

# A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS

Seyed Amir Naghibi<sup>1</sup> · Hamid Reza Pourghasemi<sup>2</sup> · Karim Abbaspour<sup>3</sup>

Received: 7 August 2016 / Accepted: 19 December 2016 / Published online: 5 January 2017  
© Springer-Verlag Wien 2017

**Abstract** Considering the unstable condition of water resources in Iran and many other countries in arid and semi-arid regions, groundwater studies are very important. Therefore, the aim of this study is to model groundwater potential by qanat locations as indicators and ten advanced and soft computing models applied to the Beheshtabad Watershed, Iran. Qanat is a man-made underground construction which gathers groundwater from higher altitudes and transmits it to low land areas where it can be used for different purposes. For this purpose, at first, the location of the qanats was detected using extensive field surveys. These qanats were classified into two datasets including training (70%) and validation (30%). Then, 14 influence factors depicting the region's physical, morphological, lithological, and hydrological features were identified to model groundwater potential. Linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), flexible discriminant analysis (FDA), penalized discriminant analysis (PDA), boosted regression tree (BRT), random forest (RF), artificial neural network (ANN), K-nearest neighbor (KNN), multivariate adaptive regression splines (MARS), and support vector machine (SVM) models were

applied in R scripts to produce groundwater potential maps. For evaluation of the performance accuracies of the developed models, ROC curve and kappa index were implemented. According to the results, RF had the best performance, followed by SVM and BRT models. Our results showed that qanat locations could be used as a good indicator for groundwater potential. Furthermore, altitude, slope, plan curvature, and profile curvature were found to be the most important influence factors. On the other hand, lithology, land use, and slope aspect were the least significant factors. The methodology in the current study could be used by land use and terrestrial planners and water resource managers to reduce the costs of groundwater resource discovery.

**Keywords** Groundwater mapping · Soft computing models · Geographic information system · Iran

## 1 Introduction

Average annual precipitation in Iran has been measured as 270 mm, which is less than one third of the world's annual average (Mahdavi 2004; Chezgi et al., 2015). Seventy percent of the precipitation falls on northern parts of the country, while other parts receive inadequate amount (Ahmadi et al. 2010). According to Mahdavi (2004), in the case of temporal distribution of the precipitation, a low share of the precipitation occurs during plan growing season. The rapid increase in the population of Iran in the last decades has caused an increasing demand for fresh water, especially for groundwater resources. Considering the mentioned facts, efficient management, conservation plans, and sustainable use of groundwater resources are necessary activities needed.

Qanat is an underground construction that collects and delivers groundwater from highland mountainous areas to

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

Seyed Amir Naghibi  
amir.naghibi2010@yahoo.com; amir.naghibi@modares.ac.ir

<sup>1</sup> Young Researchers and Elite Club, Mashhad Branch, Islamic Azad University, Mashhad, Iran

<sup>2</sup> Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran

<sup>3</sup> Eawag, Swiss Federal Institute for Aquatic Science and Technology, Ueberlandstr. 133, P.O. Box 611, 8600 Dübendorf, Switzerland

lowland residential areas for different purposes especially drinking water and farming (Perrier and Salkini 1991; Nazari Samani and Farzadmehr 2006) (Fig. 1). Qanat technology was developed by the Persian people and spread from there to more than 34 countries in Asia, Africa, and Europe (Naghibi et al. 2015). There are about 32,164 active qanat systems in Iran with approximately 9 billion cubic meters (bcm) of total discharge (Naghibi et al. 2015). More description on qanat and its structure can be found in Nazari Samani and Farzadmehr (2006) and Naghibi et al. (2015).

One new method which has provided much useful information for water resource managers is groundwater potential mapping (GPM). Different researchers have used different groundwater indicators for groundwater potential mapping such as spring, qanat, and well. In some studies, spring locations and different models were used for producing GPMs such as frequency ratio (FR) (Oh et al. 2011; Pourtaghi and Pourghasemi 2014; Davoodi Moghaddam et al. 2015); weights-of-evidence (WE) (Ozdemir 2011a); logistic regression (LR) (Ozdemir 2011a; Pourtaghi and Pourghasemi 2014); index of entropy (IE) (Naghibi et al. 2015); artificial neural network (ANN) (Lee et al. 2012a, b); analytical hierarchy process (AHP) (Rahmati et al. 2014; Razandi et al. 2015); evidential belief function (EBF) (Pourghasemi and Beheshtirad 2014); and data mining models such as random forest, classification and regression trees, and boosted regression tree (Naghibi and Pourghasemi 2015; Naghibi et al. 2016; Zabihi et al. 2016).

In the last decade, many predictive models have been used in other fields of sciences and engineering such as soil, spectroscopy, landslide susceptibility mapping, food quality, air quality, land use, and flood susceptibility mapping, including flexible discriminant analysis (FDA) (Tebaldi et al. 2002), k-

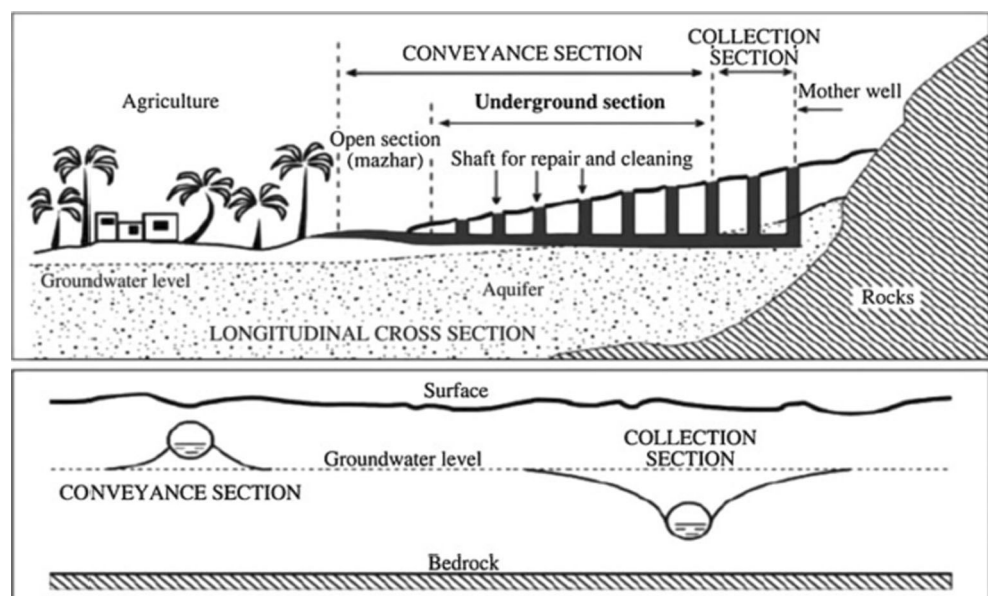
nearest neighbor classification (KNN) (Paraskevas et al. 2015), linear discriminant analysis (LDA) (Ramos-Canon et al., 2015; Eker et al. 2015), multivariate adaptive regression spline (MARS) (Felicisimo et al. 2012; Samui and Kurup 2012; Tayyebi and Pijanowski 2014), penalized discriminant analysis (PDA) (Granitto et al. 2008; Zhu and Tan 2016) quadratic discriminant analysis (QDA) (Eker et al. 2015), and SVM (Ballabio and Sterlacchini 2012; Kavzoglu et al. 2014; Hong et al. 2015; Marjanović et al. 2011; Pourghasemi et al. 2013; Tehrany et al. 2014; Tehrany et al. 2015; Tien Bui et al. 2015; Karami et al. 2015), and the results were reported satisfactory.

Considering the aforementioned literature, there are two major novelties in current study: (i) the use of the ten advanced and soft computing models for groundwater qanat potential mapping and (ii) consideration of qanats as indicator for groundwater potential.

## 2 Study area

The study area lies between 31°50'36"N and 32°34'16"N latitude and 51°26'57"E and 59°21'51"E longitude with an area of 2321 km<sup>2</sup>. Beheshtabad is located in Chaharmahal-e-Bakhtiari Province, Iran. The elevation of the study area differs from 1660 to 3560 m with an average of 2301 m. The mean yearly rainfall in Beheshtabad is 618.8 mm (Mojiri and Zarei 2006). Based on the Geological Survey of Iran (GSI 1997), the most part of the area is covered by the lithology as A in Table 1. The study area is consisted of main land use classes of agricultural (29.83%), orchard (1.37%), rangeland (66.25%), and residential (2.54%). There are 1425 springs and 228 qanats in the study area. People in the study area have a

**Fig. 1** A profile of qanat and its components (Nazari Samani and Farzadmehr 2006)



**Table 1** Lithological characteristics of Beheshtabad Watershed, Iran

Name	Lithology
A	Low-level piedmont fan and valley terraces deposit
B	Low weathering gray marls alternating with bands of more resistant shelly limestone
C	Pale-red, polygenic conglomerate, and sandstone
D	Undifferentiated metamorphic rocks, including phillite, meta-volcanics, calcschist, and crystallized limestone
E	Cream to brown-weathering, feature-forming, well-jointed limestone with intercalations of shale
F	Gray, thick-bedded, o’olitic, fetid limestone
G	Gray, thick-bedded to massive orbitolina limestone
H	High-level piedmont fan and valley terraces deposits
I	Marl and calcareous shale with intercalations of limestone
J	Polymictic conglomerate and sandstone
K	Undivided Bangestan Group, mainly limestone and shale, Albian to Campanian
L	Undivided Eocene rock
M	Un consolidated wind-blown sand deposits and back shore sand duns

high dependency to groundwater resources for water supply and other usages.

