Water Resources Management (2020) 34:139–154
https://doi.org/10.1007/s11269-019-02447-w

Decision Tree-Based Data Mining and Rule Induction for Identifying High Quality Groundwater Zones to Water Supply Management: a Novel Hybrid Use of Data Mining and GIS

Mehrdad Jeihouni[1] · Ara Toomanian[1] · Ali Mansourian[1,2]

Received: 3 May 2019 / Accepted: 12 November 2019 /
Published online: 9 December 2019

Abstract
Groundwater is an important source to supply drinking water demands in both arid and semi-arid regions. Nevertheless, locating high quality drinking water is a major challenge in such areas. Against this background, this study proceeds to utilize and compare five decision tree-based data mining algorithms including Ordinary Decision Tree (ODT), Random Forest (RF), Random Tree (RT), Chi-square Automatic Interaction Detector (CHAID), and Iterative Dichotomiser 3 (ID3) for rule induction in order to identify high quality groundwater zones for drinking purposes. The proposed methodology works by initially extracting key relevant variables affecting water quality (electrical conductivity, pH, hardness and chloride) out of a total of eight existing parameters, and using them as inputs for the rule induction process. The algorithms were evaluated with reference to both continuous and discrete datasets. The findings were speculative of the superiority, performance-wise, of rule induction using the continuous dataset as opposed to the discrete dataset. Based on validation results, in continuous dataset, RF and ODT showed higher and RT showed acceptable performance. The groundwater quality maps were generated by combining the effective parameters distribution maps using inducted rules from RF, ODT, and RT, in GIS environment. A quick glance at the generated maps reveals a drop in the quality of groundwater from south to north as well as from east to west in the study area. The RF showed the highest performance (accuracy of 97.10%) among its counterparts; and so the generated map based on rules inducted from RF is more reliable. The RF and ODT methods are more suitable in the case of continuous dataset and can be applied for rule induction to determine water quality with higher accuracy compared to other tested algorithms.

Keywords Geostatistics · Random forest · Random tree · Decision tree · Water quality

✉ Ara Toomanian
a.toomanian@ut.ac.ir

✉ Ali Mansourian
ali.mansourian@nateko.lu.se

Extended author information available on the last page of the article

1 Introduction

Groundwater is an important source of drinking water supply in arid and semi-arid areas, which generally encounter water shortages arising from climate change. Owing to climatic change and alterations in global precipitation patterns, the quantity of drought years has increased in many countries located in arid and semi-arid regions of the world. Under such circumstances, permanent rivers will change to seasonal rivers (Zarghami et al. 2011) and groundwater becomes the main source to supply water demands, especially for drinking purposes.

Determining groundwater quality and locating areas with high quality water for drinking purpose is the principle challenge in this regard. There are a variety of standards specified to determine groundwater quality, put forth by World Health Organization (WHO) and/or relevant organizations in different countries. Water quality standards in Iran are specified by the Institute of Standards and Industrial Research of Iran (ISIR). WHO and ISIR standards have specify only maximum contamination levels and admissible limits of water suitable for drinking purposes. Although parameters pertaining to underground water quality may very well be within standard limits, they lack the same quality. Furthermore, each parameter may vary along a wide range prior to reaching its admissible limit.

Employing a few parameters to determine water quality as opposed to several quality parameters (variables) in different ranges and units is one possible and rather interesting approach, which is, however, difficult and requires a *Decision support system* (DSS). Accordingly, the combination of these parameters and classification of water quality appear to be an important issue where the critical role of DSS is concerned. DSS assists decision makers in optimizing their decisions (Turban 1993). Moreover, *Geographic Information System* (GIS) is highlighted as quite an effective tool apropos of its spatial decision support role and function in developing GIS-based spatial DSS (SDSS) (Jeihouni et al. 2015).

Several studies have thus far been conducted with the objective to assess water quality parameters and specify their spatial distribution using GIS; e.g. D'Agostino et al. (1998), Hudak (2000, 2001), Gaus et al. (2003), Hudak and Sanmanee (2003), Yimit et al. (2011), Arslan (2012), Bhunia et al. (2018). These studies have solely focused on the distribution map of water quality parameters, with disregard for the combination of the layers. Nas and Berktay (2010) generated a water quality map for drinking purpose by a simple overlaying of thematic maps of pH, electrical conductivity (EC), chloride, sulfate, hardness, and nitrate. Jeihouni et al. (2015) have developed a *Multiple-criteria Spatial Decision Support System* (MC-SDSS) where they used pH, EC, chloride, sulfate, and hardness as water quality parameters and employed *Analytical Hierarchy Process* (AHP) as a *multiple-criteria decision-making* (MCDM) technique to generate water quality map based on the importance of each parameter. In this approach, the importance of each parameter was determined by experts and the significance of each parameter range, when within the permissible limit, was covered by the weight of parameter.

Data mining and knowledge discovery in databases (KDD) are suitable approaches to extracting patterns of interest from databases (Fayyad et al. 1996). The KDD is the automated extraction of valid, novel, understandable and useful patterns representing knowledge in databases (Rokach and Maimon 2005; Han et al. 2011) and Data mining is the core of the KDD process (Rokach and Maimon 2005). Data mining classifier algorithms have been developed in recent decades and have been broadly used for classification, modeling and rule induction (Peters et al. 2007; Taghizadeh-Mehrjardi et al. 2015; Rodriguez-Galiano et al.

2012, 2014; Yoo et al. 2016; Rahmati et al. 2016; Naghibi et al. 2017; Sahoo et al. 2018; Arabameri et al. 2019; Chen et al. 2019; Shahbazi et al. 2019; Miraki et al. 2019; Sherafatpour et al. 2019).

Data mining decision tree algorithms aim to develop a classification graph and predict the target class based on the input training dataset. These algorithms are able to learn the relationships between input variables and corresponding outputs, and represent each relationship by specific rules. There are several data mining algorithms based on tree induction and utilized for classification. Some of the more broadly used tree induction algorithms are; Ordinary Decision Tree (ODT), Random Tree (RT), Random Forest (RF), Iterative Dichotomiser 3 (ID3), and Chi-square Automatic Interaction Detector (CHAID). These decision tree algorithms are generated through recursive partitioning. The summary of tree-based data mining algorithms is presented in Table 1. These tree-based algorithms are frequently used in many fields and for different applications (Kim et al. 2011; Rodriguez-Galiano et al. 2012; Rahmati et al. 2016; Yoo et al. 2016; Belgiu and Drăguţ 2016; Hong et al. 2016; Naghibi et al. 2017; Heil et al. 2017; Chen et al. 2017; Robinson et al. 2018; Rayaroth and Sivaradje 2019; Al-Juboori 2019). For specific detail on the theoretical bases of data mining, its algorithms, and application, the reader is referred to the works of Rokach and Maimon (2005), Han et al. (2011), Liao et al. (2012), Rokach and Maimon (2014).

