Natural Resources Research

Novel hybrid integration approach of bagging-based Fisher's linear discriminant function for groundwater potential analysis

--Manuscript Draft--

Manuscript Number:	
Full Title:	Novel hybrid integration approach of bagging-based Fisher's linear discriminant function for groundwater potential analysis
Article Type:	Original Research
Keywords:	Groundwater; Machine learning; Fisher's linear discriminant function (FLDA); Rotation forest (RF); GIS
Corresponding Author:	Biswajeet Pradhan, PhD University of Technology Sydney Sydney, NSW AUSTRALIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Technology Sydney
Corresponding Author's Secondary Institution:	
First Author:	Wei Chen, PhD
First Author Secondary Information:	
Order of Authors:	Wei Chen, PhD
	Biswajeet Pradhan, PhD
	Shaojun Li, PhD
	Himan Shahabi, PhD
	Enke Hou, PhD
	Shengquan Wang, PhD
	Hossein Mojaddadi Rizeei, PhD
Order of Authors Secondary Information:	
Funding Information:	Chinese Academy of Science (115242KYSB20170022) — Dr Wei Chen
Abstract:	Groundwater is a vital water source in the rural and urban areas of developing and developed nations. In this study, a novel hybrid integration approach of Fisher's linear discriminant function (FLDA) with rotation forest (RFLDA) and bagging (BFLDA) ensembles was used for groundwater potential assessment at the Ningtiaota area in Shaanxi, China. A spatial database with 66 groundwater spring locations and 14 groundwater spring contributing factors was prepared; these factors were elevation, aspect, slope, plan and profile curvatures, sediment transport index, stream power index, topographic wetness index, distance to roads and streams, land use, lithology, soil and normalised difference vegetation index. The classifier attribute evaluation method based on the FLDA model was implemented to test the predictive competence of the mentioned contributing factors. The area under curve, confidence interval at 95%, standard error, Friedman test and Wilcoxon signed-rank test were used to compare and validate the success and prediction competence of the three applied models. According to the achived results, the BFLDA model showed the most prediction competence, followed by the RFLDA and FLDA models, respectively. The resulting groundwater spring potential maps can be used for groundwater development plans and land use planning.

Novel hybrid integration approach of bagging-based Fisher's linear discriminant function for groundwater potential analysis

Wei Chen[a], Biswajeet Pradhan[b,c,*], Shaojun Li[d], Himan Shahabi[e], Enke Hou[a], Shengquan Wang[a,f], Hossein Mojaddadi Rizeei[b]

[a]College of Geology & Environment, Xi'an University of Science and Technology, Xi'an, Shaanxi, 710054, China

[b]Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney, NSW 2007, Australia

[c]Department of Energy and Mineral Resources Engineering, Choongmu-gwan, Sejong University, 209 Neungdong-ro Gwangjin-gu, Seoul, 05006, Republic of Korea.

[d]State Key Laboratory of Geomechanics and Geotechnical Engineering, Institute of Rock and Soil Mechanics, Chinese Academy of Sciences, Wuhan, Hubei 430071, China

[e]Department of Geomorphology, Faculty of Natural Resources, University of Kurdistan, Sanandaj, Iran

[f]Key Laboratory of Coal Resources Exploration and Comprehensive Utilization， Ministry of Land and Resources

*Corresponding author. E-mail: Biswajeet.Pradhan@uts.edu.au, biswajeet24@gmail.com

Abstract

Groundwater is a vital water source in the rural and urban areas of developing and developed nations. In this study, a novel hybrid integration approach of Fisher's linear discriminant function (FLDA) with rotation forest (RFLDA) and bagging (BFLDA) ensembles was used for groundwater potential assessment at the Ningtiaota area in Shaanxi, China. A spatial database with 66 groundwater spring locations and 14 groundwater spring contributing factors was prepared; these factors were elevation, aspect, slope, plan and profile curvatures, sediment transport index, stream power index, topographic wetness index, distance to roads and streams, land use, lithology, soil and normalised difference vegetation index. The classifier attribute evaluation method based on the FLDA model was implemented to test the predictive competence of the mentioned contributing factors. The area under curve, confidence interval at 95%, standard error, Friedman test and Wilcoxon signed-rank test were used to compare and validate the success and prediction competence of the three applied models. According to the achived results, the BFLDA model showed the most prediction competence, followed by the RFLDA and FLDA models, respectively. The resulting groundwater spring potential maps can be used for groundwater development plans and land use planning.

Keywords: Groundwater; Machine learning; Fisher's linear discriminant function (FLDA); Rotation forest (RF); GIS

1. Introduction

Groundwater is a vital water source in rural and urban areas of developing and developed nations with various climate situations (Bera and Bandyopadhyay 2012; Waikar and Nilawar 2014). In recent years, the demand for high-quality water has increased due to the growing need in drinking, industrial, agricultural and domestic activities. Furthermore, groundwater has a low level of pollution and wide distribution, thereby attracting a large human population worldwide (Arkoprovo et al. 2012).

In the arid and semiarid regions of northwestern China, groundwater resource is significant for human lives, agriculture and industry; in some regions, groundwater is the single available water source (Yang et al. 2016). However, groundwater constrains the fragile eco-environment in this region.

The study of groundwater potential zones has received considerable attention for the implementation of an effective groundwater establishment, protection and management strategy due to the increasing demand for fresh drinking groundwater. Therefore, the assessment of groundwater potential zones is essential (e.g. measuring spring recharge) to manage groundwater quality and usage (Zabihi et al. 2016). Recently, several studies based on multitemporal datasets in groundwater spring potential mapping have been conducted by using geographic information system (GIS) tools and remote sensing datasets (U. Kumar et al. 2013; B. Kumar and Kumar 2010; Ambrish Kumar et al. 2011; Thilagavathi et al. 2015; Jha et al. 2009; Ozdemir 2011b; Elbeih 2015; Javed and Wani 2009; T. Kumar et al. 2014; Zabihi et al. 2016; Israil et al. 2006; Meijerink 1996; Gupta and Srivastava 2010; Manap et al. 2014; Rahmati et al. 2015; Naghibi et al. 2016). GIS is a powerful and useful tool, and it provides an easy approach not only for spatial data management and information analysis but also for the decision-making process in the natural sciences, such as geology and environmental management (Fedra 1993; Shahabi et al. 2014).

GIS spatial models, including statistical and bivariate algorithms, have been proposed in groundwater studies, such as frequency ratio (Ozdemir 2011a; Manap et al. 2014; Naghibi et al. 2015), analytic hierarchy process (Jandric and Srdjevic 2000; Sener and Davraz 2013; Kaliraj et al. 2014), logistic regression (Teso et al. 1996; Mair and El-Kadi 2013) and weights-of-evidence (Masetti et al. 2007; Lee et al. 2012; Ozdemir and Altural 2013; Uhan et al. 2011; Pourtaghi and Pourghasemi 2014). Recently, data mining methods, such as fuzzy logic (Nobre et al. 2007; Gemitzi et al. 2006), neurofuzzy (Dixon 2005; Safavi et al. 2013), artificial neural network (Corsini et al. 2009), decision trees (Duan et al. 2016; Lee and Lee 2015), random forest (Naghibi et al. 2016; Rahmati et al. 2016) and naive Bayesian (NB) (Aguilera et al. 2013), have been explored for groundwater spring potential mapping.

Newly, machine learning hybrid techniques and ensembles have been found to be superior to conventional techniques in various applications. An example is bagging ensemble, which can improve the prediction accuracy of a base classifier. However, the use of these techniques for groundwater potential mapping has rarely been investigated.

A literature review revealed that, although some ensemble methods focused on natural hazards, such as landslides (Tien Bui et al. 2017; Chen et al. 2017a; Althuwaynee et al. 2014; Hong et al. 2018a) and floods (Razavi Termeh et al. 2018; Tehrany et al. 2015; Tien Bui et al. 2016a; Hong et al. 2018b), few studies use machine learning ensembles in groundwater spring potential assessment. The present study aims to fill this research gap by developing a novel hybrid intelligence approach based on Fisher's linear discriminant function (FLDA) with rotation forest (RFLDA) and bagging (BFLDA) ensembles for groundwater spring potential mapping via a case study at the Ningtiaota area in Shaanxi Province, China. RFLDA and BFLDA have not been explored in groundwater spring potential mapping. Weka, ArcGIS and ENVI software are used in data analysis, model development and groundwater spring potential mapping.

2. Background of the methods used

2.1 Fisher's linear discriminant function

FLDA is one of the widespread feature recognition method in various fields (Agarwal and Chen 2010). Theoretically, m recognised classes operate the method. $X_j^{(i)}$ specifies the j-th training trial in class i. X^i denotes the average of training trials in the class i, whereas X signifies the mean of total training trials. With regard to the assumed training trials, M_b and M_w, which are scatter matrices of between-class and with-class, respectively, can be calculated as

$$M_b = \frac{1}{N}\sum_{i=1}^{m} N_i (X^i - X)(X^i - X)^T \tag{1}$$

and $$M_w = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{Ni} (X_j^{(i)} - X^i)(X_j^{(i)} - X^i)^T, \tag{2}$$

where N_i defines the number of training trials in class $i(\sum_{i=1}^{m} S_i = S)$, and N illustrates the total number of the assumed training trials.

FLDA aims to obtain a set of ideal distinguishing vectors to compose a transform of $\Phi_d = [\varphi_1, \varphi_2, ...\varphi_d]$

by maximising the Fisher criterion, indicated as (Moghaddam et al. 2007)

$$J(\Phi) \overset{def}{=} \frac{tr(\Phi_d^T M_b \Phi_d)}{tr(\Phi_d^T M_w \Phi_d)} \quad , \tag{3}$$

where T is the matrix transpose. Considering that the ideal discriminating vectors are empowered by benchmark maximisation, a vector for one input instance can be extracted, and the resulting vector can be used to classify the conforming instance (Yin et al. 2006).

2.2 Rotation forest

RF is an ensemble method that was initially proposed for classification (Rodríguez et al. 2013) and is built with independent decision trees (Ozcift and Gulten 2011). In the RF, individual tree is trained with a comprehensive dataset associated with a rotated feature space.

