



Locating groundwater artificial recharge sites using random forest: a case study of Shabestar region, Iran

Hossein Norouzi¹ · Shahram Shahmohammadi-Kalalagh²

Received: 16 April 2018 / Accepted: 21 June 2019 / Published online: 28 June 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

One of the most crucial design steps to guarantee the long life and the sustainability of artificial recharge of groundwater projects is to find the best locations. The present study focuses on identifying potential zones of groundwater artificial recharge in Shabestar region, northwest of Iran. For this purpose, random forest (RF) model, a learning method based on ensemble decision trees, was proposed for locating groundwater artificial recharge. Important factors, including slope and slope aspect, soil texture, erosion, land use, groundwater quality, permeability and geological lithology were integrated in a geographic information science (GIS). According to RF model, permeability and unsaturated zone thickness were identified as the most effective parameters for locating groundwater artificial recharge sites. Based on the proposed model, it was found that 14% of the region is located in suitable site for groundwater artificial recharge projects. The accuracy of the model was evaluated with receiver-operating characteristic (ROC) curve and the mean squared error (MSE). Low MSE and ROC curve of the model with the highest area under curve equal to 0.947, indicated high accuracy of random forest in locating groundwater artificial recharge.

Keywords Artificial recharge · GIS · Random forest · ROC curve · Shabestar region

Introduction

In recent years, excessive use of water resources in Iran, as an arid and semi-arid region, has led to perilous results. The annual rainfall in Iran, whose 90% are located in dry and semi-arid regions, is about 250 mm. While more than 50% of rainfall occurs in winter, only less than 8% occurs in summer. In other words, the majority of rainfall occurs at the time with minimum water requirement, while precipitation decreases in summer with the maximum water demands. Recent studies on water resources indicate that about 20% of the total annual precipitation in Iran is lost in the form of sudden floods, which enters Playa or Seas, and thus, is out of reach (Zhang et al. 2004). Accordingly, surface water

conservation methods have been developed for solving this problem in dry seasons.

Artificial recharge is regarded as a significant approach for storing surface water. It can be applied as a practical solution for rehabilitating aquifers and reducing groundwater drawdown. Indeed, several methods have been introduced for artificially recharging groundwater. As a case in point, basin artificial recharge method (Oaksford 1985) is one of the most effective methods for recharging groundwater. It has the following advantages over other methods (Balachandrar et al. 2010): (1) providing flood control and using water for other purposes such as agriculture, (2) higher efficiency in comparison with other artificial recharge methods, (3) more appropriate and effective in dry and semi-arid regions for controlling floods, and (4) in case one pool is dry in multi-basin systems due to dry maintenance, recharging operations can continue in other pools. Given the above-mentioned problems in the targeted region (Iran), locating groundwater sites for artificial recharge is considered as a critical issue. Several studies have been carried out for determining regions which are highly suitable for artificial recharge (Tiwari et al. 2017; Minnig et al. 2015; Zaidi et al. 2015; Mahdavi et al. 2013; Mehrabi et al. 2013; Shankar

✉ Shahram Shahmohammadi-Kalalagh
shahmohammadi_sh@yahoo.com

Hossein Norouzi
hosseinnorouzi168@yahoo.com

¹ Department of Earth Sciences, Faculty of Natural Sciences, University of Tabriz, Tabriz, Iran

² Department of Water Sciences and Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

and Mohan 2005; Bouwer 2002). There are several constantly changing parameters that affect groundwater locating sites. These parameters should be considered and integrated together. For this purpose, geographic information science (GIS) is a very useful tool (Dillon 2005; Ghayoumian et al. 2007). In the majority of the previous studies, the selected weights for locating groundwater artificial recharge sites have been only based on expert opinions; however, in the present study, in addition to experts' hydrogeological opinions, random forest method has also been used for determining the proper weights. In this way, appropriate weights with high accuracy and validity can be achieved.

Random forest (RF), developed by Breiman (2001), works more effectively than other classifiers such as support vector machines, neural networks and discriminant analysis and overcomes over-fitting problem. RF is an ensemble method which combines multiple decision tree algorithms. Decision trees are considered as types of model which are used for both classification and regression (Duda et al. 2011). Besides high prediction accuracy, RF is efficient, interpretable and non-parametric for various types of datasets. It is increasingly applied for land-cover classification from remotely sensed data (Pal 2005; Rodriguez-Galiano et al. 2012b; Sennie et al. 2008) and other fields related to the environment and water resources (Hosseini et al. 2016; Booker and Snelder 2012; Herrera et al. 2010; Huang et al. 2011).

Shabestar region is located within Urmia Lake basin, which has been recently impacted by a sharp decline in water levels due to the excessive use of groundwater and the reduction of rainfall. Artificial recharge projects around Urmia Lake can prevent reverse groundwater flow towards plains and it can also do away with salinity intrusion. Identifying potential groundwater recharge spots is considered as a prerequisite for any artificial recharge project. Due to excessive use of water resources in Shabestar region, artificial recharge has been recommended which also enriches water resources. The aim of this study is to use the random forest (RF) algorithm, a learning method based on ensemble decision trees, for locating groundwater artificial recharge sites. So far, this method has not been used in this field.

Materials and methods

Study area

Shabestar region is located in East Azerbaijan province (38°10'49"N, 45°42'10"E), southeast of Urmia Lake, and in northwestern Iran (Fig. 1). Based on Iran Meteorological Organization Data (a 30-year-old interval from 1987 to 2017), the targeted area with a cold and semi-arid climate has an average annual precipitation of about 236 mm,

average temperature of about 10 °C and average evaporation of about 1058 mm. The most important river in Shabestar plain is Daryan Chay with an average annual yield of about 0.39 m³/s. In recent years, the groundwater level and quality have been notably affected by significant precipitation reduction and notable increase in groundwater extraction rate. According to unit hydrograph of Shabestar plain, groundwater level has decreased for about 3.06 m since 2001–2002 to 2016–2017. In addition, aquifer storage has decreased about 12.01 MCM in 2015–2016 (Mousavi 2009).