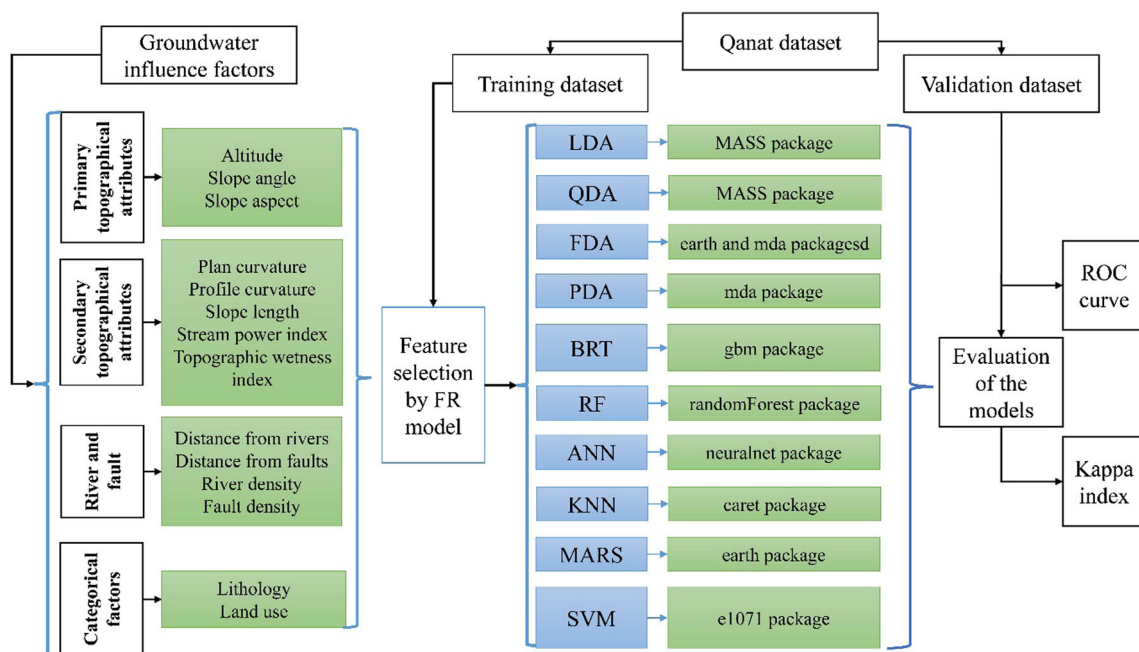
the best models using ROC curve, Cohen’s kappa, specificity, and sensitivity indices. Figure 1 lists the factors used and the processes applied in the analysis.

### 3 Methodology

The overall methodology (Fig. 2) includes the following: (1) preparing a qanat location map in the study area; (2) feature selection using frequency ratio model; (3) running ten models for groundwater qanat potential mapping; (4) validating the models; and (5) comparing different models and selection of

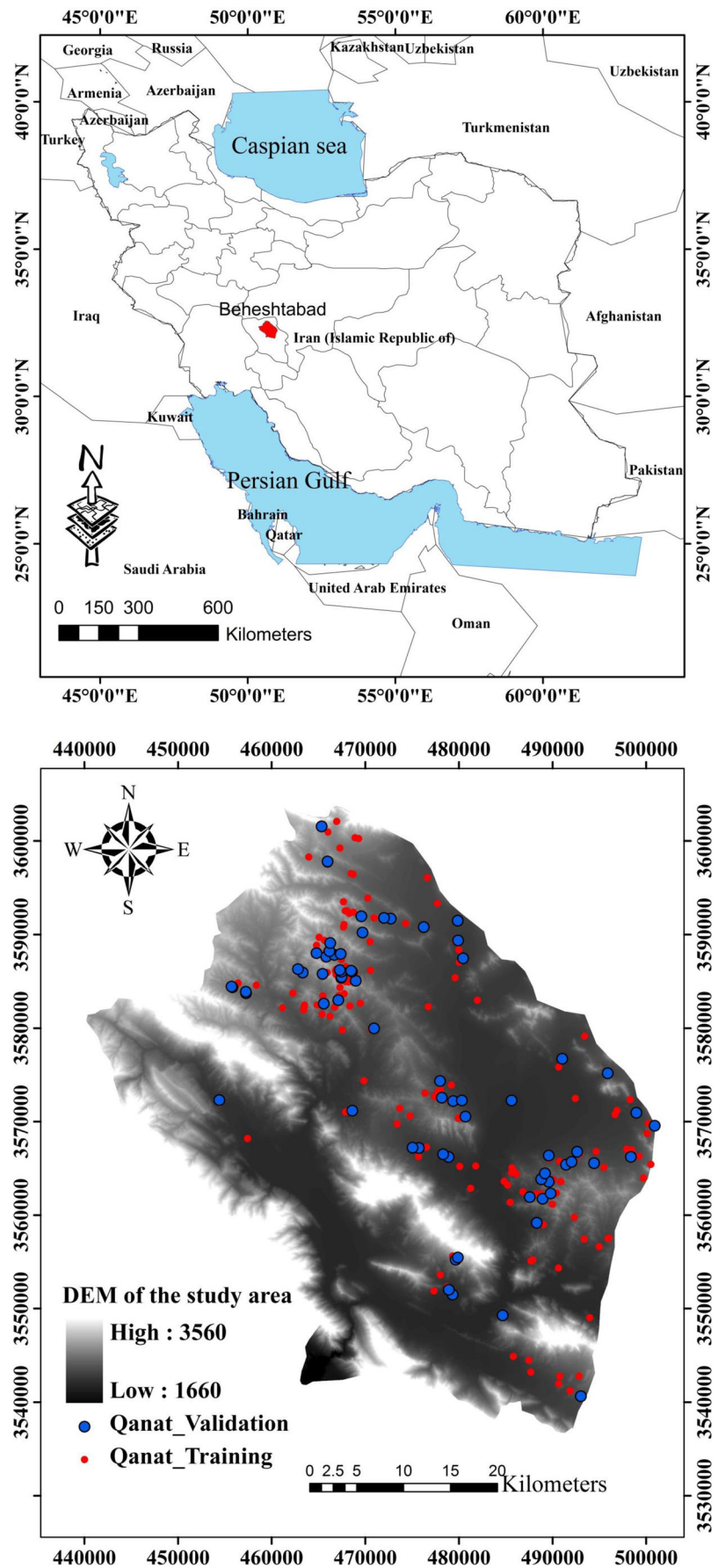
#### 3.1 Qanat dataset

In the current study, the location of qanats was detected using extensive field surveys and national reports. In total, 228 qanats were detected in the Beheshtabad Watershed and mapped at 1: 50,000 scale (Fig. 3). Water resources provided by qanat structures in the Beheshtabad area are used for



**Fig. 2** Overall methodological flow chart adopted in this study

**Fig. 3** Qanat locations with digital elevation model (DEM) map of the study area





different sections such as drinking water purposes, farming, wild life, etc. Qanats were divided into two groups: training dataset (70%, 160 qanats) and validation dataset (30%, 68 qanats) (Oh et al. 2011; Ozdemir 2011a, b; Lee et al. 2012a, b). For this, Hawth extension (Gutiérrez et al. 2009) was used to randomly classify qanat locations into two groups. Hawth extension provides tools for sampling such as generating random points and creating random selections. In addition, Hawth tools can be used to calculate and add some indices and add to the table of layers such as area, perimeter, length, etc.

### 3.2 Groundwater influence factors

To assess groundwater potentiality in this study, it is vital to consider several qanat-effective factors. Fourteen influencing groundwater factors were selected based on literature review (Naghibi et al. 2015) and data availability. These factors contain three primary topographical attributes (i.e., altitude, slope angle, slope aspect), five secondary topographical attributes (i.e., plan curvature, profile curvature, topographic wetness index, slope length, stream power index), four river and fault maps (i.e., distance from rivers and faults, rivers and faults density), and finally categorical factors (i.e., lithology and land use). These factors were then classified according to literature review (Oh et al. 2011a, b; Ozdemir, 2011a, b; Naghibi and Pourghasemi 2015).

#### 3.2.1 Primary topographical attributes maps

First, the digital elevation model (DEM) was extracted using the 1:50,000 scale topographic maps (contour lines and elevation points) in a 20-m cell size. Then, three primary factors including altitude, slope angle, and slope degree were calculated and classified. Altitude was divided into five classes (<2000, 2000–2400, 2400–2800, 2800–3200, and >3200 m) based on equal classification scheme (Fig. 4a). The slope angle map was prepared by dividing it into four classes: (0°–5°, 5°–15°, 15°–30°, and >30°) (Fig. 4b). Slope aspect was categorized into north, northeast, northwest, east, west, southeast, southwest, south, and flat based on normal or common standard classification (Fig. 4c).

#### 3.3 Secondary topographical attributes maps

Five secondary topographical attributes were used in the analysis: plan curvature, profile curvature, slope length (LS), stream power index (SPI), and topographic wetness index (TWI) (Fig. 3d–h). Plan curvature and profile curvature were prepared by using SAGA-GIS 2.8 (Fig. 4e). The plan curvature map was classified into three classes of convex, concave, and flat representing positive, negative, and zero values, respectively (Fig. 4d). Profile

curvature map was prepared and then categorized into < –0.001, –0.001 to 0.001, and >0.001 groups. LS was classified into four classes (0–20, 20–40, 40–60, and >60) (Fig. 4f). SPI was classified into four classes of (0–200, 200–400, 400–600, and >600) (Fig. 4g). TWI is a topographical index which shows the aptitude of water to gather at each point in the watershed (Moore et al. 1991). This factor was classified into (<8, 8–12, and >12) (Fig. 4h).