Here, $S = (s_1, s_2, ...s_n)$ is the vector of spring conditioning factors; $Y = (y_1, y_2)$ is the vector of spring and nonspring classes; D designates the training data; $F_1, F_2, ...F_n$ are classifiers in the ensemble structure; and T denotes a set of spring contributing parameters. T can be divided into several k subsets. The number of the contributing parameters for a subset can be calculated as $T = n / k$. For the F_i classier, T_{ij} should be the $j\text{-}th$ and $j = 1, 2, ...k$ subset of contributing parameters. E_{ij} shows the spring contributing parameters in T_{ij} from E. Basically, $E_{ij}^{'}$ is nominated randomly from E_{ij} by using the bootstrap method. At that moment, $E_{ij}^{'}$ should be transformed to achieve the constants of $ri_1^{(1)}, ri_1^{(2)}, ...ri_1^{(T1)}$, where the $r_{i,1}^{'}$ size is $T \times 1$. Ensemble RF is then created in respect to the rotation matrix that was produced by the basic classifier and transformation method (Xia et al. 2014). R_i^{α} is the rotation matrix, which is obtained by reorganising the matrix of R_i, which can be defined by Equation 4

$$R_i = \begin{bmatrix} ri_1^{(1)}, ri_1^{(2)}, \ldots ri_1^{(S1)} & 0 & \cdots & 0 \\ 0 & ri_1^{(1)}, ri_1^{(2)}, \ldots ri_1^{(S2)} & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & ri_1^{(1)}, ri_1^{(2)}, \ldots ri_1^{(Sk)} \end{bmatrix}. \quad (4)$$

Subsequently, one sparse rotation matrix called R_i is organised by the obtained coefficients that were designed for each individual class by using the average combination technique as follows:

$$\xi_k(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n} \alpha_{i,k}(\varepsilon R_i^{\alpha}), k = 1, 2, \ldots c, \quad (5)$$

where $\alpha_{i,k}(\varepsilon R_i^{\alpha})$ illustrates the probability produced by the C_i classifier to the hypothesis in which ε fits the k class. Lastly, ε is allocated to the largest confidence class.

Bagging

Bagging, which is an acronym for 'bootstrap aggregating', is created by training discrete classifiers on independent bootstrap instances that are generated with replacement from training instances (Breiman 1996a). The bagging hybrid ensemble technique was developed by Hothorn and Lausen (2005) and was referred to as 'bundling'. This approach adds the results of classifiers to the original feature for the bagging of classification trees.

Let a learning set of T comprise data $T\{(x_n, y_n), n = 1, 2, \ldots N\}$, where y is the class labels. In this study, x_n is the spring contributing parameters, and y_n is the springs and nonsprings. Here, assume that a producer for training the learning set from a predictor $h(x,T)$ is available. Then, assume that an order of learning sets $\{T_i\}$ exists, each containing of N is independent observation from the identical original distribution as T. The purpose is to practise $\{T_i\}$ to obtain a superior predictor than the single learning set of predictors $\{h(x,T)\}$.

The bagging-type hybrid ensemble technique enhances the results of each set of classifier by adding them to the original feature system for bagging the categorization procedure. Approximately one-third of the examples in the initial training system is not involved in all bootstrap trials. Breiman (1996) referred to these examples as out-of-bag samples.

2.3 Conditioning factor selection based on the FLDA method

The study of spatial relationships among groundwater spring conditioning factors is essential. However, the relationships among the spring conditioning factors have not been verified either statistically or quantitatively (Lee et al. 2017). Factors with no or negative contribution on modelling results should be eliminated to increase of model performance (Chen et al. 2018b). In this study, the classifier attribute evaluation method based on the FLDA model was used to analyse the prediction ability of contributing parameters during modelling (Witten et al. 2011).

2.4 Performance evaluation and comparison of models

2.4.1 Receiver operating characteristic curve (ROC)

ROC is the sensitivity as a function of 1-specificity (Chen et al. 2018a; Hong et al. 2018b). It plots the 1-specificity on the x axis versus the sensitivity on the y axis (Pourtaghi and Pourghasemi 2014; Al-Abadi 2015). This process considers the standard method for validating the overall performance of predicting model (Pham et al. 2017). The area under the ROC curve is one of the quantitative representation for the quality of a model; a high value of area under the receiver curve (AURC) (i.e. the maximum value of AURC is one that specifies a perfect model) shows high accuracy of the applied model (Chen et al. 2018c).

2.4.2 Friedman test

The Friedman test is a statistical test that established by Milton Friedman (Friedman 1939, 1937). This technique includes ranking of each row together and then assigning the rank's values to columns (Khosravi et al. 2018). The null hypothesis for the Friedman test is that no differences exist among the groundwater spring potential models. If the chi-square is larger than the standard value of 3.841 and the *P* value is smaller than the selected significance level (i.e. $\alpha = 0.05$), the null hypothesis would be rejected (Khosravi et al. 2018).

2.5.3 Wilcoxon signed-rank test

The Friedman test can only show if significant differences exist among the three groundwater spring potential models. Basically, this test cannot provide pairwise comparisons among the three models (Tien Bui et al. 2016b). Therefore, the Wilcoxon signed-rank test was used. The null hypothesis is that no significant difference exists among groundwater spring potential models at the significance level of $\alpha = 0.05$ (Tien Bui et al. 2016b). The z and *P* values are two statistics for this method. When z value exceed the range values (i.e. -1.96 to +1.96) and the *P* value is smaller than the significance stage ($\alpha = 0.05$), the null hypothesis would be dismissed (Tien Bui et al. 2016b; Chen et al. 2017a; Chen et al. 2017c).

3. Study area and data preparation

The Ningtiaota area of Shaanxi Province in China was selected to evaluate groundwater spring potential (Fig. 1). This territory was considered suitable because it is representative of the geomorphological, environmental and geological settings of groundwater spring processes, and the area is a part of the transition zone of the Aeolian landform and is a loess hilly region. Most of its surface area outcrops sand, loess, laterite and bedrock (Fig. 2). It covers approximately 119.77 km^2 and has a mean annual rainfall of roughly 434.1 mm. Elevation ranges from 1,118 m to 1,364 m above sea level, and land use types include farm, forest and grass lands, water body, residential and others, such as sand and bare lands. Hydrologically, the study area is situated in the Kuye River Basin, a tributary of the Yellow river. The river systems in the area include Miaogou, Kaokaowusugou and Lucaogou from north to south. The surface water in this area is greatly affected by the seasons. Generally, the rainy season occurs from March and July to September; the dry season is marked by alternating winter and spring. The general situation of the main rivers is as follows:

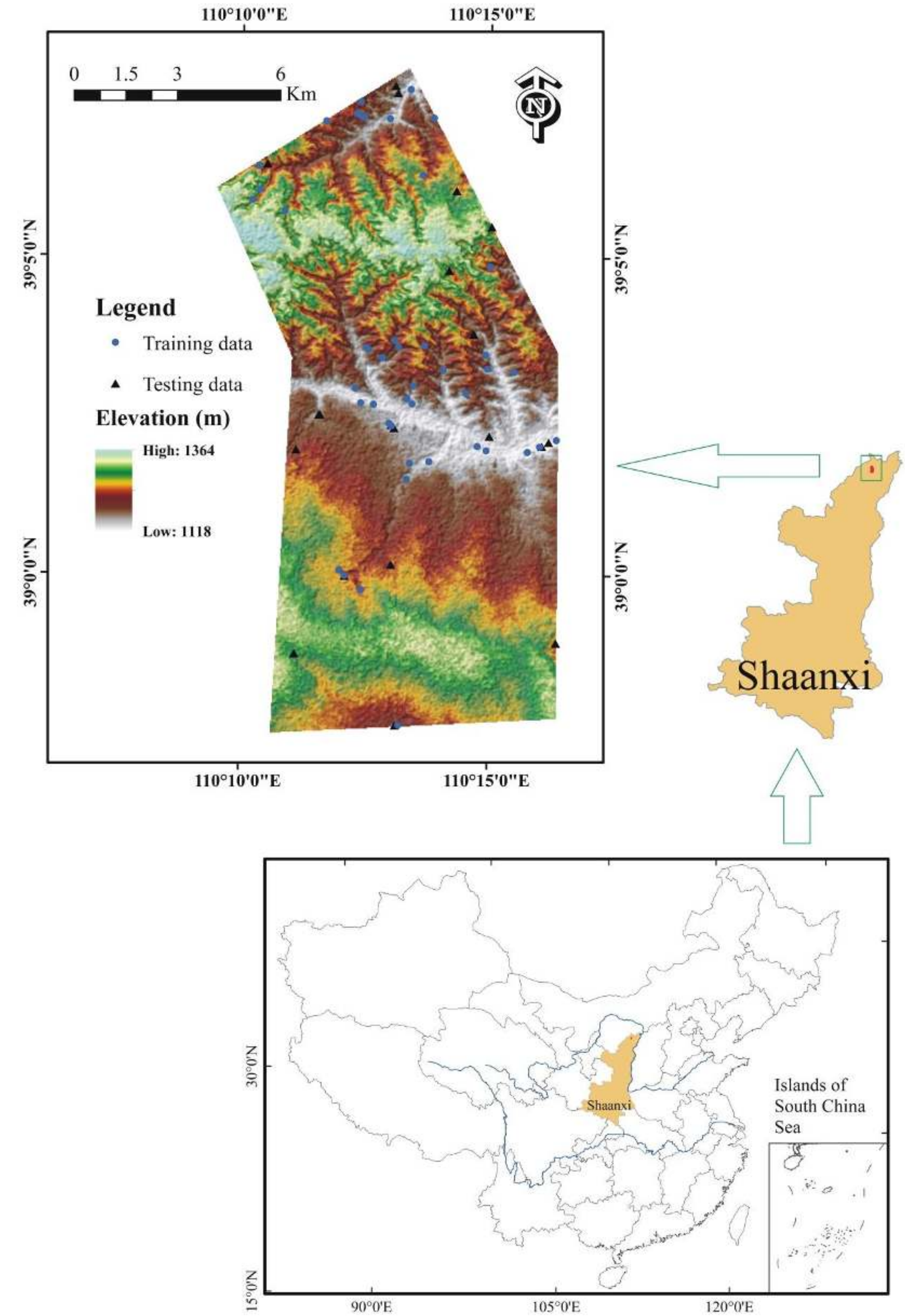


Fig. 1 Location of the study area and spring inventory

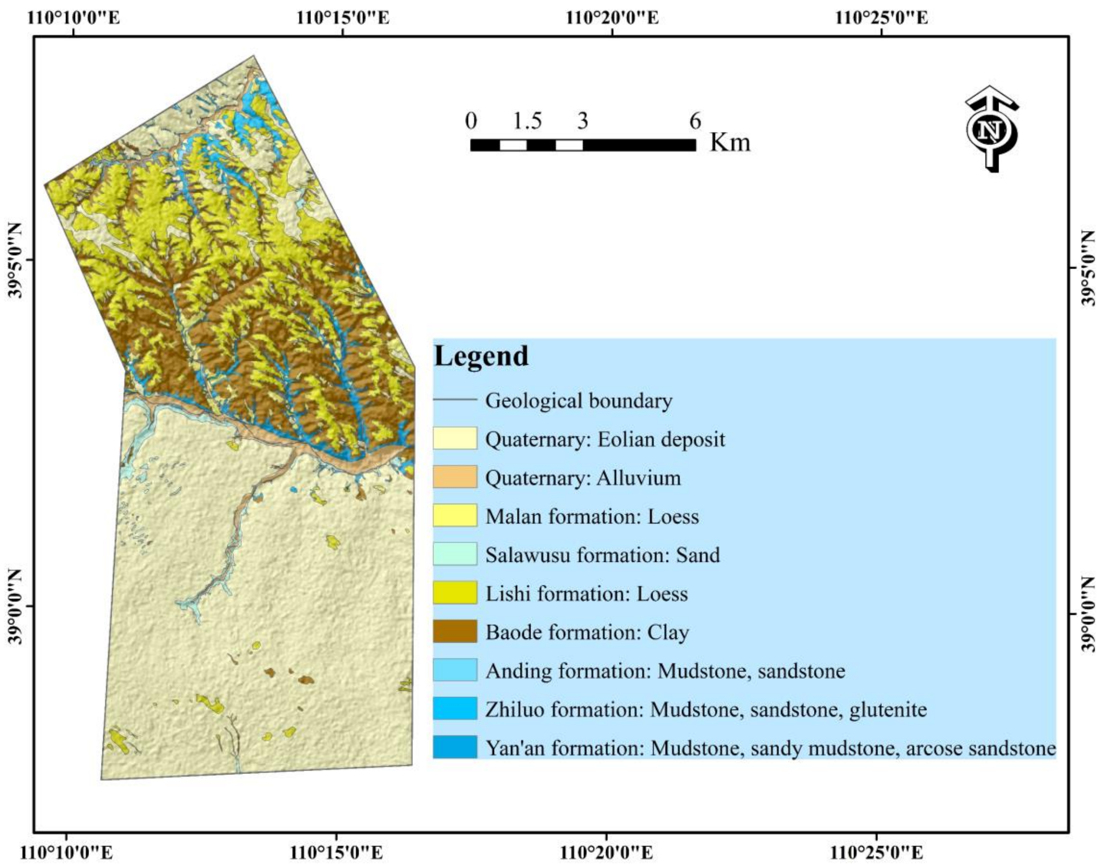


Fig. 2 Geological map of the study area

Miaogou is the second tributary of the Kuye River, which originates from Zhong'aobao in the western part of the study area and flows northeastward through the northern border of the study area. The river has a perennial flow, and the flow length within the research area is approximately 8 km.

Kaokaowusugou flows from west to east through the central part of the entire study area. The flow length is 41.9 km, with a watershed area of 259.5 km^2. According to the observations at Shaqu and Liujiashipan, the average flow discharge over the years was 0.7491 m^3/s, the maximum flow discharge was 26.0113 m^3/s and the minimum flow discharge was 0.101 m^3/s.

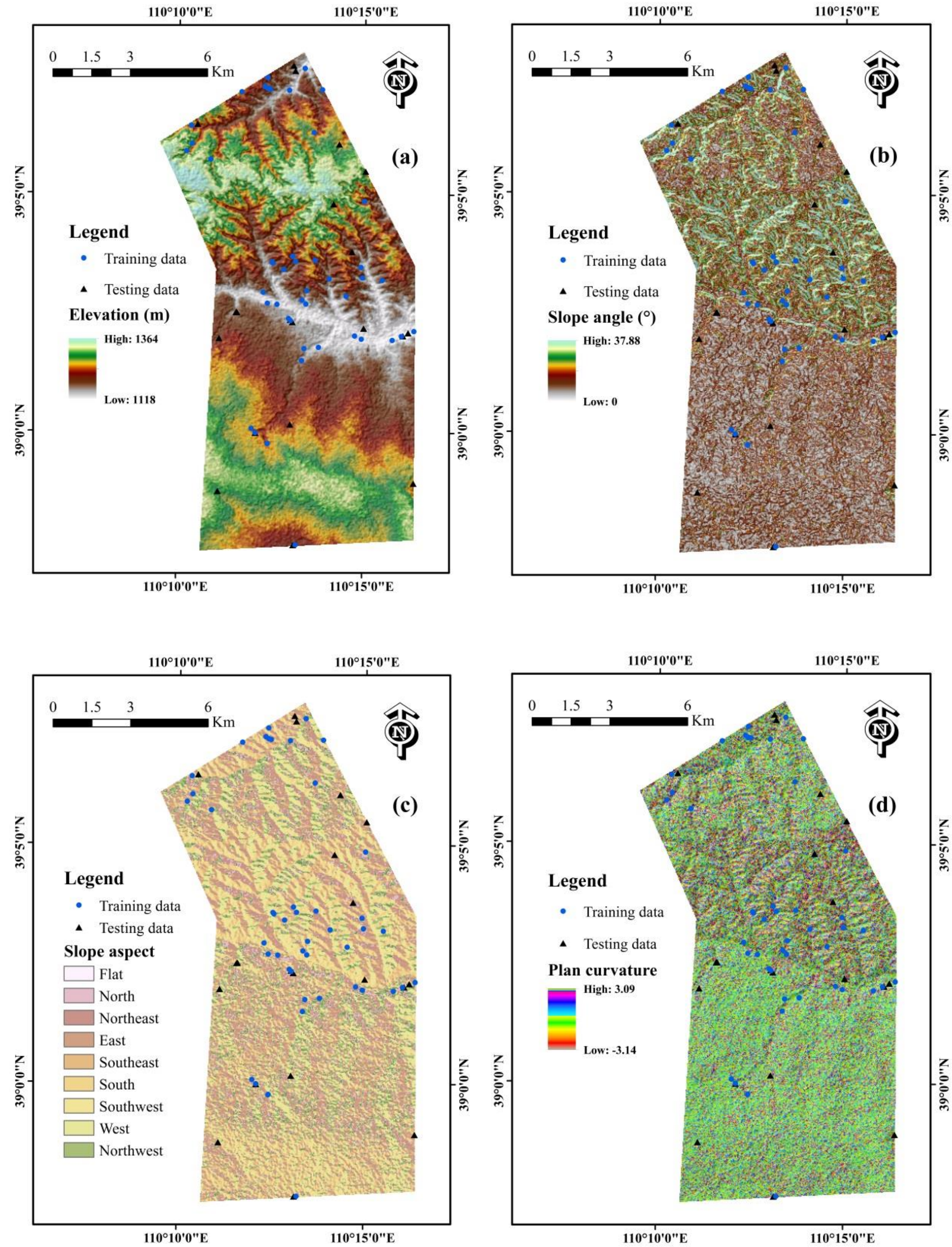
Lucaogou originates from the Qibushu in the southern part of the region and flows into Majiatagou from northwest to southeast. The flow length within the research area was 1.6 km, and the flow discharge was 13.34 L/s.

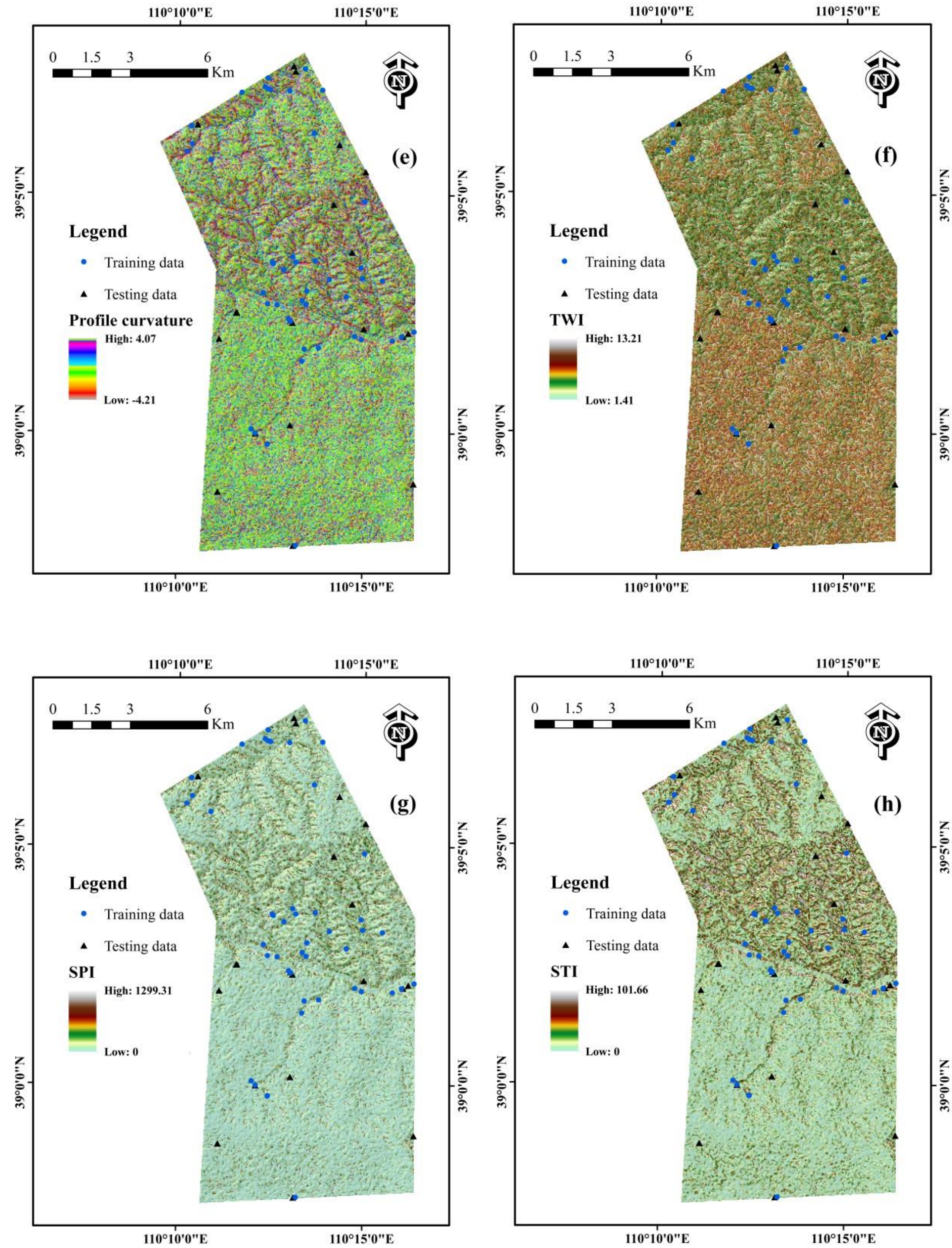
Numerous types of data about groundwater springs in the Ningtiaota area were collected from earlier reports and field surveys, including the locations of springs, types and yield. Groundwater springs (66) were divided into two datasets randomly (Fig. 1). The initial dataset included 70% of the groundwater spring locations for model training, whereas the other part included 30% that used for testing assessment.