Shabestar region consists of multiple geological formations. Lalun formation in the east and Ruteh formation in the west of the region were reported in the Permian era. While the east and northeastern part of the targeted region is covered by Shemshak, Lar formation of the Triassic and Jurassic era can be seen in the north and northwestern parts of this region. A large part of the area in the south has an alluvial deposit of the quaternary. In the west and southeast of the plain, there are limestone and marl-gypsum deposits which are related to the cretaceous period. In the north and northwest of Shabestar region, there is light gray limestone of the Triassic and Jurassic periods (Fig. 2). The aquifer of Shabestar plain is unconfined and is mainly formed by old and recent alluvial terraces, alluvial fans and fluvial sediments. As a result of drilling wells, logs and geophysical data, the western part of the plain is composed of fine-grained materials which are barely permeable. Transmissivity accounts for the ease of water movement through the entire thickness of the aquifer which is impacted by salinity development in the upper part of the aquifer. The average aquifer storage coefficient is approximately 2.5%. In general, absolute hydraulic gradient in the entries and outputs of the plains are 0.007 and 0.001, respectively (Mousavi 2009).

Important factors in locating suitable sites for groundwater artificial recharge

To find the suitable areas of the catchment for groundwater artificial recharge, some attributes, including slope and slope aspect, soil texture, erosion, land use, groundwater quality (electrical conductivity), permeability, depth to water or unsaturated zone thickness and distance from river, were considered as the effective parameters. These parameters have been used in several previous studies (Tiwari et al. 2017; Minnig et al. 2015; Zaidi et al. 2015; Mahdavi et al. 2013; Mehrabi et al. 2013; Shankar and Mohan 2005; Bouwer 2002).

Slope layer

The slope of the aquifer layers is considered to be one of the effective factors in locating appropriate places for artificially recharging groundwater. According to previous findings and

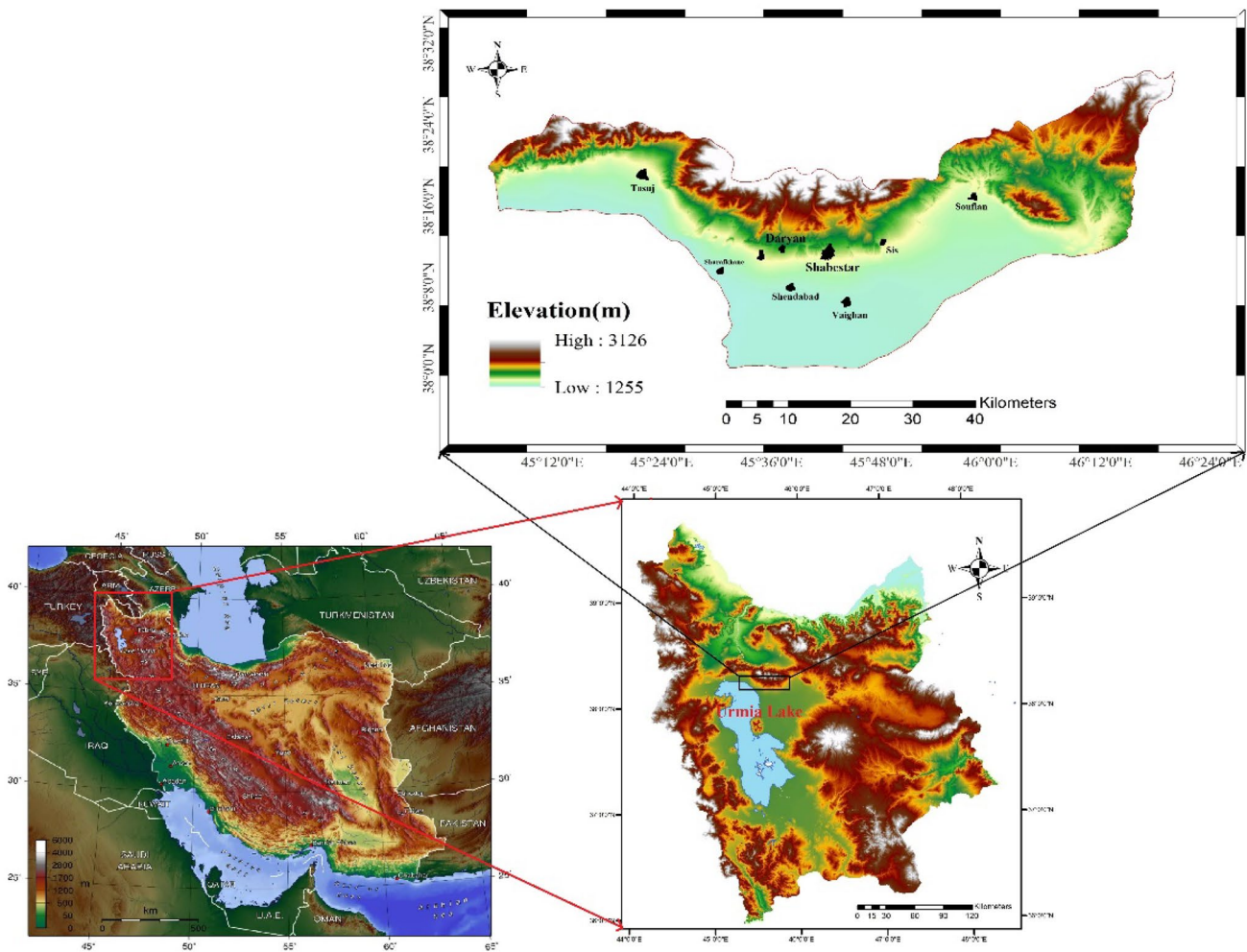


Fig. 1 Study area (Shabestar region; NW of Iran)

observations, the most appropriate slope for basin artificial recharge is about 2–8% (Balachandar et al. 2010).

