4th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Modeling of groundwater quality based on water quality index using M5P decision tree method

Shokouh Mohsenzadeh[1], Sahar Javidan [1], Mohammad Taghi Sattari [*1]

[1]University of Tabriz, Faculty of Agriculture, Water Engineering Department, Tabriz, Iran

Keywords
Average Water Quality Index
Data Mining
GIS
Groundwater

Abstract

Groundwater is recognized as one of the most important sources of fresh water for agricultural and drinking purposes in the world. About one-third of the world's population needs groundwater to supply drinking water. Therefore, groundwater plays an important role in providing water for consumption in various sectors such as industry, agriculture and drinking. In the present study, first the WQI index was calculated using data related to water quality parameters of 23 wells in Qazvin plain from 2015 to 2018. Then the performance of M5P tree model in estimating groundwater quality in Qazvin plain was evaluated based on WQI index. The results of the model were compared with the indices of Correlation Coefficient (R), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The results showed that the third scenario including TDS, EC, TH with RMSE= 5.2198, MAE= 3.7747 and R= 0.9817 was selected as the best scenario. The mean values of WQI index showed that most of the area of Qazvin plain is allocated to poor quality class.

1. Introduction

In today's world, despite the reduction of surface water resources, attention to groundwater resources has increased, so the groundwater level and their quality has changed (Asghari Moghaddam and Vadiati 2016). In Iran, due to its special location in the arid and semi-arid region, it is important to pay attention to this source and its conditions (Emami et al. 2021). The physical and chemical properties of the aquifer and its composition determine the quality of groundwater. In agricultural issues, knowing the quality of groundwater for agricultural use can help researchers and soil and water managers to make appropriate plans to preserve soil and increase water productivity.

Mokaram et al. (2017) used the Adaptive Network-based Fuzzy Inference System (ANFIS) to predict groundwater quality in the west of Fars province. Based on the results, they found that among the various forecasting models, the hybrid model in the FCM method with the highest R and the lowest error, has the highest accuracy in predicting the groundwater quality of the study area.

Hasani et al. (2018) predicted the groundwater quality class of Khan Mirza plain based on the USSL diagram using the tree decision method. The results of their study showed that only using 4 hydrochemical parameters can determine the water quality class with high accuracy.

Asadi et al. (2020) evaluated and zoned the quality of groundwater in Tabriz aquifer for drinking and agricultural uses. The results showed that about 37% (296 km^2) of groundwater in the study area has high adaptability and the remaining 63% (495 km^2) has moderate adaptation for agricultural use. The trend of WQI and IWQ indices indicates that the groundwater of the region becomes more unfavorable over time.

Trabelsi et al. (2022) used the machine learning index to evaluate the effectiveness of machine learning models to predict the suitability of groundwater in the downstream basin of the Sidi Salem Dam for irrigation purposes. The results showed that Ada Boost model is the most suitable model and ML model is the most cost-effective model. Launching a DSS based on machine learning models enhances the efficient use of water and rationalizes its use by all water stakeholders at the basin level.

The aim of this study was to evaluate the performance of the M5P tree model in estimating groundwater quality in the Qazvin plain based on the WQI index.

* **Corresponding Author**

(Shkmsn2000@gmail.com) ORCID ID 0000 – 0003 – 3982 – 3082
(javidansahar77115@gmail.com) ORCID ID 0000 – 0001 – 6739 – 8242
*(mtsattar@gmail.com) ORCID ID 0000 – 0002 – 5139 – 2118

Cite this study

Mohsenzadeh, S., Javidan, S., & Sattari, M. T. (2022). Modeling of groundwater quality based on water quality index using M5P decision tree method. 4th Intercontinental Geoinformation Days (IGD), 51-54, Tabriz, Iran

2. Method

The study area of Qazvin plain aquifer in Iran with an area of 3733.68 square kilometers. In Qazvin province, the average annual rainfall in the last 10 years is 306.3 mm and the average annual temperature is 14.9 °C, which according to De Martonne classification has a semi-arid climate. About 72% of the necessary expenditures in the agricultural sector are provided from the province's groundwater resources. Alluvial soils of Qazvin region have good permeability and any type of waste produced on the ground penetrates into the ground due to rainfall or proximity to water sources and contaminates groundwater reservoirs that are continuously and widely spread in the area.