Table 1 Contributing parameters in groundwater spring potential assessment

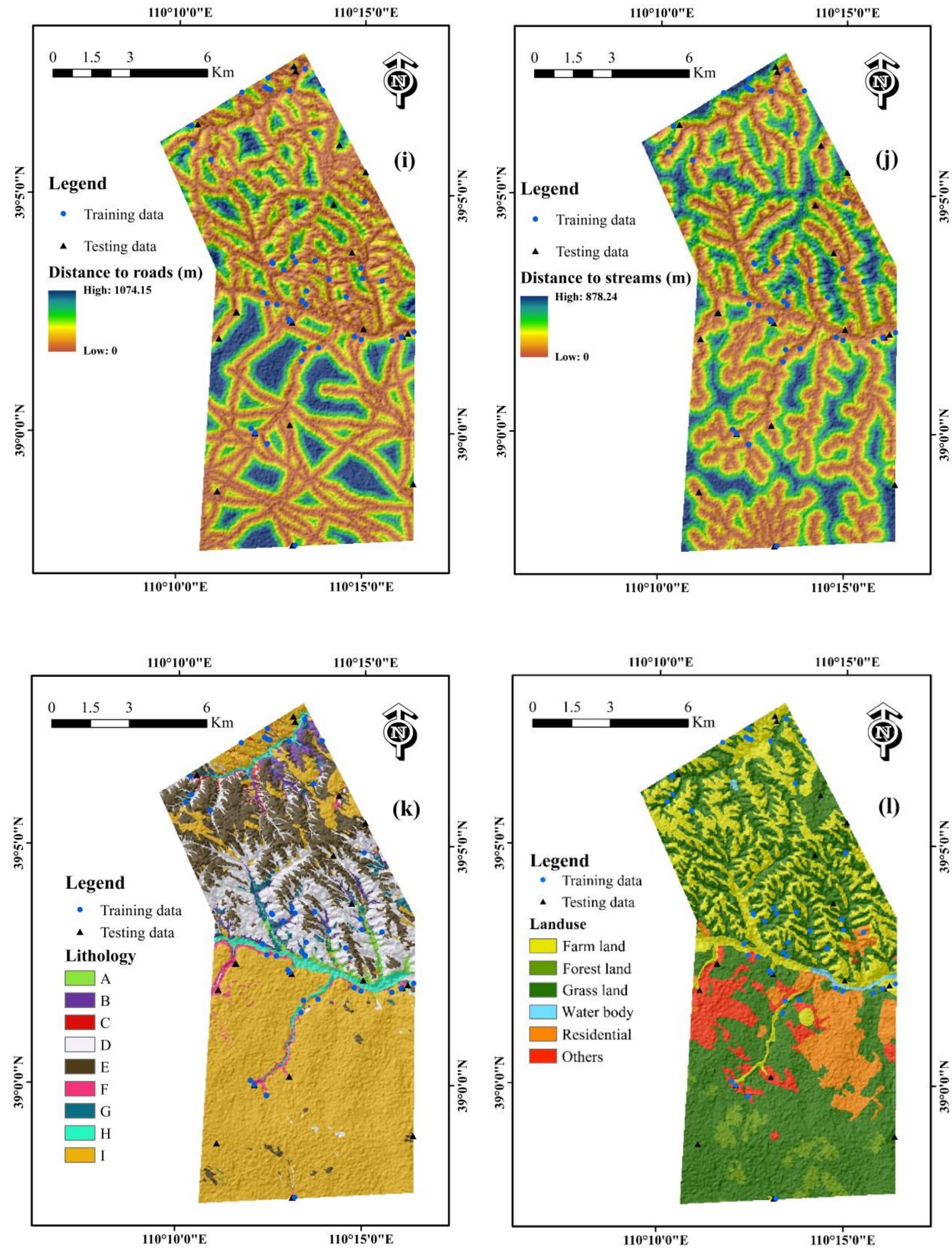
Groups	Conditioning factors	Raster type
Topographical factors	Elevation	Continuous
	Slope angle	Continuous
	Aspect	Categorical (9 class)
	Profile curvature	Continuous
	Plan curvature	Continuous
	TWI	Continuous
	SPI	Continuous
	STI	Continuous
Geological factor	Lithology	Categorical (9 class)
Environmental factors	Distance to roads	Continuous
	Distance to streams	Continuous
	Land use	Categorical (6 class)
	Soil	Categorical (4 class): *Calcari-Gypsiric Arenosols (Arc) Haplic Arenosols (ARh) Calcareous Red Clay (CMe) Luvi-Calcic Kastanozems (KSk)*
	NDVI	Continuous

Groundwater spring is affected by several topographical, geological and environmental factors. Selecting the most suitable parameters depends on the geo-environmental specification of the study area. In this study, 14 conditioning factors were selected with regard to groundwater spring potential for the modelling (Fig. 3). These conditioning factors were obtained from the compilation of ASTER GDEM with a resolution of 30 m, a geological map with a 1:10,000 scale, Landsat 8 OLI images with a resolution of 30 m and soil maps using ArcGIS and ENVI software. The conditioning factors used in groundwater spring potential assessment listed in Table 1. All the thematic maps of groundwater spring conditioning factors were discretised to be compatible with the 30 m resolution of the digital elevation model.





110°10'0"E
110°15'0"E
39°5'0"N
39°0'0"N
0 1.5 3 6 Km
N
(e)
Legend
Training data
Testing data
Profile curvature
High: 4.07
Low: -4.21
(f)
Legend
Training data
Testing data
TWI
High: 13.21
Low: 1.41
(g)
Legend
Training data
Testing data
SPI
High: 1299.31
Low: 0
(h)
Legend
Training data
Testing data
STI
High: 101.66
Low: 0


110°10'0"E
110°15'0"E
39°5'0"N
39°0'0"N
0 1.5 3 6 Km
N
(i)
Legend
Training data
Testing data
Distance to roads (m)
High: 1074.15
Low: 0
(j)
Legend
Training data
Testing data
Distance to streams (m)
High: 878.24
Low: 0
(k)
Legend
Training data
Testing data
Lithology
A
B
C
D
E
F
G
H
I
(l)
Legend
Training data
Testing data
Landuse
Farm land
Forest land
Grass land
Water body
Residential
Others

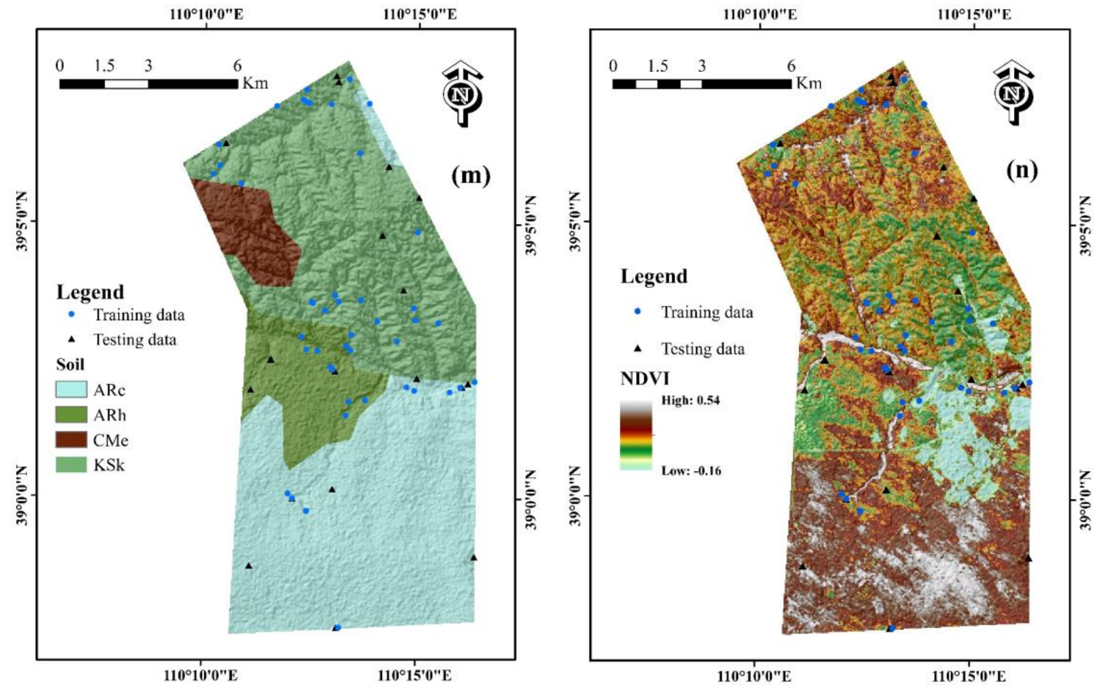


Fig. 3 Thematic maps of spring contributing parameters: (a) Elevation; (b) Slope angle; (c) Aspect; (d) Plan curvature; (e) Profile curvature; (f) TWI; (g) SPI; (h) STI; (i) Distance to roads; (j) Distance to streams; (k) Lithology; (l) Land use; (m) Soil; (n) NDVI.

4. Application of hybrid integration approaches

The procedure of this study has five main steps (Fig. 4): (i) preparation of groundwater spring locations and groundwater spring conditioning factors; (ii) selection and analysis of groundwater spring conditioning factors; (iii) groundwater spring potential modelling using FLDA, RFLDA and BFLDA models; (iv) generation of groundwater spring potential maps; and (v) model validation and comparison.

