Karst Groundwater Potential Mapping Using Machine Learning Model

H. Mohammadzadeh[1,2*], J. Hashemi[2], H. Ghalibaf Mohammadabadi [2]

[1]Groundwater and Geothermal Research Group (GRC), Water and Environment Research Institute, Ferdowsi University of Mashhad, Mashhad, Iran

[2]Department of Geology, Faculty of Science, Ferdowsi University of Mashhad, Mashhad, Iran

[*]mohammadzadeh@um.ac.ir & hashemijafar77@gmail.com

Abstract—The objective of this paper is to present a groundwater potential zoning map for the Hezar Masjid highlands, located northeast of Mashhad, using the Random Forest (RF) machine learning model. The zoning map was developed based on the locations of 1,438 springs in the area and 16 factors influencing groundwater potential. The model's performance was assessed using various statistical criteria, including the area under the receiver operating characteristic (ROC) curve (AUC = 0.93), indicating excellent accuracy.

Keyword: Machine learning, groundwater Potential, Hezar Masjed, Karst, Random Forest

INTRODUCTION

With the decline in alluvial groundwater resources, attention has increasingly turned to karst groundwater. Approximately 7 to 12 percent of the Earth's surface crust consists of karst formations [2], and karst watersheds contribute significantly to drinking and agricultural water supplies. Therefore, identifying areas with high karst groundwater potential is both important and necessary.

The findings of Nugroho et al. (2024) on groundwater potential demonstrated that the Random Forest (RF) machine learning model outperforms artificial neural networks and support vector machines [4]. Similarly, in 2024, Ragragui et al. investigated groundwater potential using machine learning and deep learning models, concluding that hybrid models deliver the best performance [5].

geographical location and springs

The study area is located in the northeastern part of Mashhad city in Khorasan Razavi province, spanning a geographical range of 36°07'35" to 37°37'30" north latitude and 58°05'00" to 61°15'35" east longitude. Springs are indicative of locations with maximum groundwater potential in a given area. Therefore, all 1,438 existing springs in the region, along with three times as many non-spring points randomly selected within the area, were used in the model.

MATERIALS & METHODS

A. Selection and analysis of factors affecting groundwater potential:

Due to the complexity of groundwater dynamics, selecting the factors influencing groundwater potential is highly challenging. However, based on the conditions of the study area and findings from previous research, 16 factors were considered Table 1.

Table 2. Groundwater Influencing Factors

Data Layers
Aspect
Slope
Convergence Index
Sediment Power Index (SPI)
Melton Ruggedness Number (MeRugNu)
Multi-Resolution Ridge Top Flatness (MRRTF)
Multi-Resolution Valley Bottom Flatness (MRVBF)
Slope Length (LS)
Lithology
Distance to Faults
Faults Density
Distance to lineaments
Lineaments density
Distance to Streams
Streams density
Normalized Difference Vegetation Index (NDVI)

In factor analysis, multicollinearity refers to the lack of independence among dependent variables in the dataset. The variance inflation factor (VIF) is used to analyse multicollinearity in the dependent variable data “Eq. (1) “.

$$\mathrm{VIF} = \frac{1}{1 - R_k^2} \qquad (1)$$

Where: R_k^2 is the squared error rate for each regression run.

B. Algorithm:

The Random Forest (RF) algorithm, which is based on a collection of Classification and Regression Trees (CART) [1], serves as a powerful tool for analyzing complex relationships between various variables in hydrogeological studies. This algorithm operates by creating multiple decision trees, each trained on a random sample of the data and a random subset of features [3].

A decision tree is an effective machine learning tool used to classify data. It employs a tree structure to break down complex decisions into a series of simpler ones. Each node in the tree poses a question, and each branch represents a possible answer to that question. Ultimately, the leaves of the tree correspond to different classes into which the data is categorized.

The RF model is trained using 75% of the data, and its performance and accuracy are evaluated with the remaining 25%.

C. Model evaluation:

The validation approach is based on calculating four parameters: true positive, true negative, false positive, and false negative. These parameters are determined by evaluating how accurately spring pixels are classified as springs or non-springs in the training and test datasets.

Statistical metrics for model comparison include accuracy, precision, false positive rate (FP-Rate), Matthews correlation coefficient (MCC), root mean square error (RMSE), mean absolute error (MAE), and the Kappa index. Higher values of sensitivity, specificity, accuracy, precision, FP-Rate, and MCC indicate better model performance, especially when RMSE and MAE values are close to zero.

A Kappa index value of 1 indicates a perfect model, whereas a value of -1 signifies an unreliable model. All equations used to calculate these parameters are provided in “Eq. (2)-(11) “.

$$\mathrm{Accuracy} = \frac{\mathrm{TN + TP}}{\mathrm{TP + FPx + TN + TP}} \qquad (2)$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{FP + TN}} \qquad (3)$$

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP + FN}} \qquad (4)$$

$$\mathrm{FPRate} = \frac{\mathrm{FP}}{\mathrm{FP + TN}} \qquad (5)$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP + FP}} \qquad (6)$$

$$\mathrm{MCC} = \frac{\mathrm{TP \times TN - FP \times FN}}{\sqrt{\mathrm{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}} \qquad (7)$$

$$\mathrm{Kappa} = \frac{\mathrm{Accuracy - B}}{1 - \mathrm{B}} \qquad (8)$$

$$\mathrm{B} = \frac{\mathrm{(TP + FN)(TP + FP) + (FP + TN)(FN + TN)}}{\sqrt{\mathrm{TP + TN + FN + FP}}} \qquad (9)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_P - X_A)^2} \qquad (10)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(X_P - X_A)| \qquad (11)$$

RESULTS

Based on the calculations, all 16 selected factors have a variance inflation factor (VIF) below 10, indicating no significant multicollinearity. Therefore, all factors were retained and used for modelling. The resulting map was categorized into five classes "Fig. 1".

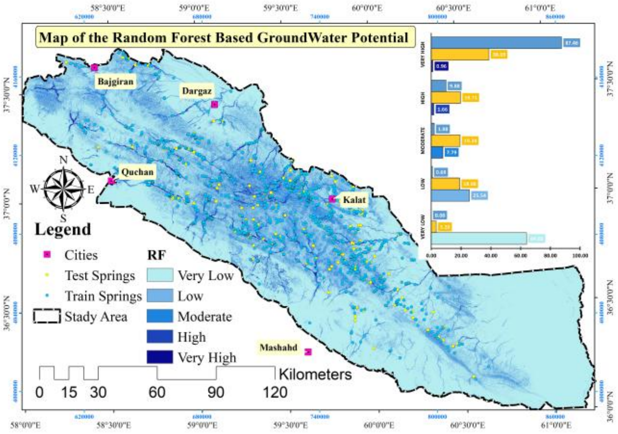

Fig. 1 Groundwater Potential Zoning Map using RF model

As shown, most of the content from the validation and training datasets falls into the high and very high classes. The validation results indicate that the model has achieved an acceptable score in the statistical parameters Table 2. Finally, in terms of the area under the curve, the RF model performed well (AUC = 0.93).

Table 3 Model Metric of RF Model

Model Metrics	RF Model
Accuracy	0.8908
Precision	0.8456
Specificity	0.9616
Sensitivity	0.6667
False Positive rate (FP-Rate)	0.0384
Matthews correlation coefficient (MCC)	0.6850
Root Mean Square Error (RMSE)	0.3304
Mean Absolute Error (MAE)	0.1092
Kappa	0.6773
Area Under Curve (AUC)	0.9300

CONCLUSION

In this study, GIS, remote sensing, and machine learning algorithms have been used to assess groundwater potential. Additionally, the novel aspect of this study is its attempt to integrate as many variables as possible that influence groundwater potential, including geological, topographic, hydrological, climatic, and land cover factors. Sixteen factors were considered, confirming the multicollinearity analysis of their influence and the applicability of these layers for potential detection. The Random Forest model was selected due to its satisfactory results in other regions worldwide. The performance and stability of the model were evaluated using several statistical criteria, which provided very good results for its application in this area. Finally, the developed methodology in this study may be useful for identifying potential groundwater areas, especially in mountainous regions with difficult access and areas where expensive exploration geophysical methods are challenging to apply over large extents.

REFERENCES

Breiman, L. (2001). Random forests. Mach. Learn, 45, 5–32.

Chen, W., Pourghasemi, H.R., & Naghibi, S.A. (2008). A Comparative Study of Landslide Susceptibility Maps Produced Using Support Vector Machine with Different Kernel Functions and Entropy Data Mining Models in China. Bull. Eng. Geol. Environ., 77, 647–664.

Knoll, L., Breuer, L., & Bach, M. (2019). Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci. Total Environ. 668, 1317–1327

Nugroho, J.T., Lestari, A.I., Gustiandi, B., Sofan, P., Prasasti, I., Rahmi, K.I.N., Noviar, H., Sari, N.M., Manalu, R.J., Arifin, S. & Taufiq, A. (2024). Groundwater Potential Mapping using Machine Learning Approach in West Java, Indonesia. Groundwater for Sustainable Development, 101382.

Ragragui, H., Aouragh, M.H., El-Hmaidi, A., Ouali, L., Saouita, J., Iallamen, Z., Ousmana, H., Jaddi, H. & El Ouali, A. (2024). Mapping and modeling groundwater potential using machine learning, deep learning and ensemble learning models in the Saiss basin (Fez-Meknes region, Morocco). Groundwater for Sustainable Development, 26, 101281