In the present study, the quality parameters of 23 wells including Total Hardness (TH), alkalinity (pH), Electrical Conductivity (EC), Total Dissolved Solid (TDS), Calcium (Ca), Sodium (Na), Magnesium (Mg), Potassium (K), Chlorine (Cl), Bicarbonate (HCO_3) and Sulfate (SO_4) have been used in the aquifer of Qazvin plain in the statistical period of 4 years (2018-2015). The spatial location of the studied wells is shown in Figure 1.

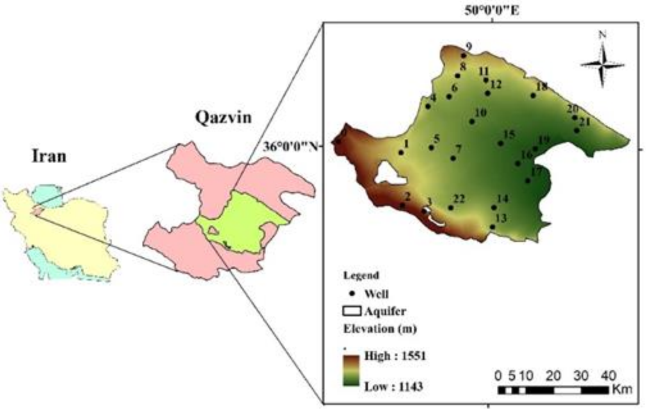

Figure 1. Location of the study area and distribution of sampling wells

The statistical characteristics of the variables used are presented in Table 1.

Table 1. Statistical characteristics of implemented parameters

Statistic	Minimum	Maximum	Mean	Skewness	Kurtosis
pH	5.60	8.70	7.63	-0.94	5.00
TDS	219.00	6296.00	1049.31	2.82	10.29
TH	120.00	1404.50	410.39	1.86	3.88
EC	351.00	9837.00	1645.50	2.80	10.17
K	0.02	0.87	0.08	7.84	82.27
Na	0.55	63.20	8.23	3.06	12.34
Mg	0.63	18.97	3.66	3.04	12.27
Ca	1.23	21.99	4.55	2.61	10.22
SO_4	0.21	26.66	5.30	1.72	3.94
Cl	0.44	58.07	7.56	2.81	9.58
HCo_3	0.50	6.45	3.78	-0.25	1.28

WQI index values were considered as target outputs. Using the correlation method, the types of input compounds (including parameters with a correlation above 0.8) were identified (Table 2). The M5P tree model was used to estimate WQI values. Of the available data, 70% were considered for calibration and 30% for validation.

Table 2. Parameters involved in each scenario

Scenario Number	Input Parameters
1	TDS
2	TDS, EC
3	TDS, EC, TH
4	TDS, EC, TH, Ca
5	TDS, EC, TH, Ca, SO_4
6	TDS, EC, TH, Ca, SO_4, Cl
7	TDS, EC, TH, Ca, SO_4, Cl, Na

2.1. Water Quality Index

Drinking water quality index was calculated using formulas 1 to 3. In these formulas, w is the weight of each parameter due to its importance in drinking and W is the relative weight of each parameter, C is the concentration of each parameter, S is the standard concentration of each parameter, q is the quality rank of each parameter and WQI is the drinking water quality index (Singh 1992).

$$W_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

$$q_i = \left(\frac{C_i}{S_i}\right) \times 100 \tag{2}$$

$$WQI = \sum_{i=1}^{n} W_i q_i \tag{3}$$

Calculated WQI values are usually divided into five categories (Table 3).

Table 3. Water quality classification based on WQI value

Classification of Drinking Water Quality		
WQI Range	Class	Type of Water
below 50	I	Excellent water
50-100	II	Good water
100-200	III	Poor water
200-300	IV	Very poor water
above 300	V	Water unsuitable for drinking

Finally, by plotting the average WQI values with the help of GIS software, it is possible to achieve changes in groundwater quality at the aquifer level. Kriging interpolation method was used to draw this map and estimate the WQI index in non-sampled points. Kriging is a statistical method that uses statistics in addition to mathematical functions to predict unknown points.