Unsaturated zone thickness

Depth to water table or the unsaturated zone thickness of aquifers is directly related to the amount of groundwater storage. That is to say, as unsaturated zone thickness of aquifers increases, the amount of groundwater storage increases and vice versa. Indeed, due to the small thickness of unsaturated zone in an aquifer, water entering the aquifer will reach the water table. Hence, the aquifer will be slowly saturated, preventing further water infiltration.

Erosion rate

Erosion controls the amount of sediments entrance to recharge basin. More erosion leads to increased entrance

sediments. Consequently, useful life and permeability of the basin are reduced.

Permeability

Permeability controls the amount of water entering the ground. In fact, it represents the water movement in the porous media. Higher permeability results in the entrance of more water in the aquifer. Another impact of permeability is evident on reducing evapotranspiration. In an aquifer with lower permeability, water remains inside the porous media and soil salinity happens after evaporation.

Land use

Given the fact that surface layers are removed in the artificial recharge design, the locations with notably little use should be considered to prevent significant damage to the environment. Furthermore, the targeted lands for artificial recharge

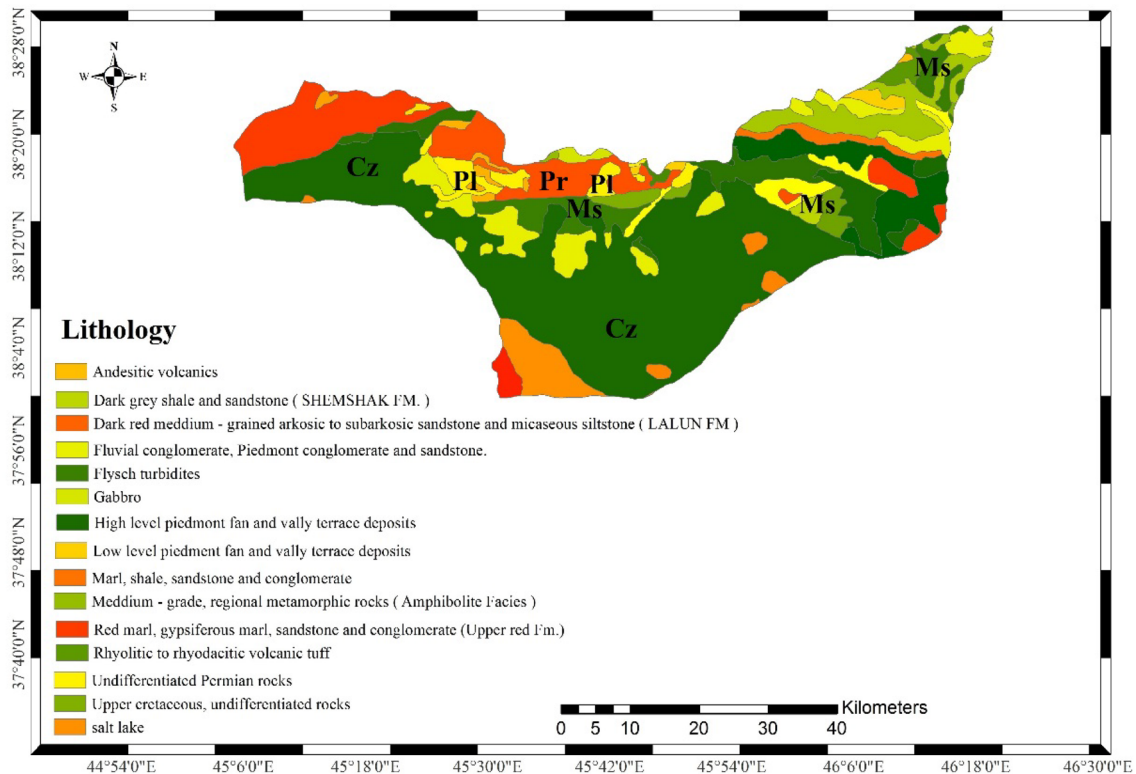


Fig. 2 Lithology map of the study area

should be poor in terms of vegetation and should not be located near agricultural areas.

Water quality

The quality of surface water is an important factor in locating groundwater artificial recharge sites. That is to say, in case the region has low quality (high electrical conductivity or total dissolved solid), artificial recharge projects cannot be practically carried out.

Distance from river

Since rivers can recharge aquifers in high-water seasons, they are regarded as one of the significant parameters in aquifer recharge. During rainfall seasons, high surface runoff occurs in rivers which can drain some of their water to the aquifer. Hence, the closeness of the river to aquifer is regarded as one of the important factors.

Weighted index overlay method

Weighted overlay is a technique for applying a common measurement scale of values to diverse and dissimilar inputs to create an integrated analysis. The weighted overlay tool applies one of the most used approaches for overlay analysis

to solve multicriteria problems such as site selection and suitability models. In a weighted overlay analysis, each of the general overlay analysis steps is followed. Geographic problems often require the analysis of many different factors. For instance, choosing the site for a new housing development means assessing such things as land cost, proximity to existing services, slope, and flood frequency. Weighted overlay only accepts integer rasters as input, such as a raster of land use or soil types. Continuous (floating point) rasters must be reclassified as integer before they can be used. Generally, the values of continuous rasters are grouped into ranges, such as for slope, or Euclidean distance outputs. Each range must be assigned a single value before it can be used in the weighted overlay tool. The reclassify tool allows for such rasters to be reclassified. Either leave the value assigned to each range (but note the range of values to which the new value corresponds) and assign weights to the cell values in the weighted overlay dialog box later, or can assign weights at the time of reclassifying. With the correct evaluation scale chosen, simply add the raster to the weighted overlay dialog box. The cells in the raster will already be set according to suitability or preference, risk, or some similarly unifying scale. The output raster's can be weighted by importance and added to produce an output raster.

