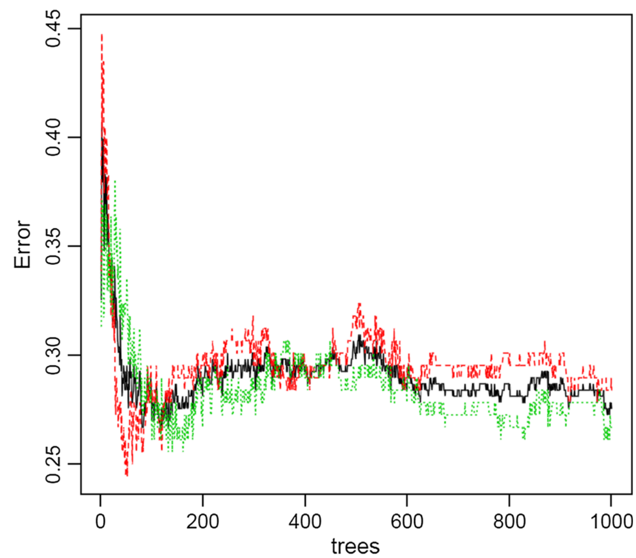


Fig. 8 The error rate of the overall RF model (*OOB* out-of-bag (*black line*), *0* spring absent (*red dash line*), and *1* spring present (*green dash line*)

Table 4 Relative influence of explanatory variables in the RF model (0 = spring absent, 1 = spring present)

Variable	Spring absent	Spring present	Mean decrease accuracy	Mean decrease Gini
Altitude (m)	13.06	19.17	22.29	14.13
Slope aspect	6.90	7.15	9.62	6.77
Slope degree (degree)	4.46	4.23	6.80	8.12
Slope-length (m)	5.38	4.90	8.00	8.35
TWI	8.89	7.90	12.00	10.7
Plan curvature (100/m)	−5.44	3.77	−0.86	6.94
Profile curvature (100/m)	−2.78	2.23	−0.33	7.16
Land use	13.30	12.71	18.49	6.74
Lithology	4.06	10.91	11.49	4.32
Distance to rivers (m)	23.78	24.36	32.75	14.78
Drainage density (km/km^2)	11.07	9.21	14.23	10.59
Distance to faults (m)	8.35	14.02	16.02	10.89
Fault density (km/km^2)	4.23	2.30	5.13	2.44

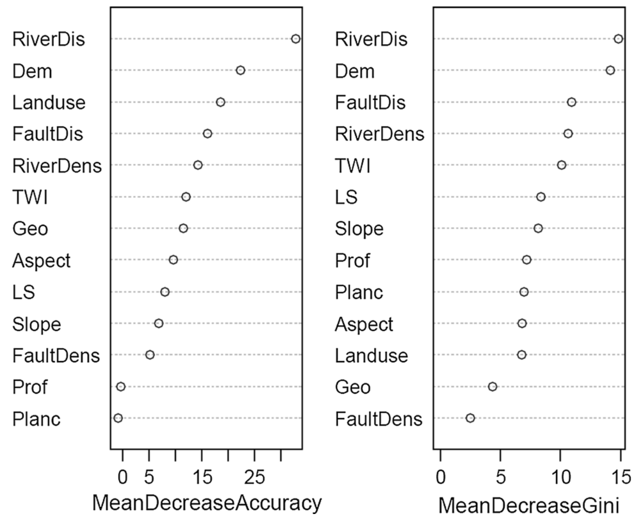


Fig. 9 The error rate of the overall RF model (*OOB* out-of-bag, *0* spring absent and *1* spring present)

Pourghasemi et al. 2013; Zare et al. 2013; Pourtaghi and Pourghasemi 2014) into low, moderate, high and very high potential categories. These results were represented in Fig. 10 and Table 6.

Application of multivariate adaptive regression spline model

The optimal MARS model presents 20 terms and includes 20 BF (the term created during the forward pass were 96), with a GCV of 0.18. Only ten of the 13 independent variables were used in the optimal model (Table 5), because MARS only uses the necessary independent variables (Gutiérrez et al. 2009). In Table 5, nsubset is an index vector specifying which cases to use, i.e., which rows in x to use (default is NULL, meaning all), gcv is generalized cross validation (GCV) of the model (aggregated over all responses) (the GCV is calculated using the penalty argument) and rss is residual sum-of-squares (RSS) of the model. So, based on Table 5, the most important variable is distance to rivers. Other important variables to explain the spatial distribution of springs in the study area are altitude, land use, slope aspect, distance to faults and TWI. In this kind of model the importance of the independent variables should be interpreted with caution (Donati and Turrini 2002; Gutiérrez et al. 2009). The groundwater spring potential map produced by the MARS model was created according to Eq. (8), and is presented in Fig. 11, and Table 6.