Equation 4 shows the general kriging relationship:

$$Z^* = \sum_{i=1}^{N} \lambda_i \, Z(x_i) \qquad (4)$$

Where n is the number of data, Z * is the estimated spatial data value, Z (x_i) is the spatial data value observed at point i, and λ_i is the sample weight x_i, indicating the importance of point i in kriging calculations, and the sum of the coefficients λ_i is 1. Will be (Meng 2020).

A semi-variable is used to determine the spatial relationship of a random variable in the statistics field. Equation 5 is the relation to the criterion of the experimental γ semivariogram (Metron 1963):

$$\gamma(h) = \frac{1}{2\,n(h)} \sum_{i=1}^{n(h)} [Z(x_i) - Z(x_i + h)^2] \qquad (5)$$

n (h) The number of pairs of points in a certain class of distance and direction, Z (X_i) and Z (x_i + h) represent the location of Z and (h) γ the values of the semi-experimental exponential change at distances h.

2.2. M5P Tree Model

The M5P algorithm is a logical reconstruction of the M5 introduced by Wang and Witten in 1997. The M5P tree model has the ability to predict numerically continuous variables from numerical traits, and the predicted results appear as multivariate linear regression models on tree leaves. The criterion of division in a node is based on the selection of the standard deviation of the output values that reach that node as a measure of error. By testing each attribute (parameter) in the node, the expected reduction in error is calculated. The reduction in standard deviation is calculated by Equation (6) (Wang and Witten 1997).

$$SDR = \frac{m}{|T|} \times \beta(i) \times \left[sd(T) - \sum\nolimits_{j \in (L,R)} \frac{|T_j|}{|T|} \times sd(T_j) \right] \qquad (6)$$

In the above relation, SDR decreases the standard deviation, T represents the series of samples that are bound, m is the number of samples that do not have missing values for this attribute, β (i) is a correction factor, TL and TR are sets that are divided by this attribute into There are.

To compare the values obtained from the M5P tree model with the values calculated from the WQI index, the criteria of evaluation of Correlation Coefficient (R), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) were used. The formulas of the above statistics are presented in Equations (7) to (9), respectively:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 . \ \sum_{i=1}^{N}(y_i - \bar{y})^2}} \qquad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \qquad (8)$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |x_i - y_i| \qquad (9)$$

In the above relations, y_i represents the estimated value of the model, x_i represents the value calculated from the qualitative index, and N represents the amount of data.

3. Results

The results obtained from the 7 input scenarios used in the M5P tree model are presented in Table 4.

Table 4. Evaluation criteria for estimating quantitative WQI values

Scenario	M5P		
	R	RMSE	MAE
1	0.9901	7.419	6.7142
2	0.9895	7.3441	6.6159
3	0.9817	5.2198	3.7747
4	0.9598	7.3118	5.2855
5	0.9694	6.9329	4.6595
6	0.9694	6.9329	4.6595
7	0.9694	6.9329	4.6595

According to Table 4, Scenario 3 including TDS, EC, TH with the lowest error rate was selected as the best scenario for estimating WQI values.

A map of the average WQI values is shown in Figure 2.

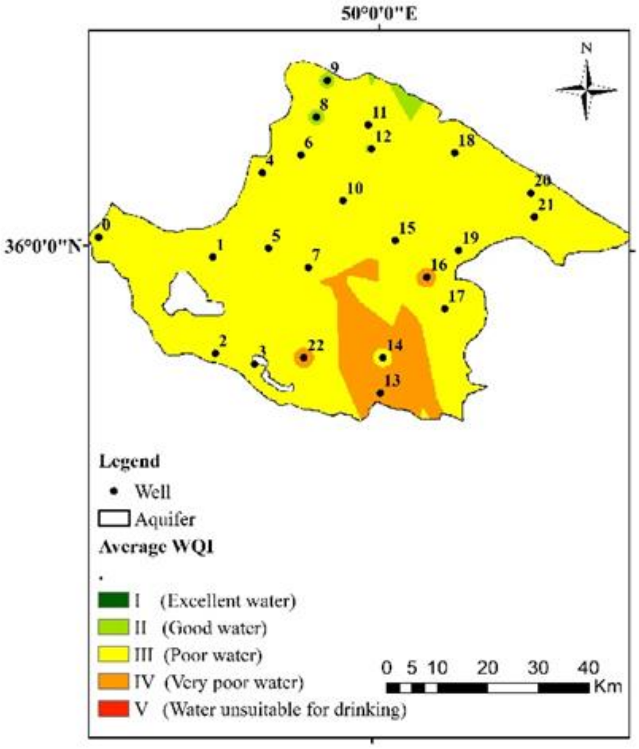

Figure 2. Map of average WQI values in Qazvin plain

According to Figure 2, it can be concluded that most of the area of Qazvin plain has water with poor quality class, also a part of the south of the plain has a very poor-quality class and only a very small part of the north of the plain has good quality water.

4. Discussion

Calculating Water Quality Index with a large number of parameters is time consuming and difficult. However, data-based methods with a very small number of parameters provide more acceptable results, and this increases the popularity of data-based methods. The results of the present study showed that the M5P tree model using the parameters TH, TDS, EC had considerable accuracy in estimating WQI values. In general, the results showed that most of the Qazvin plain has water with poor quality class.

5. Conclusion

In the present study, Water Quality Index using the parameters of Total Hardness (TH), alkalinity (pH), Electrical Conductivity (EC), Total Dissolved Solid (TDS), Calcium (Ca), Sodium (Na), Magnesium (Mg), Potassium (K), Chlorine (Cl), Bicarbonate (HCO_3) and Sulfate (SO_4) were calculated. Then, the M5P tree model was used to estimate the WQI values, considering different scenarios. The results showed that TH, TDS and EC parameters were the most effective parameters in determining the groundwater quality of Qazvin plain. Most of the plain also had a WQI index between 100 and 200, which indicates a poor-quality class.

References

Asadi, E., Isazadeh, M., Samadianfard, S., Ramli, M., Mosavi, A., Nabipour, N., Shamshirband, Sh., Hajnal, E., & Chau, K. (2020). Groundwater Quality Assessment for Sustainable Drinking and Irrigation. Sustainability, 12, 177. doi:10.3390/su12010177.

Asghari Moghaddam, A., & Vadiati, M. (2016). Sarab plain groundwater quality rating for drinking water using entropy method. Water and Soil Knowledge, 26(32), 1–13.

Emami, S., Norouzi sarkarabad, R., & Chopan, Y. (2021). Use of artificial neural network (ANN) and colonial competition algorithm to evaluate the groundwater quality of Jolfa plain for different uses. Journal of Civil Engineering Amirkabir, 53(1), 313–330.

Hasani, Z., Mirabasi najafabadi, R., & Ghasemi, A. (2018). Predicting the groundwater quality of Khanmirza plain for agricultural use using tree decision method. Hydrogeology, 3(1), 99–110.

Meng, J. (2020). Raster data projection transformation based on kriging interpolation approximate grid algorithm. Alexandria Engineering Journal. https://doi.org/10.1016/j.aej.2020.12.006.

Mokaram, M., Mokaram, M., Zarei, A., & Nejadian, B. (2017). Use of Adaptive Neural-Fuzzy Network (ANFIS) to Predict Groundwater Quality in the West of Fars Province from 2004 to 2014. Ecohydrology, 4 (2), 547-559.

Singh, D. F. (1992). Studies on the water quality index of some major rivers of Pune, Maharashtra. Proceedings of the Academy of Environmental Biology, 1, 61–66.

Trabelsi, F., & Bel Hadj Ali, S. (2022). Exploring Machine Learning Models in Predicting Irrigation Groundwater Quality Indices for Effective Decision Making in Medjerda River Basin, Tunisia. Sustainability, *14*(4), 2341. https://doi.org/10.3390/su14042341.

Wang, Y., & Witten, I. H. (1997). Inducing model trees for continuous classes, in Proceedings of the Ninth European Conference on Machine Learning. Prague, Czech Republic: Springer, 128-137.