A weighted overlay method was developed using GIS techniques for finding appropriate locations for applying

groundwater recharge. Weighted overlay depends on a number of thematic layers and the principle of multicriteria evaluation. Regarding GIS, overlay function is used for combining and integrating the layers. All data were transformed into raster's format; then, the overlay function was used for overlaying raster layers. This function is regarded as a combination of data in mathematical and weighted mode (Riad et al. 2011). Since each layer had specific weight, weight overlay index method was used for combining and integrating the layers. In weighted index overlay method, in addition to weighing the parameters in each information layer, each layer in locating groundwater artificial recharge is weighted by its value. This model has two sub-models:

1. Binary evidence maps: it is the simplest type of overlay index model; based on this model, the maps derived from the Boolean model are weighted according to their significance in locating groundwater artificial recharge.
2. Multi-class maps: in this model, in addition to weighing information layers, the units in each information layer have certain weights based on their potential. In this study, RF model was used for determining variable significance in locating groundwater artificial recharge sites. The weight range in this model depends on researcher opinion. As mentioned above, RF is regarded as a simple and convenient method for analyzing combined data which results in a better representation of geological situation in a given area. It also combines various layers based on weighing index. In this model, the value of each pixel in the output map is determined according to the following equation:

$$S = \frac{\sum_i^n S_{ij} W_i}{\sum_i^n W_i}, \quad (1)$$

where S denotes the value of each pixel in the final map; S_{ij} refers to the j unit weight of the i th map and W_i stands for the i th map weight.

Random forest

Random forest is an ensemble method which combines multiple decision tree algorithms to produce repeated predictions of the same phenomenon. Decision trees can be divided into classification trees and regression trees. A regression tree (RT) represents a set of restrictions or conditions which are hierarchically organized, and which are successively applied from a root to a terminal node or leaf of the tree (Breiman 1999; Quinlan 1990). To induce 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 repeated until a stop condition

previously specified is reached. Each of the terminal nodes, or leaves, has attached to it a simple regression model which applies in that node only. Once the tree's induction 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 taken as pruning criteria. When the RF makes a tree grow, it uses the best variable/split point within a subset of variables which has been selected randomly from the overall set of input variables. Therefore, this can decrease the strength of every single RT, but it reduces the correlation between the trees, which reduces the generalization error (Breiman 2001). Another remarkable feature of RF is that it grows its trees without pruning; in this way, too much training does not impact on the model accuracy which makes it lighter in terms of computing. Furthermore, those variables or data which are not selected in tree training are part of the sub-categories which are called out-of-bag (OOB). Such data in the RF method can be used for evaluating model performance (Peters et al. 2007). This way RF can compute an unbiased estimation of the generalization error without using an external test data subset (Breiman 2001). The generalization error converges as the number of trees increases; therefore, the RF does not overfit the data. RF also provides an assessment of the relative importance of the different variables. This aspect is useful for multisource studies, where data dimensionality is very high, and it is important to know how each predictive variable influences the prediction model to be able to select the best variables (Rodriguez-Galiano et al. 2012a, b). To assess the importance of each variable, the RF switches one of the input variables while keeping the rest constant, and it measures the decrease in accuracy which has taken place by means of the OOB error estimation (Breiman 2001). In attempting to reach a decision, a decision-maker can quickly become overwhelmed by the huge amounts of diverse types of data and information that are available on resources, conditions, and pressures within a region. A large number of variables related to the properties and behavior of the groundwater system and driving forces can exceed the ability of predictive methods to deal with it. Feature selection (FS) is an approach for selecting a subset of relevant features for building robust learning models (Guyon and Elisseeff 2003). Groundwater studies use a high number of variables related to the physical–chemical properties of the aquifer which can be highly correlated or redundant. Thus, in assessing the specific vulnerability of groundwater is important to select from a large number of possible explanatory variables (features) which are the most relevant, according to the target pollutant. FS helps increase the accuracy of predictive models by alleviating the effect of the curse of dimensionality, enhancing generalization capability, speeding up the learning process and increasing model interpretability. Random forest (RF) method is regarded as an extension of the Bagging method. RF is distinguished from