$$
\begin{aligned}
\mathbf{Spm}_{\mathbf{MARS}} = {} & [(+1.039548) - 0.268 * \max(0,\ \text{Landuse-5}) + 0.129 * \max(0,\ \text{FaultDens-6.098443}) \\
& - 0.00007 * \max(0, 14689.25\text{-FaultDis}) - 0.029 * \max(0,\ \text{5-Landuse}) * \max(0,\ \text{6-Aspect}) \\
& - 0.047 * \max(0,\ \text{5-Landuse}) * \max(0,\ \text{FaultDens-2.266543}) \\
& + 0.000009 * \max(0,\ \text{Landuse-2}) * \max(0, 14689.25\ \text{-FaultDis}) \\
& - 0.00009 * \max(0,\ \text{5-Landuse}) * \max(0,\ \text{FaultDis-10642.18}) \\
& + 0.004 * \max(0,\ \text{Aspect-6}) * \max(0,\ \text{RiverDis-323.1099}) \\
& - 0.015 * \max(0,\ \text{Aspect-6}) * \max(0,\ \text{Slope-14.22205}) \\
& + 0.001 * \max(0,\ \text{1754-Dem}) * \max(0, 2.09658\text{-RiverDens}) \\
& - 0.0005 * \max(0,\ \text{1754-Dem}) * \max(0, 3.109633\text{-Slope}) \\
& - 0.00009 * \max(0,\ \text{1754-Dem}) * \max(0, 17.7175\text{-TWI}) \\
& + 0.00005 * \max(0, 6.098443\text{-FaultDens}) * \max(0, \text{FaultDis-13788.58}) \\
& - 0.0000003 * \max(0, 14689.25\text{-FaultDis}) * \max(0,\ \text{RiverDis-161.5549}) \\
& + 0.00001 * \max(0, 8813.518\text{-FaultDis}) * \max(0, 13.54917\text{-LS}) \\
& + 0.0000004 * \max(0, 14689.25\text{-FaultDis}) * \max(0,\ \text{LS-3.2082}) \\
& - 0.013 * \max(0, \text{RiverDens-2.123947}) * \max(0, 13.54917\text{-LS}) \\
& - 0.047 * \max(0, 2.123947\text{-RiverDens}) * \max(0, 13.54917\text{-LS}) \\
& + 0.002 * \max(0, 60\text{-RiverDis}) * \max(0, 13.2875\ \text{TWI}).
\end{aligned}
\tag{8}
$$

