Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 2064575, 11 pages
http://dx.doi.org/10.1155/2016/2064575

Research Article

Assessment of Groundwater Potential Based on Multicriteria Decision Making Model and Decision Tree Algorithms

Huajie Duan, Zhengdong Deng, Feifan Deng, and Daqing Wang

PLA University of Science and Technology, Nanjing 210007, China

Correspondence should be addressed to Zhengdong Deng; dengzdong@sina.com

Received 17 July 2016; Revised 1 November 2016; Accepted 7 November 2016

Academic Editor: Alessandro Lo Schiavo

Groundwater plays an important role in global climate change and satisfying human needs. In the study, RS (remote sensing) and GIS (geographic information system) were utilized to generate five thematic layers, lithology, lineament density, topology, slope, and river density considered as factors influencing the groundwater potential. Then, the multicriteria decision model (MCDM) was integrated with C5.0 and CART, respectively, to generate the decision tree with 80 surveyed tube wells divided into four classes on the basis of the yield. To test the precision of the decision tree algorithms, the 10-fold cross validation and kappa coefficient were adopted and the average kappa coefficient for C5.0 and CART was 90.45% and 85.09%, respectively. After applying the decision tree to the whole study area, four classes of groundwater potential zones were demarcated. According to the classification result, the four grades of groundwater potential zones, "very good," "good," "moderate," and "poor," occupy 4.61%, 8.58%, 26.59%, and 60.23%, respectively, with C5.0 algorithm, while occupying the percentages of 4.68%, 10.09%, 26.10%, and 59.13%, respectively, with CART algorithm. Therefore, we can draw the conclusion that C5.0 algorithm is more appropriate than CART for the groundwater potential zone prediction.

1. Introduction

Increasing population and water scarcity have raised the importance of groundwater zones, as they are a major source of freshwater. Integrated remote sensing and GIS are widely used in groundwater mapping. Locating potential groundwater targets is becoming more convenient and cost-effective with the advent of a number of satellite imageries. Remotely sensed based groundwater exploration has made it feasible to explore the areas with limited human access, for the wide visual range, short time cycle, and increasing spatial resolution.

A lot of work has been done on the delineation of groundwater potential zones, including in tropical humid regions, such as Tirnavos area, Greece [1], Timor Leste, Indonesia [2], SW, Nigeria [3], and New Delhi [4], and in mid-latitude semiarid areas, such as Boryeong and Pohang, Korea [5, 6], Sultan Mountains (Konya, Turkey) [7, 8], Udaipur, India [9], Beheshtabad watershed, and Chaharmahal-and-Bakhtiari Province, Iran [10]. Through the analysis of the characteristics and factors within the typical regions, we found out basic principles and methods for factor selection, which provided reference and basis for the study area.

Various researchers have effectively implemented multicriteria decision model (MCDM) for accurately identifying the groundwater potential zones [1–4]. The major factors that influence the groundwater potential are lithology, rainfall, slope, drainage density, lineament density, and so forth. The factors' values are mostly continuous; however, the predecessors mostly divided the continuous factors into several discrete levels according to the relationship with groundwater potential, causing a great loss of the original information. In the previous research, we have established the fuzzy membership functions to analyze each factor's impact on groundwater enrichment from the perspective of continuity [11].

The MCDM is based on either manual decision method like analytic hierarchy process (AHP) [1–4, 9] or machine learning method, such as artificial neural network fitting [5], frequency ratio, weights of evidence and logistic regression [7, 8], boosted regression tree, classification and regression

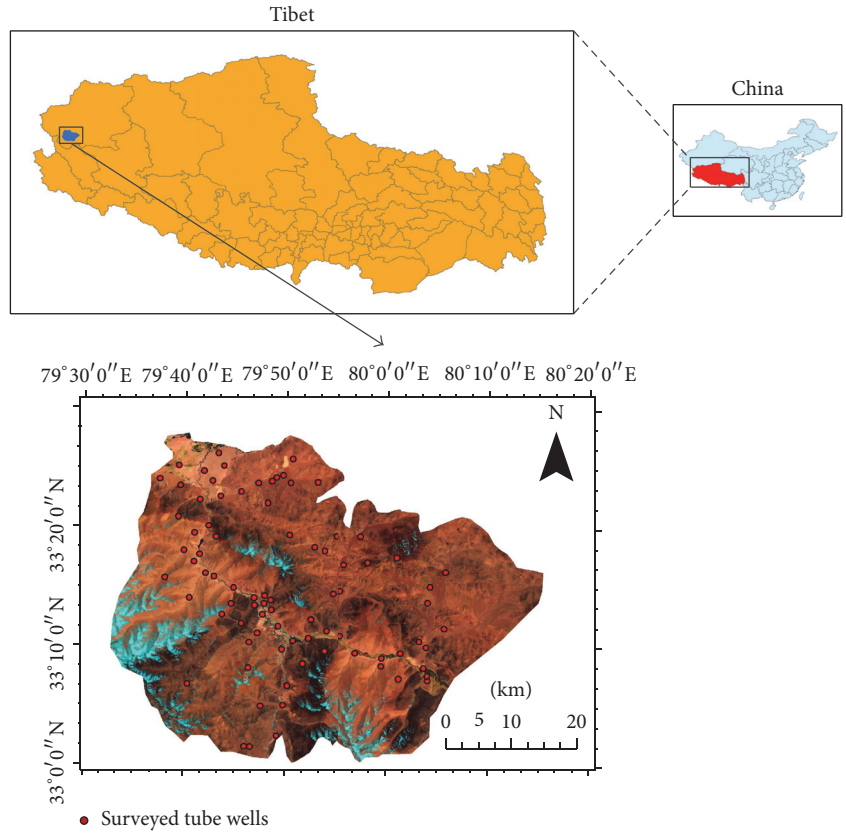

Figure 1: Location map of the study area (Landsat 8 OLI 7/5/3).

tree (CART), random forest [10], chi-squared automatic interaction detector, or the quick, unbiased, and efficient statistical tree algorithms [12, 13]. The flexibility of the AHP method allows the revision of the weights and rating of parameters in order to be suitable for other regions according to their specific characteristics [1]. However, regarding assigning weights to the different thematic layers, personal judgment reduced the objectivity of the model [9]. Decision tree techniques provide a multivariate method, which is known as a successful automatic classification scheme [14, 15]. CART algorithm [16] showed more application than other decision tree algorithms. C5.0 algorithm [16–19], as one of the decision tree techniques, serves as an enhancement of C4.5 and shows relatively better classification result, for it can simultaneously handle continuous and categorical variables, with the unbiased processing. The k-fold cross validation method [20] is an effective way to improve the precision of the decision tree model, with the basic idea randomly dividing the samples into the training set and validation set, and circling for k times to ensure the robustness of the model.

In the current study, considering the groundwater potential relating factors, lithology, lineament density, topology, slope, and river density, decision tree algorithms, C5.0 and CART, respectively, with the 10-fold cross validation were applied to generate and test the decision tree models, for predicting the groundwater potential grade in the whole study area.

2. Materials and Methods

2.1. Study Area and Materials. The study area (shown in Figure 1) is located in the southwestern part of Ritu county, Ali city, with the extent of 79°30′–80°20′E and 33°–33°30′N and surrounded by Kailash mountain in the south and Karakoram mountain in the north. The county is in the southeast of Ban Gong Lake Basin across the Ban Gong Lake-Nu Jiang fault zone. The area belongs to the wide valley in the plateau and mountain lake basin and provides runoff to Ban Gong Lake in the southeast direction. The study area is dominated by subfrigid monsoon climate. The temperature ranges from −22.1 to 13.6 Celsius degrees, with the annual average temperature of 0.5 Celsius degrees. The annual sunshine period is 3370.9 hours, with the frost-free period of 95 days. The annual rainfall is quite low being only 75 mm and having high evaporation of 2456.3 mm. The Ban Gong Lake has a maximum depth of 41.3 m and lies in east-west direction, having salt water in the midwest and freshwater in the east. The original plant Ban Gong willow growing along the lake valley helps in soil and water conservation. The lowest depression for catchment travels along the Ban Gong Lake-Nu Jiang fault and the beaded lake basin depression exists between the surrounding mountains. The elevation ranges between 4196 and 6240 m in the study zone, having an area of about 2240 km^2.

The available data sources include geology map with the scale of 1 : 250000 purchased from the National Geological

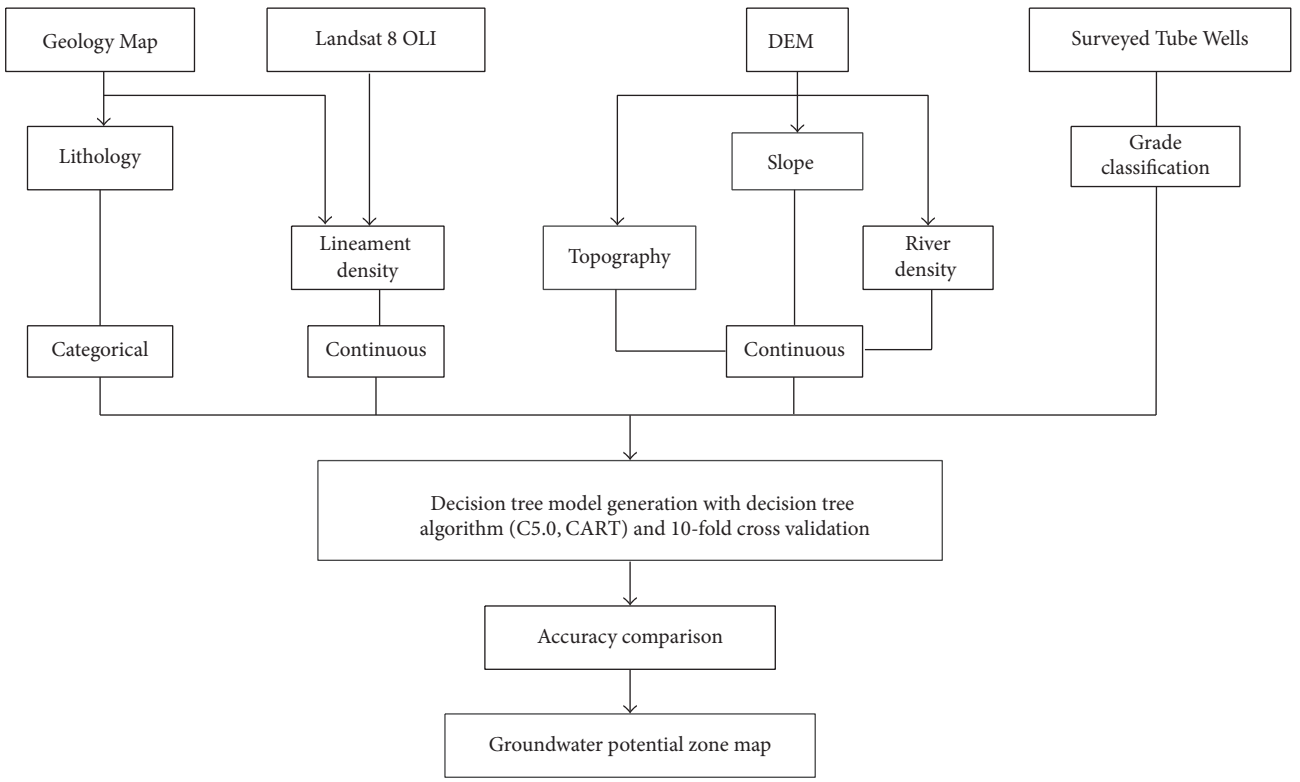

Figure 2: Flow chart for mapping groundwater potential.

Table 1: Groundwater potential grade criteria on the tube well yield.

Grade	Very good	Good	Moderate	Poor
Yield (t/d)	>800	300–800	100–300	<100
Number	21	25	18	16

Library, DEM from ASTER satellite with the horizon distribution of 30 m and Landsat 8 OLI image with the acquisition date 5/22/2013 downloaded from https://www.usgs.gov/ and cloud cover, 2.43%, sun elevation, 68.06, sun azimuth, 119.95, and tube wells yield data with field survey. 80 investigated tube wells were divided into four grades according to the yield (Table 1) [21].

2.2. Methodological Framework. The study on the groundwater potential zone delineation is carried out from the perspective of hydrogeology, considering the occurrence space and supply condition. Lithology and lineament density are chosen as the occurrence space factor; topology, slope, and river density are related to the groundwater supply condition. Lithology is the categorical variable, and the rest of the variables are continuous. After the analysis between the factors and groundwater potential grade, C5.0 and CART algorithms were, respectively, applied to generate the decision tree, and the 10-fold cross validation was adopted to test the classification accuracy. The specific technical route is shown in Figure 2.

2.3. Factors Related to Groundwater Potential. Lithology [22] influences the water-holding capacity of aquifer and directly affects the occurrence and distribution of groundwater. The lithology thematic map was derived through digitizing the 1 : 250,000 scale geology map from the National Geological Library as shown in Figure 3. The Quaternary (Q) including three kinds of sedimentary type, alluvial, diluvia, and lacustrine, with the distribution in low-lying area belongs to the loose sediment for the melting snow water that flows from the high mountains. The Jurassic (J) layers are distributed widely along the east-west direction, with the modern (J3), middle (J2), and early age (J1), respectively, lying in the middle, south, and mid-north. The Modern Cretaceous (K2) spreads along the north-south direction. The Early Cretaceous (K1) is mostly distributed in the north along the east-west. The Paleogene (E) lies mostly in the west, with the scatter distribution in the central and eastern part.

The linear faults, accompanied by the cranny, provide space for the occurrence of groundwater [23]. In the stratum with the same lithology, the intersection of the faults leads to development of the cranny, which tends to be the groundwater enrichment zone with the connectivity enhancement. The linear structures are extracted from the satellite image based on the discontinuity of the color from the surrounding areas. Orthographical correction is applied to Landsat 8 OLI image with DEM to eliminate the shadow's influence on the visual interpretation in the study area. Combination of bands 7/5/3 (SWIR 2/near infrared/green) proved to be the most suitable for the extraction with the geology map in ENVI 5.1.

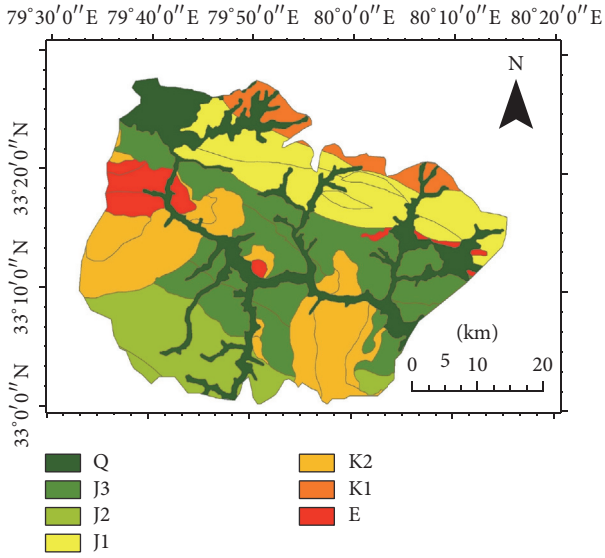

Figure 3: Lithology map of the study area.

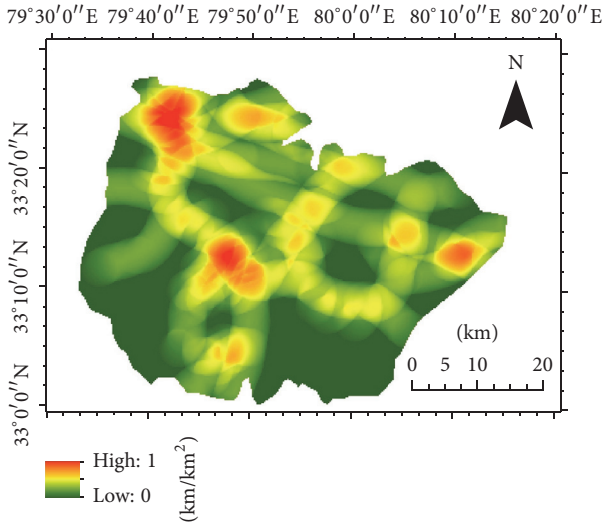

Figure 4: Lineament density map of the study area.

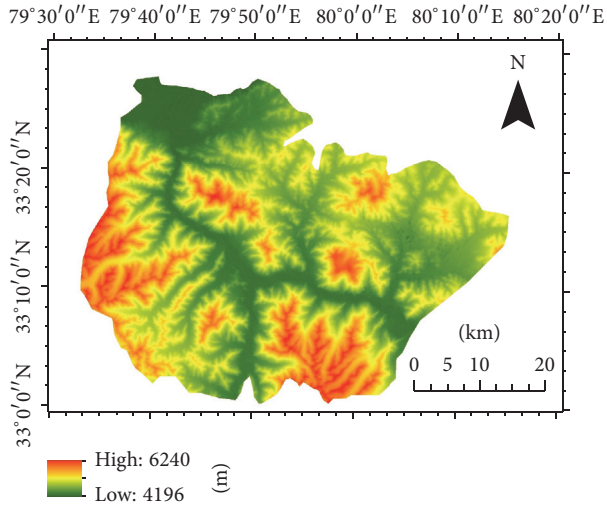

Figure 5: Topography map of the study area.

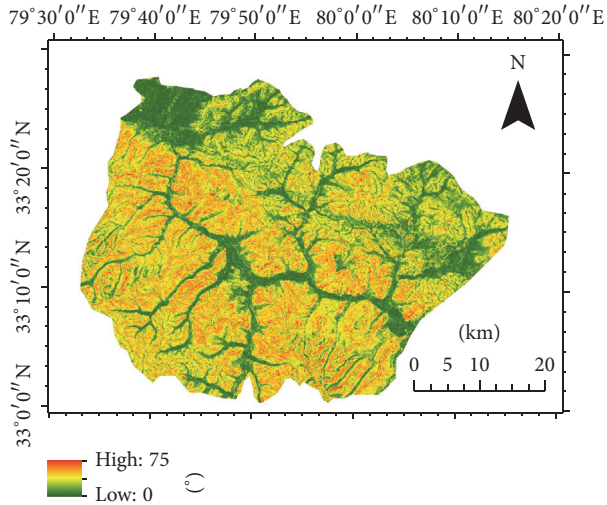

Figure 6: Slope map of the study area.

The lineament density shown in Figure 4 was calculated in ArcGIS 10.1 with "line density" command.

Topography controls the groundwater supply conditions. The mountainous region provides better runoff conditions and most of the precipitation is accounted for in the surface runoff with minimum infiltration to the groundwater. On the other hand, precipitation in plains provides slower runoff and facilitates groundwater recharge. Topography map is shown in Figure 5. The topography map with 30 m spatial resolution was extracted from DEM data in ArcGIS 10.1.

Groundwater flow is usually driven by surface force, and the boundary of the terrain is mostly the boundary of the shallow aquifer. Slope [24] is important in analyzing the terrain, as it can affect the groundwater in terms of its storage, flow, and discharge, especially in mountainous areas. Slope was extracted from the DEM in ArcGIS 10.1 and is shown in Figure 6. In general, slopes control the infiltration and flow ability of the surface water. Usually, the steep slope indicates greater water velocity. Therefore, it is observed that in the areas of steeper relief the runoff increases while minimizing the groundwater recharge. On the contrary, on the relatively gentle sloping terrain, the groundwater potentiality increases due to greater infiltration. Thus, lower slope results in greater recharge.

Flow accumulation reflects the upstream flow quantity. In the study area, the supply source is mainly the melting snow. Flow accumulation was derived from DEM for generating a stream network. It can be seen from Figure 7 that the study area has a dendritic pattern for drainage. The dendritic network is usually found in region underlain with homogeneous surface without abrupt changes in geological conditions.

The river density [25] represents the recharge conditions to quantify the influence caused by surface water, where higher density provides better recharge conditions. Based on

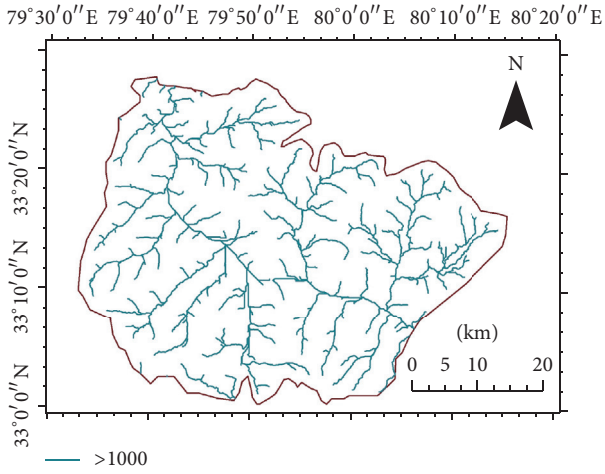

Figure 7: Flow accumulation map of the study area.

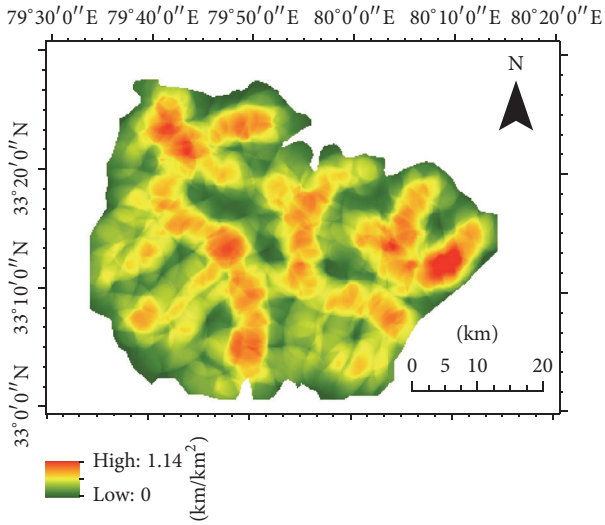

Figure 8: River density map of the study area.

the flow accumulation, the river density was calculated using the "line density" command in ArcGIS 10.1, as shown in Figure 8.

2.4. Research Method. The decision tree algorithms [16] are suitable for the multifactor classification problem. For the mixture of the continuous and discrete factors in the groundwater potential evaluation model, C5.0 and CART decision tree algorithms were adopted together with the 10-fold cross validation method to improve the classification accuracy and unbiasedness.

2.4.1. C5.0. The C5.0 decision tree algorithm is rooted in ID3 and C4.5. The ID3 algorithm [26] with the maximum gain as the division standard aims to achieve the maximum of the information gain in every node, which often gives priority to the variables with more classes. To make up for the defect of ID3, C4.5 adopts the gain-ratio criterion [19, 27, 28]. With the gain-ratio criterion, the binary nodes divide the continuous variable. The processing method for categorical variables is firstly to refer each of the categories as a branch and then merge each two branches iteratively until the two branches remain. However, the heuristic search may not find the best point for the categorical variables division. For continuous variables, C5.0 algorithm can easily find the division point [29]. To get rid of the biasedness on the continuous variables, the algorithm improves the gain of the continuous variables. Besides, C5.0 algorithm can simplify the originally complex decision tree to the equivalent tree, for the easy understanding. With more splitting layers than other algorithms, it can ensure the high purity of the result nodes. C5.0 algorithm achieves the self-correction after several iterations with boosting technology [30]. The artificial methods lay emphasis on the sole impact of various factors; however, C5.0 algorithm can consider all the factors to analyze the comprehensive influence based on data statistics. C5.0 algorithm has been widely applied to the multivariate classification, for its unbiasedness and precision towards the continuous and categorical variables compared to other decision tree algorithms [17, 31–37]. To improve the classification accuracy, boosting technology was applied in C5.0 decision tree algorithm and could adjust the decision tree according to the fault samples until reaching a high precision [38]. Besides, C5.0 algorithm is more applicable to the large data samples.

Assume the splitting node T is expected to separate the $|T|$ samples into K target categories [27]. The symbol $p_T(i)$ stands for the percentage of category i at node T, $i = 1, 2, \ldots, K$. The inclusive information at node T can be expressed as

$$\text{Info}(T) = -\sum_{i=1}^{K} p_T(i) \times \log_2 \left[p_T(i)\right]. \quad (1)$$

For the decision tree branch, the samples are divided into $T_1, T_2, \ldots, T_n$ by the node T, with the subsamples of $|T_1|$, $|T_2|$, and $|T_n|$, respectively. Then, the information can be expressed as

$$\text{Info}(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \text{Info}(T_i). \quad (2)$$

After combining the above information formulas, the information gain can be expressed as

$$\text{Gain}(X,T) = \text{Info}(T) - \text{Info}(X,T). \quad (3)$$

To improve the application of the information gain concept, information gain ratio was put forward:

$$\text{GainRatio}(X,T) = \frac{\text{Gain}(X,T)}{\text{SplitInfo}(X,T)}, \quad (4)$$

where

$$\text{SplitInfo}(X,T) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2 \left(\frac{|T_i|}{|T|}\right). \quad (5)$$

The practice showed that the information gain ratio was more preferred to the continuous variable; therefore, the information gain ratio for the continuous variable with n distinct values should be expressed as [16]

$$\text{Gain}(X,T) = \text{Info}(T) - \text{Info}(X,T) - \log_2 \frac{(n-1)}{|T|}. \quad (6)$$

Table 2: The precision of each loop based on the 10-fold cross validation (%).

	1	2	3	4	5	6	7	8	9	10	Average accuracy
C5.0	83.33	100.00	81.25	89.86	90.25	87.29	94.50	94.21	87.50	96.28	90.45
CART	80.57	89.21	92.23	85.34	81.23	100	80.12	78.24	81.57	82.34	85.09

Table 3: The importance of each factor based on C5.0 and CART algorithm.

	Lithology	Topology	River density	Slope	Lineament density
C5.0	0.363	0.331	0.159	0.117	0.031
CART	0.355	0.308	0.024	0.010	0.312

2.4.2. CART. CART decision tree algorithm [39] can divide the sample set into two subsample sets, making the root and intermediate nodes with two branches based on the recursively binary segmentation technology. CART can handle both continuous and discrete variables, with the impurity level-Gini coefficient as the discriminant basis, considering the probability distribution under the division node.

Assume a total of K classes, variable X, and node for T; then the Gini index is defined as [16]

$$\text{Gini}(T) = \sum_{i=1}^{K} p_T(i)\left(1 - p_T(i)\right) = 1 - \sum_{i=1}^{K} p_T^2(i). \quad (7)$$

When the classes show the equal probability in the node T, the Gini index achieves the maximum $1 - 1/K$; with only one kind in node T, the Gini index achieves the minimum. The Gini index increases with the impurity; therefore, the subnodes should be added to lower the impurity. When taking the misclassification cost matrix into consideration, the Gini index formula becomes

$$\text{Gini}(T) = \sum_{j \neq i} C(i \mid j)\, p_T(i)\, p_T(j), \quad (8)$$

where $C(i \mid j)$ represents the cost for misclassifying the case category j as i. Assuming that the subnode S added to node T, the Gini index can be expressed as

$$\text{Gini}(S,T) = \text{Gini}(T) - p_L \text{Gini}(T_L) - p_R \text{Gini}(T_R), \quad (9)$$

where p_L, p_R stand for the proportion of cases in node T classified into T_L and T_R.

2.4.3. 10-Fold Cross Validation. The basic idea for the k-fold cross validation [40, 41] is to equally divide the surveyed samples into k parts, of which $k - 1$ parts served as the training samples with the remaining one part for validation with k circulations. The method can guarantee every sample acted as the training sample and the verification sample for one time in the circulations. For the k-fold cross validation, the commonly used value for k is 10, called 10-fold cross validation [42].

3. Results and Discussion

In this study, five groundwater relating factors were used in the analysis and the factors except lithology were continuous. 80 tube wells were utilized for training with C5.0 and CART, respectively, in the statistical analysis software SPSS Clementine 12.0 and MATLAB. According to the confusion matrix [43] constructed by the verification result, the kappa coefficient [44, 45] is used to evaluate the accuracy. Based on the 10-fold cross validation with C5.0 and CART algorithm, respectively, the precision of each loop is shown in Table 2. The importance of each factor was calculated firstly to determine the selection order of the factors for classification as shown in Table 3. For the C5.0 algorithm, the importance was 0.363, 0.331, 0.159, 0.117, and 0.031, respectively, for lithology, topology, river density, slope, and lineament density. For the CART algorithm, the importance was 0.355, 0.308, 0.024, 0.010, and 0.312, respectively. After the ten loops, the decision tree with the higher classification accuracy was chosen as optimal. Figure 9 shows the optimal decision tree generated by C5.0 algorithm, with 6 layers, 21 nodes, 10 internal nodes, and 11 terminal nodes. Table 4 shows the rules for the optimal decision tree by C5.0. The optimal decision tree generated by CART algorithm is shown in Figure 10, with 8 layers, 21 nodes, 11 internal nodes, and 10 terminal nodes, and the rules are shown in Table 5.

$$\text{Kappa} = \frac{N \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} x_{i+} x_{+i}}{N^2 - \sum_{i=1}^{r} x_{i+} x_{+i}}, \quad (10)$$

where r is the number of the total categories; N is the total number for verification; x_{ii} is the number of the correct classifications; x_{i+} is the number of samples mistaken from the category i for others; x_{+i} is the number of samples mistaken from other categories for i.

According to the classification rules, we can see that not every rule takes all the variables into account; therefore, for the area lack of the detailed information, the typical factors can help to predict the groundwater potential classification grade. The decision trees show that both the two algorithms can determine the dividing point of the variables, especially for the continuous variables based on the training data, which makes the division of the interval more scientific and reduces the segmentation error, compared with the artificial division.

According to the decision tree result generated by C5.0 algorithm, we did deep analysis. Topology was divided by six nodes: 4196–4301–4316.5–4357–4400.5–6240; and the groundwater was distributed in the low-lying areas. Slope contained two ranges: 0–2.43–75; and the flat areas were beneficial to the surface water infiltration. River density was

Table 4: Rules for groundwater potential grade based on C5.0.

Grade	Rule
Very good	Lithology, Q; 4196 < topology ≤ 4301; lineament density ≤ 0.558
	Lithology, Q; 4196 < topology ≤ 4316.5; lineament density > 0.558
	Lithology, Q; 4316.5 < topology ≤ 4400.5; river density > 0.784; slope ≤ 2.43
Good	Lithology, Q; 4301 < topology ≤ 4316.5; lineament density ≤ 0.558
	Lithology, Q; 4316.5 < topology ≤ 4357; river density ≤ 0.784
	Lithology, Q; 4357 < topology ≤ 4400.5; river density > 0.784; slope > 2.43
	Lithology, Q; 4400.5 < topology; river density > 0.650
Moderate	Lithology, others except Q; river density > 0.506
	Lithology, Q; 4400.5 < topology; river density ≤ 0.650
	Lithology, Q; 4357 < topology ≤ 4400.5; river density ≤ 0.784
Poor	Lithology, others except Q; river density ≤ 0.5

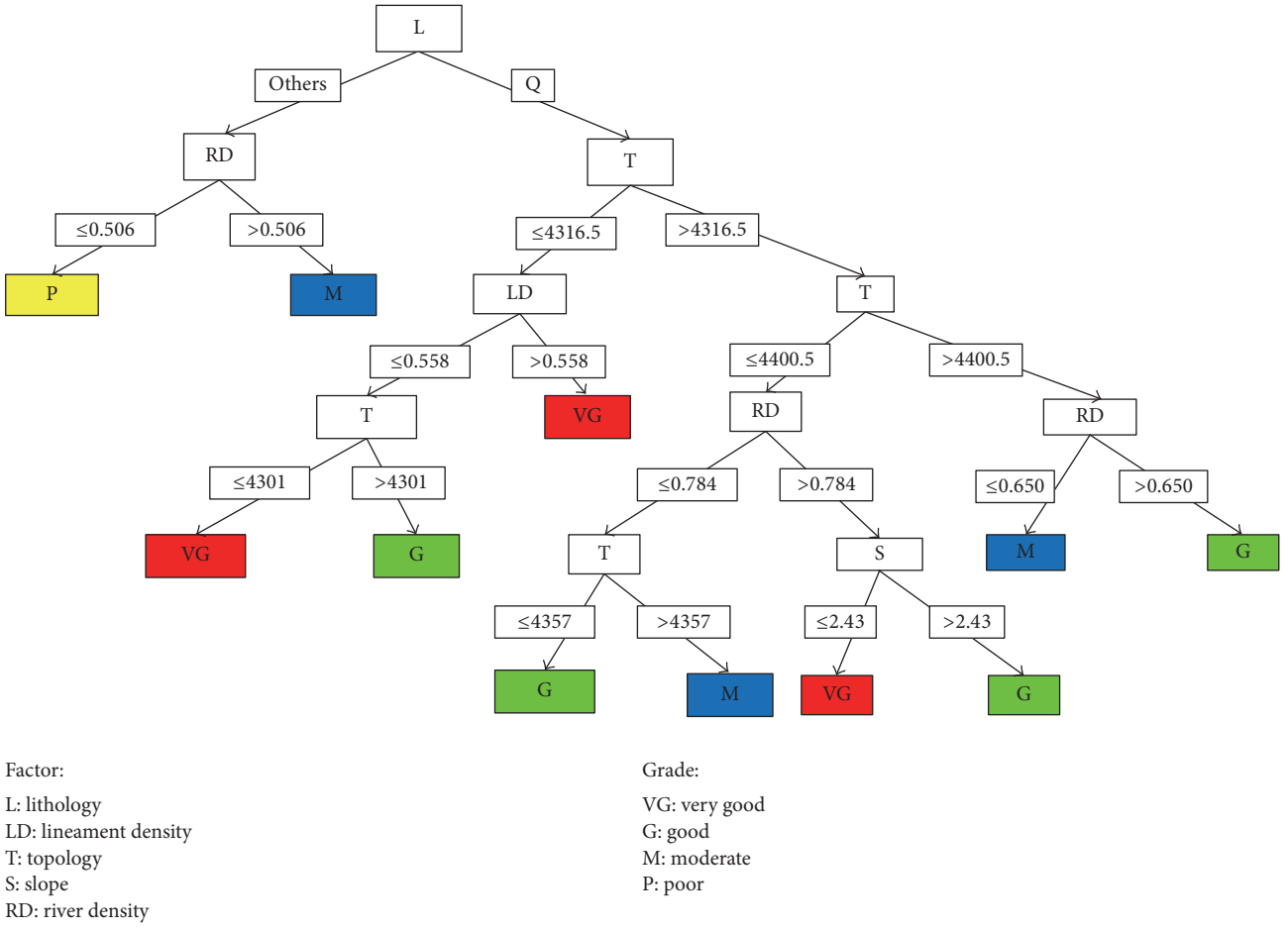

Figure 9: The optimal decision tree generated by C5.0 algorithm.

classified into four intervals: 0–0.506–0.650–0.784–1.14 and reflected the flow capacity in the region, and the higher, the better for groundwater enrichment. Lineament density was divided into two intervals: 0–0.558–1, and the more occurrence space for groundwater existed with the higher lineament density. However, for the CART algorithm, topology was divided into three intervals: 4196–4304–4400–6240; slope contained two ranges: 0–2.386–75; river density was classified into three intervals: 0–0.685–0.703–1.14; lineament density was divided into four intervals: 0–0.296–0.335–0.713–1. After applying the optimal decision trees, respectively, to the whole study area [6, 41], the groundwater potential zone maps were derived and shown in Figures 11 and 12.

According to the results generated by the decision trees, the "very good" area is mostly located in the broad plain zone with a patchy distribution, covering 103.25 km^2 about

Table 5: Rules for groundwater potential grade based on CART.

Grade	Rule
Very good	River density > 0.703; 4304 < topology ≤ 4400; lineament density > 0.713
	Lineament density > 0.296; topology ≤ 4304
	Lineament density > 0.335; 4304 < topology ≤ 4400; river density > 0.703; slope ≤ 2.386
Good	River density < 0.703; 4304 < topology ≤ 4400; lineament density > 0.335
	0.335 < lineament density ≤ 0.713; 4304 < topology ≤ 4400; river density > 0.703; slope > 2.386
	Lineament density > 0.296; topology > 4400; river density > 0.685
Moderate	Lineament density ≤ 0.296; lithology, Q
	0.296 < lineament density ≤ 0.335; 4304 < topology ≤ 4400
	Lineament density > 0.296; topology > 4400; river density ≤ 0.685
Poor	Lineament density ≤ 0.296; lithology, others except Q

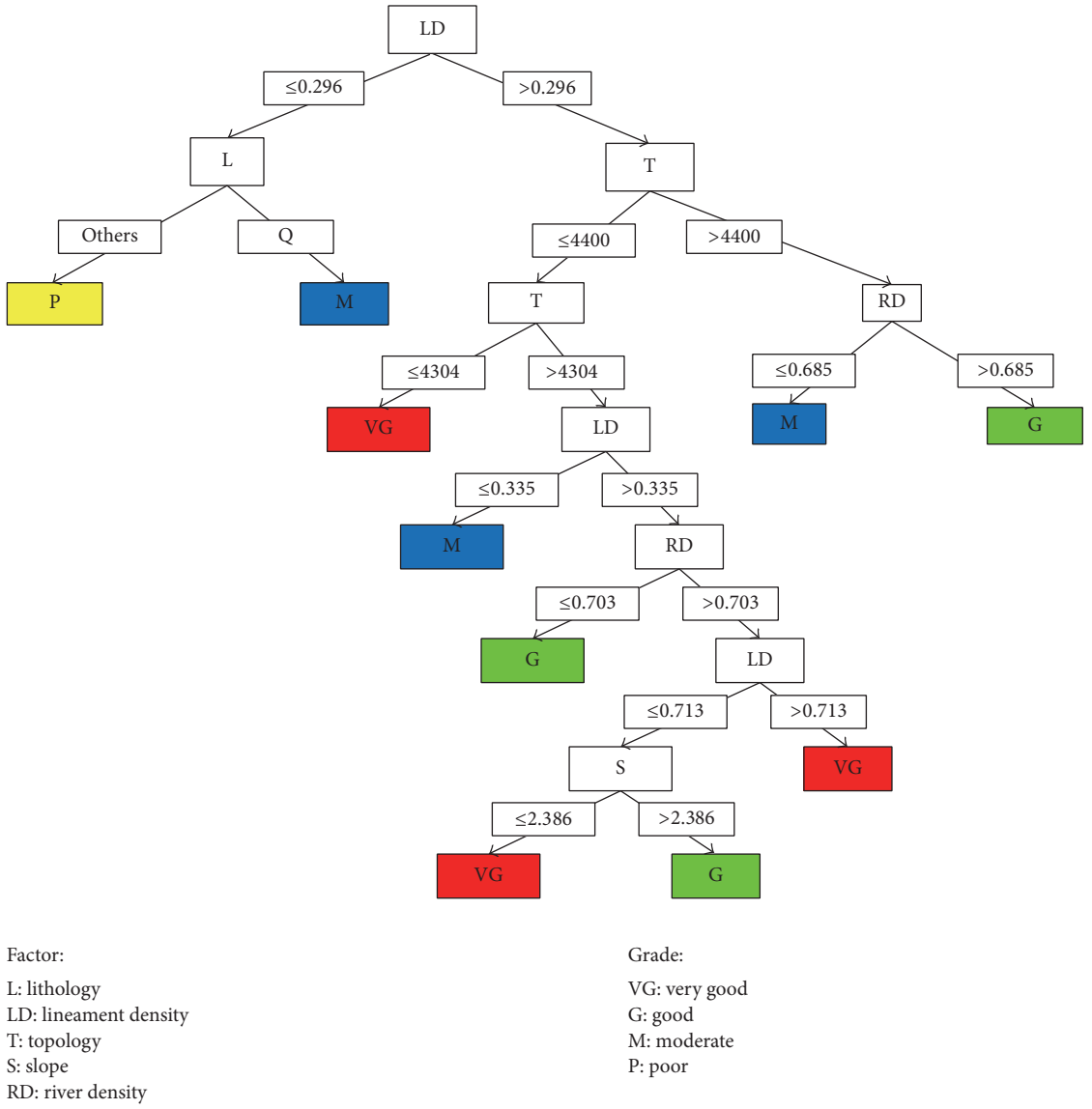

Figure 10: The optimal decision tree generated by CART algorithm.

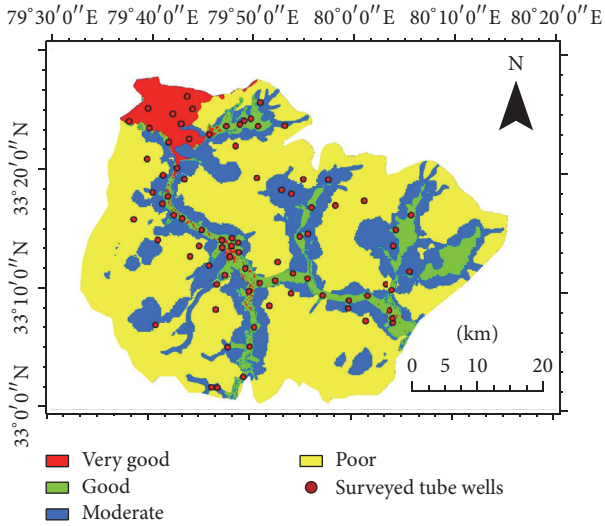

Figure 11: Groundwater potential map of the study area based on C5.0.

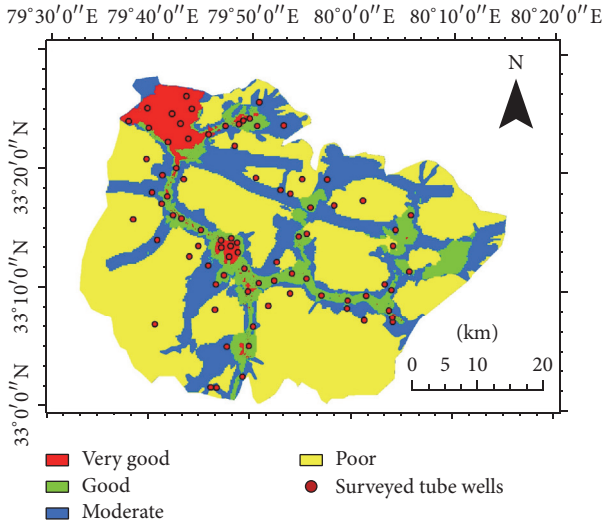

Figure 12: Groundwater potential map of the study area based on CART.

4.61% of the study area for C5.0 algorithm and 105.05 km^2 about 4.68% of the study area for CART algorithm, for the water infiltration into the underground with sufficient time and space. The study shows that low-lying areas with good flow condition, well-developed stratigraphic gap, and strong connectivity like the southwest beaches of Ban Gong Lake can be the first target for groundwater resources. The "good" area was distributed along the river with 192.10 km^2, covering 8.58% for C5.0 algorithm, and 226.02 km^2, about 10.09% of the study area for CART algorithm, just like a long strip. The zones mainly had a planar distribution on the bottom of the diluvia fan on the low mountain and hilly terrain and a banded distribution on the mountain watershed of pluvial valleys, situated upstream and on the periphery of "very good" area, which can serve as the candidate target for groundwater exploration. The "moderate" zone with 595.51 km^2, occupying 26.59% for C5.0 algorithm, and 584.53 km^2, about 26.10% of the study area for CART algorithm, spreads around the upstream tributaries. The zone is located on both sides of the river valleys and the top of the diluvia fan. The "poor" area occupies 60.23%, with the area of 1349.14 km^2 for C5.0 algorithm, and 1324.40 km^2, about 59.13% of the study area for CART algorithm, which is mostly a mountainous region with a high altitude.

4. Conclusions

In this study, C5.0 and CART algorithms were applied for the decision tree generation to predict the groundwater potential zone with the five relating factors and the 10-fold cross validation method was adopted to verify the classification result with the kappa coefficient. From this paper, we can draw some conclusions as follows.

(1) In the study area, the five groundwater relating factors, lithology, topology, slope, river density, and lineament density were appropriate for the groundwater potential grade prediction and the importance based on C5.0 algorithm was 0.363, 0.331, 0.159, 0.117, and 0.03, respectively; for CART algorithm, the importance was 0.355, 0.308, 0.024, 0.010, and 0.312, respectively.

(2) Based on the 10-fold cross validation, both C5.0 and CART could be applied for MCDM with the categorical and continuous variables simultaneously, with the average accuracy of 90.45% and 85.09%, respectively; however, C5.0 algorithm showed higher classification accuracy than CART algorithm.

(3) After applying the optimal decision trees to the whole study area, respectively, the groundwater potential zone map was delineated and the four grades of groundwater potential zones, "very good," "good," "moderate," and "poor," occupied the area of 103.25 km^2, 192.10 km^2, 595.51 km^2, and 1349.14 km^2, with the percentages of 4.61%, 8.58%, 26.59%, and 60.23%, respectively, for C5.0, and for CART the area of 105.05 km^2, 226.02 km^2, 584.53 km^2, and 1324.40 km^2, with the percentages of 4.68%, 10.09%, 26.10%, and 59.13%, respectively.

The study result can provide reference for groundwater exploration and in the future work we will consider more relating factors and survey more wells to enrich the model. The integration of decision tree algorithms and MCDM in our study applies only to the qualitative assessment for the lack of the prior knowledge in the large area; therefore, the extra analysis is needed for the specific point investigation. The accuracy demonstrates that the 10-fold cross validation is suitable for training and verifying the decision tree; however, the tested dataset is limited and more tube wells should be investigated to validate the stability of the model.

Competing Interests

The authors declare that there is no conflict of interests in this research work.

Acknowledgments

This work was supported by Development Program of China: Groundwater Exploration Technology in the Water Shortage Region (863 Program 2012AA062601).

References

[1] D. Oikonomidis, S. Dimogianni, N. Kazakis, and K. Voudouris, "A GIS/Remote Sensing-based methodology for groundwater potentiality assessment in Tirnavos area, Greece," *Journal of Hydrology*, vol. 525, pp. 197–208, 2015.

[2] D. Pinto, S. Shrestha, M. S. Babel, and S. Ninsawat, "Delineation of groundwater potential zones in the Comoro watershed, Timor Leste using GIS, remote sensing and analytic hierarchy process (AHP) technique," *Applied Water Science*, 2015.

[3] O. A. Fashae, M. N. Tijani, A. O. Talabi, and O. I. Adedeji, "Delineation of groundwater potential zones in the crystalline basement terrain of SW-Nigeria: an integrated GIS and remote sensing approach," *Applied Water Science*, vol. 4, no. 1, pp. 19–38, 2014.

[4] J. Mallick, C. K. Singh, H. Al-Wadi et al., "Geospatial and geostatistical approach for groundwater potential zone delineation," *Hydrological Processes*, vol. 29, no. 3, pp. 395–418, 2015.

[5] S. Lee, K.-Y. Song, Y. Kim, and I. Park, "Regional groundwater productivity potential mapping using a geographic information system (GIS) based artificial neural network model," *Hydrogeology Journal*, vol. 20, no. 8, pp. 1511–1527, 2012.

[6] S. Lee and C.-W. Lee, "Application of decision-tree model to groundwater productivity-potential mapping," *Sustainability*, vol. 7, no. 10, pp. 13416–13432, 2015.

[7] A. Ozdemir, "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*, vol. 405, no. 1-2, pp. 123–136, 2011.

[8] A. Ozdemir, "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*, vol. 411, no. 3-4, pp. 290–308, 2011.

[9] D. Machiwal, M. K. Jha, and B. C. Mal, "Assessment of groundwater potential in a semi-arid region of india using remote sensing, GIS and MCDM techniques," *Water Resources Management*, vol. 25, no. 5, pp. 1359–1386, 2011.

[10] S. A. Naghibi and H. R. Pourghasemi, "A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping," *Water Resources Management*, vol. 29, no. 14, pp. 5217–5236, 2015.

[11] D. Zheng-Dong, Y. Xin, L. Fan et al., "Construction and investigation of groundwater remote sensing fuzzy assessment index," *Chinese Journal of Geophysics*, vol. 56, no. 11, pp. 3908–3916, 2013.

[12] O. F. Althuwaynee, B. Pradhan, H.-J. Park, and J. H. Lee, "A novel ensemble decision tree-based CHi-squared Automatic Interaction Detection (CHAID) and multivariate logistic regression models in landslide susceptibility mapping," *Landslides*, vol. 11, no. 6, pp. 1063–1078, 2014.

[13] I. Park and S. Lee, "Spatial prediction of landslide susceptibility using a decision tree approach: a case study of the Pyeongchang area, Korea," *International Journal of Remote Sensing*, vol. 35, no. 16, pp. 6089–6112, 2014.

[14] C. Baker, R. Lawrence, C. Montagne, and D. Patten, "Mapping wetlands and riparian areas using landsat ETM+ imagery and decision-tree-based models," *Wetlands*, vol. 26, no. 2, pp. 465–474, 2006.

[15] R. Kohavi and J. R. Quinlan, "Data mining tasks and methods: classification: decision-tree discovery," in *Handbook of Data Mining and Knowledge Discovery*, pp. 267–276, Oxford University Press, New York, NY, USA, 2002.

[16] N. Lin, D. Noe, and X. He, "Tree-based methods and their applications," in *Springer Handbook of Engineering Statistics*, pp. 551–570, Springer, London, UK, 2006.

[17] I. Klein, U. Gessner, and C. Kuenzer, "Regional land cover mapping and change detection in Central Asia using MODIS time-series," *Applied Geography*, vol. 35, no. 1-2, pp. 219–234, 2012.

[18] J. R. Quinlan, "Improved use of continuous attributes in C4. 5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77–90, 1996.

[19] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques," *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.

[20] K. Polat and S. Güneş, "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems," *Expert Systems with Applications*, vol. 36, no. 2, pp. 1587–1592, 2009.

[21] M. K. Jha, V. M. Chowdary, and A. Chowdhury, "Groundwater assessment in Salboni Block, West Bengal (India) using remote sensing, geographical information system and multi-criteria decision analysis techniques," *Hydrogeology Journal*, vol. 18, no. 7, pp. 1713–1728, 2010.

[22] V. B. Rekha, A. P. Thomas, M. Suma, and H. Vijith, "An integration of spatial information technology for groundwater potential and quality investigations in Koduvan Ár Sub-Watershed of Meenachil River Basin, Kerala, India," *Journal of the Indian Society of Remote Sensing*, vol. 39, no. 1, pp. 63–71, 2011.

[23] O. Rahmati, A. Nazari Samani, M. Mahdavi, H. R. Pourghasemi, and H. Zeinivand, "Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS," *Arabian Journal of Geosciences*, vol. 8, no. 9, pp. 7059–7071, 2014.

[24] K. R. Preeja, S. Joseph, J. Thomas, and H. Vijith, "Identification of groundwater potential zones of a tropical river basin (Kerala, India) using remote sensing and GIS techniques," *Journal of the Indian Society of Remote Sensing*, vol. 39, no. 1, pp. 83–94, 2011.

[25] K. Narendra, K. Nageswara Rao, and P. Swarna Latha, "Integrating remote sensing and GIS for identification of groundwater prospective zones in the Narava basin, Visakhapatnam region, Andhra Pradesh," *Journal of the Geological Society of India*, vol. 81, no. 2, pp. 248–260, 2013.

[26] M. Ture, F. Tokatli, and I. Kurt, "Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2017–2026, 2009.

[27] N. A. AL-Fakhry, "Summarizing data by using data mining techniques a comparative by using C4.5 and C5.0 algorithms," *International Education and Research Journal*, vol. 2, no. 4, 2016.

[28] R. Kohavi and J. R. Quinlan, "Data mining tasks and methods: classification: decision-tree discovery," in *Handbook of Data Mining and Knowledge Discovery* , pp. 267–276, Oxford University Press, 2002.

[29] U. M. Fayyad and K. B. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine Learning*, vol. 8, no. 1, pp. 87–102, 1992.

[30] J. R. Quinlan, "Bagging, boosting, and C4. 5," *AAAI/IAAI*, vol. 1, pp. 725–730, 1996.

[31] S. Pang and J. C. Gong, "C5. 0 classification algorithm and application on individual credit evaluation of banks," *Systems Engineering-Theory & Practice*, vol. 29, no. 12, pp. 94–104, 2009.

[32] M. A. Friedl, C. E. Brodley, and A. H. Strahler, "Maximizing land cover classification accuracies produced by decision trees at continental to global scales," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 2, pp. 969–977, 1999.

[33] C. J. Moran and E. N. Bui, "Spatial data mining for enhanced soil map modelling," *International Journal of Geographical Information Science*, vol. 16, no. 6, pp. 533–549, 2002.

[34] X. Li and A. G.-O. Yeh, "Data mining of cellular automata's transition rules," *International Journal of Geographical Information Science*, vol. 18, no. 8, pp. 723–744, 2004.

[35] P. H. Williams, R. Eyles, and G. Weiller, "Plant microRNA prediction by supervised machine learning using C5.0 decision trees," *Journal of Nucleic Acids*, vol. 2012, Article ID 652979, 10 pages, 2012.

[36] N. Patil, R. Lathi, and V. Chitre, "Customer card classification based on C5. 0 & CART algorithms," *International Journal of Engineering Research and Applications*, vol. 2, no. 4, pp. 164–167, 2012.

[37] M. M. Javidi and E. F. Roshan, "Speech emotion recognition by using combinations of C5.0, neural network (NN), and support vector machines (SVM) classification methods," *Journal of Mathematics and Computer Science*, vol. 6, pp. 191–200, 2013.

[38] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: bagging, boosting, and variants," *Machine Learning*, vol. 36, no. 1-2, pp. 105–139, 1999.

[39] Y. Shao and R. S. Lunetta, "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 70, pp. 78–87, 2012.

[40] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2004.

[41] C. Chen, B. He, and Z. Zeng, "A method for mineral prospectivity mapping integrating C4.5 decision tree, weights-of-evidence and m-branch smoothing techniques: a case study in the eastern Kunlun Mountains, China," *Earth Science Informatics*, vol. 7, no. 1, pp. 13–24, 2014.

[42] M. van der Gaag, T. Hoffman, M. Remijsen et al., "The five-factor model of the positive and negative syndrome scale II: a ten-fold cross-validation of a revised model," *Schizophrenia Research*, vol. 85, no. 1–3, pp. 280–287, 2006.

[43] M. Al Saud, "Mapping potential areas for groundwater storage in Wadi Aurnah Basin, western Arabian Peninsula, using remote sensing and geographic information system techniques," *Hydrogeology Journal*, vol. 18, no. 6, pp. 1481–1495, 2010.

[44] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sensing of Environment*, vol. 37, no. 1, pp. 35–46, 1991.

[45] F. K. Hoehler, "Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity," *Journal of Clinical Epidemiology*, vol. 53, no. 5, pp. 499–503, 2000.