Data mining and rule induction techniques are able to extract rules from data and predict previously unknown events (Yoo et al. 2016). Decision tree-based techniques have a high capability for rule induction and extracting relationship between variables, in order to categorize them into meaningful classes. Given the similarities between the determining water quality based on several parameters and classification, data mining classification algorithms and rule induction techniques can be used to generate water quality maps. The mentioned capability of data mining and rule induction in classification of water quality has not been addressed in any research as of yet. This capability will be tested, for the first time, on groundwater quality mapping.

The prime objectives of this study include: (1) specification of key water quality parameters influencing the determination of water quality, (2) evaluation of the capability of decision tree based rule induction methods such as ODT, RF, RT, ID3 and CHAID in identifying high quality groundwater zones for drinking purposes, and (3) generation of groundwater quality maps in Tabriz City and its township based on key parameters and extracted rules from data mining techniques.

This paper proposes a novel approach for water quality mapping by hybrid use of data mining and GIS. The innovation of the study lies in inducing rules from database to combine the layers of the effective water quality parameters to generate the groundwater quality map.

2 Materials and Methods

2.1 Study Area

The study area is Tabriz City, the capital of the East Azerbaijan province, and its township in northwestern Iran (Fig. 1). The area is located between latitudes 37° 56′ and 38° 11′ N and longitudes 46° 2′ and 46° 36′ E, with an approximate area of 1200 km^2. Tabriz is located in a semi-arid region and has a high population circa 1.5 million. Tabriz water demands are mainly

Table 1 The summary of tree-based mining algorithms (Rokach and Maimon 2005, 2014; Quinlan 1986; Pudumalar et al. 2017)

Algorithm	Data type		Class label	Description
	Nominal	Numerical		
ODT	*	*	Nominal	Typical basic type of decision tree algorithms
RT	*	*	Nominal	The training sample set is projected to subsets and for each split the algorithm uses only a random subset of attributes to generate a tree
RF	*	*	Nominal	The RF creates a set of random trees by repeating the RT process
ID3	*		Nominal	ID3 is very simple decision tree algorithm and generates tree based on a fixed sample set by using a greedy search and the tree model generated without pruning
CHAID	*		Nominal	Employs chi-squared based criterion instead of the gain ratio or information gain this algorithm generate tree model without pruning. It is a non-parametric algorithm and uses frequencies instead of mean and variance

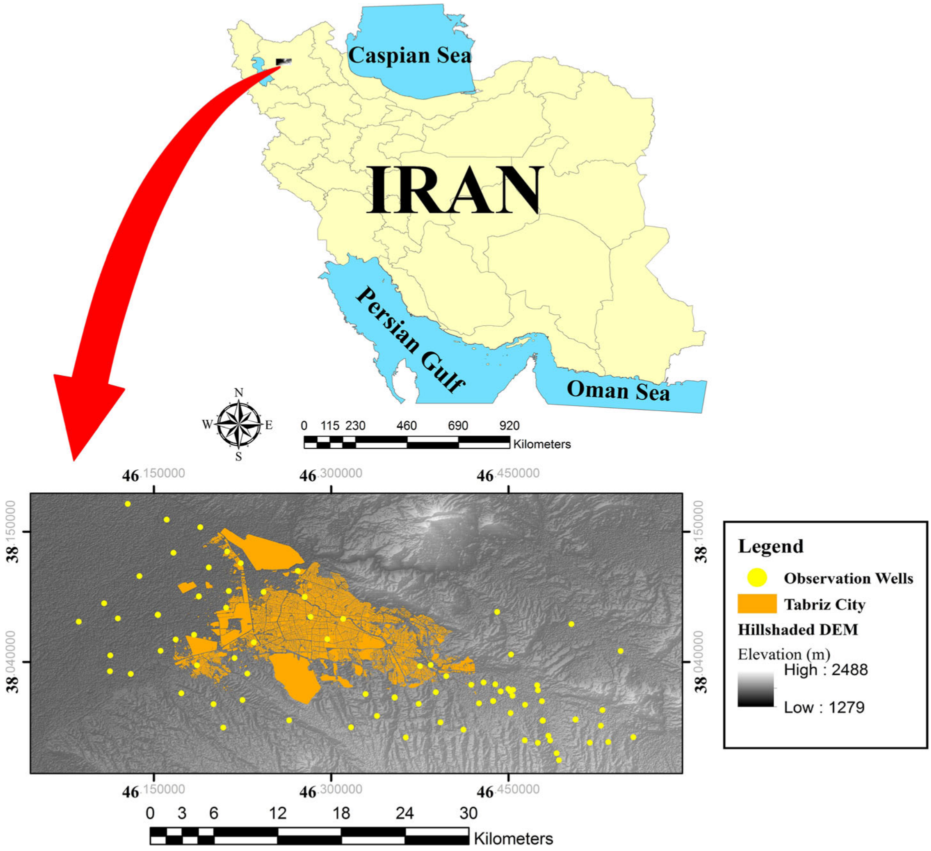

Fig. 1 Study area and observation points

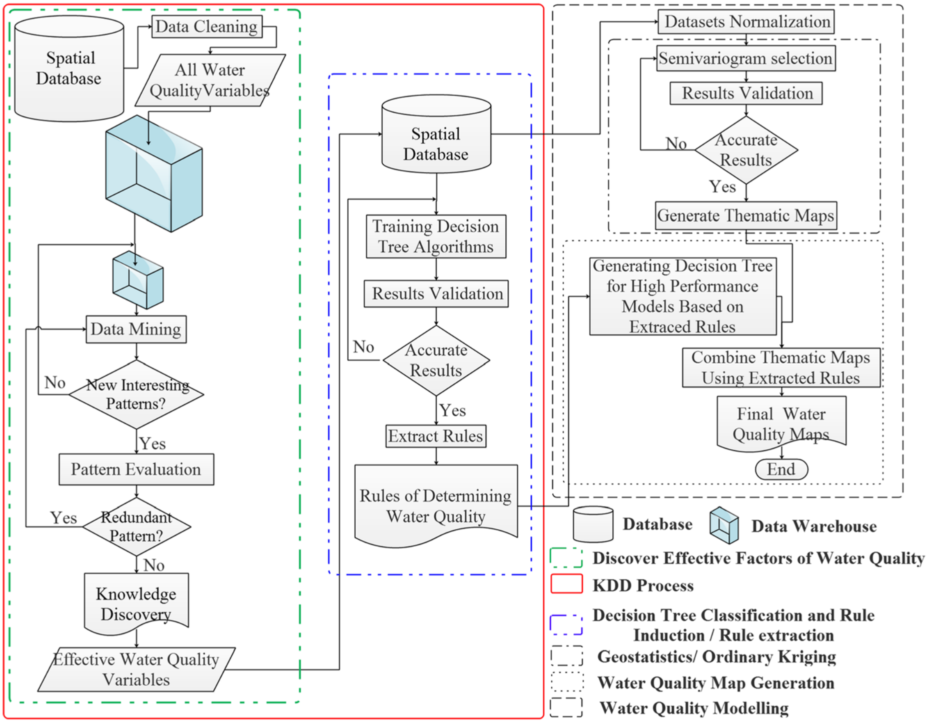
Spatial Database
Data Cleaning
All Water QualityVariables
Data Mining
No
New Interesting Patterns?
Yes
Pattern Evaluation
Yes
Redundant Pattern?
No
Knowledge Discovery
Effective Water Quality Variables
Spatial Database
Training Decision Tree Algorithms
Results Validation
No
Accurate Results
Yes
Extract Rules
Rules of Determining Water Quality
Datasets Normalization
Semivariogram selection
Results Validation
No
Accurate Results
Yes
Generate Thematic Maps
Generating Decision Tree for High Performance Models Based on Extraced Rules
Combine Thematic Maps Using Extracted Rules
Final Water Quality Maps
End
Database
Data Warehouse
Discover Effective Factors of Water Quality
KDD Process
Decision Tree Classification and Rule Induction / Rule extraction
Geostatistics/ Ordinary Kriging
Water Quality Map Generation
Water Quality Modelling