#### 3.3.1 River and fault maps

Four maps were extracted from river and fault layers including drainage density, distance from rivers, fault density, and distance from faults. Distance from river and drainage density was created using topographic maps of Beheshtabad watershed. In order to produce fault-related groundwater factors, distance from fault and fault density maps were calculated implementing a geological map of Beheshtabad watershed. Distance from river map was classified into (<100, 100–200, 200–300, 300–400, and >400 m) (Fig. 4i). Distance from fault map was classified into <250, 250–500, 500–750, 750–1000, and >1000 m (Fig. 4j). For drainage density and fault density, natural breaks were used for classification (Fig. 4k, l).

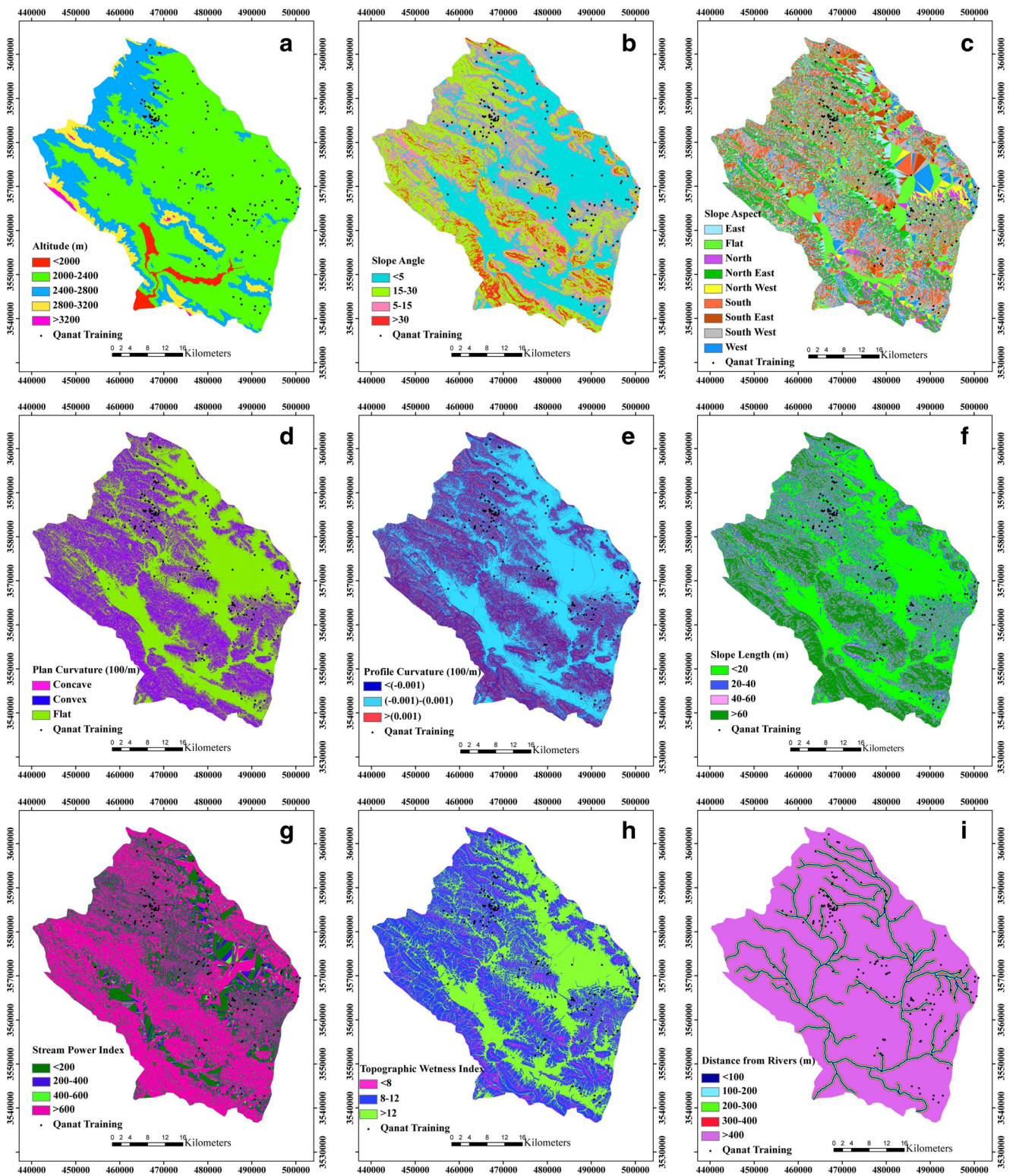
#### 3.3.2 Categorical factors (land use and lithology)

Using Landsat images, the land use map was created. Supervised classification and maximum likelihood algorithms were used to produce land use map of the study area. Four land use classes were identified including agricultural, residential, orchard, and rangeland (Fig. 4m).

The lithology map of the study area that has 13 lithological classes was created from a geological map obtained from Geology Survey of Iran (GSI) (1997). In the study area, there are different lithology classes which were classified from A to M (Table 1 and Fig. 4n).

### 3.4 Feature selection

FR is defined as the probability of happening of an exact event (Bonham-Carter, 1994). An FR value of greater than 1 for an influence factor indicates a high correlation (Oh et al. 2011). FR also shows the relationship between qanat occurrence and groundwater influence factors. Although it ignores the interactions between influence factors, it was used as an indicator to show whether the factors are useful alone or not. In this step, FR values were calculated to determine whether the factors are influence or not. For this, FR was calculated for each of the factors' classes by  $FR = \left(\frac{q}{Q} / \frac{p}{P}\right)$ . In this equation,  $q$  is



**Fig. 4** Topographical parameter maps of the study area: **a** altitude, **b** slope angle, **c** slope aspect, **d** plan curvature, **e** profile curvature, **f** slope length, **g** stream power index, **h** topographic wetness index, **i** distance from rivers, **j** distance from faults, **k** drainage density, **l** fault density, **m** land use, **n** lithology

the number of pixels with a qanat for each factor,  $Q$  shows the number of total qanats in study area,  $p$  is the number of pixels

in the class area of the factor, and  $P$  is the total number of pixels in the area.



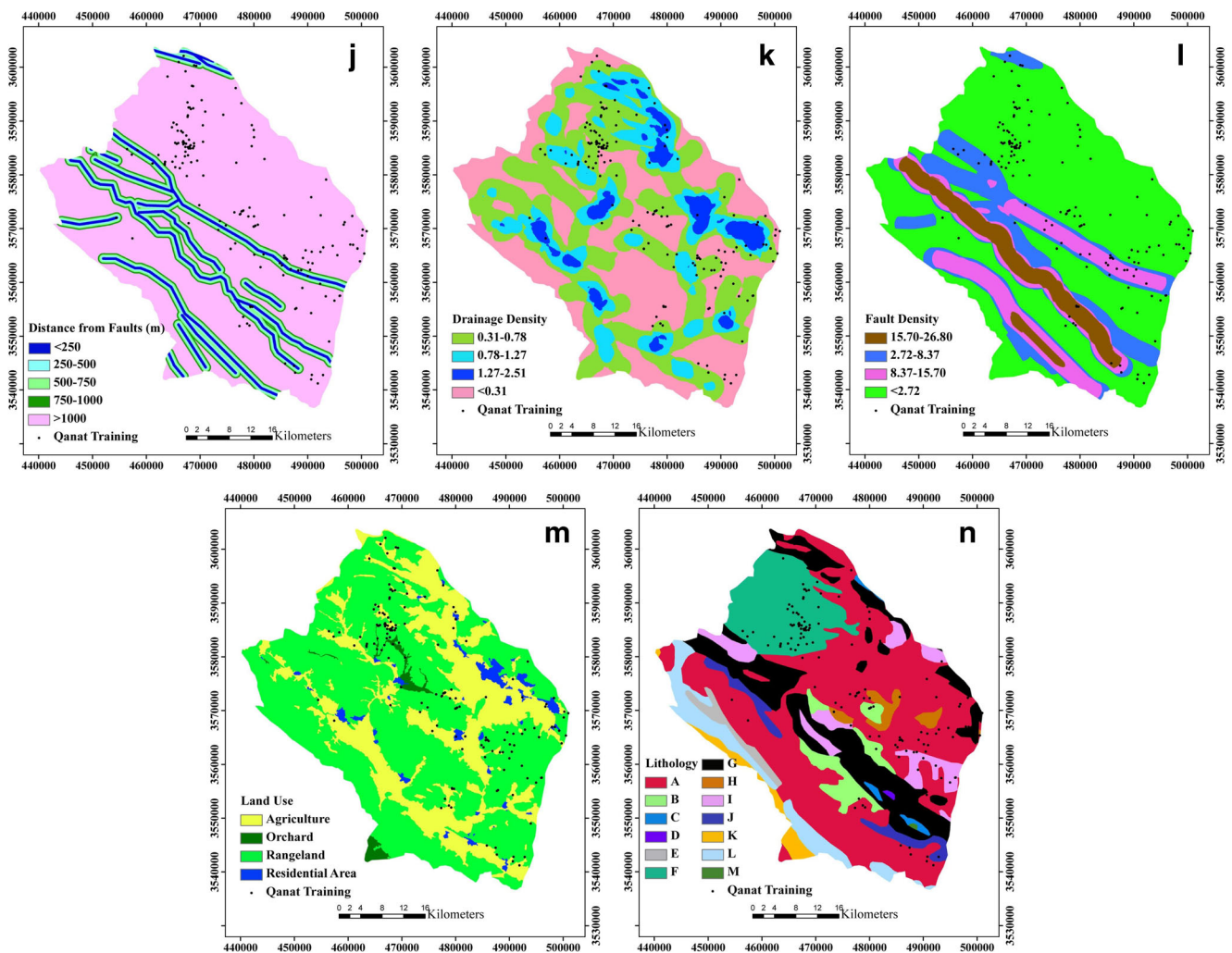


Fig. 4 (continued)

### 3.5 Soft computing models for groundwater qanat potential assessment

In this study, four discriminant analysis methods were implemented for modeling groundwater potential including LDA, QDA, FDA, and PDA models.