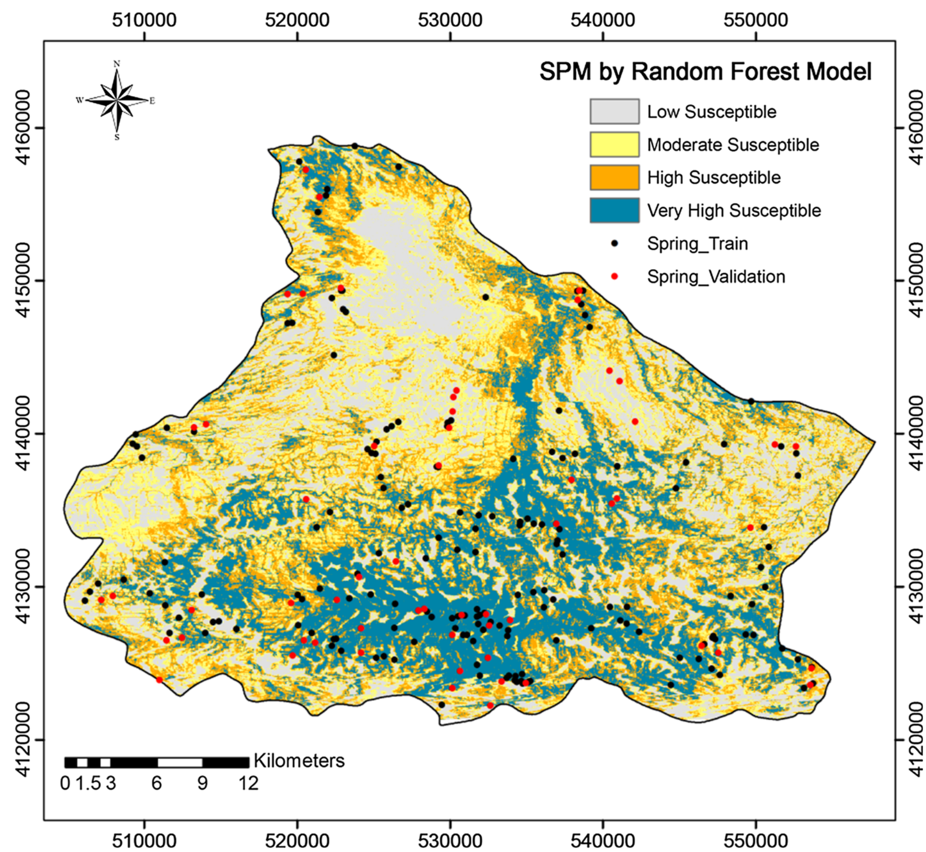


Fig. 10 Spring potential map produced by the random forest (RF) model

Table 5 The distribution of the explanatory variables and areas with respect to the spring occurrence potential zones

Factor	nsubset	GCV	RSS
Distance to rivers (m)	19	100.00	100.00
Altitude (m)	18	83.30	88.80
Land use	17	76.20	83.30
Slope aspect	17	76.20	83.30
Distance to faults (m)	17	76.20	83.30
TWI	17	76.20	83.30
Slope-length (m)	16	69.80	78.30
Slope degree (degree)	15	63.30	73.40
Drainage density (km/km^2)	14	56.20	68.30
Fault density (km/km^2)	12	46.00	60.10

Validation of groundwater spring potential models

A key step in statistical modeling is an assessment of its quality is validation (Chung-Jo and Fabbri 2003). In this study, the spring potential model quality was validated using an independent dataset that was not used for constructing and building of model. From the 234 springs identified, 176 (70 %) locations were employed to produce spring potential maps, while the remaining 58 (30 %) cases were withheld for model validation. To determine the accuracy of applied models (MARS and RF), two verification methods—success rate and prediction rate curves—were used by comparing the existing spring locations with the two spring potential maps (Figs. 12, 13).

One method to represent the quality of deterministic and probabilistic models is the receiver operating characteristic (ROC) curve (Swets 1988). The area under the ROC curve (AUC) shows the forecast model quality by describing the model's capability to forecast correctly the occurrence or non-occurrence of pre-defined "events" (Negnevitsky 2002). The ROC curve draws the false positive rate on the X axis and the true positive rate on the Y axis and evaluates the trade-off between the two rates (Negnevitsky 2002). To obtain values for each prediction pattern, the calculated index values of all cells in the study area were sorted in descending order (Pradhan et al. 2010a, b). If the area under the ROC curve (AUC) is close to 1.0, the result of the test would be excellent. On the contrary, AUC of 0.5 indicates performance equivalent to random chance. Using the spring potential map grid cells in the training dataset, the success-rate results were calculated. The success rate curves were gained using the 70 % training dataset (176 spring locations). Figure 12a, b illustrates the ROCs for the two spring potential maps in this study. The FR and MARS models have nearly the same area under the curve (AUC)