Fig. 2 Flowchart for the proposed methodology

supplied from two dams located at 30 and 200 Km distances from Tabriz, with smaller contributions from certain wells located at a distance of 25Km from Tabriz (Jeihouni et al. 2015). Recent drought years have threatened the water supply from permanent rivers for

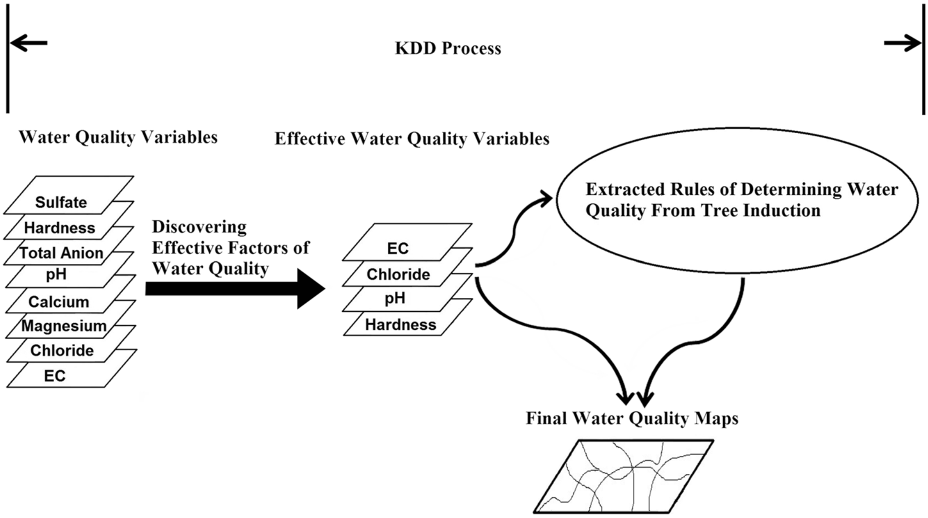
KDD Process
Water Quality Variables
Effective Water Quality Variables
Sulfate
Hardness
Total Anion
pH
Calcium
Magnesium
Chloride
EC
Discovering Effective Factors of Water Quality
EC
Chloride
pH
Hardness
Extracted Rules of Determining Water Quality From Tree Induction
Final Water Quality Maps

Fig. 3 A brief schema of the proposed method

Table 2 Statistical evaluation of groundwater quality variables

Parameter	Min	Max	Mean	Median	SD	Skewness	Kurtosis	Transformation[b]
Sulfate(mg/L)	0.15	19	4.15	2.67	3.76	1.20	4.55	–
Hardness(°F)[a]	4.27	144.19	34.40	23.76	28.96	1.39	4.75	Lognormal
Total anion	1.95	115.85	21.77	12.83	22.85	1.87	4.88	–
pH[a]	7.05	9.1	8.20	8.2	0.36	−0.23	3.09	Not needed
Calcium(mg/L)	0.9	27.22	5.68	4.47	4.53	2.36	7.69	–
Magnesium(mg/L)	0.44	19.3	4.89	2.65	4.74	1.26	3.78	–
Chloride(mg/L)[a]	0.22	100	12.50	3.45	19.44	2.43	7.39	Lognormal
EC (μS/cm)[a]	199	11,590	2181.6	1287.3	2286.1	1.87	6.61	Lognormal

[a] The effective factors extracted during the feature selection phase

[b] Transformations applied on selected factors for generation of distribution maps using OK

Tabriz. The importance of locating high quality groundwater resources at near distances of Tabriz has been thoroughly discussed by Jeihouni et al. (2015).

2.2 Data and Data Preprocessing

Groundwater quality variables such as sulfate, hardness, total anion, pH, calcium, magnesium, chloride, and EC were measured at 80 observation wells of Tabriz urban groundwater and its township (Fig. 1). The dataset has been collected by the Iranian Ministry of Energy (IMOE). Then, based on laboratory result for each parameter, experts of water quality analysis were asked to rank each water sample from class A to I (A being the optimum quality class and I the poor-quality class). The final dataset was eventually prepared for further data mining analysis.

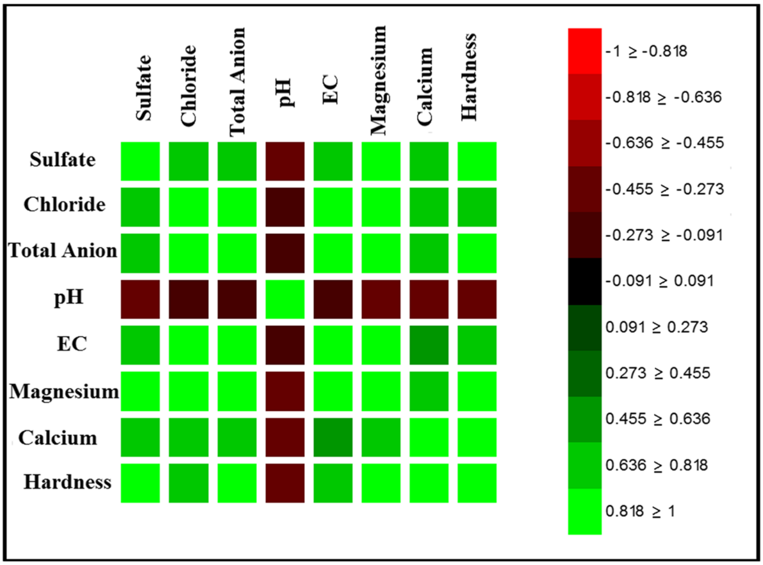

Fig. 4 Correlation matrix map for all groundwater quality variables

Table 3 Performance report of the employed methods capability in full range (continuous) datasets