Fisher (1936) introduced a linear discriminant analysis model which can separate two classes an object by seeking for a linear combination of variables (Eker et al. 2015). In the LDA, the estimated values (i.e.,  $N$ ) are determined using a linear combination of a set of explanatory variables (i.e., effective factors) such as  $N = fX + q$  ( $q = \text{constant}$ ), which best differentiates the group of a case by finding the  $f$  coefficients (Eker et al. 2015). LDA has two assumptions including the existence of normal distribution in effective factors and close covariance values. For running LDA model, MASS package was used in R 3.0.2.

In quadratic discriminant analysis approach, a quadratic surface will be determined which could separate the group of a case. QDA searches for a group membership consisting of a matrix ( $m \times m$ ) ( $m = \text{number of effective factors}$ ) and a linear combination of these factors such that  $Q = x^T Vx + K^T x + f$  (Eker et al. 2015), where  $V$  shows the  $m \times m$  matrix of coefficients,  $K$  shows the linear combination coefficient, and  $f$  is a constant (Eker et al. 2015). In the QDA, the assumption of the normal distribution still exists, but there is no need for covariance values to be close. The QDA was analyzed using “MASS” package in R 3.0.2 as well.

FDA was developed to combine non-parametric regression models with LDA to get more flexibility in the decision boundaries (Mallet et al. 1996). This goal could be achieved by casting regression methods and classification methods into one framework (Mallet et al. 1996). On the other hand, FDA is the application

of LDA on the matrix achieved with regression methods and on the transformed class matrix (Reyn et al. 2006).

The penalized discriminant analysis developed by Hastie et al. (2001) is a regularized version of the traditional Fisher's LDA (Ripley 1996). PDA is suggested to be more appropriate for problems in which LDA has overfitting problem (Mallet et al. 1996). PDA, first, recasts LDA classification problem as a regularized linear regression problem. Then, it applies one of the many famous techniques available (Granitto et al. 2008).

Boosted regression tree uses both statistical and machine learning (SL and ML) techniques (Youssef et al. 2015). BRT has a different approach from traditional regression methods that produce a single best model (Elith et al. 2006; Leathwick et al. 2006). BRT combines a large number of simple tree models to improve the performance of prediction using boosting technique (Elith et al. 2006; Leathwick et al. 2006). In addition to boosting, the BRT also implements regression trees in the modeling process. Regression trees are categorized from the classification and regression tree approaches which are from decision tree group of models. An R script (gbm and dismo) was used to run BRT.

Random forest is a new non-parametric method and a very accurate classifier and robust against noise (Breiman, 2001). The algorithm extracts binary trees selected randomly that implement a sub-dataset of the feature observations through bootstrapping method. Then, RF selects a sub-dataset from the whole dataset to build the model (Zabihi et al. 2016). The data which is not included in the sub-dataset are called out-of-bag (OOB) (Breiman 2001; Catani et al. 2013). RF grows multiple decision trees on random sub-datasets from training dataset and related variables. Small changes in the training data cause a high variance in single classification trees and often results in rather low prediction accuracies (Breiman 1996). RF was fitted in R (R Development Core Team 2005) version 3.0.2, implementing the randomForest package (Ridgeway 2006). Artificial neural network comprises several layers of nodes (neurons) (Lee et al. 2012a, b). These layers of nodes exchange messages with each other (Lee et al. 2012a, b). Every node is connected to the other nodes in the next layer. Input layer includes influence factors and the output layer delivers one or more predictive values for the response variable(s), which in this case is the probability of qanat occurrence (Aertsen et al. 2010). Between them, there are one or more hidden layers and the network is trained by implementing an iterative method for determining the weights of the connections between the layers (Aertsen et al. 2010). In the current study, the back propagation (BP) algorithm with "mlpe" package was tested for groundwater qanat potential mapping and the results were compared with other soft computing models.

KNN classification is a non-parametric model for classification and regression problems (Chirici et al. 2015). Rote

classifier, first, memorizes the whole training data. Then, the mentioned Rote classifier classifies only when the features of the test object (i.e., qanat) match with one of the training objects (Wu et al. 2008). However, there is a drawback in this approach. Since many test objects do not match with any of the training records, they cannot be categorized (Wu et al. 2008). The best selection of the  $k$  value depends on the data. Everitt et al. (2011) mentioned that larger values of  $k$  results in a reduction in the effect of noise on the classification. However, it makes boundaries among classes less clear (Everitt et al. 2011). The KNN model fit in R 3.0.2 was evaluated using the "rminer" package.

MARS is a method that is implemented in order to fit the relationship between input (in this case, groundwater influence factors) and output variables (in this case, qanat occurrence) (Friedman, 1991). MARS combines three techniques to build a new model. These techniques are (1) constructing splines mathematically, (2) binary recursive partitioning (BRP), and (3) linear regression (LR) (Friedman 1991). This model determines the type of relationship between the response factor (qanat occurrence) and the groundwater influence factors which could be linear or non-linear (Hastie et al. 2001). The MARS model was fitted using the "earth" package in R 3.0.2.

Support vector machines (SVMs) are a set of machine learning techniques based on the concept of optimal separating hyperplane which are developed by Vapnik (1995). SVM can be thought as non-linear classifiers which aim to find the most extensive margin between two classes in feature space (Ballabio and Sterlacchini 2012). The SVM approach aims to (1) reduce the error test and (2) reduce the model complexity (Ballabio and Sterlacchini 2012). In other words, SVM depends on data which means that the model capacity is calibrated to match data complexity which is called structural risk minimization, a paradigm on which SVM is based (Vapnik 1995; Cherkassky and Mulier 2007). In this study, SVM was fit in R 3.0.2 using the rminer package.

### 3.6 Data-based sensitivity analysis

Data-based sensitivity analysis (DSA) was used in order to define the importance of the groundwater influence factors in the ANN, QDA, and SVM models. The sensitivity analysis (SA) dataset is consisted of  $N_s$  random sub-datasets which are selected from the initial dataset. In the SA dataset, all  $x_z$  values are altered by  $x_{ak}$  and the responses are gathered (Cortez and Embrechts 2013), where  $x_z: a \in (1, 2, 3, \dots, B)$  and  $x_{ak}$  is the first input level, and  $B$  represents input variables. In the next step, the previous function is repeated implementing a different  $j$  value ( $j \in \{1, 2, 3, \dots, L\}$ ,  $L = \text{level}$ ) (Cortez and Embrechts 2013). The mentioned process will be repeated



for all groundwater influence factors and results in a complexity of the order  $\vartheta(M * L * N_s * P)$ , where  $N_s$  is the length of the training samples (Cortez and Embrechts 2013).

### 3.7 Validation and comparison of the groundwater potential maps

For evaluating the models, ROC, sensitivity, specificity, kappa, and qanat density indices were used. Sensitivity is proportion of qanats (in this study) which are correctly estimated as qanat (Negnevitsky 2002). On the other hand, specificity is called to the proportion of the non-qanats (in this study) that are correctly predicted as non-qanat (Negnevitsky 2002). Tradeoff between these indices is called ROC curve (Negnevitsky 2002; Hong et al. 2016; Karimi et al. 2016) which was also calculated and used in this study. Kappa index was also calculated as below:

$$k = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

where,  $P_o$  is  $TP + TN/n$ , and  $P_e = (TP + FN)(TP + FP) + (FP + TN)(FN + TN)/\sqrt{N}$ . In these equations, TP is true positive, TN is called the true negative, FP is false positive, FN shows false negative, and  $N$  represents the total number of observations (Moosavi and Niazi 2015; Naghibi and Moradi Dashtpajardi 2016).

Qanat density (QD) was also computed for the implemented models as below:

$$QD = \frac{P_Q}{P_P} \quad (2)$$

where  $QD$  represents qanat density,  $P_Q$  depicts percentage of the qanats in each class of the GPMs, and  $P_P$  percentage of the pixels in each class of the GPMs.

## 4 Results

### 4.1 Results of the feature selection

According to Table 2, for slope angle, lower slopes had larger correlation with qanat occurrence as the class of 0–5 had the highest FR with a value of 1.71. For slope aspect, flat and southeast aspects had the highest FR values (1.87 and 1.34, respectively). For altitude, the class of 2000–2400 has the highest FR value (1.27). For plan curvature, the flat class had the highest FR with a value of 1.43, indicating a high probability qanat occurrence in this class. For the profile curvature class of  $-0.001$  to  $0.001$  had the highest FR value (1.20). Qanats are concentrated in areas with LS  $<20$  (FR value of 1.44), in areas with an SPI ( $<200$ ) (1.40), and in areas

with TWI  $>12$  (1.69). For distance from rivers, the class of  $<100$  had the highest FR value (2.63), while for distance from faults, the class of  $>1000$  m had the highest FR value (1.11). The results of river density showed that the 0.31–0.78 class had a high density of qanats with FR value of 1.30. In the case of fault density, the  $<2.72$  and 8.37–15.70 classes had the highest values (FR = 1.13, and 1.12, respectively). In respect of land use factor, orchard and agriculture had the highest qanat concentration with frequency values of 1.82 and 1.44, respectively. Finally, in the case of lithology, F and I had the highest FR values (2.72 and 1.30, respectively).

Considering the FR values, it can be concluded that all of the factors, in one or some classes, have higher FR values than 1 which implies that the factors can be used as influence factors on groundwater potential.