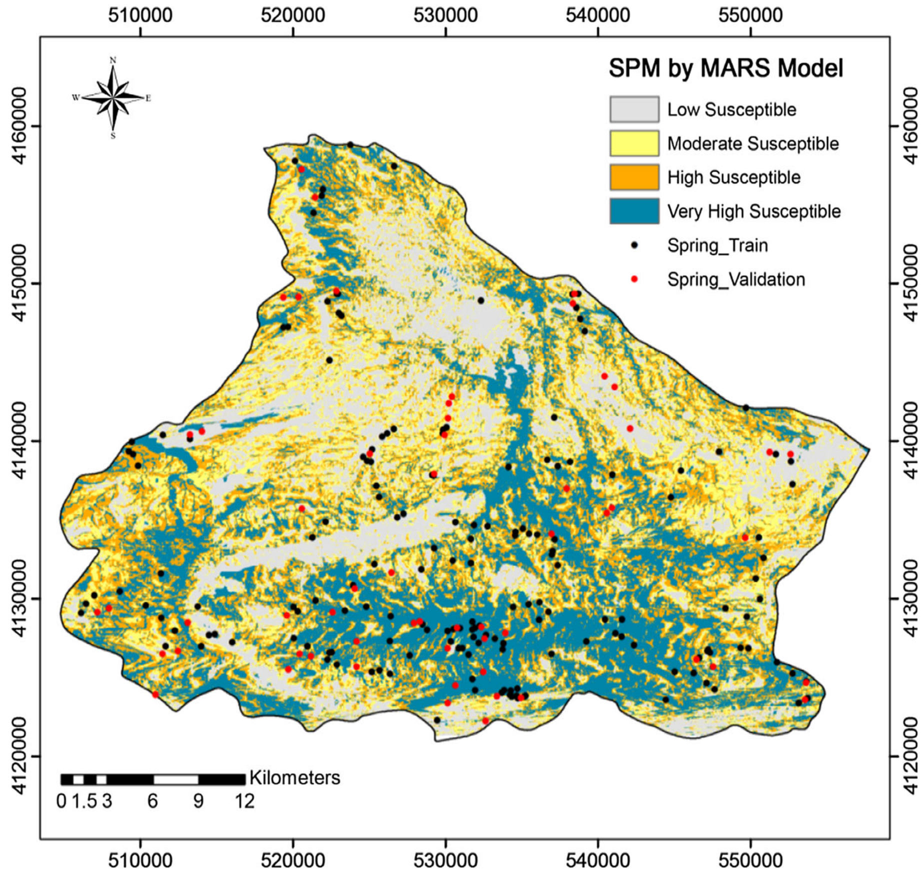


Fig. 11 Spring potential map produced by the multivariate adaptive regression spline (MARS) model

Table 6 The relative distribution of spring occurrence categories obtained by the two modeling approaches

Spring potential mapping	RF Area (%)	MARS Area (%)
Low	24.38	23.63
Moderate	25.28	22.99
High	25.43	25.57
Very high	24.92	27.81

values (0. 79). The success rate alone is not a suitable technique for judging the models prediction power (Tien Bui et al. 2012), however, because the success rate technique utilize the training spring pixels that have already been employed for constructing the spring models. However, the prediction rate method may help to understand how well the resulting spring potential maps have classified the areas of existing spring (Tien Bui et al. 2012). The prediction rate describes how well the model and predictor variables predict the spring (Lee and Pradhan 2007; Tien Bui et al. 2012; Pourghasemi et al. 2012). The results of the ROC curve test or prediction rate are shown in Fig. 13a, b. These curves show that the MARS model has relatively higher prediction performance (AUC = 0.7326) than the RF model (AUC = 0.7098).

Discussion and conclusion

Groundwater occurrence and movement are most basically controlled by the aquifer's permeability and the lithology of the underlying strata (Shahid et al. 2000; Ozdemir 2011). Especially in a fractured bedrock aquifer, movement of groundwater is governed by many other factors including topography, lithology, geological structures, fractures (density, aperture and connectivity), secondary porosity, groundwater recharge, drainage pattern, land-forms, land cover, and climatic conditions (Oh et al. 2011). Assessment of spring occurrence potential has become a valuable subject for water resource management authorities, and for regional land-use planning and environmental preservation. In the past, various methods have been applied to this task. In this study, groundwater potential maps were identified using MARS and RF models, predicting spring occurrence based on mappable explanatory variables. At first, using compiled information of Iranian Department Water Resources Management and extensive field investigations, a spring inventory map was prepared. Then, 13 data layers (altitude, slope aspect, slope degree, slope length, TWI, plan curvature, profile curvature, land use, lithology, distance to rivers, drainage density, distance to faults and fault

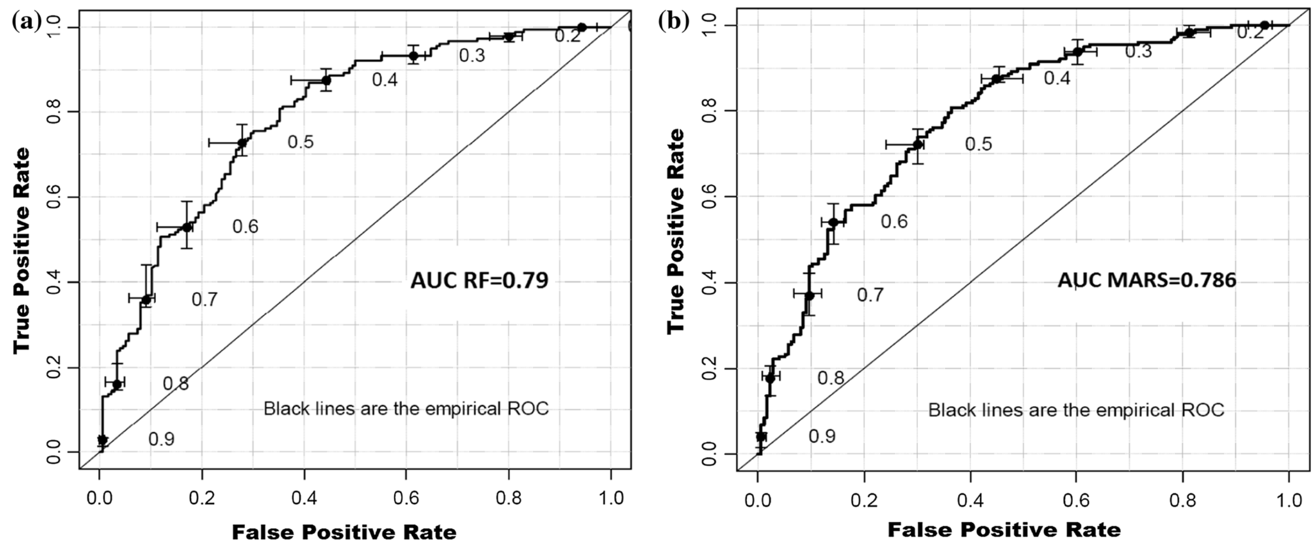


Fig. 12 Receiver operating characteristic (ROC) *curve* for the spring potential maps produced by **a** RF and **b** MARS model

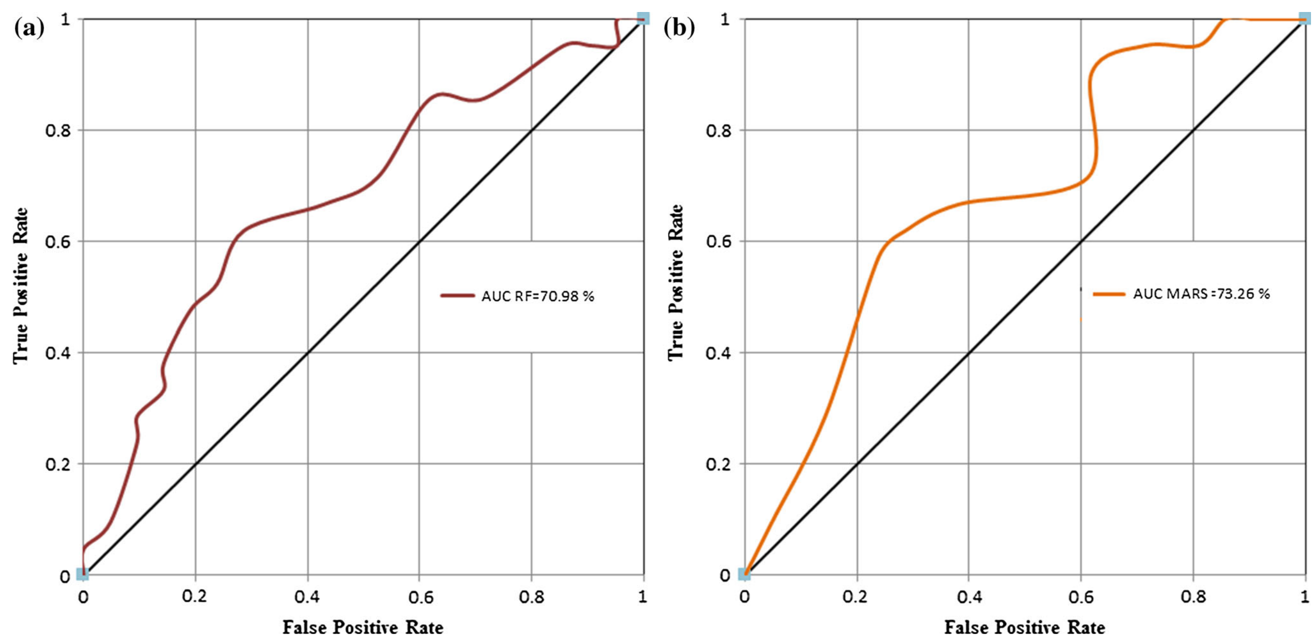


Fig. 13 Prediction Receiver operating characteristic (ROC) *curve* for the spring potential maps produced by **a** RF and **b** MARS model

density) were derived from the spatial database for use as explanatory variables. Using these explanatory variables, groundwater spring potential maps were produced using statistical modeling techniques random forest (RF) and multivariate adaptive regression spline (MARS). Carranza and Hale (2002) noted that expert knowledge is required to divide the dataset into training and validation data. For this reason, of 234 observed spring locations, 176 (70 %) cases were used as training data and the remaining 58 (30 %) was used for validation. AUC curves were prepared for the two models to test their accuracy. The validation results indicated that the MARS model has rather better predication accuracy (73.26 %) than the RF (70.98 %) model. The RF technique has several the advantages that growing large numbers of trees does not overfit the data, and random predictor selection keeps bias low, providing better models for prediction (Prasad et al. 2006). However, the RF model is prone to over fitting for some very noisy datasets and it do perform well when a majority of input variables are irrelevant (Breiman 2001). The MARS technique has advantages over traditional regression-based analyses. MARS picks only the most important explanatory variables from a user specified order. That is, the user may choose to include multiple variables at the beginning of the analysis and MARS will select out only the most important ones to include in the final result. This pruning process omits variables that have limited efficacy in the prediction of the outcome measure (Kennison and Cox 2013).

These techniques can be used in other areas, but they must be tuned to regions with similar characteristics to reflect the diversity of settings in which spring occur. However, other statistical modeling techniques may be suitable and more comparison would help guide selection of the best technique for a given. As a final conclusion, groundwater spring potential maps can be useful for planners and engineers in water-resource management and land-use planning. These spring potential maps can be applied to groundwater management and groundwater resource exploration, and the ability to create the maps using statistical modeling techniques shows great promise in wider application of spring potential mapping.

Acknowledgments The authors would like to thank Dr. Michael Fienen at the USGS Wisconsin Water Science Center for revising of language of manuscript. Also, we gratefully acknowledge of Editor-in-Chief Prof. James W. LaMoreaux and the two anonymous reviewers for their helpful comments on the previous version of the manuscript.

References

Balashi MS, McGuirez AD, Duffy P, Flannigan M, Walsh J, Melillo J (2009) Assessing the response of area burned to changing climate in western boreal North America using a Multivariate Adaptive Regression Splines (MARS) approach. Glob Change Biol 15:578–600. doi:10.1111/j.1365-2486.2008.01679.x

Bera K, Bandyopadhyay J (2012) Ground water potential mapping in Dulung watershaed using remote sensing and GIS techniques, West Bangal, India. Int J Sci Res Publ 2(12):1–7

Beven K, Kirkby MJ (1979) A physically based, variable contributing area model of basin hydrology. Hydrol Sci Bull 24:43–69

Breiman L (2001) Random forests. Mach Learn 45(l):5–32

Breiman L, Cutler A (2006) Random Forests. http://stat-www.berkeley.edu/users/breiman/RandomForests/cchome.htm

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Chapman & Hall/CRC

Calle ML, Urrea V (2010) Letter to the editor: stability of random forest importance measures. Brief Bioinform 12(1):86–89

Carranza EJM, Hale M (2002) Evidential belief functions for data-driven geologically-constrained predictive mapping of gold potential, Baguio district, Philippines. Ore Geol Rev 22:117–132

Catani F, Lagomarsino D, Segoni S, Tofani V (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci 13:2815–2831

Chung CF, Leclerc Y (1994) A quantitative technique for zoning landslide hazard. International Association for Mathematical Geology Annual Conference, Quebec, pp 87–93

Chung-Jo F, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. Nat Hazards 30:451–472

Conoscenti CH, Ciaccio M, Caraballo-Arias NA, Gómez-Gutiérrez A, Rotigliano E, Agnesi V (2014) 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. doi:10.1016/j.geomorph.2014.09.020

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:317–403

Davoodi Moghaddam D, Rezaei M, Pourghasemi HR, Pourtaghie ZS, Pradhan B (2013) Groundwater spring potential mapping using bivariate statistical model and GIS in the Taleghan Watershed, Iran. Arab J Geosci. doi:10.1007/s12517-013-1161-5

Donati L, Turrini MC (2002) An objective method to rank the importance of the factors predisposing to landslides with the GIS methodology: application to an area of the Apennines (Valnerina; Perugia Italy). Eng Geol 63:277–289

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–14

Ganapuram S, Vijaya Kumar GT, Murali Krishna IV, Kahya E, Demirel MC (2009) Mapping of groundwater potential zones in the Musi basin using remote sensing data and GIS. Adv Eng Softw 40:506–518

Geology Survey of Iran (GSI) (1997) http://www.gsi.ir/Main/Lang_en/index.html

Godebo TR (2005) Application of remote sensing and GIS for geological investigation and groundwater potential zone identification, Southeastern Ethiopian Plateau, Bale Mountains and the surrounding areas. M.Sc. Thesis. Addi Ababa University, p. 89

Gutiérrez AG, Schnabel S, Contador JFL (2009) Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. Ecol Model 220:3630–3637

Heckerman D (1986) Probabilistic interpretation of MYCIN's certainty factors. In: Kanal LN, Lemmer JF (eds) Uncertainty in artificial intelligence. Elsevier, New York, pp 298–311

Israil M, Al-hadithi M, Singhal DC, Kumar B, Rao MS, Verma K (2006) Groundwater resources evaluation in the Piedmont zone of Himalaya, India, using isotope and GIS technique. J Spatial Hydrol 6(1):34–38

Jaiswal RK, Mukherjee S, Krishnamurthy J, Saxena R (2003) Role of remote sensing and GIS techniques for generation of groundwater prospect zones towards rural development: an approach. Int J Remote Sens 24:993–1008

Jha MK, Chowdhury A, Chowdary VM, Peiffer S (2007) Groundwater management and development by integrated remote sensing and geographic information systems: prospects and constraints. Water Resour Manage 21:427–467

Kaliraj S, Chandrasekar N, Magesh NS (2013) Identification of potential groundwater recharge zones in Vaigai upper basin, Tamil Nadu, using GIS-based analytical hierarchical process (AHP) technique. Arab J Geosci. doi:10.1007/s12517-013-0849-x

Kanungo DP, Sarkar S, Sharma Sh (2011) Combining neural network with fuzzy, certainty factor and likelihood ratio concepts for spatial prediction of landslides. Nat Hazards 59(3):1491–1512

Kennison RF, Cox J (2013) Health and functional limitations predict depression scores in the health and retirement study; results straight from MARS. Calif J Health Promot 11(1):97–108

Lee S, Pradhan B (2006) Probabilistic landslide hazards and risk mapping on Penang Island, Malaysia. J Earth Syst sci 115(6):661–667

Lee S, Pradhan B (2007) Landslide hazard mapping at Selangor, using frequency ratio and logistic regression models. Landslides 4:33–41

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2:18–22

Mair A, El-Kadi AI (2013) Logistic regression modeling to assess groundwater vulnerability to contamination in Hawaii, USA. J Contam Hydrol 153:1–23. doi:10.1016/j.jconhyd.2013.07.004
Micheletti N, Foresti L, Robert S, Leuenberger M, Pedrazzini A, Jaboyedoff M, Kanevski M (2014) Machine learning feature selection methods for landslide susceptibility mapping. Math Geosci 46:33–57
Milborrow S (2012) Derived from mda: MARS by Trevor Hastie and Rob Tibshirani: multivariate Adaptive Regression Spline Models. R package version 3.2-2. http://CRAN.R-project.org/package=earth
Moore ID, Burch GJ (1986) Sediment transport capacity of sheet and rill flow: application of unit stream power theory. Water Resour 22:1350–1360
Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modeling: a review of hydrological, geomorphological and biological applications. Hydrol Pro 5:3–30
Murugesan B, Thirunavukkarasu R, Senapathi V, Balasubramanian G (2012) Application of remote sensing and GIS analysis for groundwater potential zone in Kodaikanal Taluka, South India. Earth Sci 7(1):65–75
Naghibi A, Pourghasemi HR (2015) A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods for groundwater potential mapping in Iran. Water Resour Manage 29(14):5217–5236. doi:10.1007/s11269-015-1114-8
Naghibi SA, Pourghasemi HR, Pourtaghi ZS, Rezaei A (2014) Groundwater qanat potential mapping using frequency ratio and Shannon's entropy models in the Moghan watershed, Iran. Earth Sci Inform. doi:10.1007/s12145-014-0145-7
Naghibi SA, Pourghasemi HR, Dixon B (2016) Groundwater spring potential using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ Monit Assess. doi:10.1007/s10661-015-5049-6
Negnevitsky M (2002) Artificial Intelligence: a guide to intelligent systems. AddisonWesley/Pearson Education, Harlow, p 394
Oh HJ, Lee S (2010) Cross-validation of logistic regression model for landslide susceptibility mapping at Geneoung areas, Korea. Disaster Adv 3(2):44–55
Oh HJ, Kim YS, Choi JK, Lee S (2011) GIS mapping of regional probabilistic groundwater potential in the area of Pohang City, Korea. J Hydrol 399:158–172
Ozdemir A (2011) GIS-based groundwater spring potential mapping in the Sultan Mountains (Konya, Turkey) using frequency ratio, weights of evidence and logistic regression methods and their comparison. J Hydrol 411:290–308
Pourghasemi HR, Pradhan B, Gokceoglu C (2012) Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. Nat Hazards 63(2):965–996
Pourghasemi HR, Pradhan B, Gokceoglu C, Mohammadi M, Moradi HR (2013) Application of weights-of-evidence and certainty factor models and their comparison in landslide susceptibility mapping at Haraz watershed, Iran. Arab J Geosci 6(7):2351–2365
Pourtaghi ZS, Pourghasemi HR (2014) GIS-based groundwater spring potential assessment and mapping in the Birjand Township, southern Khorasan Province, Iran. Hydrogeol J 2(3):643–662
Pradhan B, Lee S, Buchroithner MF (2010a) A GIS-based back-propagation neural network model and its cross-application and validation for landslide susceptibility analyses. Comput Environ Urban Syst 34(3):216–235
Pradhan B, Lee S, Buchroithner MF (2010b) Remote sensing and GIS-based landslide susceptibility analysis and its cross-validation in three test areas using a frequency ratio model. Photogramm Fernerkund Geo Inform 1:17–32. doi:10.1127/14328364/2010/0037
Prasad A, Iverson L, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9(2):181–199
Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaurmann, SanMateo
Rahmati O, Pourghasemi HR, Melesse A (2016) Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran Region, Iran. Catena 137:360–372. doi:10.1016/j.catena.2015.10.010
Rahmati O, Samani AN, Mahdavi M, Pourghasemi HR, Zeinivand H (2015) Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS. Arab J Geosci 8 (9):7059–7071
Rodriguez-Galiano V, Mendes MP, Garcia-Soldado MJ, Chica-Olmo M, Ribeiro L (2014) Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain). Sci Total Environ 476–477:189–206
Saha D, Dhar YR, Vittala SS (2010) Delineation of groundwater development potential zones in parts of marginal Ganga Alluvial Plain in South Bihar, Eastern India. Environ Monit Assess 165:179–191
Samui P, Kothari DP (2012) A multivariate adaptive regression spline approach for prediction of maximum shear modulus (G_{max}) and minimum damping ratio. Eng J 16(5):69–77
Sarkar S, Kanungo DP (2004) An integrated approach for landslide susceptibility mapping using remote sensing and GIS. Photogram Eng Remote Sens 70(5):617–625
Shahid S, Nath SK, Roy J (2000) Groundwater potential modeling in a soft rock area using a GIS. Int J Remote Sens 21(9):1919–1924
Shortliffe EH, Buchanan GG (1975) A model of inexact reasoning in medicine. Math Biosci 23:351–379
Sidle RC, Ochiai H (2006) Landslides: processes, prediction, and land use. American Geophysical Union, Washington, DC **312 pp**
Solomon S, Quiel F (2006) Groundwater study using remote sensing and geographic information systems (GIS) in the central highlands of Eritrea. Hydrol J 14:729–741
Sorichetta A, Ballabio C, Masetti M, Robinson GR Jr, Sterlacchini S (2013) A comparison of data-driven groundwater vulnerability assessment methods. Ground Water 51(6):866–879. doi:10.1111/gwat.12012
Swets JA (1988) Measuring the accuracy of diagnostic systems. Sciene 240:1285–1293
Talebi A, Uijlenhoet R, Troch PA (2007) Soil moisture storage and hillslope stability. Nat Hazards Earth Syst Sci 7:523–534
Tien Bui D, Pradhan B, Lofman O, Revhaug I, Dick OB (2012) Spatial prediction of landslide hazards in Hoa Binh province (Vietnam): a comparative assessment of the efficacy of evidential belief functions and fuzzy logic models. Catena 96:28–40
Waikar ML, Nilawar AP (2014) Identification of Groundwater Potential Zone using Remote Sensing and GIS Technique. Int J Innov Res Sci Eng Technol 3(5):1264–1274
Williams G (2011) Data mining with rattle and R (The art of excavating data for knowledge discovery series), 1st edn. Springer-Verlag, New York. doi:10.1007/978-1-4419-9890-3
Yao D, Yang J, Zhan X (2013) A novel method for disease prediction: hybrid of random forest and multivariate adaptive regression splines. J comput 8(1):170–177
Yesilnacar EK (2005) The application of computational intelligence to landslide susceptibility mapping in Turkey. Ph.D Thesis. Department of Geomatics the University of Melbourne, p. 423

Youssef AM, Pourghasemi HR, Pourtaghi Z, Al-Katheeri MM (2015) Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir region, Saudi Arabia. Landslides, doi:10.1007/s10346-015-0614-1

Zare M, Pourghasemi HR, Vafakhah M, Pradhan B (2013) Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multi-layer perceptron (MLP) and radial basic function (RBF) algorithms. Arab J Geosci 6(8):2873–2888