Method	Accuracy (%)	Classification error (%)	Kappa	Spearman rho	Kendall_tau	Correlation
ODT	91.30	8.70	0.885	0.902	0.862	0.916
RT	79.71	20.29	0.731	0.911	0.822	0.845
RF	97.10	2.90	0.962	0.996	0.987	0.983

2.3 Feature Selection

Feature selection is an important phase in modelling, which decreases the dimensionality of data and increases model efficiency (Taghizadeh-Mehrjardi et al. 2016). A variety of heuristic and classical methods have thus far been frequently used for feature selection such as expert knowledge (Sharififar et al. 2019), stepwise linear regression (SR) (Wang et al. 2018) and genetic algorithms (GA) (Wang et al. 2018; Yang et al. 2019). In this study the effective water quality parameters were selected during the KDD process and pattern evaluation phase for extracting effective factors. In this phase, the RF algorithm was used to discover patterns and select the most effective parameters based on extracted patterns.

2.4 Methodology

A three step method was proposed to generate water quality maps. The first step consisted of basic preprocessing, such as creation of a spatial database, data cleaning and extraction of water quality variables. The data were then relocated to an operational repository called an "operational data store" for further processing, integration and additional operations to verify data quality, prior to implementing the data warehouse (DW), the core and the central repository for integrated data. The DW stores data applicable for decision making and creating analytical reports, upon which a professional DW called Data Mart is generated based on the required datasets and parameters before employing data mining pattern recognition techniques. The RF algorithm is then used to extract patterns and discover relevant parameters that affect water quality. The next step includes the use of the most effective parameters as inputs to discover the relationships between parameters and water quality classes in both full range/continuous (numerical) and classified/discrete (nominal) datasets by ODT, RT, RF, ID3 and CHAID methods to induct rules that determine the quality of groundwater for drinking purposes. In the final step, to generate water quality maps, the inducted rules from three high accuracy models were used to combine the spatial distribution (thematic) maps of effective parameters generated using ordinary kriging (OK) as a geostatistical technique. The brief methodology and schema of the proposed method are shown in Figs. 2 and 3.

Table 4 Groundwater quality classification for EC, chloride, hardness (Ducci 1999) and pH (Jeihouni et al. 2015)

Quality	Class	EC (µS/cm)	pH	Hardness(°F)	Chloride (mg/L)
Optimum	A	<1000	7—7.5	<30	<50
Medium	B	1000—2000	7.5—8	30—50	50—200
Poor	C	>2000	>8	>50	>200

3 Results and Discussion

The statistical evaluation results of the groundwater quality variables; sulfate, hardness, total anions, pH, calcium, magnesium, chloride and EC are shown in Table 2, with the corresponding correlation matrix map presented in Fig. 4. The effective parameters for water quality, as determined using data mining processes and pattern extraction, were: hardness, pH, chloride and EC, all of which are highlighted in Table 2. The mentioned factors were subsequently used as input to generate decision trees and perform rule induction.

The tree generation algorithms were evaluated on two dataset types; numerical and nominal, so as to assess their rule induction capabilities. In the case of the continuous technique (for numerical dataset), the numerical data were used and trees were generated based on OD, RT and RF methods. The inducted rules were assessed based on statistical criteria such as accuracy, classification error, Kappa coefficient, Spearman rho, Kendall_tau and correlation. The validation results are shown in Table 3.

As evident in Table 3, the prediction capability of RF method was higher than ODT and RT method based on the model evaluation criteria. RF had the lowest classification error and highest accuracy, kappa, Spearman rho and correlation values of 97.10, 2.9, 0.962, 0.996, 0.987 and 0.983, respectively, among other methods, which prove the ability and superiority of this model. Moreover, the accuracy of ODT was higher than RT based on the validation criteria.

In the next step, to generate the decision trees and rule induction for nominal dataset, each effective parameter was classified into three classes based on Ducci (1999) and Jeihouni et al. (2015). The classification thresholds are shown in Table 4. The decision trees were then generated by employing ODT, RT, RF, ID3, and CHAID methods. The rule induction accuracy was assessed based on statistical criteria. Table 5 lists the validation results.

As indicated by Table 5, the accuracies of inducted rules in the nominal dataset were very poor in comparison with the numerical dataset. The maximum Kappa coefficient (0.617) was observed for the ID3 method. ID3 also had the highest accuracy in nominal datasets owing to its capability and compatibility in handling nominal datasets. Despite of ID3's high ability for rule induction from the nominal datasets, this method failed to attain a high accuracy.

The results in Tables 3 and 5 indicate the superiority of numerical decision tree generation algorithms and rule induction methods (ODT, RT, and RF), in terms of accuracy and ability in the case of the continuous datasets. Accordingly, the numerical approach was more suitable for handling, mining and inducing rules for water quality classification.

Based on the results presented in Tables 3 and 5, given the high performances of the applied algorithms in numerical datasets, the water quality decision trees were only generated for ODT, RT, and RF models. An instance of a classification tree generated by the ODT technique is shown in Fig. 5.

Table 5 Performance report of the employed methods capability in classified (discrete) datasets

Method	Accuracy (%)	Classification error (%)	Kappa	Spearman rho	Kendall_tau	Correlation
ODT	70.00	30.00	0.598	0.792	0.726	0.840
RT	61.43	38.57	0.438	0.780	0.692	0.786
RF	60.00	40.00	0.41	0.799	0.692	0.813
ID3	71.43	28.57	0.617	0.796	0.733	0.845
CHAID	70.00	30.00	0.597	0.795	0.731	0.845

