Characterizing Groundwater Potential Using GIS-Based Machine Learning Model in Chihe River Basin, China

Dejian Wang

Hefei University of Technology

Jiazhong Qian (✉ qianjiazhong@hfut.edu.cn)

Hefei University of Technology

Lei Ma

Hefei University of Technology

Weidong Zhao

Hefei University of Technology

Di Gao

Hefei University of Technology

Xiaoliang Hou

Hefei University of Technology

Haichun Ma

Hefei University of Technology

Research Article

Keywords: Groundwater potential, GIS, Logistic regression, Deep neural networks, Random forest

Posted Date: December 3rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1044219/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Mapping of groundwater potential over space, built by synergizing environmental variables and machine learning models, was of great significance for regional water resources management. Taking the Chihe River basin in Anhui province as an example, thirteen influence factors were used to predict the spatial distribution of groundwater, including elevation, slope, aspect, plan curvature, profile curvature, topographic wetness index (TWI), drainage density, distance to rivers, distance to faults, lithology, soil type, land use, and normalized difference vegetation index (NDVI). The potential of groundwater resource in this region was predicted using GIS-based machine learning models, including logistic regression (LR), deep neural networks (DNN), and random forest (RF) model. Then, the accuracy of prediction results was evaluated by calculating the RMSE, MAE and R evaluation index. The results show that there is no collinearity among the 13 environmental impact factors, which can provide corresponding environmental variables for the evaluation of regional groundwater potential. Machine learning models show that groundwater potential is concentrated in moderate to high potential areas. Among them, the moderate to the high potential of this area accounted for 81.14% in the LR model, 90.36% and 87.55% in the DNN model and the RF model, respectively. According to the result of these evaluation indexes, the three models all have high prediction accuracy, among which the LR model performs more prominently. The good prediction capabilities of these machine learning technologies can provide a reliable scientific basis for spatial prediction of groundwater potential and management of water resources.

Introduction

With the development of social economy and population growth, the contradiction between water resource supply and demand has become increasingly acute. As the demand for groundwater increasing, division of dominant potential areas of groundwater has become an important tool for the implementation of groundwater measurement, protection and management (Ozdemir et al. 2011; Kordestani et al. 2019). Groundwater potential assessment is a vital work for regional development, since groundwater supplies necessities water source of local resident life, industrial production and agricultural irrigation (Wang et al. 2018).

Groundwater potential analysis is to determine the best area for groundwater development by studying multiple factors that affect the presence of groundwater in a certain area (Díaz-Alcaide and Martínez-Santos 2019). Traditional methods are used to determine the best development area for groundwater, including field surveys, geophysical methods and drilling project, etc. The common prediction methods for groundwater potential are geostatistics, which are often combined with geographic information systems (GIS) (Al-Fugara et al. 2020; Manap et al. 2014). These methods can process a large spatial database, and can be applied to environment and geology fields (Hyun-Joo et al. 2011; Rahmati et al. 2016; Saro et al. 2012). In the past 10 years, with the development of data mining technology, groundwater potential analysis combined with spatial statistical models and GIS technology has become a research hotspot. Alireza et al. (2019) used a database of 122 wells to map the groundwater potential of Shahroud plain in northern Iran through a GIS-based multivariate and bivariate model. Corsini et al.

(2009) used machine learning model to analyze the potential of groundwater and found that the combination of weight of evidence (WOE) and artificial neural network (ANN) model has good objectivity, and the results of these two models are easy to implement in ArcGIS, combined with index factor analysis to achieve better prediction results.

However, the application effects of various models are also different due to regional geological environment, climatic factors, and regional scales. In addition, the selection of index factors is restricted by objective conditions with no fixed standard. Machine learning models have good applicability in processing multi-dimensional, nonlinear mass data, and improving the generalization ability of model (Kermani et al. 2021). It has been applied to research in many fields. Especially in the area of groundwater spatial variability and spatial prediction, machine learning methods that collaborate with multi-source environmental variables show great potential. Common models include random forest (RF) (Chen et al. 2017; Naghibi et al. 2016), weight of evidence (WOE) (Chen et al. 2018), support vector machine (SVM) (Rahmati et al. 2018), decision tree (Chen et al. 2020), neural networks (NN) (Lei et al. 2021), K-nearest neighbor (KNN), and statistical methods such as frequency ratio (FR) (Razandi 2015) and logistic regression (LR) (Zandi et al. 2016).

With the continuous research on machine learning methods, the use of multiple models has achieved good prediction results in terms of groundwater potential, geological hazards, soil organic matter, and soil stability (Norouzi et al. 2019; Arabameri et al. 2020; Tien Bui et al. 2012; Ding et al. 2017). Devkota et al. (2013) used the certainty coefficient method, logistic regression model, weight of evidence method, and frequency ratio method to evaluate landslide hazard susceptibility of Mugling-Narayanghat road area in Nepal. They analyzed and compared the evaluation results of each model method. Pourtaghi and Pourghasemi (2014) conducted groundwater potential analysis in southern Khorasan province, Iran. Evaluation of groundwater potential based on frequency ratio (FR), weights-of-evidence (WoE) and logistic regression (LR) models, and the evaluation results show that the predictions of three models have high accuracy. The groundwater spring potential map can be useful for planners and engineers in water resources management and land-use planning. Hao et al. (2016) found that deep learning algorithms have more levels of non-linear operations than shallow learning methods such as neural networks and support vector machines. Hengl et al. (2015) found that random forest can avoid overfitting, are insensitive to multiple linearities. This model is easy to handle missing data, and the RF algorithm consistently outperforms the linear regression algorithm.

Hence, groundwater potential mapping can be applied to the development and planning of regional water resources management system (Naghibi et al. 2019). It also reveals the relationship between groundwater resources and human activities, and helps to understand the vulnerability of ecosystem and over-exploitation. This work can also be used to formulate groundwater sustainable management strategies for water resource planners to determine the suitable location of production wells in the Chihe River basin (Rizeei et al. 2019).

In this area, the spatial distribution of groundwater is still unclear. The work on groundwater potential evaluation is absent, and it is difficult to provide an intuitive basis for groundwater development and management in the study area. Rather than using traditional learning approaches, we leverage the correlations between various environmental impact factors and the presence of groundwater in machine learning models that analyzes the relationship between the various data sets and uses them in a predictive pattern.

The objectives of this study are to map the groundwater potential of this area, evaluate the accuracy of LR, DNN and RF model using RMSE, MAE and R evaluation index, and find a suitable method for groundwater prediction performance. Datasets used for training, validation and prediction are easily available in GIS while the processing of data in model can contribute further to disclose the groundwater potential. Thus, a new methodological regime using logistic regression (LR), deep neural networks (DNN) and random forest (RF) model of Chihe River basin was developed to improve groundwater potential mapping. The groundwater potential maps are expected to provide necessary data support for groundwater assessment and water resources management in the Chihe River basin.

Description Of The Study Area

The study area is located in the eastern part of China with longitude between 117°26'13.5" and 118°11'50" E and latitude between 32°17'47" and 32°37'46" N, and covers an area of about 4008 km² (Fig. 1). The height above sea level of the Chihe River basin varies between 11 m and 383 m. The Chihe River system runs through the whole area. This area is in the north of the Jianghuai drainage divide, which belongs to the Huaihe River network. Its east and west sides are higher than the central region.

The fresh-water for human activities in this area is mainly taken from groundwater, which is generally lacking. Most of the soil is yellow-brown clay with poor water-holding capacity, and the precipitation is difficult to infiltrate and surface runoff is fast. In addition, there are few ponds, dams, and reservoirs in this area with poor storage capacity of water. However, a large area of red sandstone is covered under the soil layer (Fig. 2). The inflow of most wells in this area is less than 1 m³/h. In the rainy season, most of the precipitation flows into the surface water system, and flows out of this area along the Chihe River into Huaihe River. Therefore, the resources of regional groundwater are very limited, and the exploitation of groundwater is difficult and costly.

Groundwater in the study area is mainly storage in pores and fractures of bedrocks. There are widely distributed in pore-fracture of loose rock aquifers in this area, which mainly consist of phreatic water and weak-confined water. Lithology is mainly silty-fine sand, with uneven thickness. The distribution of water richness in the aquifer is uneven in time and space. The mean inflow of a single well in this area varies between 0.3 m³/d and 1.5 m³/d, some can reach 5 m³/d and 20 m³/d. Groundwater quantities in this area are generally poor, and showing a band distribution.

Materials And Methods

Database

With the help of various methods and techniques, groundwater potential mapping with high reliability and accuracy could be built (Moghaddam et al. 2015). In the current study, spatial data and materials were prepared, including geology map and hydrogeology map. The digital elevation model (DEM) with a spatial resolution of 30×30 m was used to extract a set of influence factors in the study area. All impact factors were processed in the ArcGIS 10.2 software. In order to evaluate the potential of regional groundwater, the whole area was divided into 400501 grids with a size of 100×100 m based on the prediction accuracy of models and the geological conditions of this area.

In terms of groundwater data, a total of 245 wells were identified based on field survey by using a handheld GPS and historical hydrogeological materials. For this analysis, these wells were randomly divided into two groups, of which 172 wells (70%) were used for training datasets and 73 wells (30%) for validation (Fig. 1).

Selection And Analysis Of Influence Factors

The presence of groundwater is closely related to various environmental geological factors (Cantonati et al. 2016). Through quantifying environmental geological parameters to realize the analysis of groundwater potential in this area, and it is concluded that there is a functional relationship between factors and groundwater, to evaluate the presence or absence of groundwater by using the machine learning method. There are no fixed guidelines in selecting of the groundwater potential influence factors (Oh et al. 2011).

Based on the results of the field geological survey, this study analyzed the relevant geological materials and previous literature. Thirteen influence factors were selected to predict the spatial distribution of groundwater, including elevation, slope, aspect, plan curvature, profile curvature, topographic wetness index (TWI), drainage density, distance to rivers, distance to faults, lithology, soil type, land use, normalized difference vegetation index (NDVI) (Fig. 3a-m). The elevation was often used as an important factor in finding the presence of groundwater (Wang et al. 2015). It was extracted from the DEM to show the undulations of the terrain. This study divides elevation into 8 categories according to an equal-interval classification scheme, including: <20 m, 20-40, 40-60, 60-80, 80-100, 100-120, 120-140, and >140 m.

Based on the side slope unit of terrain segmentation, the slope was adopted, which can control the flow of groundwater directly. The slope unit was extracted with 30 m resolution from DEM as the basic data, and the hydrological analysis module in ArcGIS 10.2 was used to extract the regional slope. The value of the slope was divided into 7 categories according to the natural breakpoint method, including: <0.5°, 0.5-1°, 1-1.5°, 1.5-2°, 2-5°, 5-10°, and >10°. To a certain extent, the slope can indicate the direction of groundwater flow (Naghibi et al. 2016). The slope aspect was the inclination direction of slope, which controls the flow of precipitation, wind direction and plant photosynthesis (Zabihi et al. 2016). Compared with shady slopes, sunny slopes had longer sunshine time, and their surface had stronger weathering and evaporation. The aspect was extracted from the DEM and was divided into 9 categories according to the

different directions, including: Flat, North, Northeast, East, Southeast, South, Southwest, West, and Northwest.

Plan curvature was the change rate of slope at any point on the surface, which was formed by the intersection of a horizontal plane and the surface (Arabameri et al. 2020). This morphological feature will affect the convergence and divergence of surface runoff, and can reflect the degree of contour curvature. After completing the aspect extraction in ArcGIS 10.2, the slope was extracted from this aspect and was shown as a plan curvature map. The plan curvature was divided into 7 categories using an equal-interval classification scheme, including: <20, 20-30, 30-40, 40-50, 50-60, 60-70, and >70. Profile curvature represents the change rate of the surface slope at any given point. Like the plan curvature, profile curvature map was generated by calculating the slope of DEM twice in succession, and the values were reclassified into 7 categories, including: <0.4, 0.4-0.8, 0.8-1.2, 1.2-2, 2-5, 5-10, and >10.

The topographic wetness index (TWI) is a physical indicator of the influence of regional topography on groundwater flow direction and convergence (Moore et al. 1991). This index is a function of the slope and upstream contribution area. It is defined as Eq. (1) follows:

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

1

where α is the upstream area, and β is the slope of each point. According to the different TWI values, five classes were created: <6, 6-8, 8-10, 10-12, and >12.

Drainage density is the total length of rivers per unit regional area. It is closely related to the precipitation, difference elevation, and moisture retention capacity of soil. The drainage density in this area is binned into four classes: <0.15, 0.15-0.3, 0.3-0.45, and >0.45 km/km². Distance to rivers is a key factor affecting the potential of groundwater. Rivers are an important source of groundwater recharge. This area is mainly covered with Chihe River systems. The distance values between wells and rivers have an important influence to this research. Based on the hydrological conditions of this area, the buffer zones on the borderlands of the river system were divided into 5 classes: <100, 100-200, 200-300, 300-400, and >400 m.

Faults can control the flow and storage of regional groundwater. Regional faults were extracted from the geological map and were reclassified into five groups based on distance: <500, 500-1000, 1000-1500, 1500-2000, and >2000 m. Lithology of aquifer is the basis of groundwater flow and storage, and it determines the porosity and permeability of the aquifer (Ayazi et al. 2010). Lithology categories were extracted from the regional geological map. Fourteen types of lithology were divided, as shown in Table 1.

Table 1
Regional lithology information of study area

Number	Unit	Lithology	Period
1	K ₂ x	Sandstone, argillaceous, siltstone	Cretaceous
2	K ₂ z	Sandstone, glutenite	
3	K ₁ g	Sandstone, glutenite, mixed shale	
4	E ₃ s	Mudstone, siltstone, glutenite	Paleogene
5	β ₆	Basalt	Cenozoic
6	E ₂ t	Sandstone, glutenite, mudstone, mixed basalt	Paleogene
7	Ar ₂ xz	Gneiss, feldspar quartzite, schist	Archaeozoic Eon
8	Є ₂	Limestone, dolomites	Cambrian
9	O ₁	Medium-thick limestone, dolomites	Ordovician
10	Zz ₁ z	Banded siliceous dolomite, marl	Ediacaran
11	δO ₅	Quartz diorite	Late Jurassic-Cretaceous
12	Qh	Clay rock, siltstone, glutenite, mudstone	Holocene
13	Xy ₂	Moyite	Early Jurassic-Cretaceous
14	Qp	Clay rock, sandstone, mudstone	Pleistocene

Soil type has a vital influence on the infiltration of surface water and recharge of groundwater. It determines the distribution of groundwater potential to a certain extent (Razandi et al. 2015). This factor was roughly divided into three groups from the regional geological map according to different types, including rock outcrops, sandy soil, and cohesive soil. Different land-use types affect the quantity of groundwater resources and quality of groundwater. The relationship between human activities and natural systems was revealed and the distribution of groundwater potential was reflected (Chen et al. 2018). The land use map was classified into water bodies, residential areas, forests, and agriculture. For the normalized difference vegetation index (NDVI), it refers to the percentage of the vertical projection area of vegetation on the ground to the total statistical area (Pourghasemi et al. 2013). The NDVI map was calculated and mapped by the ArcGIS software. Eight groups were reclassified based on a natural break method: <0.125, 0.125-0.25, 0.25-0.375, 0.375-0.5, 0.5-0.625, 0.625-0.75, 0.75-0.875, and >0.875.

Multicollinearity Analysis Of Factors

A multicollinearity analysis of factors was performed to select factors, which have a significant relationship with groundwater distribution. Multicollinearity refers to a certain extent of linear correlation

between independent factors, which will affect the contribution of factors to the model (Pourtaghi and Pourghasemi 2014). If there was collinearity between two factors, it was difficult to distinguish the effect of each factor on the results, and the regression model lacks stability. Then, two statistical parameters to determine the multicollinearity problem between each factor were proposed, namely tolerance (TOL) and variance inflation factor (VIF). The values of $TOL > 0.1$ or $VIF < 10$ suggest independence between each factor.

Description Of Models

Logistic regression (LR)

A nonlinear dynamic response relationship between a dependent variable and several corresponding independent variables was established by training and testing the known samples in the logistic regression (LR) model, and then predicts or evaluates the probability of an event in unknown samples (Lombardo et al. 2018). When evaluating groundwater potential, each influencing factor was taken as an independent variable, and the presence or absence of groundwater was taken as a dependent variable. In this study, P is the probability of the presence of groundwater with a range $[0, 1]$, $1-P$ is the probability of the absence of groundwater, $P/(1-P)$ is the ratio of probability, which is often taken as its natural logarithm. The LR model equation is expressed as follow:

$$\ln \frac{P}{1-P} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

2

where, $X_1, X_2, \dots, X_n$ represents the independent factors; $\beta_0, \beta_1, \dots, \beta_n$ represents the regression coefficients. The probability P of groundwater potential can be obtained from Eq. 3:

$$P = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}}$$

3

In order to calculate the groundwater potential index of the Chihe River basin by using the LR model, 245 known well locations (with potential index 1) and 245 randomly selected non-well location units (with potential index 0) were randomly divided into two parts, of which 70% was used for model training and the others was used for model validation. Groundwater potential indexes of each grid unit calculated by the LR model ranges from 0.001 to 0.999, the high value with greater the groundwater potential. The data processing of the LR model was completed by SPSS 22.0 statistical software, and the P values of groundwater were imported into ArcGIS 10.2 to generate the groundwater potential map.

Deep neural networks (DNN)

In deep neural networks, sample data was processed through multiple layers, the initial low-level feature layer was gradually converted into a high-level feature layer with more abstract. The distributed characteristics of regional data were mined, which was more conducive to the visualization of

classification or characteristics. Multiple layers of deep structure with nonlinear factors in the processing of machine learning of deep neural network (DNN) can accomplish complex function approximation. Through the summary and analysis of DNN model, a multi-layer perceptron model was established with the MATLAB software to conduct nonlinear prediction research on groundwater potential in this area (Jiang et al. 2019).

The function of the activation layer in the DNN model is to use the function to activate output result of the entire connected layer. The activation function is a non-linear transformation function that can imitate threshold activation characteristics of brain neurons. It introduces non-linear features into DNN and strengthens the expressive ability of this model. Two activation functions Relu and Sigmoid are used in this model. Relu function is a piecewise function widely used. It can improve the speed of convergence and increase the sparsity of the network. The sigmoid function is actually the expression of logistic regression function, which maps the input of this layer to [0, 1]. It is suitable for second-class discrimination in the last layer of this network. Two functions can be represented as Eq. 4 and Eq. 5.

Relu function:
$$f(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (4)$$

Sigmoid function:
$$f(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

In model training, the loss function adopts two kinds of cross entropy, and the Adam method is used to find the optimum value. The initial learning rate is set to 0.005, and the iteration of model is 2500.

DNN algorithm was continuously optimized by using the training samples (70%) and the validation samples (30%). Then, predicted data was substituted into the model for training to obtain the predicted value, and input the values with location information into ArcGIS to generate the groundwater potential map.

Random forest (RF)

Random forest (RF) is the most representative algorithm based on bagging integrated learning (Norouzi et al. 2019). It uses random sampling to perform integrated learning on multiple decision trees, and finally makes predictions through a majority voting mechanism. By extracting several samples from the original datasets with samples returned, using the decision tree algorithm to train these extracted samples, and then combining these decision trees together (Naghbi et al. 2016). After voting, the final classification result was the one with the most votes.

These main steps of the RF model are as follows: first, sampling many times from the original sample set and return the sample every time, each time to form a training dataset. Then, a decision tree is generated. If the sample has X groups of features in total, n features are randomly selected from the X groups as the split feature set of each internal node of a decision tree. Subsequently, the node is split using an optimal split method of classification feature set. Classification and regression tree (CART) algorithms are used to generate decision trees. Finally, integrate all decision trees to form a random forest model to classify and

predict unknown data. Voting on the results of each tree produce the result and the most vote is the classification result.

This study uses the RF package in R 4.1.0 software to fit the model. Takes each groundwater potential influence factor as an auxiliary dataset, the final model variables are screened out by error in the outside of the RF package, variables are eliminated one by one, and then the change of error in the outside of the RF package is observed. If the error increases, the variable is retained, otherwise eliminate the variable.

Model Comparison

The accuracy of model evaluation is controlled by many factors. This study evaluates the prediction accuracy of three models by calculating the mean absolute error (MAE), root mean square error (RMSE) and correlation coefficient (R) between measured values and predicted values of the validation data (Singha et al. 2020). RMSE can evaluate the variation degree of data, and its equation is shown in Eq. 6. The smaller value of the RMSE is, the higher prediction accurate of the model is. MAE is the average of absolute errors, which can better reflect the actual error of the predicted value. The smaller value of the MAE is, the more accurate prediction of the model is, as shown in Eq. 7. R value is a statistical indicator that can reflect the degree of correlation between variables, which can be calculated as Eq. 8. The value of R is closer to 1, the stronger of correlation between two variables.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (A_i - B_i)^2}$$

6

$$MAE = \frac{1}{n} \sum_{i=1}^n |A_i - B_i|$$

7

$$R = \frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^n (A_i - \bar{A})^2} \sqrt{\sum_{i=1}^n (B_i - \bar{B})^2}}$$

8

where A_i is measured values; B_i is predicted values; $\bar{A}$ and $\bar{B}$ represent the average value of measured values and predicted values.

Results

Multicollinearity Analysis

Thirteen influence factors in this study were checked for multicollinearity in SPSS 22.0 software. Results showed that the highest value of VIF is 3.597 and the lowest value of TOL is 0.278 (Table 2). It indicates that these factors are independent of each other. The analysis result of multicollinearity is shown in Figure 4.

Table 2
Multicollinearity analysis of each factor

Factor	Collinearity statistics	
	TOL	VIF
Elevation	0.774	1.293
Slope	0.278	3.597
Aspect	0.635	1.575
Plan curvature	0.564	1.771
Profile curvature	0.480	2.082
TWI	0.417	2.397
Drainage density	0.772	1.295
Distance to rivers	0.786	1.273
Distance to faults	0.817	1.224
Lithology	0.621	1.609
Soil type	0.612	1.633
NDVI	0.913	1.096
Land use	0.902	1.109

Model Results

Logistic regression (LR)

Comprehensive coefficient of the Hosmer-Lemeshow test in groundwater potential index was calculated, and its significance value is 0.141, showing good model fitting. The variable coefficients and odds ratio (OR) values in the LR model were shown in Table 3. The larger regression coefficient value and OR value of each input variable have the greater influence on groundwater potential. Grid data from the area was put into Eq. 2 and Eq. 3, and then the distribution of groundwater potential index of Chihe River basin was calculated. As shown in Fig. 5 and Table 4, the groundwater potential map was reclassified into five categories according to the natural break method, which was very high (29.5%), high (30.2%), moderate (21.44%), low (15.92%), and very low (2.94%).

Table 3
Related variables in LR equations

Factor	Regression coefficient	Odds ratio (OR)
Elevation	-0.044	0.957
Slope	-0.124	0.889
Aspect	0.007	1.007
Plan curvature	0.003	1.003
Profile curvature	-0.052	0.949
TWI	0.188	1.207
Drainage density	2.358	10.567
Distance to rivers	-0.009	0.991
Distance to faults	0.000	1.000
Lithology	-0.079	0.924
Soil type	0.244	1.276
NDVI	-0.054	0.948
Land use	-0.231	0.794
Intercept	6.013	-

Table 4
Statistical comparison results of groundwater potential of each model

Models	LR		DNN		RF	
	Area/ km ²	Ratio/ %	Area/ km ²	Ratio/ %	Area/ km ²	Ratio/ %
Very high	1182.4	29.5	1475.72	36.82	1689.33	42.15
High	1210.33	30.2	1533.22	38.25	679.51	16.95
Moderate	859.25	21.44	612.86	15.29	1140.26	28.45
Low	638.27	15.92	209.03	5.22	237.14	5.92
Very low	117.75	2.94	177.17	4.42	261.75	6.53

Deep neural networks (DNN)

DNN model learns from training data and validation data. After many tests, adjustments and optimizations, a four-layer perceptron model (13-6-6-1) was built, and the error between predicted values

and validation values was less than 0.15, indicating relatively accurate prediction of this model. When the number of iterations exceeds 200 times in this model, the training result tends to be stable. Then, the prediction data was imported into the DNN model to obtain the groundwater potential index. Finally, the forecast index was imported into ArcGIS software for mapping (Fig. 6), and the results were divided into 5 categories according to the natural break method, as followed by very high (36.82%), high (38.25%), moderate (15.29%), low (5.22%), and very low (4.42%) (Table 4).

Random forest (RF)

The trained RF model confirmed the predictive performance of the regional groundwater potential. Seventy percent of the data was used for training RF model, and the other thirty percent data for verifying this model. Through experiment, the proper number of decision trees (*ntree*) and node value (*mtry*) had values of 500 and 3, respectively (Hengl et al. 2015). Finally, the optimized RF model was used to calculate the potential prediction value based on the raster data, and the result is shown in Fig. 7 and Table 4. This map was classified into five categories of very high (42.15%), high (16.95%), moderate (28.45%), low (5.92%), and very low (6.53%) using the natural break method.

Prediction Accuracy Evaluation

In this study, the values of RMSE, MAE and R evaluation index were calculated using validation dataset. As shown in Table 5, it can be seen that the RMSE value of LR model is the smallest, R value is the largest, and the value of MAE is slightly higher than that of RF model. It shows that the LR model has significantly higher prediction accuracy than DNN and RF model. The RMSE value of DNN model is lower than RF model; the R value is significantly higher than RF model. It shows that the DNN model has higher significantly prediction accuracy than RF model. Overall, the results of evaluation index indicate a good prediction accuracy of three models (Kayhomayoon et al. 2021). However, the LR model has a better performance in evaluation and prediction of regional groundwater potential.

Table 5
Evaluation index of groundwater potential of different models

Model	RMSE	MAE	R
LR	0.273	0.215	0.683
DNN	2.43	0.744	0.558
RF	7.504	0.172	0.014

Discussion

This study on the potential of groundwater is very important for the development, utilization and protection of water resources in a region (Díaz-Alcaide et al. 2019). Over the years, there have been different methods for water resources research, which can provide effective guides. However, many

methods often have low accuracy in prediction and evaluation, and it is difficult to give detailed and systematic research results. Recently, with the gradual deepening of interdisciplinary subjects, various methods have gradually been applied to the field of hydrogeology. With the rise of the big data period, the potential of machine learning methods in regional groundwater spatial prediction is gradually revealed (Pourghasemi et al. 2020). The development of information science and data science has made it possible for people to obtain accurate hydrogeology information. Many machine learning models that cooperate with multi-source environmental variables have been widely used in revealing the spatial variability of groundwater and spatial mapping (Majumdar et al. 2020).

Spatial variation of groundwater is the result of the interaction between natural environment and human activities. Natural environmental factors have widely used as modeling indicators due to being easily spatialized. However, human activity factors do not have spatial continuity, which increases the difficulty of model training. How to quantify this discontinuity index is of great significance for improving the accuracy prediction of different models (Kalantar et al. 2019). Chen et al. (2018) applied weights of evidence (WoE), logistic regression (LR), and functional tree (FT) model to predict regional groundwater potential in the transition zone between Aeolian landforms and a loess hilly region. The validation clearly highlights the efficacy of the integrated LR, WOE and FT models employed for proper regional groundwater resource evaluation. Rizeei et al. (2019) developed and validated a novel ensemble multi-adaptive boosting logistic regression (MABLR) model that exhibited the highest performance for predicting groundwater aquifer potential.

Through model analysis, training and validation, the relationship between environmental impact factors and groundwater potential was extracted. Then, the LR, DNN and RF model for regional spatial prediction of groundwater were selected. Compared with the traditional interpolation analysis, this process can systematically obtain the presence probability of groundwater, which effectively improves the prediction accuracy. It provides more accurate guidance for the management of regional groundwater resources.

In this study, the prediction accuracy of LR model is better than the other two models (DNN and RF model). Nonlinear relationship between the groundwater and environmental variables was well captured by LR model. In this model, independent variables can be discrete or continuous, and the calculation result was expressed by a regression formula. As shown in the LR model, the regression coefficients and OR values of TWI, drainage density and soil type were higher, indicating that it had a great influence on regional groundwater. It can be seen from Table 3 that drainage density was the dominant factor affecting groundwater potential. Elevation, slope and river distance were negatively correlated with groundwater potential, which should be related to the low degree of topographic relief.

The DNN and RF model has accuracy prediction results of groundwater potential in this area, which is consistent with the prediction result of the LR model. Among them, RF model can well capture the nonlinear relationship between groundwater potential and environmental variables. It also can simulate the high-order interaction between variables. Norouzi and shahmohammadi-kalalagh (2019) used RF model to accurately locate the artificial recharge source of groundwater in shabestar region, Iran.

Arabameri et al. (2020) used the integrated machine learning method to map the groundwater potential of Bastam watershed, Iran. RF modeling recovered that LU / LC, lithology, and elevation were the most important factors for predicting groundwater potential and production. Integrated machine learning method will provide an accurate and effective reference for the groundwater potential (Lee et al. 2020).

From Table 4, there was a slight difference in the prediction results of regional groundwater potential from the three models, but overall, this area covers high potential of the groundwater. As shown in Fig. 5 to Fig. 7, the highest groundwater potential in this region was mainly distributed around the Chihe River. Groundwater potential in the east and west side hilly areas is low, and the central plain is mainly composed of sand rocks with the relatively higher water potential. The distribution of groundwater resources in this area was basically controlled by these factors. These factors basically control the distribution of groundwater resources in the region. Whereas each factor provides partial information of the regional hydrological conditions, the combination of environmental factors allows obtaining a comprehensive understanding of the distribution and hydrogeology information of this complex system. Geological factors, such as fault distance and lithology factors have less affected on groundwater. Land use is closely related to human activities, so this factor also has a certain impact on groundwater potential.

Conclusions

Taking the Chihe River basin as the study area, this work not only applied GIS-based machine learning models to identify areas of groundwater potential, but also compare and discuss the applicability and accuracy of these models.

The relationship between geological environment, human activities and topographic features of the Chihe River basin were analyzed, and the presence laws of groundwater were summarized. Thirteen independent variables of environmental impact for groundwater potential analysis were selected, including elevation, slope, aspect, plan curvature, profile curvature, TWI, drainage density, distance to rivers, distance to faults, lithology, soil type, land use and NDVI. This study applied LR, DNN, and RF model to evaluate and partition areas of groundwater potential. The applicability of three models were evaluated through the evaluation indicators MAE, RMSE and R. On the whole, it is clear that the three models indicate good accuracy. It also shows that these models can objectively evaluate the groundwater potential of Chihe River basin.

Groundwater potential is divided into five categories according to the natural break point method. Results show that this area is mainly concentrated in the moderate to high groundwater potential. Among them, the moderate to high potential of this area accounted for 81.14%, 90.36% and 87.55% for the LR model, the DNN model, and the RF model, respectively. Overall, the regional groundwater has good potential.

The result of research in this paper shows that the combination between machine learning models, hydrological databases, DEM data and thematic information map can better avoid the subjectivity of experts. According to different models, the relationship between environmental factors and groundwater

is analyzed, and a more reliable regional groundwater potential map is generated. Compared with traditional methods, this combination technologies provides some new ideas. It can give an important reference for regional hydrogeological surveys.

Declarations

Acknowledgements

This study was supported by the National Natural Science Foundation of China (No. 41831289; No. 41772250; No. 41877191) and the Public welfare geological survey program of Anhui Province (Grant No. 2015-g-26).

References

1. Al-Fugara A, Pourghasemi HR, Al-Shabeeb AR, Habib M, Al-Adamat R, Al-Amoush H, Collins AL (2020) A comparison of machine learning models for the mapping of groundwater spring potential. *Environ Earth Sci* 79(10):1–19. <https://doi.org/10.1007/s12665-020-08944-1>
2. Alireza A, Khalil R, Artemi C, Luigi L, Jess R (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(1):160–177. <https://doi.org/10.1016/j.scitotenv.2018.12.115>
3. Arabameri A, Lee S, Tiefenbacher JP, Ngo PTT (2020) Novel ensemble of MCDM-artificial intelligence techniques for groundwater-potential mapping in arid and semi-arid regions (Iran). *Remote Sens-Basel* 12(3):490. <https://doi.org/10.3390/rs12030490>
4. Ayazi MH, Pirasteh S, Arvin AKP, Pradhan B, Nikouravan B, Mansor S (2010) Disasters and risk reduction in groundwater: Zagros mountain southwest Iran using geoinformatics techniques. *Dis Adv* 3(1):51–57. <https://doi.org/10.5194/tc-4-621-2010>
5. Cantonati M, Segadelli S, Ogata K, Tran H, Sanders D, Gerecke R, Rott E, Filippini M, Gargini A, Celico F (2016) A global review on ambient Limestone-Precipitating Springs (LPS): hydrogeological setting, ecology, and conservation. *Sci Total Environ* 568:624–637. <https://doi.org/10.1016/j.scitotenv.2016.02.105>
6. Chen W, Li H, Hou EK, Wang SQ, Wang GR, Panahi M, Li T, Peng T, Guo C, Niu C (2018) GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. *Sci Total Environ* 634(1):853–867. <https://doi.org/10.1016/j.scitotenv.2018.04.055>
7. Chen W, Xie XS, Wang JL, Pradhan B, Hong HY, Bui DT, Duan Z, Ma JQ (2017) 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. <https://doi.org/10.1016/j.catena.2016.11.032>

8. Chen W, Zhao XA, Tsangaratos PC, Shahabi HDE, Ilia IC, Xue WF, Wang XA, Ahmad BB (2020) Evaluating the usage of tree-based ensemble methods in groundwater spring potential mapping. *J Hydrol* 583:124602. <https://doi.org/10.1016/j.jhydrol.2020.124602>
9. 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. <https://doi.org/10.1016/j.geomorph.2008.03.015>
10. Devkota KC, Regmi AD, Pourghasemi HR, Yoshida KC, Pradhan BE, Ryu IC, Dhital MR, Althuwaynee OF (2013) Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling-Narayanghat road section in Nepal Himalaya. *Nat Hazards* 65(2):135–165. <https://doi.org/10.1007/s11069-012-0347-6>
11. Ding QF, Chen W, Hong HY (2017) Application of frequency ratio, weights of evidence and evidential belief function models in landslide susceptibility mapping. *Geocarto Int* 32(6):619–639. <https://doi.org/10.1080/10106049.2016.1165294>
12. Díaz-Alcaide S, Martínez-Santos P (2019) Review: Advances in groundwater potential mapping. *Hydrogeol J* 27(7):2307–2324. <https://doi.org/10.1007/s10040-019-02001-3>
13. Hao X, Zhang G, Ma S (2016) Deep learning. *International Journal of Semantic Computing* 10(3):417–439. <https://doi.org/10.1142/S1793351X16500045>
14. Hengl T, Heuvelink GB, Kempen B, Leenaars JG, Walsh MG, Shepherd KD, Sila A, MacMillan RA, Mendes de Jesus J, Tamene L, Jérôme ET (2015) Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PLoS ONE* 10(6):e0125814. <https://doi.org/10.1371/journal.pone.0125814>
15. Hyun-Joo O, Yong-Sung K, Jong-Kuk C, Eungyu P, Saro L (2011) GIS mapping of regional probabilistic groundwater potential in the area of Pohang City. *Korea J Hydrol* 399(3–4):158–172. <https://doi.org/10.1016/j.jhydrol.2010.12.027>
16. Jiang ZJ, Mallants D, Peeters L, Gao L, Soerensen C, Mariethoz G (2019) High-resolution paleovalley classification from airborne electromagnetic imaging and deep neural network training using digital elevation model data. *Hydrol Earth Syst Sc* 23(6):2561–2580. <https://doi.org/10.5194/hess-23-2561-2019>
17. Kalantar B, Al-Najjar HAH, Pradhan B, Saeidi V, Halin AA, Ueda N, Naghibi SA (2019) Optimized conditioning factors using machine learning techniques for groundwater potential mapping. *Water-Sui* 11(9):1909. <https://doi.org/10.3390/w11091909>
18. Kayhomayoon Z, Azar NA, Milan SG, Moghaddam HK, Berndtsson R (2021) Novel approach for predicting groundwater storage loss using machine learning. *J Environ Manage* 296:113237. <https://doi.org/10.1016/j.jenvman.2021.113237>
19. Kermani MZ, Batelaan O, Fadaee M, Hinkelmann RP (2021) Ensemble machine learning paradigms in hydrology: A review. *J Hydrol* 598:126266. <https://doi.org/10.1016/j.jhydrol.2021.126266>
20. Kordestani MD, Naghibi SA, Hashemi H, Ahmadi K, Kalantar B, Pradhan B (2019) Groundwater potential mapping using a novel data-mining ensemble model. *Hydrogeol J* 27(1):211–224.

<https://doi.org/10.1007/s10040-018-1848-5>

21. Lee S, Hyun Y, Lee S, Lee MJ (2020) Groundwater potential mapping using remote sensing and GIS-based machine learning techniques. *Remote Sens-Basel* 12(7):1200.
<https://doi.org/10.3390/rs12071200>
22. Lei XX, Chen W, Panahi M, Falah F, Rahmati O, Uuemaa E, Kalantari Z, Ferreira CSS, Rezaie F, Tiefenbacher JP, Lee S, Bian HY (2021) Urban flood modeling using deep-learning approaches in Seoul, South Korea. *J Hydrol* 601:126684. <https://doi.org/10.1016/j.jhydrol.2021.126684>
23. Lombardo L, Opitz T, Huser R (2018) Point process-based modeling of multiple debris flow landslides using INLA: an application to the 2009 Messina disaster. *Stoch Env Res Risk A* 32(7):2179–2198.
<https://doi.org/10.1007/s00477-018-1518-0>
24. Majumdar S, Smith R, Butler JJ Jr, Lakshmi V (2020) Groundwater withdrawal prediction using integrated multitemporal remote sensing data sets and machine learning. *Water Resour Res* 56(11):e2020WR028059. <https://doi.org/10.1029/2020WR028059>
25. Manap MA, Nampak H, Pradhan B, Lee S, Sulaiman WN, Ramli MF (2014) Application of probabilistic-based frequency ratio model in groundwater potential mapping using remote sensing data and GIS. *Arab J Geosci* 7(2):711–724. <https://doi.org/10.1007/s12517-012-0795-z>
26. Moghaddam DD, Rezaei M, Pourghasemi HR, Pourtaghie ZS, Pradhan B (2015) Groundwater spring potential mapping using bivariate statistical model and GIS in the Taleghan watershed, Iran. *Arab J Geosci* 8(2):913–929. <https://doi.org/10.1007/s12517-013-1161-5>
27. Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrol Process* 5(1):3–30.
<https://doi.org/10.1002/hyp.3360050103>
28. Naghibi SA, Dolatkordestani M, Rezaei A, Amouzegari P, Heravi MT, Kalantar B, Pradhan B (2019) Application of rotation forest with decision trees as base classifier and a novel ensemble model in spatial modeling of groundwater potential. *Environ Monit Assess* 191(4):248.
<https://doi.org/10.1007/s10661-019-7362-y>
29. Naghibi SA, Pourghasemi HR, Dixon B (2016) GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ Monit Assess* 188(1):44–70. <https://doi.org/10.1007/s10661-015-5049-6>
30. Norouzi HA, Shahmohammadi-Kalalagh SB (2019) Locating groundwater artificial recharge sites using random forest: a case study of Shabestar region, Iran. *Environ Earth Sci* 78(13):380.
<https://doi.org/10.1007/s12665-019-8381-2>
31. Oh HJ, Pradhan B (2011) Application of a neuro-fuzzy model to landslide-susceptibility mapping for shallow landslides in a tropical hilly area. *Comput Geosci* 37(9):1264–1276.
<https://doi.org/10.1016/j.cageo.2010.10.012>
32. Ozdemir A (2011) Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). *J Hydrol* 405(1–2):123–136.
<https://doi.org/10.1016/j.jhydrol.2011.05.015>

33. Pourghasemi HR, Moradi HR, Aghda SMF, Gokceoglu C, Pradhan B (2013) GIS-based landslide susceptibility mapping with probabilistic likelihood ratio and spatial multicriteria evaluation models (North of Tehran, Iran). *Arab J Geosci* 7(5):1857–1878. <https://doi.org/10.1007/s12517-012-0825-x>
34. Pourghasemi HR, Sadhasivam NB, Yousefi SC, Tavangar SD, Ghaffari Nazarlou HE, Santosh M (2020) Using machine learning algorithms to map the groundwater recharge potential zones. *J Environ Manage* 265:110525. <https://doi.org/10.1016/j.jenvman.2020.110525>
35. Pourtaghi ZS, Pourghasemi HR (2014) GIS-based groundwater spring potential assessment and mapping in the Birjand Township, southern Khorasan Province, Iran. *Hydrogeol J* 22(3):643–662. <https://doi.org/10.1007/s10040-013-1089-6>
36. Rahmati O, Naghibi SA, Shahabi H, Bui DT, Pradhan B, Azareh A, Rafiei-Sardooi E, Samani AN, Melesse AM (2018) Groundwater spring potential modelling: comprising the capability and robustness of three different modeling approaches. *J Hydrol* 565:248–261. <https://doi.org/10.1016/j.jhydrol.2018.08.027>
37. 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. <https://doi.org/10.1016/j.catena.2015.10.010>
38. Razandi Y, Pourghasemi HR, Neisani NS (2015) Application of analytical hierarchy process, frequency ratio, and certainty factor models for groundwater potential mapping using GIS. *Earth Sci Inform* 8(4):867–883. <https://doi.org/10.1007/s12145-015-0220-8>
39. Rizzei HM, Pradhan B, Saharkhiz MA, Lee S (2019) Groundwater aquifer potential modeling using an ensemble multi-adoptive boosting logistic regression technique. *J Hydrol* 579:124172. <https://doi.org/10.1016/j.jhydrol.2019.124172>
40. Saro L, Yong-Sung K, Hyun-Joo O (2012) Application of a weights-of-evidence method and GIS to regional groundwater productivity potential mapping. *J Environ Manage* 96(1):91–105. <https://doi.org/10.1016/j.jenvman.2011.09.016>
41. Singha S, Pasupuleti S, Singha SS, Kumar S (2020) Effectiveness of groundwater heavy metal pollution indices studies by deep-learning. *J Contam Hydro* 235:103718. <https://doi.org/10.1016/j.jconhyd.2020.103718>
42. Tien Bui D, Pradhan B, Lofman O, Revhaug I, Dick O (2012) Landslide susceptibility mapping at Hoa Binh province (Vietnam) using an adaptive neuro-fuzzy inference system and GIS. *Comput Geosci-UK* 45:199–211. <https://doi.org/10.1016/j.cageo.2011.10.031>
43. Wang Q, Li W, Chen W, Bai H (2015) GIS-based assessment of landslide susceptibility using certainty factor and index of entropy models for the Qianyang County of Baoji city, China. *J Earth Syst Sci* 124(7):1399–1415. <https://doi.org/10.1007/s12040-015-0624-3>
44. Wang XJ, Zhang JY, Shahid S, Xie W, Du CY, Shang XC, Zhang X (2018) Modeling domestic water demand in Huaihe River Basin of China under climate change and population dynamics. *Environ Dev Sustain* 20(2):911–924. <https://doi.org/10.1007/s10668-017-9919-7>

45. Zabihi M, Pourghasemi HR, Pourtaghi ZS, Behzadfar M (2016) GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. *Environ Earth Sci* 75(8):1–19. <https://doi.org/10.1109/CCECE.2016.7726731>
46. Zandi J, Ghazvinei PT, Hashim R, Yusof KBW, Ariffin J, Motamedi S (2016) Mapping of regional potential groundwater springs using Logistic Regression statistical method. *Water Resour+* 43(1):48–57. <https://doi.org/10.1134/S0097807816010097>

Figures

Figure 1

Location of the study area

Figure 2

(a) Regional geological map and hydrogeological profile; (b) drilling column diagram and lithology

Figure 3

Regional groundwater influence factors: (a) elevation, (b) slope, (c) aspect, (d) plan curvature, (e) profile curvature, (f) TWI, (g) drainage density, (h) distance to rivers, (i) distance to faults, (j) lithology, (k) soil type, (l) land use, (m) NDVI

Figure 4

Histogram of multicollinearity results of regional influence factor

Figure 5

Groundwater potential distribution established by the LR model

Figure 6

Groundwater potential distribution established by the DNN model

Figure 7

Groundwater potential distribution established by the RF model