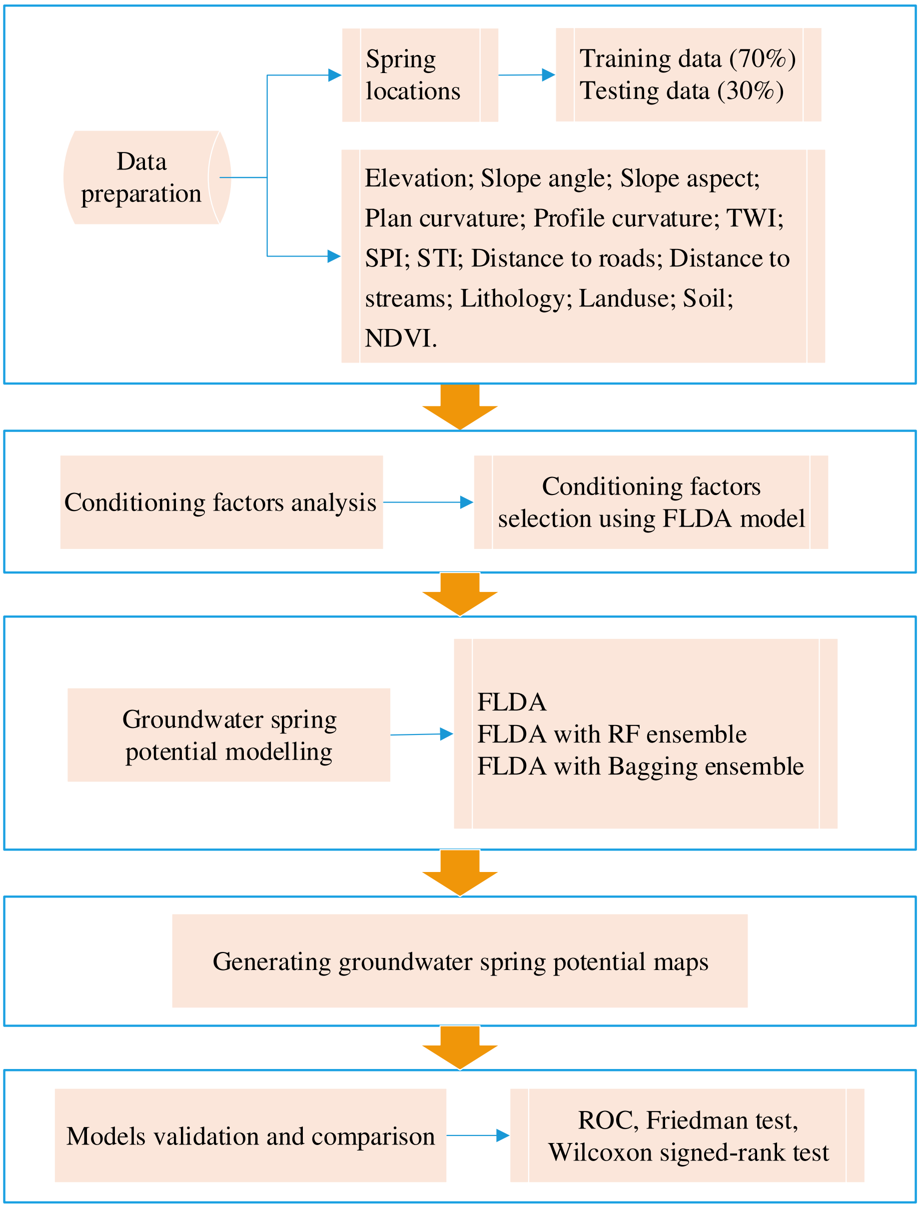


Fig. 4 Schematic of the study procedure

5. Results and discussions

5.1 Selection of conditioning factors

The classifier attribute evaluation technique based on the FLDA model with the average merit (AM) and its standard deviation was utilised using 10-fold cross validation system. The most effective parameters have higher AM values (Chen et al. 2017a). Results of feature selection indicate that lithology (0.562) is the most significant parameter for groundwater spring potential modelling, followed by elevation (0.503), distance to roads (0.481), distance to streams (0.440), SPI (0.433), STI (0.419), soil (0.416), aspect (0.395), slope (0.379), TWI (0.335), profile curvature (0.297), plan curvature (0.286), NDVI (0.200) and land use (0.092). Therefore, all 14 groundwater spring contributing parameters have positive contributions to the model and were incorporated in the training and testing datasets for further analysis (Table 2).

Table 2 Predictive capabilities of spring contributing parameters using the FLDA method

Number	Conditioning factors	Average merit	Standard deviation
1	Lithology	0.562	± 0.008
2	Elevation	0.503	± 0.013
3	Distance to roads	0.481	± 0.010
4	Distance to streams	0.440	± 0.017
5	SPI	0.433	± 0.025
6	STI	0.419	± 0.012
7	Soil	0.416	± 0.010
8	Aspect	0.395	± 0.014
9	Slope	0.379	± 0.019
10	TWI	0.335	± 0.009
11	Profile curvature	0.297	± 0.009
12	Plan curvature	0.286	± 0.007
13	NDVI	0.200	± 0.012
14	Land use	0.092	± 0.007

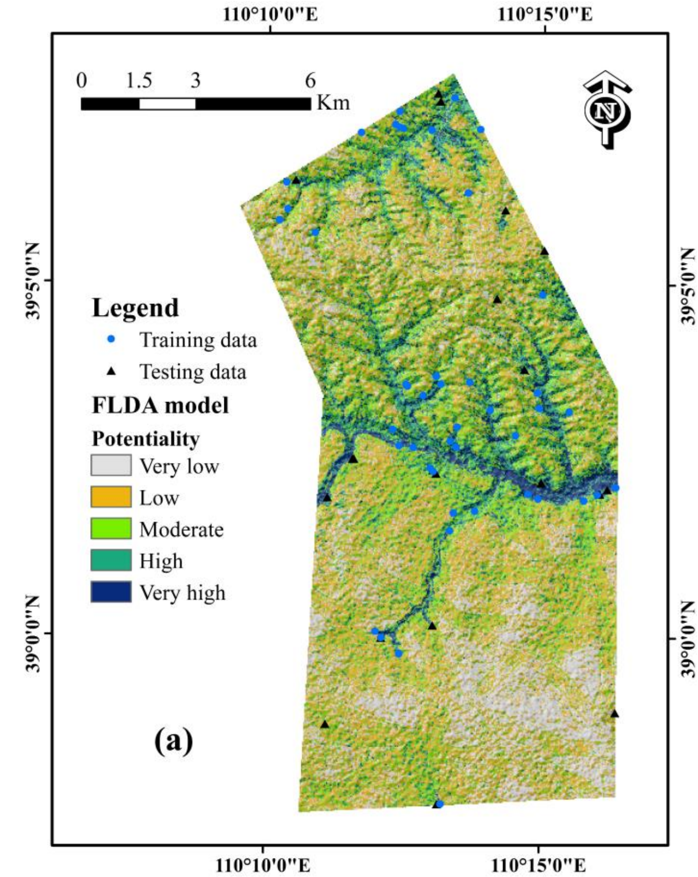
5.2 Model construction

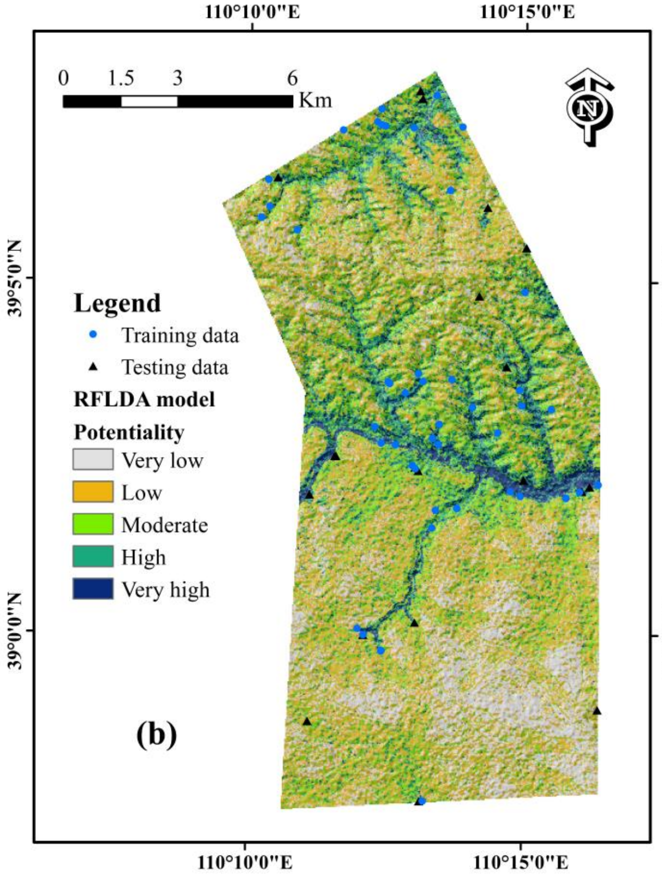
In groundwater spring potential assessment, the dependent factors is considered a binary variable (spring and nonspring). Consequently, the spring and nonspring sample points are essential for groundwater spring potential mapping. The testing and training datasets contained equal numbers of spring and nonspring. Thus, the equal number of nonspring points was designated randomly from groundwater spring-free locations and was randomly divided into 70% and 30% for training and testing, respectively. The generating and splitting process, which is a randomisation approach, was repeated 30 times. Then, the predictive competencies of individual applied for the three models were assessed by using the AURC technique to discover the ideal combination of spring and nonspring samples. A 10-fold cross-validation method, in which parameters are selected in the training dataset to avoid the overfitting problem and to decrease variability, was employed for all tests to obtain an unbiased estimate of AURC values (Chen et al. 2017b; Jiang and Chen 2016; Alkhasawneh et al. 2014).

With the use of the training dataset, three models were constructed for groundwater spring potential assessment, and certain parameters were determined to obtain high prediction accuracy. The FLDA model used default ridge values (the ridge penalty factor for the output layer) of 1.0E-6. The RFLDA model used 10 for the number of iterations, 5 for seeds, 1 for the number of execution slots (threads) to use for the construction of the ensemble, principal components for projection filter and 50 for the percentage of instances to be removed. The BFLDA model used 10 for the number of iterations, 6 for seeds and 1 for the number of execution slots (threads) for the construction of the ensemble.