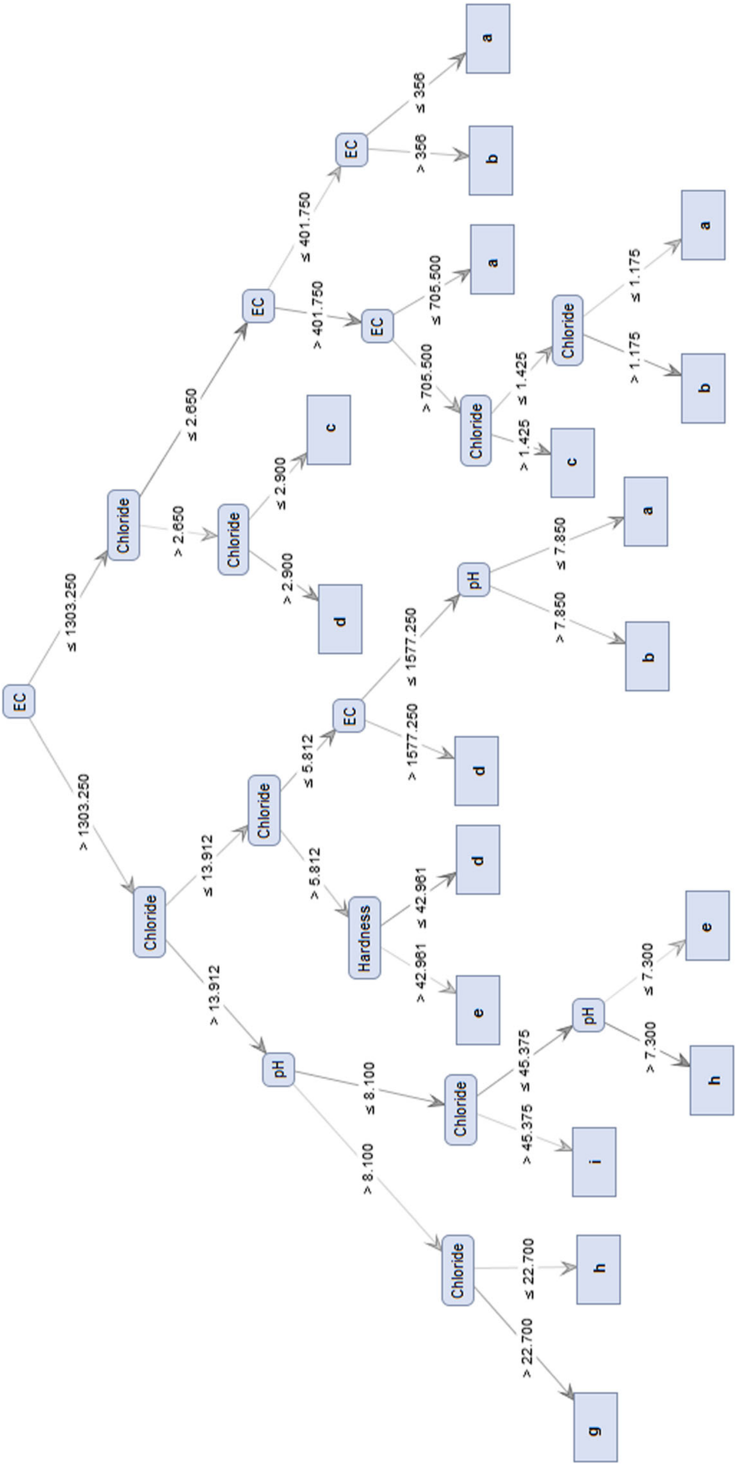

Fig. 5 Classification tree generated by ODT technique

The generated trees and inducted IF-THEN rules were implemented on water quality parameters distribution maps to generate final water quality maps based on ODT, RT and RF methods. To generate spatial distribution maps of each effective parameter a geostatistical approach of OK was employed.

The implementation of the OK method involves an initial checking of the spatial autocorrelation of effective parameters through Moran's I (index). Figure 6 illustrates a graphical representation of the Moran's I report. As can be observed in Fig. 6, all parameters have cluster patterns and could be modeled by OK. Due to the high performance of the OK in normally distributed datasets (Jeihouni et al. 2018), the normality of effective parameters datasets were assessed through their histograms and statistical criteria of skewness and kurtosis, which are presented in Table 2. pH had a normal distribution, whereas EC, hardness and chloride datasets were not normally distributed. For this reason, the non-normally distributed datasets were normalized by lognormal transformation (Table 2).

The basis of the OK method is to find the best fitted variogram in order to obtain an accurate estimation of each effective factor at unsampled locations. Accordingly, the 11 commonly used semi-variogram models (e.g. Circular, Spherical, Tetraspherical, Pentaspherical, Exponential, Gaussian, Rational Quadratic, Hole effect, K-Bessel, J-Bessel, and Stable) were tested for hardness, pH, chloride, and EC datasets. The fitted semi-variograms for all parameters are presented in Fig. 7 and their corresponding

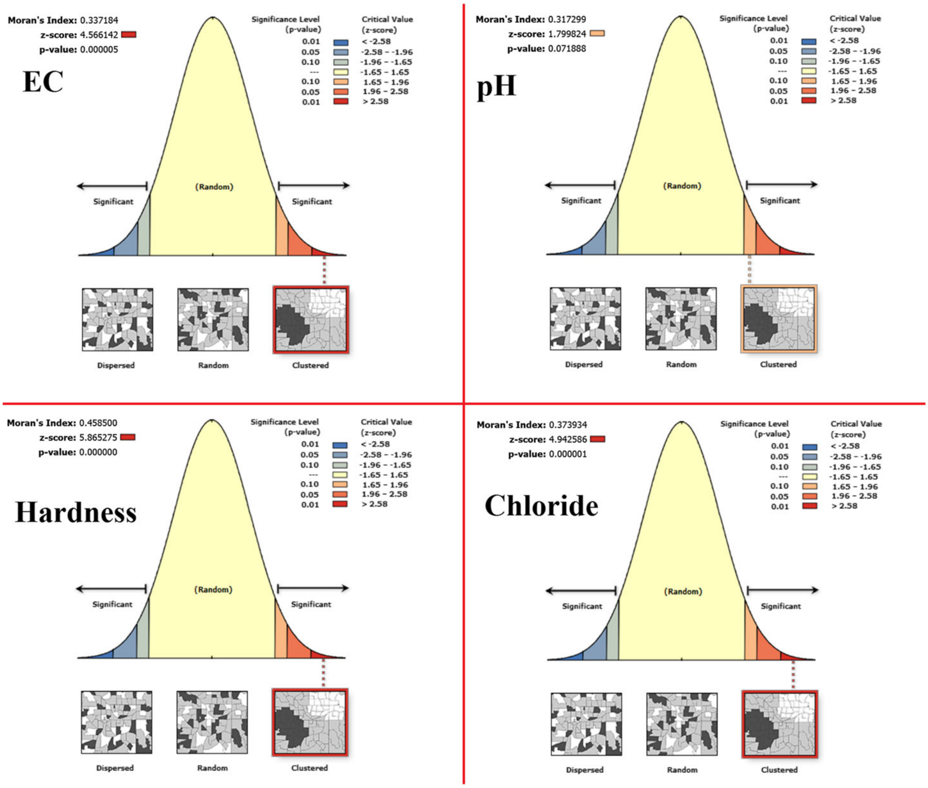

Fig. 6 The spatial autocorrelation report based on Moran's I (index) for effective parameters