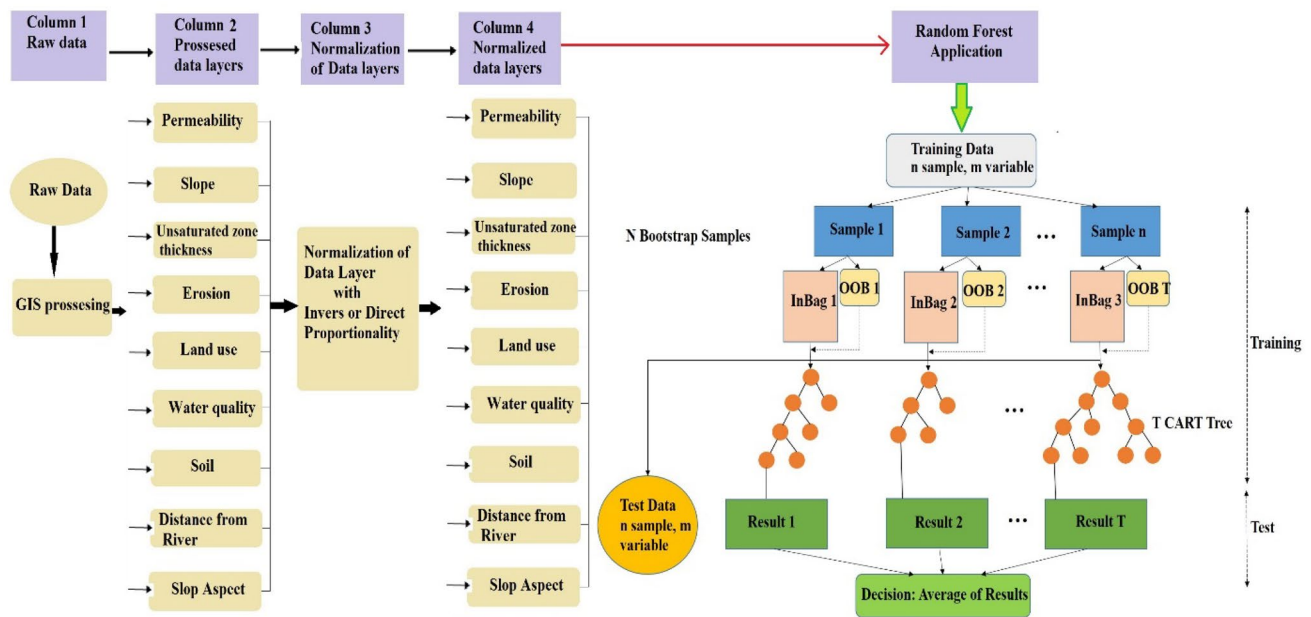


Fig. 3 A general methodological flowchart of random forest

bagging method with respect to random feature selection. Bagging is an algorithm which can change weak learners into strong learners. Since RF method can handle continuous, categorical, missing values and binary data, it is appropriate for modeling high-dimensional data. The bootstrapping and ensemble make RF strong enough to overcome over-fitting problems. This model does not need to prune trees. As mentioned before, in addition to high prediction accuracy, RF is efficient, interpretable and non-parametric for various types of datasets (Breiman 2001). Prediction error is measured based on out-of-bag samples according to Eq. 2. Furthermore, OOB error, also known as out-of-bag estimate, is a method for measuring prediction error of random forests, boosted decision trees and other machine-learning models. It uses bootstrap aggregating to sub-sample data samples for training. OOB refers to mean prediction error on each training sample x_i using only those trees which do not have x_i in their bootstrap sample (Gareth et al. 2013).

$$\text{MSE}^{\text{OOB}} = n^{-1} \sum_{i=1}^n [\hat{y}(x_i) - y_i]^2, \quad (2)$$

where $\hat{y}(x_i)$ denotes average OOB predictions for the i th observation. In ensemble methods, for reaching a decision through diverse data types, decision-maker can be quickly influenced (Rodríguez-Galiano et al. 2012b). A large number of variables related to system properties and behavior can exceed the ability of predictive methods for dealing with it. Although more information might be useful in modeling, the increasing number of input features may introduce additional

complexity in terms of computational time and dimension (Bellman 2003). High dimensionality in the dataset related to the inclusion of additional features can overwhelm the expected increase in prediction accuracy. Feature selection (FS) is an approach for selecting a subset of relevant features for building robust learning models (Blum and Langley 1997; Guyon and Elisseeff 2003; Saeys et al. 2007). FS enhances the accuracy of prediction models by accelerating training process, increasing generalizability, decreasing the impact of dimensions and increasing the capability of interpretation. In this study, RF algorithms, a learning method based on ensemble of decision trees, was proposed for locating groundwater artificial recharge based on a hydrogeological framework. RF technique is assumed as an accurate prediction modeling which outperforms other methods in terms of high prediction accuracy, ability to learn nonlinear relationships, ability to determine important variables in the prediction and less sensitivity against trapping (Quinlan 2014). The general RF algorithm and the flowchart of RF application, with respect to locating artificial recharge, are shown in Fig. 3.

Results and discussion

Thematic layers used in the weighted overlay index method

All the thematic layers, shown in Fig. 4 were integrated in ArcGIS 10.1 platform to prepare a map depicting suitable

areas for artificial groundwater recharge. According to Fig. 4, the slope of the region decreases toward the South; it approaches zero and becomes almost flat in the South and near Urmia Lake. The majority of alluvial deposits in this region are sandy-loam sediments and clay sediments are located between alluvium and Urmia Lake. Moreover, the permeability level becomes low towards the plain output and is very low in the southern part of the region. The groundwater quality decreases towards the outputs of the plain due to high resistance time and high evaporation; also, the electrical conductivity increases.

Locating map of groundwater artificial recharge

The information or maps used in this study, including relative weight and final weight which were assigned to each layer, are listed in Table 1. As stated in Table 1, permeability of the region and the slope aspect are considered as the most and least significant factors in groundwater artificial recharge, respectively. The first step in creating a prediction model is tuning parameters of the model. To achieve convergence in error values and obtain a reliable estimate, the model was made up of 1–100 trees. As decision trees exceed 100, error rate is not reduced. Hence, 100 trees are adequate. The results of the model were evaluated by estimating OOB error. Moreover, for reducing the dimension and enhancing the accuracy and interpretability of the model, feature selection method was used and important variables were also identified in the prediction. Although more information might be useful in modeling, the increasing number of input features may introduce additional complexity in terms of computational time and dimension (Bellman 2003). For this reason, the most important and effective parameters were selected for locating groundwater artificial recharge sites.

In this research study, weighted overlay method was used. Nevertheless, there are different methods which can be used to combine the layers. Weighted overlay, having a high flexibility in data combinations, is considered to be a simple and appropriate method for combined analysis of spatial data (Riad et al. 2011). In this method, for determining the relative significance of each parameter in comparison with other parameters, their locating effectiveness is taken into consideration. After the appropriate weights were selected based on expert opinions and random forest model, the weighted index overlay was used for overlaying the layers in GIS. For selecting weights using random forest, the weights for each parameter were first assumed based on expert opinion and were introduced into the model. Then, using Salford Predictive Modeler (SPM), a random forest model was run based on the presented weights and RF algorithm. Hence, the effective degree of each parameter in locating groundwater artificial recharge sites was determined and optimum weights were selected. The most important

parameter is the largest weight. The weighted index overlay method uses fewer data and weights are only based on expert opinion, so this method does not consider correlations and relationships between different parameters. However, the RF method can be implemented with a large number of data along with applying the expert opinion and determine the relative importance of the variables. In this study, due to the uncertainty of the data and the number of them, at the first based on the weighted index overlay method overlaying map was obtained, and then the output value of this map was used as the target value in the RF method. In fact, the RF method was used to weight optimization. In this way, parameters including slope, slope aspect, soil texture, erosion, land use, groundwater quality (electrical conductivity), permeability, depth to water or unsaturated zone thickness and distance from the river, were used as predictors for this target value. Then, the relative importance of these parameters and their weight were calculated based on the RF method. Finally, the output map was obtained based on overlaying the parameters with the new weight. Permeability, slope, depth to water level and erosion were identified as the most important parameters. Overlaying of layers was done based on predicted weights by random forest model. Random forest model should be used in such studies for achieving the desired results based on the hydrogeological framework; otherwise, the model would not be effective. Locating map of groundwater artificial recharge obtained from the composition of the layers is shown in Fig. 5. According to the proposed model (Fig. 5), 14% of the region was located in suitable sites for artificially recharging groundwater.

Model error evaluation

It can be argued that the stability of a given model depends on the accuracy of prediction, the number of necessary variables, and the data availability accuracy. For evaluating model error, ROC¹ curve and mean squared error (MSE) value should be taken into consideration. ROC curve is plotted by changing the threshold value against the expected output. Each TPR–FPR result and a series of such pairs are used for drawing ROC curve. TPR and FPR are known as sensitivity and specificity, respectively. The area under the ROC curve is also used as an error estimation method. According to the results, as AUC² value is closer to one, it can be argued that the proposed model is more accurate. As shown in Fig. 6, by increasing decision trees, MSE value of the model is reduced to the lowest amount; consequently, the accuracy of the model is enhanced. ROC curve of the model with the highest AUC equal to 0.947 indicates high

¹ Receiver-operating characteristic.

² Area under curve.

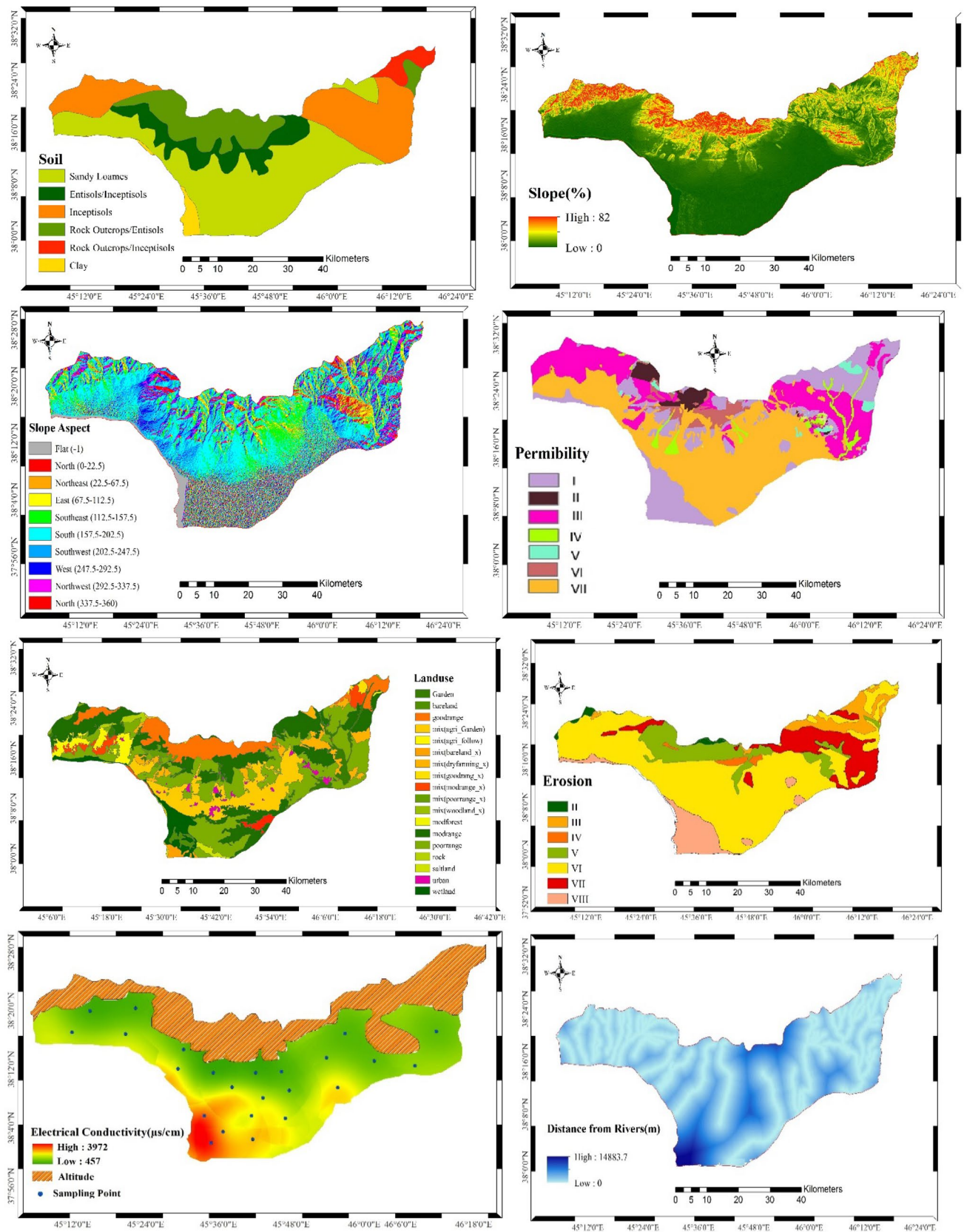


Fig. 4 Important thematic layers used in the weighted overlay index method

Table 1 Assigned weights and the properties used in layers

Layer	Properties	Relative weight	Variable importance using RF	Final weight
Permeability	I Igneous rocks, diabase, dense basalt	1	100	0.23
	II Metamorphic rocks	2		
	III Dense sandstones	3		
	IV Sandstone and limestone	4		
	V Old sediments	5		
	VI Conglomerate and river sediments	6		
	VII Young alluvial terraces	9		
Slope	2 >	1	82	0.2
	2–7	9		
	7–15	6		
	15 <	1		
Depth to water or unsaturated zone thickness	4–20	1	80	0.18
	20–35	9		
	35–50	7		
	50–70	4		
	70–135	1		
Erosion	Very low	9	52	0.17
	Low	6		
	Moderate	3		
	High	1		
Land use	Areas with poor vegetation	8	46	0.12
	Rangelands have moderate vegetation and lands within the river	5		
	Good vegetation ranges	2		
	Forest, agricultural and industrial areas, urban and lake areas	1		
Water quality	300–1000	9	32	0.1
	1000–2000	7		
	2000–3000	4		
	3000 <	1		
Soil	Sandy loam	5	23	0.06
	Entisols	4		
	Inceptisols	3		
	Clay	2		
	Rock outcrop	1		
Distance from river	0–500	5	12	0.04
	500–1000	4		
	1000–3000	2		
	3000 <	1		
Slop aspect	Flat	1	3	0.01
	North	1		
	South	3		
	East	2		
	West	2		

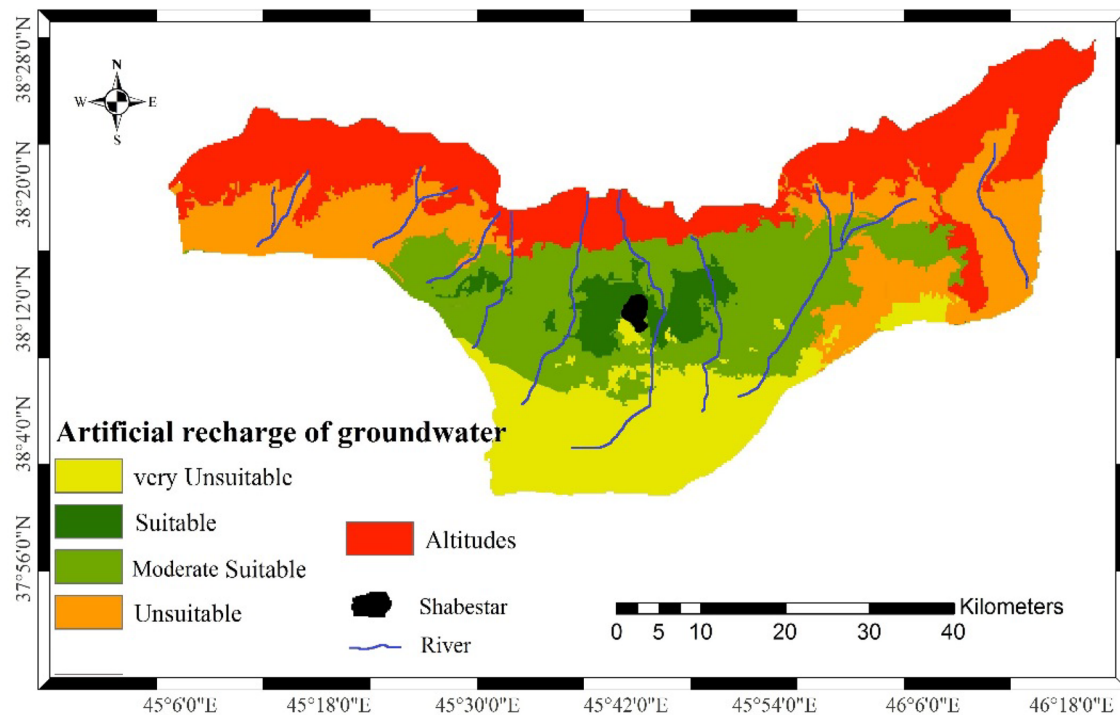


Fig. 5 Locating map of groundwater artificial recharge

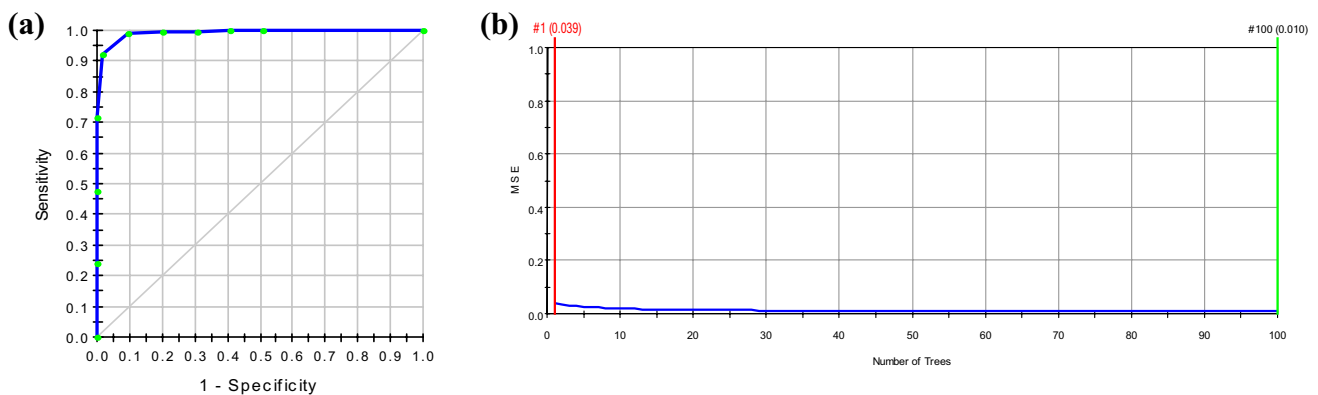


Fig. 6 Model error evaluation; **a** ROC curve and **b** MSE value

accuracy of random forest. In the study area, two groundwater artificial recharge project have been done and these regions are located in the suitable site on the locating map of groundwater artificial recharge (Fig. 5). Therefore, this can confirm the validity of this research.

Conclusion

Identifying potential groundwater recharge areas is considered as a pre-requisite for artificial recharge projects. For implementing artificial recharge projects, besides

comprehensive geological, hydrological, hydrogeological and hydro-chemical studies of a given region, some basic criteria should be identified for finding suitable areas in artificial recharge projects. In this research study, we identified appropriate sites for groundwater artificial recharge using the parameters of permeability, slope, slope aspect, erosion, unsaturated zone thickness, land use, soil, distance from river and water quality layers. Moreover, weighted index overlay method and combining different layers in GIS software were applied. Along with this purpose, random forest (RF) algorithm, a learning method based on ensemble of decision trees, was proposed for

locating groundwater artificial recharge based on a hydrogeological framework. The results of this study revealed that RF can be considered as an accurate prediction model which outperforms many other methods thanks to its high prediction accuracy, ability to learn nonlinear relationships, ability to determine significant variables in the prediction and low sensitivity against trapping. ROC curve and MSE value were used for evaluating model error. It was found that with increasing decision trees, the MSE value of the model decreases to its lowest amount. As a result, the accuracy of the model increases. Moreover, given the highest AUC equal to 0.947, ROC curve of the model confirmed high accuracy of random forest.

References

- Balachandar D, Alaguraja P, Sundaraj P, Rutharvelmurthy K, Kumaraswamy K (2010) Application of remote sensing and GIS for artificial recharge zone in Sivaganga district, Tamilnadu, India. *Int J Geomat Geosci* 1(1):84
- Bellman R (2003) *Dynamic programming*. Dover Publications, Mineola
- Blum AL, Langley P (1997) Selection of relevant features and examples in machine learning. *J Artif Intell* 97(1–2):245–271
- Booker DJ, Snelder TH (2012) Comparing methods for estimating flow duration curves at ungauged sites. *J Hydrol* 434–435:78–94
- Bouwer H (2002) Artificial recharge of groundwater: hydrogeology and engineering. *J Hydrol* 10:121–142
- Breiman L (1999) Random forests. *UC Berkeley TR567*
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Dillon P (2005) Future management of aquifer recharge. *Hydrogeol J* 13:313–316
- Duda RO, Hart PE, Stork DG (2011) *Pattern classification*. Wiley, New York
- Gareth J, Daniela W, Trevor H, Robert T (2013) *An introduction to statistical learning: with applications in R*. Springer, New York, pp 316–321
- Ghayoumian J, Mohseni-Saravi M, Feiznia S, Nouri B, Malekian A (2007) Application of GIS techniques to determine areas most suitable for artificial groundwater recharge in a coastal aquifer in southern Iran. *J Asian Earth Sci* 30:346–374
- Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182
- Herrera M, Torgo L, Izquierdo J, Pérez-García R (2010) Predictive models for forecasting hourly urban water demand. *J Hydrol* 387(1–2):141–150
- Hosseini N, Asghari MA, Nadiri A (2016) Determining vulnerable areas of Malikan plain aquifer for nitrate using random forest method. *J Environ Stud* 41:923–942
- Huang J, Xu J, Liu X, Liu J, Wang L (2011) Spatial distribution pattern analysis of groundwater nitrate nitrogen pollution in Shandong intensive farming regions of China using neural network method. *J Math Comput Model* 54(3–4):995–1004
- Mahdavi A, Tabatabaei SH, Mahdavi R, Nouri Emamzadei MR (2013) Application of digital techniques to identify aquifer artificial recharge sites in GIS environment. *Int J Digit Earth* 6(6):589–609
- Mehrabi H, Zeinivand H, Hadidi M (2013) Site selection for groundwater artificial recharge in Silakhor rangelands using GIS technique. *J Rangel Sci* 2(4):687–695
- Minnig M, Moecka Ch, Radnya D, Schirmer M (2015) Impact of urbanization on groundwater recharge rates in Dübendorf, Switzerland. *J Hydrol* 563:1135–1146
- Mousavi A (2009) Update of quantitative model and providing quality model of shabestar plain aquifer. Dissertation, University of Tabriz (in Persian)
- Oaksford ET (1985) Artificial recharge: methods, hydraulics and monitoring. In: Asano T (ed) *Artificial recharge of groundwater*. Butterworth, Boston, pp 69–127
- Pal M (2005) Random Forest classifier for remote sensing classification. *Int J Remote Sens* 26(1):217–222
- Peters J, De Baets B, Verhoest NE, Samson R, Degroevae S, De Becker P, Huybrechts W (2007) Random forests as a tool for ecohydrological distribution modelling. *J Ecol Model* 207(2–4):304–318
- Quinlan JR (1990) Learning logical definitions from relations. *Mach Learn* 5(3):239–266
- Quinlan JR (2014) *C4.5: programs for machine learning*. Elsevier
- Riad PH, Billib M, Hassan AA, Salam MA, El Din MN (2011) Application of the overlay weighted model and Boolean logic to determine the best locations for artificial recharge of groundwater. *J Urb Environ Engin* 5(2):57–66
- Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012a) An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J Photogramm Remote Sens* 67:93–104
- Rodriguez-Galiano VF, Chica-Olmo M, Abarca-Hernandez F, Atkinson PM, Jeganathan C (2012b) Random forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sens Environ* 121:93–107
- Saeyes Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinform* 23(19):2507–2517
- Sesnie SE, Gessler PE, Finegan B, Thessler S (2008) Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. *Remote Sens Environ* 112(5):2145–2159
- Shankar MR, Mohan G (2005) A GIS based hydrogeomorphic approach for identification of site-specific artificial-recharge techniques in the Deccan Volcanic Province. *J Earth Syst Sci* 114(5):505–514
- Tiwari AK, Lavy M, Amanzio G, De Maio M, Singh PK, Mahato MK (2017) Identification of artificial groundwater recharging zone using a GIS-based fuzzy logic approach: a case study in a coal mine area of the Damodar Valley, India. *Appl Water Sci* 7(8):4513–4524
- Zaidi FK, Nazzal Y, Ahmed I, Naeema M, Jafria MK (2015) Identification of potential artificial groundwater recharge zones in north-western Saudi Arabia using GIS and Boolean logic. *J Afr Earth Sc* 111:156–169
- Zhang GH, Fei YH, Xing KA (2004) A study on artificial underground reserve function in the aquifer with river channels in Taihang piedmont plain of China: a case study in the alluvial plain of Hutuo River. *J Arid Land Resour Environ* 18(1):42–48

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.