5.3 Generation of groundwater spring potential maps

After the construction of the three models, the built groundwater potential models were validated by using the testing dataset and then applied through the entire area to create groundwater spring potential maps. The calculated groundwater spring potential indices for the whole study area by using the three models ranged from 0.000 to 1. Subsequently, all calculated groundwater spring potential indices were applied to prepare the groundwater spring potential maps by using the ArcGIS software. Finally, these groundwater spring potential maps were reclassified into five different intervals by using the Jenks natural breaks classification process, which is one of the most popular classification methods for creating classification maps (Fig. 5) (Naghibi et al. 2017; Akshay Kumar and Krishna 2018). The area percentages are 20.47%, 35.60%, 25.58%, 13.65% and 4.70%, which denote very low, low, moderate, high and very high classes with the FLDA model, respectively. For the RFLDA model, the area percentages are 23.05%, 35.45%, 24.40%, 12.77% and 4.33%, whereas the area percentages are 26.22%, 35.65%, 24.01%, 10.65% and 3.47% for the BFLDA model (Fig. 6).




110°10'0"E
110°15'0"E
0 1.5 3 6
Km
N
39°5'0"N
39°0'0"N
Legend
Training data
Testing data
RFLDA model
Potentiality
Very low
Low
Moderate
High
Very high
(b)

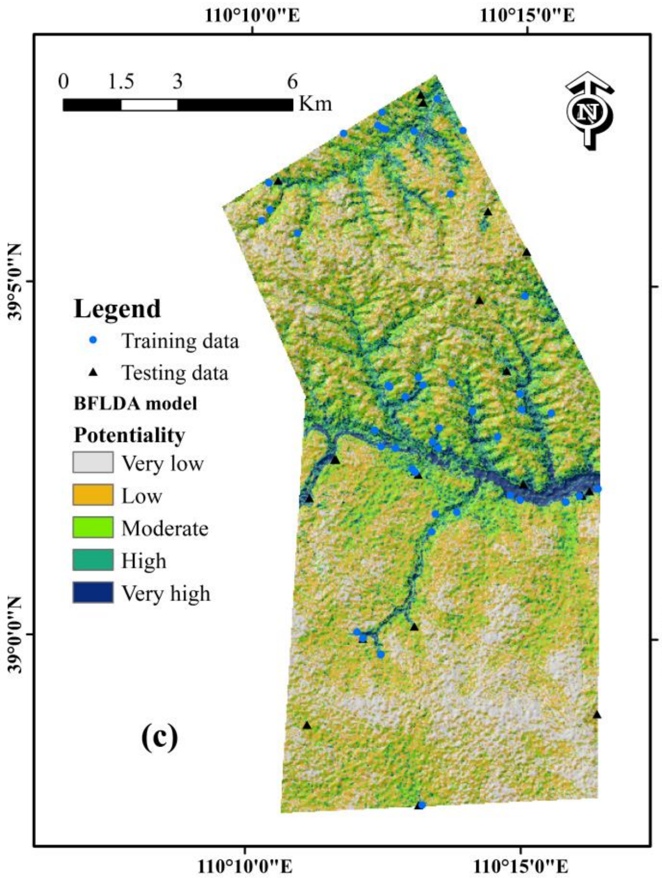
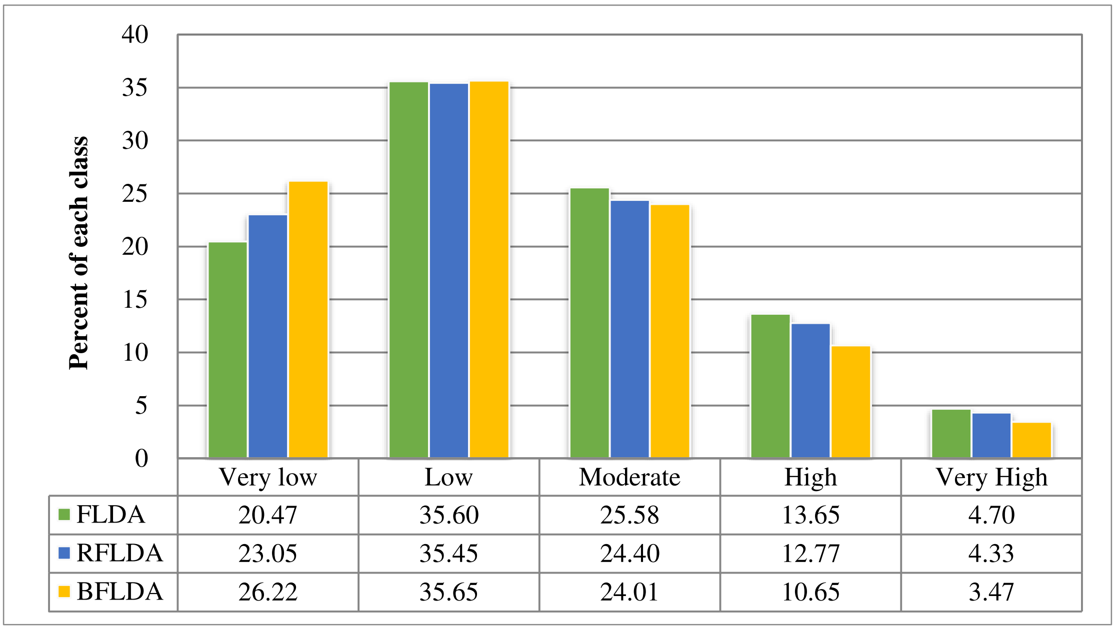

110°10'0"E
110°15'0"E
0 1.5 3 6
Km
N
39°5'0"N
39°0'0"N
Legend
Training data
Testing data
BFLDA model
Potentiality
Very low
Low
Moderate
High
Very high
(c)

Fig. 5 Groundwater spring potential maps: (a) FLDA, (b) RFLDA and (c) BFLDA models



	Very low	Low	Moderate	High	Very High
FLDA	20.47	35.60	25.58	13.65	4.70
RFLDA	23.05	35.45	24.40	12.77	4.33
BFLDA	26.22	35.65	24.01	10.65	3.47

Fig. 6 Area percentages of groundwater spring potential classes

5.4 Model validation

The predictive capability of the three models was evaluated using evaluation statistics, including AURC, confidence interval (CI) and standard error at 95%. Results of the success rate curve using the training dataset are shown in Fig. 7 and Table 3. The BFLDA model showed the best performance, with the top AURC value of 0.892, the lowest standard error of 0.025 and the finest CI of 0.843–0.941, whereas the RFLDA and FLDA models obtained slightly lower values for all the aforementioned criteria. Results of the prediction rate curve are shown in Fig. 8 and Table 4. The BFLDA model also showed the best performance, with the peak AURC value of 0.746, the least standard error of 0.067 and the finest CI of 0.614–0.877. In general, considering the training and testing datasets, all three models showed acceptable goodness-of-fit; however, the BFLDA model presented the best performance among all.

In addition, results of the Friedman test are presented in Tables 5 and 6, which illustrate the mean rank for the FLDA, RFLDA and BFLDA models are 1.49, 2.05 and 2.46. The *P* value and chi-square for this test are 0.000 and 87.254, which are far from the standard values of 3.841 and 0.05. Therefore, in this test, the primary statement was true, and the null hypothesis was rejected. Results of the Wilcoxon signed-rank test are shown in Table 7. The significant z and P values are far from the standard values (i.e. -1.96 and +1.96) and 0.05, individually. Therefore, all the three groundwater spring potential models are significantly different.

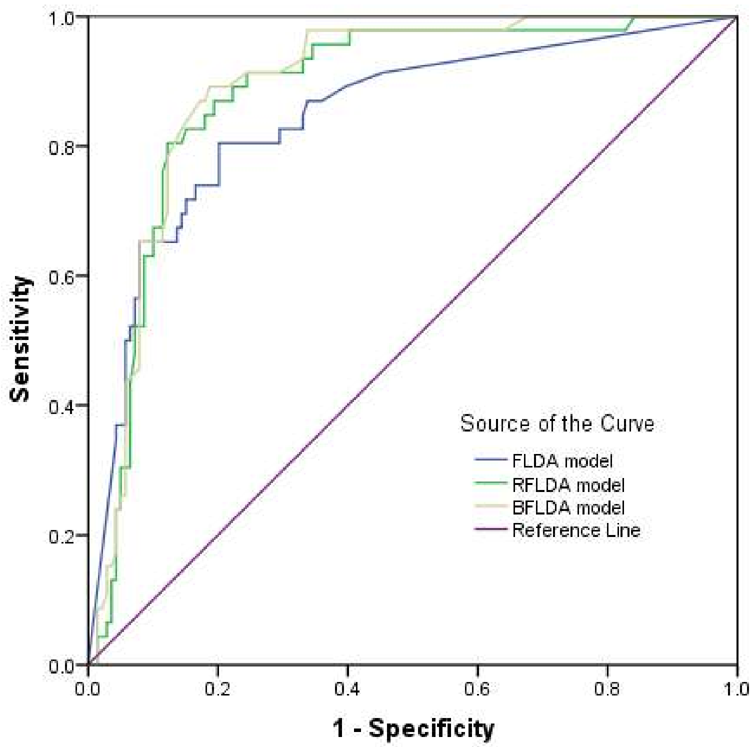


Fig. 7 the training ROC curves for three models

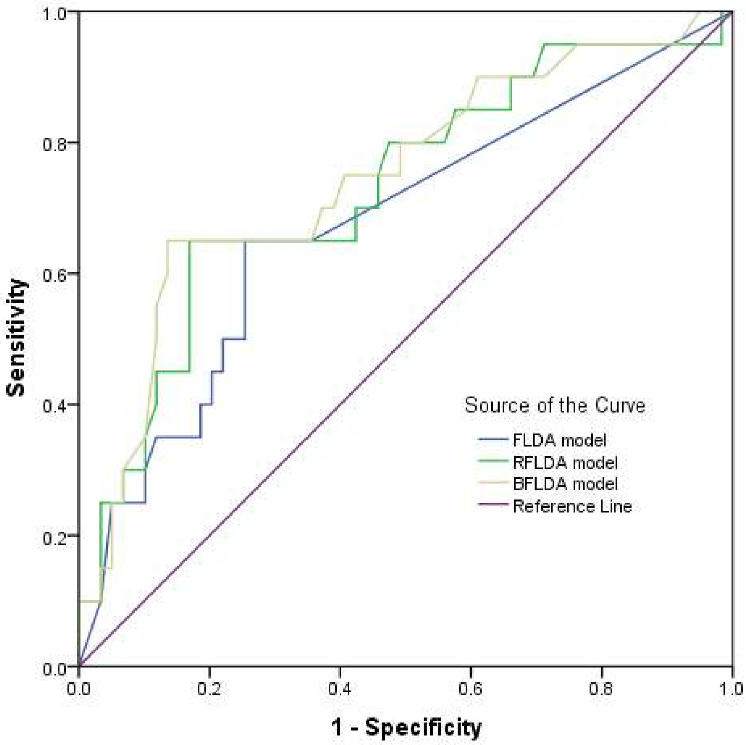


Fig. 8 the testing ROC curves for three models

Table 3 Parameters of ROC curves of training analysis

Test Result Variable	Area	Std. Error	95% Confidence interval	
			Lower bound	Upper bound
FLDA	0.845	0.035	0.776	0.914
RFLDA	0.882	0.028	0.828	0.936
BFLDA	0.892	0.025	0.843	0.941

Table 4 Parameters of ROC curves of testing analysis

Test Result Variable	Area	Std. Error	95% Confidence interval	
			Lower bound	Upper bound
FLDA	0.675	0.073	0.532	0.819
RFLDA	0.728	0.069	0.593	0.863
BFLDA	0.746	0.067	0.614	0.877

Table 5 Average ranking of the three models

Models	Mean Rank
FLDA	1.49
RFLDA	2.05
BFLDA	2.46

Table 6 Results of the Friedman test for the three models with α = 0.05

Chi-Square	87.254
df	2
P.	0.000

Table 7 Pairwise model comparison based on the Wilcoxon signed-rank test

Pairwise Comparison	Z value	P value	Significance
RFLDA vs. FLDA	-6.963	0.000	Yes
BFLDA vs. FLDA	-7.205	0.000	Yes
BFLDA vs. RFLDA	-5.033	0.000	Yes

6. Conclusions

In current study, a novel hybrid integration method of FLDA with RF and bagging ensembles was applied and evaluated for groundwater spring potential mapping at the Ningtiaota area in Shaanxi Province, China. Sixty-six groundwater springs and 14 groundwater spring contributing parameters were initially selected for this study; these 14 parameters were elevation, slope angle, aspect, plan curvature, profile curvature, TWI, SPI, STI, distance to roads, distance to streams, lithology, land use, soil and NDVI. The predictive capability of these contributing parameters was tested by using the classifier attribute evaluation method based on the FLDA model. All 14 groundwater spring conditioning factors were incorporated in the training and testing datasets for further analysis. The applied models were validated and compared using ROC, Std. Error, CI at 95% and the Friedman and Wilcoxon signed-rank tests. The BFLDA model, which has the highest AURC values, smallest Std. Error and narrowest CI, is considered a promising technique for groundwater spring potential mapping.

Acknowledgements

This research was financially supported by the International Partnership Program of the Chinese

Academy of Sciences (Grant No. 115242KYSB20170022), the National Natural Science Foundation of China (Grant No. 41807192, 41472234), the China Postdoctoral Science Foundation (Grant No. 2018T111084, 2017M613168), the Shanxi Province Postdoctoral Science Foundation (Grant No. 2017BSHYDZZ07) and the Open Fund of the Key Laboratory of Coal Resources Exploration and Comprehensive Utilization, Ministry of Land and Resources (Grant No. ZZ2016-1).

References

Abdulkareem, J. H., Pradhan, B., Sulaiman, W. N. A., & Jamil, N. R. (2018). Quantification of Runoff as Influenced by Morphometric Characteristics in a Rural Complex Catchment. Earth Systems and Environment, 2(1), 145-162.

Abdulkareem, J. H., Sulaiman, W. N. A., Pradhan, B., & Jamil, N. R. (2018). Long-Term Hydrologic Impact Assessment of Non-point Source Pollution Measured Through Land Use/Land Cover (LULC) Changes in a Tropical Complex Catchment. Earth Systems and Environment, 1-18.

Roy, D. K., & Datta, B. (2018). A Review of Surrogate Models and Their Ensembles to Develop Saltwater Intrusion Management Strategies in Coastal Aquifers. Earth Systems and Environment, 1-19.

Rojas, R., Commander, P., McFarlane, D., Ali, R., Dawes, W., Barron, O., ... & Charles, S. (2018). Groundwater Resource Assessment and Conceptualization in the Pilbara Region, Western Australia. Earth Systems and Environment, 1-21.

Agarwal, D., & Chen, B. C. fLDA (2010). matrix factorization through latent dirichlet allocation. In *International Conference on Web Search and Web Data Mining, WSDM 2010, New York, Ny, Usa, February, 2010* (pp. 91-100)

Aguilera, P. A., Fernández, A., Ropero, R. F., & Molina, L. (2013). Groundwater quality assessment using data clustering based on hybrid Bayesian networks. *Stochastic environmental research and risk assessment, 27*(2), 435-447.

Al-Abadi, A. M. (2015). Groundwater potential mapping at northeastern Wasit and Missan governorates, Iraq using a data-driven weights of evidence technique in framework of GIS. [journal article]. *Environmental Earth Sciences, 74*(2), 1109-1124, doi:10.1007/s12665-015-4097-0.

Alkhasawneh, M. S., Ngah, U. K., Tay, L. T., Mat Isa, N. A., & Al-Batah, M. S. (2014). Modeling and Testing Landslide Hazard Using Decision Tree. *Journal of Applied Mathematics,2014,(2014-2-4), 2014*(1), 568-575.

Althuwaynee, O. F., Pradhan, B., Park, H. J., & Lee, J. H. (2014). A novel ensemble decision tree-based CHi-squared Automatic Interaction Detection (CHAID) and multivariate logistic regression models in landslide susceptibility mapping. *Landslides, 11*(6), 1063-1078.

Arkoprovo, B., Adarsa, J., & Prakash, S. S. (2012). Delineation of groundwater potential zones using satellite remote sensing and geographic information system techniques: a case study from Ganjam district, Orissa, India. *Research Journal of Recent Sciences, 2277*, 2502.

Bera, K., & Bandyopadhyay, J. (2012). Ground Water Potential Mapping in Dulung Watershed using Remote Sensing & GIS techniques, West Bengal, India. *International Journal of Scientific and Research Publications, 2*(12), 1-7.

Breiman, L. (1996a). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (1996b). Out-Of-Bag Estimation.

Chen, W., Peng, J., Hong, H., Shahabi, H., Pradhan, B., Liu, J., et al. (2018a). Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. *Science Of the Total Environment, 626*, 1121-1135, doi:https://doi.org/10.1016/j.scitotenv.2018.01.124.

Chen, W., Shahabi, H., Shirzadi, A., Li, T., Guo, C., Hong, H., et al. (2018b). A Novel Ensemble Approach of Bivariate Statistical Based Logistic Model Tree Classifier for Landslide Susceptibility Assessment. *Geocarto International*, 1-32, doi:10.1080/10106049.2018.1425738.

Chen, W., Shirzadi, A., Shahabi, H., Ahmad, B. B., Zhang, S., Hong, H., et al. (2017a). A novel hybrid artificial intelligence approach based on the rotation forest ensemble and naïve Bayes tree classifiers for a landslide susceptibility assessment in Langao County, China. *Geomatics, Natural Hazards and Risk, 8*(2), 1955-1977, doi:10.1080/19475705.2017.1401560.

Chen, W., Xie, X., Peng, J., Shahabi, H., Hong, H., Tien Bui, D., et al. (2018c). GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method. *Catena, 164*, 135-149.

Chen, W., Xie, X., Peng, J., Wang, J., Duan, Z., & Hong, H. (2017b). GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models. *Geomatics, Natural Hazards and Risk, 8*(2), 950-973, doi:10.1080/19475705.2017.1289250.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Tien Bui, D., et al. (2017c). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena, 151*, 147-160.

Corsini, A., Cervi, F., & Ronchetti, F. (2009). Weight of evidence and artificial neural networks for potential groundwater spring mapping: an application to the Mt. Modino area (Northern Apennines, Italy). *Geomorphology, 111*(1–2), 79-87.

Dixon, B. (2005). Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: a GIS-based sensitivity analysis. *Journal of Hydrology, 309*(1), 17-38.

Duan, H., Deng, Z., Deng, F., & Wang, D. (2016). Assessment of Groundwater Potential Based on Multicriteria Decision Making Model and Decision Tree Algorithms. *Mathematical Problems in Engineering, 2016*.

Elbeih, S. F. (2015). An overview of integrated remote sensing and GIS for groundwater mapping in Egypt. *Ain Shams Engineering Journal, 6*(1), 1-15.

Fedra, K. (1993). GIS and environmental modeling. *Environmental modeling with GIS*, 35-50.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Publications of the American Statistical Association, 32*(200), 675-701.

Friedman, M. (1939). A correction: The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Publications of the American Statistical Association, 34*(205), 109.

Gemitzi, A., Petalas, C., Tsihrintzis, V. A., & Pisinaras, V. (2006). Assessment of groundwater vulnerability to pollution: a combination of GIS, fuzzy logic and decision making techniques. *Environmental Geology, 49*(5), 653-673.

Gupta, M., & Srivastava, P. K. (2010). Integrating GIS and remote sensing for identification of groundwater potential zones in the hilly terrain of Pavagarh, Gujarat, India. *Water International, 35*(2), 233-245.

Hong, H., Liu, J., Bui, D. T., Pradhan, B., Acharya, T. D., Pham, B. T., et al. (2018a). Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *Catena, 163*, 399-413, doi:https://doi.org/10.1016/j.catena.2018.01.005.

Hong, H., Panahi, M., Shirzadi, A., Ma, T., Liu, J., Zhu, A. X., et al. (2018b). Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. *Science Of the Total Environment, 621*, 1124-1141, doi:https://doi.org/10.1016/j.scitotenv.2017.10.114.

Hothorn, T., & Lausen, B. (2005). Bundling classifiers by bagging trees. *Computational Statistics & Data Analysis, 49*(4), 1068-1078.

Israil, M., Al-Hadithi, M., & Singhal, D. (2006). Application of a resistivity survey and geographical information system (GIS) analysis for hydrogeological zoning of a piedmont area, Himalayan foothill region, India. *Hydrogeology Journal, 14*(5), 753-759.

Jandric, Z., & Srdjevic, B. Analytic hierarchy process in selecting best groundwater pond. In *31st International Geological Congress, 2000* (pp. 1-9)

Javed, A., & Wani, M. H. (2009). Delineation of groundwater potential zones in Kakund watershed, Eastern Rajasthan, using remote sensing and GIS techniques. *Journal of the Geological Society of India, 73*(2), 229-236.

Jha, M. K., Bongane, G. M., Chowdary, V. M., Cluckie, I. D., Chen, Y., Babovic, V., et al. Groundwater potential zoning by remote sensing, GIS and MCDM techniques: a case study of eastern India. In *Symposium JS.4 at the IAHS & IAH Convention, 2009* (pp. 432-441)

Jiang, P., & Chen, J. (2016). Displacement prediction of landslide based on generalized regression neural networks with K -fold cross-validation. *Neurocomputing, 198*(198), 40-47.

Kaliraj, S., Chandrasekar, N., & Magesh, N. (2014). Identification of potential groundwater recharge zones in Vaigai upper basin, Tamil Nadu, using GIS-based analytical hierarchical process (AHP) technique. *Arabian Journal of Geosciences, 7*(4), 1385-1401.

Khosravi, K., Pham, B. T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., et al. (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Science Of the Total Environment, 627*, 744-755, doi:https://doi.org/10.1016/j.scitotenv.2018.01.266.

Kumar, A., & Krishna, A. P. (2018). Assessment of groundwater potential zones in coal mining impacted hard-rock terrain of India by integrating geospatial and analytic hierarchy process (AHP) approach. *Geocarto International, 33*(2), 105-129, doi:10.1080/10106049.2016.1232314.

Kumar, A., Sharma, H. C., & Kumar, S. (2011). Planning for replenishing the depleted groundwater in upper Gangetic plains using RS and GIS. *Indian Journal of Soil Conservation*.

Kumar, B., & Kumar, U. (2010). Integrated approach using RS and GIS techniques for mapping of ground water prospects in Lower Sanjai Watershed, Jharkhand. *International Journal of Geomatics & Geosciences, 1*(3), 587-598.

Kumar, T., Gautam, A. K., & Kumar, T. (2014). Appraising the accuracy of GIS-based Multi-criteria decision making technique for delineation of Groundwater potential zones. *Water resources management, 28*(13), 4449-4466.

Kumar, U., Kumar, B., & Mallick, N. (2013). Groundwater Prospects Zonation Based on RS and GIS Using Fuzzy Algebra in Khoh River Watershed, Pauri-Garhwal District, Uttarakhand, India. *Global Perspectives on Geography, volume 1*(3), 37-45.

Lee, S., Hong, S. M., & Jung, H. S. (2017). GIS-based groundwater potential mapping using artificial neural network and support vector machine models: the case of Boryeong city in Korea. *Geocarto International*, 1-33.

Lee, S., Kim, Y.-S., & Oh, H.-J. (2012). Application of a weights-of-evidence method and GIS to regional groundwater productivity potential mapping. *Journal of environmental management, 96*(1), 91-105.

Lee, S., & Lee, C.-W. (2015). Application of Decision-Tree Model to Groundwater Productivity-Potential Mapping. *Sustainability, 7*(10), 13416-13432.

Mair, A., & El-Kadi, A. I. (2013). Logistic regression modeling to assess groundwater vulnerability to contamination in Hawaii, USA. *J Contam Hydrol, 153*, 1-23.

Manap, M. A., Nampak, H., Pradhan, B., Lee, S., Sulaiman, W. N. A., & Ramli, M. F. (2014). Application of probabilistic-based frequency ratio model in groundwater potential mapping using remote sensing data and GIS. *Arabian Journal of Geosciences, 7*(2), 711-724.

Masetti, M., Poli, S., & Sterlacchini, S. (2007). The use of the weights-of-evidence modeling technique to estimate the vulnerability of groundwater to nitrate contamination. *Natural Resources Research, 16*(2), 109-119.

Meijerink, A. (1996). Remote sensing applications to hydrology: groundwater. *Hydrological sciences journal, 41*(4), 549-561.

Moghaddam, B., Weiss, Y., & Avidan, S. (2007). Spectral method for sparse linear discriminant analysis. US.

Naghibi, S. A., Moghaddam, D. D., Kalantar, B., Pradhan, B., & Kisi, O. (2017). A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. *Journal of Hydrology, 548*, 471-483.

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental monitoring and assessment, 188*(1), 1-27.

Naghibi, S. A., Pourghasemi, H. R., Pourtaghi, Z. S., & Rezaei, A. (2015). Groundwater qanat potential mapping using frequency ratio and Shannon's entropy models in the Moghan watershed, Iran. *Earth Science Informatics, 8*(1), 171-186.

Nobre, R., Rotunno Filho, O., Mansur, W., Nobre, M., & Cosenza, C. (2007). Groundwater vulnerability and risk mapping using GIS, modeling and a fuzzy logic tool. *J Contam Hydrol, 94*(3), 277-292.

Ozcift, A., & Gulten, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Computer Methods & Programs in Biomedicine, 104*(3), 443-451.

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. *Journal of Hydrology, 411*(3–4), 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). *Journal of Hydrology, 405*(1), 123-136.

Ozdemir, A., & Altural, T. (2013). A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan Mountains, SW Turkey. *Journal of Asian Earth Sciences, 64*, 180-197.

Pham, B. T., Tien Bui, D., & Prakash, I. (2017). Landslide Susceptibility Assessment Using Bagging Ensemble Based Alternating Decision Trees, Logistic Regression and J48 Decision Trees Methods: A Comparative Study. *Geotechnical and Geological Engineering, 35*(6), 2597-2611, doi:10.1007/s10706-017-0264-2.

Pourtaghi, Z. S., & Pourghasemi, H. R. (2014). GIS-based groundwater spring potential assessment and mapping in the Birjand Township, southern Khorasan Province, Iran. *Hydrogeology Journal, 22*(3), 643-662.

Rahmati, O., Pourghasemi, H. R., & Melesse, A. M. (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.

Rahmati, O., Samani, A. N., Mahdavi, M., Pourghasemi, H. R., & Zeinivand, H. (2015). Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS. *Arabian Journal of Geosciences, 8*(9), 7059-7071.

Razavi Termeh, S. V., Kornejady, A., Pourghasemi, H. R., & Keesstra, S. (2018). Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. *Science Of the Total Environment, 615*, 438-451, doi:https://doi.org/10.1016/j.scitotenv.2017.09.262.

Rodríguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2013). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 28*(10), 1619-1630.

Safavi, H. R., Chakraei, I., Kabiri-Samani, A., & Golmohammadi, M. H. (2013). Optimal reservoir operation based on conjunctive use of surface water and groundwater using neuro-fuzzy systems. *Water resources management, 27*(12), 4259-4275.

Sener, E., & Davraz, A. (2013). Assessment of groundwater vulnerability based on a modified DRASTIC model, GIS and an analytic hierarchy process (AHP) method: the case of Egirdir Lake basin (Isparta, Turkey). *Hydrogeology Journal, 21*(3), 701-714.

Shahabi, H., Keihanfard, S., Ahmad, B. B., & Amiri, M. J. T. (2014). Evaluating Boolean, AHP and WLC methods for the selection of waste landfill sites using GIS and satellite images. *Environmental Earth Sciences, 71*(9), 4221-4233.

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2015). Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stochastic Environmental Research & Risk Assessment, 29*(4), 1149-1165.

Teso, R. R., Poe, M. P., Younglove, T., & McCool, P. M. (1996). Use of logistic regression and GIS modeling to predict groundwater vulnerability to pesticides. *Journal of Environmental Quality, 25*(3), 425-432.

Thilagavathi, N., Subramani, T., Suresh, M., & Karunanidhi, D. (2015). Mapping of groundwater potential zones in Salem Chalk Hills, Tamil Nadu, India, using remote sensing and GIS techniques. *Environmental Monitoring & Assessment, 187*(4), 1-17.

Tien Bui, D., Pradhan, B., Nampak, H., Bui, Q.-T., Tran, Q.-A., & Nguyen, Q.-P. (2016a). Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS. *Journal of Hydrology, 540*, 317-330, doi:https://doi.org/10.1016/j.jhydrol.2016.06.027.

Tien Bui, D., Tuan, T. A., Hoang, N.-D., Thanh, N. Q., Nguyen, D. B., Van Liem, N., et al. (2017). Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization. [journal article]. *Landslides, 14*(2), 447-458, doi:10.1007/s10346-016-0711-9.

Tien Bui, D., Tuan, T. A., Klempe, H., Pradhan, B., & Revhaug, I. (2016b). 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. [journal article]. *Landslides, 13*(2), 361-378, doi:10.1007/s10346-015-0557-6.

Uhan, J., Vižintin, G., & Pezdič, J. (2011). Groundwater nitrate vulnerability assessment in alluvial aquifer using process-based models and weights-of-evidence method: Lower Savinja Valley case study (Slovenia). *Environmental Earth Sciences, 64*(1), 97-105.

Waikar, M., & Nilawar, A. P. (2014). Identification of groundwater potential zone using remote sensing and GIS technique. *Int J Innov Res Sci Eng Technol, 3*(5), 1264-1274.

Witten, I. H., Frank, E., & Mark, A. H. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Third edition. Morgan Kaufmann, Burlington, USA.

Xia, J., Du, P., He, X., & Chanussot, J. (2014). Hyperspectral Remote Sensing Image Classification Based on Rotation Forest. *IEEE Geoscience & Remote Sensing Letters, 11*(1), 239-243.

Yang, Z. Y., Wang, W. K., Wang, Z., Jiang, G. H., & Li, W. L. (2016). Ecology-oriented groundwater resource assessment in the Tuwei River watershed, Shaanxi Province, China. *Hydrogeology Journal, 24*(8), 1-14.

Yin, H., Fu, P., & Meng, S. (2006). Sampled FLDA for face recognition with single training image per person. *Neurocomputing, 69*(16-18), 2443-2445.

Zabihi, M., Pourghasemi, H. R., Pourtaghi, Z. S., & Behzadfar, M. (2016). GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. *Environmental Earth Sciences, 75*(8), 1-19.