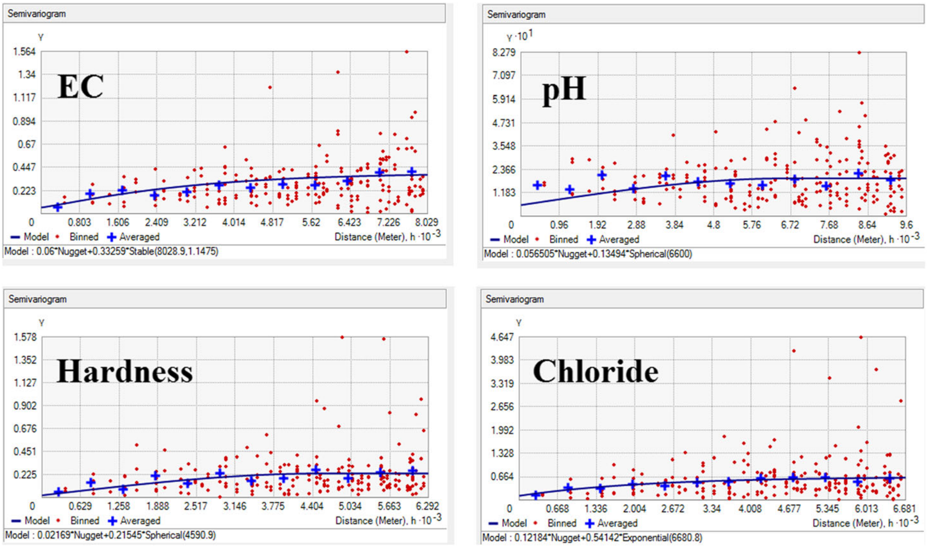

Fig. 7 The fitted semi-variogram models for effective parameters

model parameters are summarized in Table 6. The minimum and maximum ranges indicates the highest and the lowest spatial variability (Jeihouni et al. 2018), belonged to hardness and EC, respectively. Moreover, the factors nugget to sill ratio (%) was evaluated as a criterion for classifying the spatial dependence (Caro et al. 2013). The ratio less than 25% indicates the strong spatial dependence, the ratio between 25 and 75% indicates the moderate spatial dependence, and the ratio greater than 75% indicates the weak spatial dependence of the variable (Cambardella et al. 1994). According to this criterion, EC, hardness and chloride have strong spatial dependence, while pH has a moderate spatial dependence. Spatial distribution maps of the effective factors were then generated based on the best fitted semi-variogram model for each factor (Fig. 8).

In the final stage, the distribution maps were combined based on the inducted rules from ODT, RT, and RF methods to generate final groundwater quality maps (Figs. 9, 10, and 11).

Groundwater quality maps generated using inducted rules from ODT, RT and RF (Figs. 9, 10, and 11) confirm the overall groundwater quality patterns, albeit some inconsistencies were apparent in certain areas. Referring to the final groundwater quality maps, the groundwater quality decrease from south to north and from east to west of the region, indicating quality gradients. All quality maps indicate the southern areas as the optimal choice for drinking water supply, whereas northern,

Table 6 Semi-variogram models parameters for groundwater quality factors

Factor	Model	Nugget	Sill	Range (m)	Nugget/Sill (ratio %)
EC	Stable	0.060	0.332	8028	18.07
pH	Spherical	0.056	0.135	6600	41.48
Hardness	Spherical	0.021	0.215	4590	9.76
Chloride	Exponential	0.12	0.541	6680	22.18

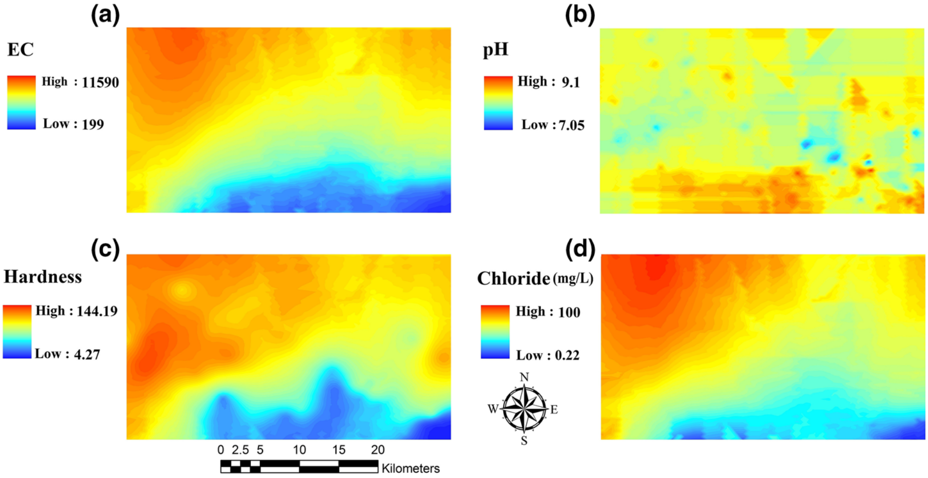

Fig. 8 Spatial distribution maps of water quality variables in the study area: **a** EC, **b** pH, **c** Hardness, and **d** Chloride

northwestern and western parts of the study region were of poor quality. The groundwater resources located to the south and south east of the Tabriz are charged by the Sahand Mountains. Differences between maps hint of differences among rules inducted from the employed algorithms, which resulted in dissimilar classes.

Regarding Fig. 9, the ODT method identified high quality water sources in northeastern, eastern, southeastern, southern, southwestern and central parts of the study area. The RT method (Fig. 10), however identified regions to the south and southeast as optimum quality water sources, while sectors in the northeast, southwest and central areas were labeled as medium quality and northwest of the study area as poor quality. The performance and capability of the RT method in determining optimum and poor-quality zones was lower. Based on RF map (Fig. 11), only the southern and southeastern parts of the study area were

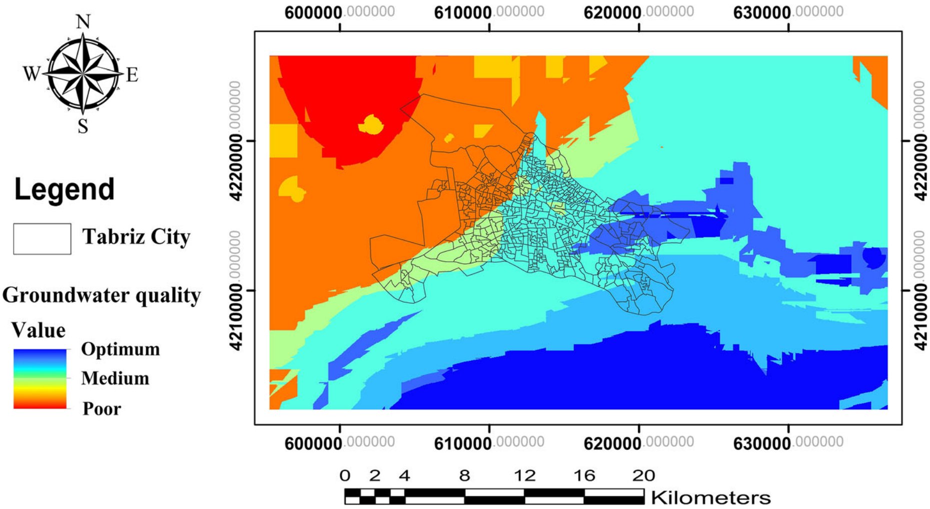

Fig. 9 Water quality map using the inducted rules from ODT (optimal in blue and poor in red)