### 4.2 Results of the models

The results of the LDA indicated that GPM could be calculated as below:

$$\begin{aligned} GQPI_{LDA} = & -0.139 * \text{slope degree} - 0.071 * \text{slope aspect} \\ & + 0.002 * \text{altitude} - 32.299 * \text{plan curvature} \\ & - 22.828 * \text{profile curvature} + 0.011 * \text{slope length} \\ & - 0.0000006 * \text{stream power index} \\ & - 0.0004 * \text{topographic wetness index} - 0.0004 * \\ & \text{distance from rivers} + 0.00002 * \text{distance from faults} \\ & - 1.164 * \text{river density} \\ & - 0.015 * \text{fault density} - 0.031 * \text{land use} - 0.025 * \text{lithology} \end{aligned} \quad (3)$$

In the GPM obtained by using LDA, low potential, moderate potential, high potential, and very high potential classes cover 21.29, 25.56, 33.58, and 19.57% of the study area, respectively (Fig. 5a).

In the case of QDA, the GPM was grouped into four classes of low potential, moderate potential, high potential, and very high potential (Fig. 5b). The moderate potential GPM class derived using the QDA model covers 25.27% of the area; 24.69, 17.22, and 32.80% of the area are assigned to low potential, high potential, and very high potential GPM zones, respectively (Table 3). In this method, altitude, profile curvature, and plan curvature had the highest importance (Fig. 6a).

FDA resulted in a final model included degree value of 1 and nprune of 10. In this case, nprune depicts number of terms and degree represents product degree. In the GPM obtained by using the FDA model, high and very high classes covered 23.12 and 17.12% of the study area, while low potential and moderate potential classes of potentiality covered 32.25 and 24.52% of the area (Fig. 5c).

**Table 2** Spatial relationship between each effective factor and qanat locations using frequency ratio (FR) model

Factor	Class	No. of pixel in domain	No. of qanats	Frequency ratio (FR)
Slope angle (degree)	0–5	2,298,984	109	1.71
	5–15	1,462,498	37	0.91
	15–30	1,679,643	14	0.30
	>30	361,705	0	0
Slope aspect	Flat	735,195	38	1.87
	North	557,953	11	0.71
	Northeast	855,674	18	0.76
	East	632,044	13	0.74
	Southeast	593,011	22	1.34
	South	679,745	13	0.69
	Southwest	774,993	23	1.07
	West	545,134	15	0.99
Altitude (m)	Northwest	429,081	7	0.59
	<2000	138,471	0	0
	2000–2400	3,982,550	140	1.27
	2400–2800	1,338,961	20	0.54
	2800–3200	321,421	0	0
Plan curvature (100/m)	>3200	21,427	0	0
	Concave	1,305,957	28	0.77
	Flat	3,077,923	122	1.43
Profile curvature (100\m)	Convex	1,418,950	10	0.25
	<(-0.001)	1,423,662	40	1.01
	(-0.001) to (0.001)	3,197,653	106	1.20
Slope length (m)	>(0.001)	1,181,515	14	0.42
	<20	2,552,771	102	1.44
	20–40	994,761	21	0.76
	40–60	726,026	9	0.44
Stream power index	>60	1,529,272	28	0.66
	<200	1,159,782	45	1.40
	200–400	730,882	15	0.74
	400–600	536,393	14	0.94
Topographic wetness index	>600	3,375,773	86	0.92
	<8	407,887	0	0
	8–12	3,059,222	51	0.60
Distance from rivers (m)	>12	2,335,721	109	1.69
	<100	330,835	24	2.63
	100–200	290,402	14	1.74
	200–300	285,638	4	0.50
	300–400	278,972	9	1.17
Distance from faults (m)	>400	4,616,983	109	0.85
	<250	386,142	8	0.75
	250–500	377,923	5	0.47
	500–750	353,617	5	0.51
	75–1000	320,728	8	0.90
River density (km/km <sup>2</sup> )	>1000	4,364,420	134	1.11
	<0.31	2,203,090	48	0.79
	0.31–0.78	2,246,028	81	1.30
	0.78–1.27	998,728	21	0.76
	1.27–2.51	354,984	10	1.02

**Table 2** (continued)

Factor	Class	No. of pixel in domain	No. of qanats	Frequency ratio (FR)
Fault density (km/km <sup>2</sup> )	<2.72	3,757,240	118	1.13
	2.72–8.37	901,771	17	0.68
	8.37–15.70	676,037	21	1.12
	15.70–26.80	467,782	4	0.31
Land use	Agriculture	1,731,384	69	1.44
	Orchard	79,370	4	1.82
	Rangeland	3,844,777	83	0.78
	Residential	147,299	4	0.98
Lithology	A	2,659,654	63	0.85
	B	234,579	6	0.92
	C	31,616	0	0
	D	5095	0	0
	E	94,800	0	0
	F	757,669	57	2.72
	G	1,005,012	19	0.68
	H	102,749	3	1.05
	I	305,502	11	1.30
	J	151,436	1	0.23
	K	104,632	0	0
	L	346,945	0	0
	M	3141	0	0

Penalized discriminant analysis was fitted and the final model had shrinkage penalty coefficient ( $\lambda$ ) value of 0.1 with accuracy value of 0.61 and kappa value of .023. In the case of GPM obtained by using PDA, moderate potential class covered 26.11% of the study area and low potential, high potential, and very high potential classes contained 21.43, 33.01, and 19.45% of the area (Fig. 4d).

In the case of BRT, the five most influential variables were altitude (18.42%), TWI (13.91%), distance from faults (12.52%), slope aspect (9.99%), and slope degree (9.70%), respectively (Fig. 6b). GPM obtained by using BRT model is illustrated in Fig. 5b. We found that low potential, moderate potential, high potential, and very high potential classes cover 24.44, 25.44, 25.06, and 25.05% of the study area, respectively (Fig. 4e).

RF provides two indices to determine the importance of the input variables, i.e., mean decrease in accuracy and mean decrease in Gini (Immitzer et al. 2012). Altitude, TWI, slope, and distance from faults had the highest importance between effective factors, respectively (Fig. 6c). On the other hand, fault density, land use, and aspect had the lowest power. Low potential, moderate potential, high potential, and very high potential classes cover 31.27, 30.74, 23.10, and 14.89% of the study area, respectively (Fig. 5f).

In the case of ANN, the final multilayer perceptron (MLP) was applied with 14 input layers, 7 hidden layers,

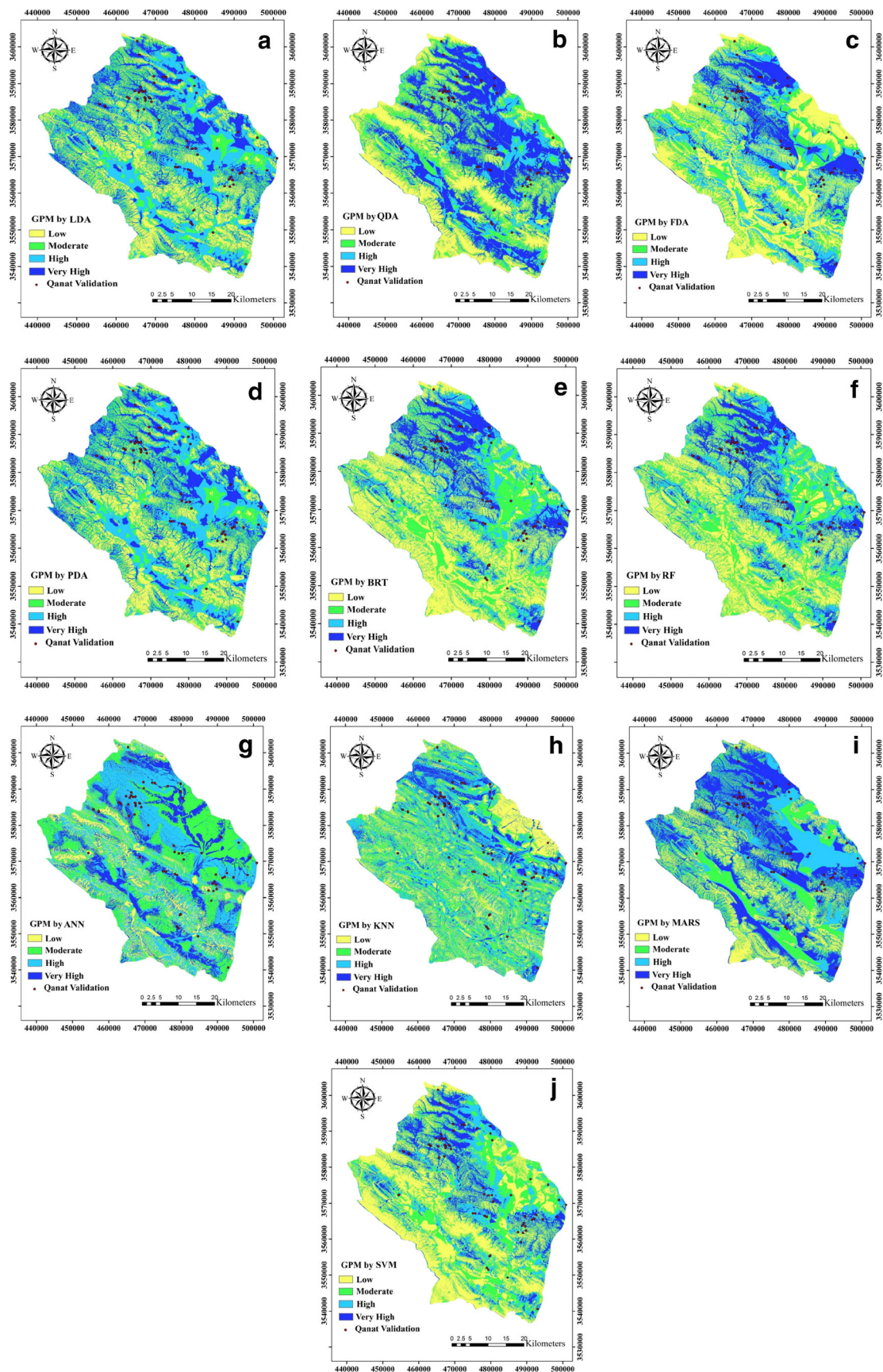
and 1 output layer as a 14-7-1 network. A groundwater qanat potential map was calculated using artificial neural network (Fig. 5g). The range of classes and percentage of each class are presented in Table 3. According to the results, low potential, moderate potential, high potential, and very high potential classes cover 13.00, 34.87, 32.57, and 19.54% of the study area, respectively. According to Fig. 6d, TWI, slope degree, and river density had the most important influence factors, while land use, slope aspect, and fault density had the lowest importance.

Groundwater qanat potential map was calculated using KNN by a  $k = 3$  in the study area (Fig. 5h). Based on this method, low potential, moderate potential, high potential, and very high potential classes cover 21.72, 34.57, 31.05, and 12.66% of the study area, respectively.

The main equation calculated by multivariate adaptive regression spline model is represented as

$$\begin{aligned}
 \text{MARS} = & 0.928 - 0.018 * \max(0, \text{Slope} - 8.324) - 0.00004 * \max \\
 & (0, \text{Fault Dis} - 6596.97) + 67.984 * \max(0, \text{Profile} - 0.0009) \\
 & * \text{Landuse2} - 0.061 * \max(0, 14.489 - \text{TWI}) * \text{Landuse2} \\
 & + 0.005 * \max(0, 2280 - \text{Altitude}) * \max(0, 0.353 - \text{River Dens}) \\
 & - 0.0002 * \max(0, 2280 - \text{Altitude}) * \max(0, 8.797 - \text{Fault Dens}) \\
 & - 0.014 * \max(0, \text{Profile} - 0.004) * \max(0, 6596.969 - \text{Fault Dis})
 \end{aligned} \quad (4)$$





**Fig. 5** Groundwater potential maps produced by **a** LDA, **b** QDA, **c** FDA, **d** PDA, **e** BRT, **f** RF, **g** ANN, **h** KNN, **i** MARS, **j** SVM

The groundwater potential map produced by MARS is shown in Fig. 5i. In this method, low potential, moderate potential, high potential, and very high potential classes cover 16.21, 23.88, 27.22, and 32.67% of the study area, respectively.

In the case of SVM, C-SVM (SVM type 1) with a Gaussian radial basic kernel function, hyperparameter sigma = 0.076, number of support vectors = 249, objective function value = -189.32, and training error = 0.2 was fitted. Finally, groundwater potential map produced by using SVM model is shown in Fig. 5j. The range and classes are represented in Table 3. According to the results, low potential, moderate potential, high potential, and very high potential classes cover 36.34, 29.49, 20.58, and 13.57% of the study area, respectively. According to Fig. 6e, altitude, slope degree, and plan curvature had the highest importance between effective factors, while lithology, land use, and slope aspect had the lowest importance.

### 4.3 Validation of qanat potential maps

The results of the sensitivity showed that QDA and ANN were the best models for predicting qanat locations (Table 4). On the other hand, MARS and KNN were the weakest models in this case. The results of the specificity depicted good performance of BRT and RF models and weak performance of QDA and ANN models (Table 4). Results of the AUC-ROCs indicate that RF, SVM, and BRT had the highest ROC values (0.846, 0.83, and 0.829, respectively) indicating better performance by these models in groundwater potential mapping, while ANN and KNN had the lowest values of ROC (0.632, and 0.703, respectively, indicating weak performances; Table 4). In the case of Kappa index, RF, BRT, FDA, and SVM showed better results, while ANN, KNN, and MARS showed weak performances (Table 4). In addition, QD was also calculated and represented in Table 5. According to the results, PDA, SVM, BRT, and RF had the highest values of 18.15, 2.76, 1.98, and 1.93, respectively. Overall, it can be concluded that RF, SVM, and BRT were the most successful method for groundwater modeling in this study.

**Table 3** Distribution of the qanat potential values and areas with respect to the groundwater qanat occurrence potential zones

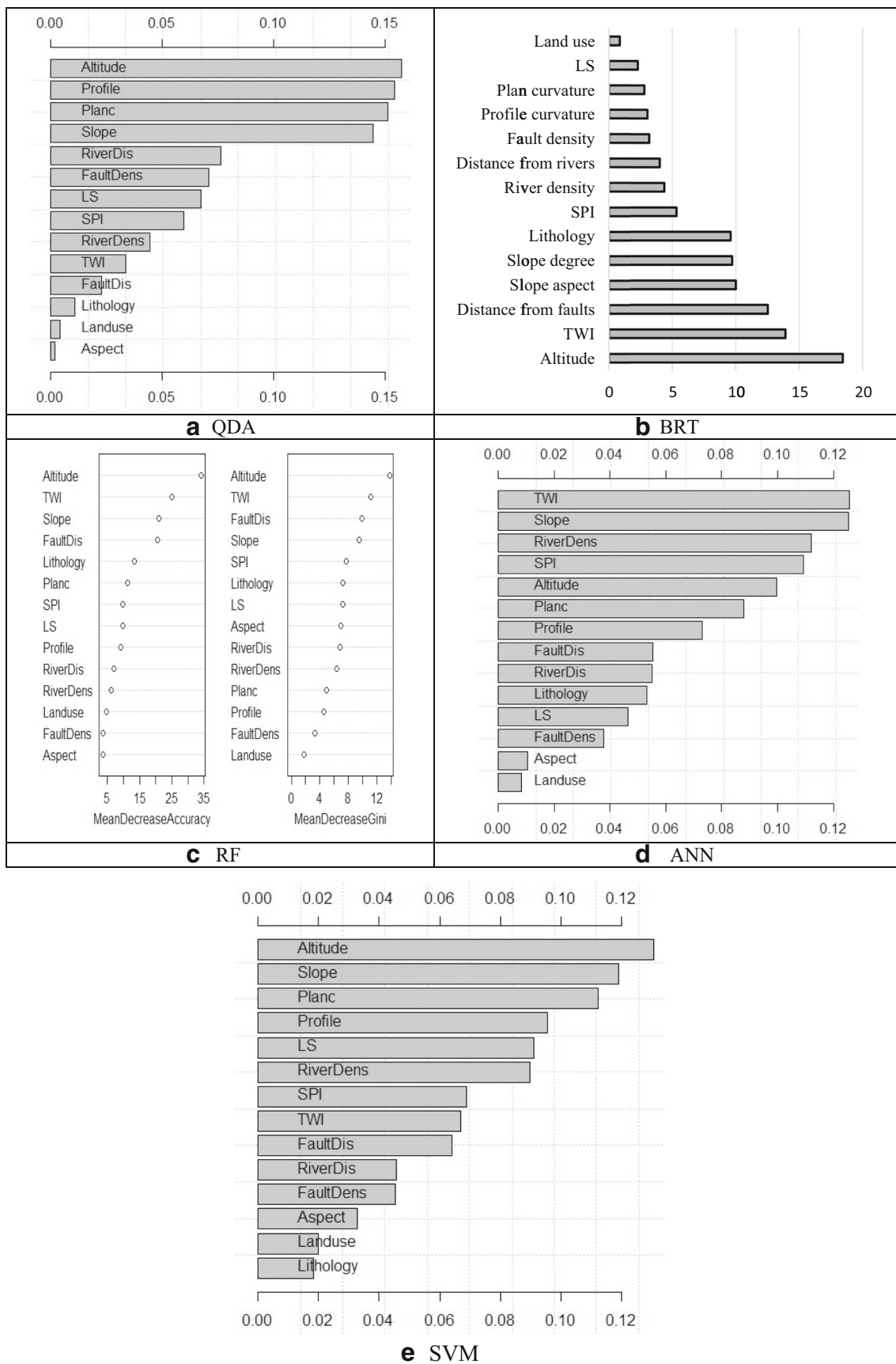
GQP classes (%)	LDA	QDA	FDA	PDA	RF	SVM	BRT	ANN	KNN	MARS
Low	21.29	24.69	32.25	21.43	31.27	36.34	24.44	13.00	21.72	16.21
Moderate	25.56	25.27	24.52	26.11	30.74	29.49	25.44	34.87	34.57	23.88
High	33.58	17.22	23.12	33.01	23.10	20.58	25.06	32.57	31.05	27.22
Very high	19.57	32.80	17.12	19.45	14.89	13.57	25.05	19.54	12.66	32.67

## 5 Discussion

In this section, the results were discussed in three parts including the performance of models and their comparison, the comparison between qanat and spring as indicators of groundwater potential, and the importance of variables in groundwater modeling.

### 5.1 The performance of models and their comparison

According to the results, RF had the best performance in groundwater modeling, followed by SVM and BRT models. The better performance of RF can be due to its ability to run on large databases and capability to handle thousands of input variables without variable deletion. Stumpf and Kerle (2011) mentioned that random forests take advantage of the high variance among individual trees which lets each tree vote for the class membership. Then, RF determines the respective class based on the most number of votes. In addition, RF is able to deal with interactions and non-linearities between effective factors (Catani et al. 2013). Also, RF showed suitable performance in different fields of study including wild-fire, ecology, groundwater spring potential mapping, and landslide susceptibility mapping (Peters et al. 2007; Oliveira et al. 2012; Vorpahl et al. 2012; Naghibi and Pourghasemi 2015). The SVM model has the advantage of handling complex, non-linear relationships, and is very robust to noise (Ballabio and Sterlacchini 2012; Tien Bui et al. 2016). It is shown that the SVM model performs well in different fields of study including flood susceptibility mapping (Tehrany et al. 2014; Tehrany et al. 2015) and landslide susceptibility mapping (Ballabio and Sterlacchini 2012; Pourghasemi et al. 2013). It was seen that BRT model had suitable performance and attained third rank in groundwater potential mapping among the implemented models. BRT model includes strong features of tree-based models; it can model different types of factors and cope with missing data. BRT is also able to handle interaction effects among effective factors (inputs) (Elith et al. 2008). In addition, BRT showed good performance in different fields of study such as ecology (Abeare 2009; Aertsen et al. 2011), groundwater spring potential mapping



**Fig. 6** The importance of influence factors in QDA (a), BRT (b), RF (c), ANN (d), and SVM (e) models



**Table 4** Prediction rate curve for the qanat potential maps produced by different models implemented in the current study

Models	AUC-ROC	Sensitivity (true positive rate)	Specificity (false positive rate)	Cohen's Kappa index
RF	0.846	75.0	79.4	54.4
SVM	0.830	70.6	76.5	47.1
BRT	0.829	75.0	79.4	54.0
FDA	0.775	76.5	72.1	48.5
MARS	0.760	64.7	63.2	27.9
QDA	0.757	92.6	36.8	29.4
LDA	0.735	69.1	60.3	29.4
PDA	0.722	69.1	60.3	29.4
KNN	0.703	66.2	61.8	27.9
ANN	0.632	82.4	44.1	26.5

(Naghibi and Pourghasemi 2015), and landslide susceptibility mapping (Youssef et al. 2015). Overall, tree-based models have some advantages such as feature selection and pruning. Feature selection leads to selection of the most important factors which can be used for splitting and making decision. On the other hand, TB models employ pruning features which makes them more general and makes the results more acceptable. Among discriminant analysis models, it was seen that FDA had the best performance, followed by QDA, LDA, and PDA models. In the case of LDA, it is very sensitive to outliers, and no dependent factor may be definitely correlated to a linear combination of other variables. ANN is prone to overfitting and KNN performs poorly on high-dimensionality datasets (Tien Bui et al. 2012).

## 5.2 The comparison between qanat and spring as indicators of groundwater potential

Using qanat as an indicator for groundwater potential was a novelty in the current study. Naghibi and Pourghasemi (2015) used spring data to map groundwater potential by five models of BRT, RF, CART, EBF, and GLM in the Beheshtabad watershed, Iran. According to their results, BRT and RF models had ROC values of 86.12 and 86.05, respectively. In the current study, BRT and RF models showed ROC values of 84.90 and 86.31, respectively. The results obtained by spring and qanat as groundwater potential indicators are very similar.

**Table 5** Qanat density (QD) on groundwater potential maps developed from different machine learning models

Class	LDA	QDA	PDA	FDA	BRT	RF	ANN	KNN	MARS	SVM
Low	0.66	0.48	0.62	0.58	0.60	0.56	0.68	0.71	0.32	0.49
Moderate	0.86	0.87	0.93	0.78	0.86	0.96	0.72	0.94	0.95	1.07
High	0.90	1.15	0.82	1.18	1.04	1.05	1.35	1.02	0.86	0.64
Very HIGH	1.73	1.41	18.15	1.93	1.98	1.93	1.13	1.63	1.49	2.76

So, it can be concluded that qanat is also a good indicator for groundwater potential.

## 5.3 The importance of variables in groundwater modeling

Altitude, slope degree, plan curvature, and profile curvature were found to be more significant factors compared to others. On the other hand, lithology, land use, and slope aspect were the least important influence factors. Therefore, it can be concluded that two primary and two secondary parameters had the highest importance and contribution in the modeling process. The four mentioned primary and secondary topographical parameters affect water flow concentration as well as its infiltration in any part of the watershed. As slope degree increases, flow speed increases and a decrease in infiltration can be observed. Lower altitudes contain more developed drainage system which leads to more water flow and subsequently higher available water. Considering the topographical condition of the study area which is mountainous, higher importance of the primary and secondary topographical factors could be justified. In addition, the quality of the input factors could influence the results of the models and subsequently the factors' importance. Thus, it can be concluded that the importance of the factors could be affected by general topographical condition of the area and its quality. In another study, Naghibi and Pourghasemi (2015) found altitude, distance from faults, SPI, and fault density to be more important for groundwater modeling. Hence, there are some differences in the sensitivity

of different factors in the use of qanat versus spring for groundwater modeling.

## 6 Conclusion

This study evaluated the performance of ten models in the groundwater potential mapping using qanat locations as indicator in the Beheshtabad Watershed, Iran. For this purpose, 14 influence factors were prepared and used in the modeling. The ROC, sensitivity, specificity, kappa, and qanat density indices were used to evaluate the performance of the models. The results showed that the RF, SVM, and BRT models were more suitable for groundwater potential mapping using qanat location. Overall, it was seen that tree-based models (BRT and RF) and SVM had better performance than discriminant models, ANN, MARS, and KNN. In addition, among discriminant analysis models, FDA had the best performance. Furthermore, the suitability of qanat locations as groundwater potential indicator was verified. Therefore, in many countries, these constructions can be used as groundwater indicator for mapping groundwater potential. In addition, it was concluded that altitude, slope degree, plan curvature, and profile curvature were more effective compared to other factors. The results of the current study can be used by land use planners and water resource managers in order to reduce the costs of groundwater resource discovery and exploitation.

**Acknowledgements** The authors would like to thank two anonymous reviewers and editorial positive comments.

## References

- Abeare SM (2009) Comparisons of boosted regression tree, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico Lonline Fishery. Master's Thesis, Louisiana State University
- Aertsen W, Kint V, Van Orshoven J, Özkan K, Muys B (2010) Comparison and ranking of different modeling techniques for prediction of site index in Mediterranean mountain forests. *Ecol Model* 221:1119–1130
- Aertsen W, Kint V, Van Orshoven J, Muys B (2011) Evaluation of modeling techniques for forest site productivity prediction in contrasting eco-regions using stochastic multi-criteria acceptability analysis (SMAA). *Environ Model Softw* 26(7):929–937
- Ahmadi H, Nazari Samani A, Malekian A (2010) The qanat: a living history in Iran. *Water and sustainability in arid regions*. Springer, Netherlands, pp 125–138
- Ballabio C, Sterlacchini S (2012) Support vector machines for land-slide susceptibility mapping: the Staffora River basin case study, Italy. *Math Geosci* 44:47–70
- Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Bonham-Carter GF (1994) *Geographic information systems for geoscientists: modelling with GIS*. Computer Methamphetamine Geos, vol. 13. Pergamon, New York
- 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
- Cherkassky V, Mulier F (2007) *Learning from data: concepts, theory, and methods*. Wiley, New York
- Chezgi J, Pourghasemi HR, Naghibi SA, Moradi HR, Kheirkhah Zarkesh M (2015) Assessment of a spatial multi-criteria evaluation to site selection underground dams in the Alborz Province. *Iran Geocarto Int* 31:628–646. doi:10.1080/10106049.2015.1073366
- Chirici G, Mura M, McInerney D, Py N, Tomppo EO, Waser LT, McRoberts RE (2015) A meta-analysis and review of the literature on the k-nearest neighbors technique for forestry applications that use remotely sensed data. *Draft* 176:282–294. doi:10.1016/j.rse.2016.02.001
- Cortez P, Embrechts MJ (2013) Using sensitivity analysis and visualization techniques to open black box data mining models. *Inf Sci* 225: 1–17. doi:10.1016/j.ins.2012.10.039
- Davoodi Moghaddam D, Rezaei M, Pourghasemi HR, Pourtaghie ZS, Pradhan B (2015) Groundwater spring potential mapping using bivariate statistical model and GIS in the Taleghan watershed, Iran. *Arab J Geosci* 8(2):913–929
- Eker AM, Dekmen M, Cambazoglu S, Duzgun SHB, Akgun H (2015) Evaluation and comparison of landslide susceptibility mapping methods: a case study for the Ulus district, Bartın, northern Turkey. *Int J Geogr Inf Sci* 29(1):132–158
- Elith J, Graham CH, Anderson RP et al (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29:129–151
- Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *J Anim Ecol* 77(4):802–813
- Everitt BS, Landau S, Leese M, Stahl D (2011) *Miscellaneous clustering methods, in cluster analysis*, 5th edn. John Wiley & Sons Ltd., Chichester
- Felicísimo A, Cuartero A, Remondo J, Quiros E (2012) Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. *Landslides* 10:175–189
- Friedman JH (1991) Multivariate adaptive regression splines. *Annual Statistics* 19:1–141
- Geology Survey of Iran (GSI) (1997) [http://www.gsi.ir/Main/Lang\\_en/index.html](http://www.gsi.ir/Main/Lang_en/index.html)
- Granitto PM, Biasioli F, Endrizzi I, Gasperi F (2008) Discriminant models based on sensory evaluations: single assessors versus panel average. *Food Qual Prefer* 19(6):589–595. doi:10.1016/j.foodqual.2008.03.006
- Gutiérrez ÁG, Schnabel S, Lavado Contador JF (2009) Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. *Ecol Model* 220(24):3630–3637
- Hastie T, Tibshirani R, Friedman J (2001) *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York
- Hong H, Pradhan B, Xu C, Tien Bui D (2015) Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. *Catena* 133:266–281
- Hong H, Naghibi SA, Pourghasemi HR, Pradhan B (2016) GIS-based landslide spatial modeling in Ganzhou City, China. *Arab J Geosci* 9(2):112. doi:10.1007/s12517-015-2094-y
- Immitzer M, Atzberger C, Koukal T (2012) Eignung von WorldView-2 Satellitenbildern für die Baumartenklassifizierung unter besonderer Berücksichtigung der vier neuen Spektralkanäle. *Photogramm Fernerkun*:573–588
- Karami A, Khoorani A, Noohegar A, Shamsi SR, Moosavi V (2015) Gully erosion mapping using object-based and pixel-based image classification methods. *Environ Eng Geosci* 21(2):101–110
- Karimi SE, Emami SN, Tahmasebipour N, Pourghasemi HR, Naghibi SA, Arami SA, Pradhan B (2016) Assessment and comparison of

- combined bivariate and AHP models with logistic regression for landslide susceptibility mapping in the Chaharmahal-e-Bakhtiari Province, Iran. *Arab J Geosci* 9(3):201. doi:10.1007/s12517-015-2258-9
- Kavzoglu T, Sahin E, Colkesen I (2014) Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* 11:425–439
- Leathwick JR, Elith J, Francis MP, Hastie T, Taylor P (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Mar Ecol Prog Ser* 321:267–281
- Lee S, Song KY, Kim Y, Park I (2012a) Regional groundwater productivity potential mapping using a geographic information system (GIS) based artificial neural network model. *Hydrogeol J* 20: 1511–1527
- Lee S, Park I, Choi J-K (2012b) Spatial prediction of ground subsidence susceptibility using an artificial neural network. *Environ Manag* 49(2):347–358
- Mahdavi M (2004) Applied hydrology, vol 2, 5th edn. University of Tehran Press, Tehran
- Mallet Y, Coomans D, deVel O (1996) Recent developments in discriminant analysis on high dimensional spectral data. *Chemometrics Intell. Lab Syst* 35(2):157–173
- Marjanović M, Kovačević M, Bajat B, Voženilek V (2011) Landslide susceptibility assessment using SVM machine learning algorithm. *Eng Geol* 123:225–234
- Mojiri HR, Zarei AR (2006) The investigation of precipitation condition in the Zagros area and its effects on the central plateau of Iran. The 2nd Conference of Water Resource Management. Tehran, Iran
- Moosavi V, Niazi Y (2015) Development of hybrid wavelet packet-statistical models (WP-SM) for landslide susceptibility mapping. *Landslides*. doi:10.1007/s10346-014-0547-0
- Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modeling: a review of hydrological, geomorphological and biological applications. *Hydrol Process* 5:3–30
- Naghbi SA, Pourghasemi HR (2015) 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*. doi:10.1007/s11269-015-1114-8
- Naghbi SA, Pourghasemi HR, Pourtaghi ZS, Rezaei A (2015) Groundwater qanat potential mapping using frequency ratio and Shannon's entropy models in the Moghan watershed, Iran. *Earth Sci Inf* 8(1):171–186
- Naghbi SA, Pourghasemi HR, Dixon B (2016) GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ Monit Assess* 188:44. doi:10.1007/s10661-015-5049-6
- Naghbi SA, Moradi Dashtpajardi M (2016) Evaluation of four supervised learning methods for groundwater spring potential mapping in Khalkhal region (Iran) using GIS-based features. *Hydrogeol J*. doi:10.1007/s10040-016-1466-z
- Nazari Samani A, Farzadmehr J (2006) Qanat as a traditional and advantageous approach for water supply in Iran. In: Proceedings of the International Symposium on Water and Management for Sustainable Irrigated Agriculture, Adana, Turkey
- Negnevitsky M (2002) Artificial intelligence: a guide to intelligent systems. Addison-Wesley/Pearson, Harlow, p 394
- Oh HJ, Kim YS, Choi JK, Park E, Lee S (2011) GIS mapping of regional probabilistic groundwater potential in the area of Pohang City, Korea. *J Hydrol* 399:158–172
- Oliveira S, Oehler F, San-Miguel-Ayanz J (2012) Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest. *For Ecol Manag* 275:117–129
- Ozdemir A (2011a) 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
- Ozdemir A (2011b) Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). *J Hydrol* 405:123–136
- Paraskevas T, Constantinos L, Dimitrios R, Ioanna L (2015) Landslide susceptibility assessments using the k-nearest neighbor algorithm and expert knowledge. Case study of the basin of Selinounda river, Achaia County, Greece. *SafeChania 2015, The knowledge triangle in the Civil Protection Service Center of Mediterranean Architecture, Chania, Crete, Greece, 10–14 June 2015*
- Peters J, De Baets B, Verhoest NEC, Samson R, Degroeve S, De Becker P, Huybrechts W (2007) Random forests as a tool for ecohydrological distribution modelling. *Ecol Model* 207:304–318
- Perrier ER, Salkini AB, editors (1991) Supplemental irrigation in the near East and North Africa. Proceedings of a Workshop on Regional Consultation on Supplemental Irrigation. ICARDA and FAO, 1987 Dec 7–9; Rabat, Morocco: Kluwer Academic Publishers; p. 611
- Pourghasemi HR, Jirandeh AG, Pradhan B, Xu C, Gokceoglu C (2013) Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran. *J Earth Syst Sci* 122:349–369
- Pourghasemi HR, Beheshtirad M (2014) Assessment of a data-driven evidential belief function model and GIS for groundwater potential mapping in the Koohrang Watershed, Iran. *Geocarto Int*. doi:10.1080/10106049.2014.966161
- Pourtaghi ZS, Pourghasemi HR (2014) GIS-based groundwater spring potential assessment and mapping in the Birjand township, southern Khorasan Province, Iran. *Hydrogeol J* 22(3):643–662
- R Development Core Team (2005) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, <http://www.R-project.org>.
- Rahmati O, Nazari Samani A, Mahdavi M, Pourghasemi HR, Zeinivand H (2014) Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS. *Arab J Geosci*. doi:10.1007/s12517-014-1668-4
- Ramos-Canon AM, Prada-Sarmiento LF, Trujillo-Vela MG, Macias JP, Santos-R AC (2015) Alfonso Mariano Ramos-Cañón I Luis Felipe Prada-Sarmiento I Mario Germán Trujillo-Vela I Juan Pablo Macias I Ana Carolina Santos-R Linear discriminant analysis to describe the relationship between rainfall and landslides in Bogotá, Colombia. *Landslides*, DOI: 10.1007/s10346-015-0593-2
- Razandi Y, Pourghasemi HR, Samani Neisani N, Rahmati O (2015) Application of analytical hierarchy process, frequency ratio, and certainty factor models for groundwater potential mapping using GIS. *Earth Sci Inf*. doi:10.1007/s12145-015-0220-8
- Reyn SC, Sabatier R, Molinari N (2006) Choice of B-splines with free parameters in the flexible discriminant analysis context. *Comput Stat Data Anal* 51(3):1765–1778. doi:10.1016/j.csda.2005.11.018
- Ridgeway G (2006) gbm: generalized boosted regression models. R package version 1.5–5, URL <http://CRAN.R-project.org/>.
- Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
- Samui P, Kurup P (2012) Multivariate adaptive regression spline (MARS) and least squares support vector machine (LSSVM) for OCR prediction. *Soft Comput* 16(8):1347–1351
- Stumpf A, Kerle N (2011) Object-oriented mapping of landslides using random forests. *Remote Sens Environ* 115(10):2564–2577
- Tayyebi A, Pijanowski BC (2014) Modeling multiple land use changes using ANN, CART and MARS: comparing tradeoffs in goodness of fit and explanatory power of data mining tools. *Int J Appl Earth Obs Geoinf* 28:102–116
- Tebaldi C, Nychka D, Brown BG, Shanman R (2002) Flexible discriminant techniques for forecasting clear-air turbulence. *Environmetrics* 13(8):859–878. doi:10.1002/env.562



- Tehrany MS, Pradhan B, Jebur MN (2014) Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *J Hydrol* 512:332–343
- Tehrany MS, Pradhan B, Mansor S, Ahmad N (2015) Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena* 125:91–101
- Tien Bui D, Pradhan B, Lofman O, Revhaug I (2012) Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and naive bayes models. *Math Probl Eng*. doi:10.1155/2012/974638
- Tien Bui D, Tuan TA, Klempe H, Pradhan B, Revhaug I (2015) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*. doi:10.1007/s10346-015-0557-6
- Tien Bui D, Pham BT, Nguyen QP, Hoang N-D (2016) Spatial prediction of rainfall-induced shallow landslides using hybrid integration approach of least-squares support vector machines and differential evolution optimization: a case study in Central Vietnam. *Int J Digital Earth* 8947:1–21. doi:10.1080/17538947.2016.1169561
- Vapnik V (1995) *The nature of statistical learning theory*. Springer, New York
- Vorpahl P, Elsenbeer H, Märker M, Schroder B (2012) How can statistical models help to determine driving factors of landslides? *Ecol Model* 239:27–39
- Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, Steinberg D (2008) Top 10 algorithms in data mining. *Knowl Inf Syst* 14. 10.1007/s10115-007-0114-2
- Youssef AM, Pourghasemi HR, Pourtaghi ZS, 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
- Zhu Y, Tan TL (2016) Penalized discriminant analysis for the detection of wild-grown and cultivated *Ganoderma lucidum* using Fourier transform infrared spectroscopy. *Spectrochim Acta A Mol Biomol Spectrosc* 159:68–77
- Zabihi M, Pourghasemi HR, Pourtaghi ZS, Behzadfar M (2016) GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. *Environ Earth Sci* 75(8):665. doi:10.1007/s12665-016-5424-9