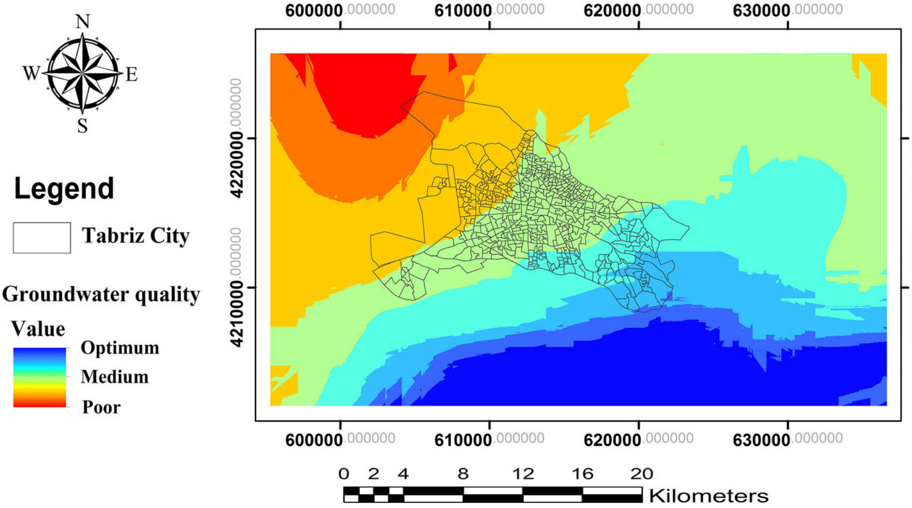

Fig. 10 Water quality map using the inducted rules from RT (optimal in blue and poor in red)

highlighted as optimum quality. These areas are surrounded by a thin border of medium quality water, while other zones have poor or near poor quality for drinking purpose.

Based on the accuracy assessment tests and validation results, all three algorithms had high performances and dependable results; however, the map generated using the RF method was the most reliable and can be used as a base for managerial decision-making procedures. This study highlighted the ability of a hybrid approach, incorporating data mining and GIS, in order to determine groundwater quality and locate high quality groundwater sources. Data mining can convert experts' ideas to tangible IF-THEN rules based on the utilized decision tree methods, which can be utilized by non-experts for water supply management.

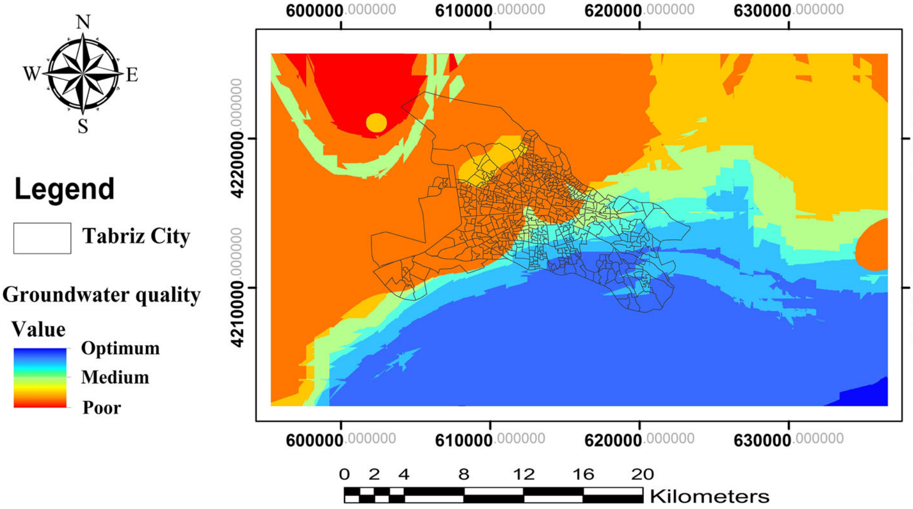

Fig. 11 Water quality map using the inducted rules from RF (optimal in blue and poor in red)

4 Conclusion

This study sought to evaluate five common tree-based data mining algorithms (ODT, RT, RF, ID3 and CHAID) with respect to their performance and capabilities in rule induction for identifying high quality groundwater zones for drinking purposes. The main goals of the current study were to induct rules for determining water quality and utilize them to generate water quality maps over the Tabriz city. To achieve these goals, the most relevant water quality parameters were extracted from among eight water quality variables, and were then used as inputs to different data mining algorithms in both continuous (numerical) and discrete (nominal) datasets. The algorithms performances were assessed by specific statistical analysis approaches. The results indicate the superiority of inducted rules from ODT, RT and RF in numerical dataset and an under-performance of all five algorithms in nominal dataset, even ID3 and CHAID, which are generally suitable for nominal datasets. The inducted rules from ODT, RT and RF have superior performance and accuracy, and were consequently used to generate water quality maps. Spatial distribution maps for the key parameters were subsequently generated using OK method and combined based on the inducted rules. The final generated quality maps demonstrate the groundwater quality gradient, wherein water quality decreases from south to north and from east to west of the study region. Decision tree-based data mining algorithms and rule induction approaches showed a relatively high capability in classifying water quality based on a limited dataset. The RF method had the highest performance and the generated quality map was more reliable. Finally, it is recommended that the RF and ODT methods be used to induce rules for water quality determination in numerical datasets.

Funding Information Open access funding provided by Lund University.

References

Al-Juboori AM (2019) Generating monthly stream flow using nearest river data: assessing different trees models. Water Resour Manag 33(9):3257–3270

Arabameri A, Rezaei K, Cerda A, Lombardo L, Rodrigo-Comino J (2019) GIS-based groundwater potential mapping in Shahroud plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. Sci Total Environ 658:160–177

Arslan H (2012) Spatial and temporal mapping of groundwater salinity using ordinary kriging and indicator kriging: the case of Bafra plain, Turkey. Agric Water Manag 113:57–63

Belgiu M, Drăguț L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31

Bhunia GS, Keshavarzi A, Shit PK, Omran ESE, Bagherzadeh A (2018) Evaluation of groundwater quality and its suitability for drinking and irrigation using GIS and geostatistics techniques in semiarid region of Neyshabur, Iran. Appl Water Sci 8(6):168–116. https://doi.org/10.1007/s13201-018-0795-6

Cambardella CA, Moorman TB, Parkin TB, Karlen DL, Novak JM, Turco RF, Konopka AE (1994) Field-scale variability of soil properties in Central Iowa soils. Soil Sci Soc Am J 58(5):1501–1511

Caro A, Legarda F, Romero L, Herranz M, Barrera M, Valiño F et al (2013) Map on predicted deposition of Cs-137 in Spanish soils from geostatistical analyses. J Environ Radioactiv 115:53–59

Chen G, Long T, Xiong J, Bai Y (2017) Multiple random forests modelling for urban water consumption forecasting. Water Resour Manag 31(15):4715–4729

Chen W, Tsangaratos P, Ilia I, Duan Z, Chen X (2019) Groundwater spring potential mapping using population-based evolutionary algorithms and data mining methods. Sci Total Environ 684:31–49

D'Agostino V, Greene E, Passarella G, Vurro M (1998) Spatial and temporal study of nitrate concentration in groundwater by means of coregionalization. Environ Geol 36:285–295

Ducci D (1999) GIS techniques for mapping groundwater contamination risk. Nat Hazards 20(2–3):279–294

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Commun ACM 39:27–34

Gaus I, Kinniburgh D, Talbot J, Webster R (2003) Geostatistical analysis of arsenic concentration in groundwater in Bangladesh using disjunctive kriging. Environ Geol 44:939–948

Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier

Heil J, Michaelis X, Marschner B, Stumpe B (2017) The power of random forest for the identification and quantification of technogenic substrates in urban soils on the basis of DRIFT spectra. Environ Pollut 230: 574–583

Hong H, Pourghasemi HR, Pourtaghi ZS (2016) Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. Geomorphology 259:105–118

Hudak P (2000) Regional trends in nitrate content of Texas groundwater. J Hydrol 228:37–47

Hudak PF (2001) Water hardness and sodium trends in Texas aquifers. Environ Monit Assess 68:177–185

Hudak PF, Sanmanee S (2003) Spatial patterns of nitrate, chloride, sulfate, and fluoride concentrations in the woodbine aquifer of north-Central Texas. Environ Monit Assess 82:311–320

Jeihouni M, Toomanian A, Alavipanah SK, Hamzeh S, Pilesjö P (2018) Long term groundwater balance and water quality monitoring in the eastern plains of Urmia Lake, Iran: a novel GIS based low cost approach. J Afr Earth Sci 147:11–19

Jeihouni M, Toomanian A, Alavipanah SK, Shahabi M, Bazdar S (2015) An application of MC-SDSS for water supply management during a drought crisis. Environ Monit Assess 187:396

Kim K, Yoo K, Ki D, Son IS, Oh KJ, Park J (2011) Decision-tree-based data mining and rule induction for predicting and mapping soil bacterial diversity. Environ Monit Assess 178:595–610

Liao S-H, Chu P-H, Hsiao P-Y (2012) Data mining techniques and applications–a decade review from 2000 to 2011. Expert Syst Appl 39:11303–11311

Miraki S, Zanganeh SH, Chapi K, Singh VP, Shirzadi A, Shahabi H, Pham BT (2019) Mapping groundwater potential using a novel hybrid intelligence approach. Water Resour Manag 33(1):281–302

Naghibi SA, Ahmadi K, Daneshi A (2017) Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resour Manag 31(9): 2761–2775

Nas B, Berktay A (2010) Groundwater quality mapping in urban groundwater using GIS. Environ Monit Assess 160:215–227

Peters J, De Baets B, Verhoest NE, Samson R, Degroeve S, De Becker P, Huybrechts W (2007) Random forests as a tool for ecohydrological distribution modeling. Ecol Model 207:304–318

Pudumalar S, Ramanujam E, Rajashree RH, Kavya C, Kiruthika T, Nisha J (2017) Crop recommendation system for precision agriculture. In: 2016 Eighth International Conference on Advanced Computing (ICoAC). IEEE, pp 32–36

Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106

Rahmati O, Pourghasemi HR, Melesse AM (2016) Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran region, Iran. Catena 137:360–372

Rayaroth R, Sivaradje G (2019) Random bagging classifier and shuffled frog leaping based optimal sensor placement for leakage detection in WDS. Water Resour Manag 33(9):3111–3125

Robinson G, Moutari S, Ahmed AA, Hamill GA (2018) An advanced calibration method for image analysis in laboratory-scale seawater intrusion problems. Water Resour Manag 32(9):3087–3102

Rodriguez-Galiano V, Mendes MP, Garcia-Soldado MJ, Chica-Olmo M, Ribeiro L (2014) Predictive modeling of groundwater nitrate pollution using random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (southern Spain). Sci Total Environ 476:189–206

Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens 67:93–104

Rokach L, Maimon O (2005) The data mining and knowledge discovery handbook: a complete guide for researchers and practitioners. Springer, New York

Rokach L, Maimon O (2014) Data mining with decision trees: theory and applications. Second edition. World scientific

Sahoo M, Kasot A, Dhar A, Kar A (2018) On predictability of groundwater level in shallow wells using satellite observations. Water Resour Manag 32(4):1225–1244

Shahbazi F, McBratney A, Malone B, Oustan S, Minasny B (2019) Retrospective monitoring of the spatial variability of crystalline iron in soils of the east shore of Urmia Lake, Iran using remotely sensed data and digital maps. Geoderma 337:1196–1207

Sharififar A, Sarmadian F, Malone BP, Minasny B (2019) Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma 350:84–92

Sherafatpour Z, Roozbahani A, Hasani Y (2019) Agricultural water allocation by integration of hydro-economic modeling with Bayesian networks and random forest approaches. Water Resour Manag 33(7):2277–2299

Taghizadeh-Mehrjardi R, Nabiollahi K, Kerry R (2016) Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma 266:98–110

Taghizadeh-Mehrjardi R, Nabiollahi K, Minasny B, Triantafilis J (2015) Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253:67–77

Turban E (1993) Decision support and expert systems: management support systems. Prentice Hall PTR

Wang B, Waters C, Orgill S, Cowie A, Clark A, Li Liu D et al (2018) Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. Ecol Indic 88:425–438

Yang Y, Rossel RAV, Li S, Bissett A, Lee J, Shi Z et al (2019) Soil bacterial abundance and diversity better explained and predicted with spectro-transfer functions. Soil Biol Biochem 129:29–38

Yimit H, Eziz M, Mamat M, Tohti G (2011) Variations in groundwater levels and salinity in the Ili River irrigation area, Xinjiang, Northwest China: a geostatistical approach. Int J Sust Dev World 18:55–64

Yoo K, Shukla SK, Ahn JJ, Oh K, Park J (2016) Decision tree-based data mining and rule induction for identifying hydrogeological parameters that influence groundwater pollution sensitivity. J Clean Prod 122: 277–286

Zarghami M, Abdi A, Babaeian I, Hassanzadeh Y, Kanani R (2011) Impacts of climate change on runoffs in East Azerbaijan, Iran. Glob Planet Change 78:137–146

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

Affiliations

Mehrdad Jeihouni[1] • **Ara Toomanian**[1] • **Ali Mansourian**[1,2]

[1] Deparment of Remote Sensing and GIS, Faculty of Geography, University of Tehran, Azin Alley. 50, Vesal Str., Tehran, Iran

[2] GIS Center, Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden