Science of The Total Environment
Volume 644, Issue December, 2018, pp. 954-962
DOI: 10.1016/j.scitotenv.2018.07.054

A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination

Farzaneh Sajedi-Hosseini[a], Arash Malekian*[a], Bahram Choubin*[b], Omid Rahmati[c], Sabrina Cipullo[d], Frederic Coulon[d], Biswajeet Pradhan[e,f]

[a] Department of Reclamation of Arid and Mountainous Regions, University of Tehran, Karaj, 31585-3314, Iran.

[b] Department of Watershed Management, Sari Agricultural Sciences and Natural Resources University, P.O. Box 737, Sari, Iran.

[c] Department of Watershed Management, Lorestan University, Iran.

[d] School of Water, Energy and Environment, Cranfield University, Cranfield, MK43 0AL, UK

[e] School of Systems, Management, and Leadership, Faculty of Engineering and IT, University of Technology Sydney, New South Wales, Australia

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

* Corresponding authors, Emails: Malekian@ut.ac.ir; Bahram.choubin@ut.ac.ir

Abstract

This study aimed to develop a novel framework for risk assessment of nitrate groundwater contamination by integrating chemical and statistical analysis for an arid region. A standard method was applied for assessing the vulnerability of groundwater to nitrate pollution in Lenjanat plain, Iran. Nitrate concentration were collected from 102 wells of the plain and used to provide pollution occurrence and probability maps. Three machine learning models including boosted regression trees (BRT), multivariate discriminant analysis (MDA), and support vector machine (SVM) were used for the probability of groundwater pollution occurrence. Afterwards, an ensemble modeling approach was applied for production of the groundwater pollution occurrence probability map. Validation of the models was carried out using area under the receiver operating characteristic curve method (AUC); values above 80 percent were selected to contribute in ensembling process. Results indicated that accuracy for the three models ranged from 0.81 to 0.87, therefore all models were considered for ensemble modeling process. The resultant groundwater pollution risk (produced by vulnerability, pollution, and probability maps) indicated that the central regions of the plain have high and very high risk of nitrate pollution further confirmed by the exiting landuse map. The findings may provide very helpful information in decision making for groundwater pollution risk management especially in semi-arid regions.

Keywords: Groundwater pollution; Nitrate; Probability; Risk; Vulnerability; GIS

1. Introduction

Groundwater is one of the most valuable natural resources especially in arid regions due to negligible rainfall and the scarcity of surface water resources (Neshat et al., 2014; Choubin and Malekian, 2017). Groundwater provides about 63% of drinking water for population of Iran (IMOF, 2014), and it is the single source of drinking water for some large cities and many rural communities. In 2014, groundwater

accounted for about 5% of water withdrawn for public use for cities and about 6% of water withdrawn by self-supplied systems for domestic supply (IMOF, 2014).

A variety of chemicals, including nitrate, can pass through the soil and potentially contaminate groundwater (Hutchins et al., 2018). Beneath the agricultural lands, nitrate is the primary form of nitrogen. It is soluble in water and can easily pass through soil to the groundwater table. Nitrate can remain in groundwater for decades and accumulate to high levels as more nitrogen is used to the land surface every year. Knowing where and what type of risks to groundwater exist can alert water resource managers to protect water supplies.

A number of different approaches including interpolation methods, statistical models, index methods, and process-based models have been applied to assess the status of pollution and vulnerability of groundwater around the world. The first method is geostatistical based techniques which use interpolation methods, such as Kriging methods (Stigter et al., 2006; Narany et al., 2014), to assess the contamination risk in groundwater. These approaches require very dense sampling points and always faced with high uncertainties. The second approach is based on statistical models such as linear and non-linear regressions (Johnson and Belitz, 2009). These methods are able to model the pollution through correlation between pollutant's concentration and various causative parameters (McLay et al., 2001). However, correlation does not imply causality and these models need experts knowledge to make accurate and meaningful predictions. The third group is called index methods, which devote a weight to each factor mostly based on expert's knowledge. Some of these expert methods include susceptibility index (SI) (Van Beynen et al., 2012), DRASTIC method (Aller et al., 1987; Neshat et al., 2014; Majolagbe et al., 2016), GOD method (Foster, 1987), and DRAV model (Zhou et al., 2010). The fourth and most complex approach is process based models such as ground-water flow model (MODFLOW) (Nobre et al., 2007), water flow and nitrate transport global model (WNGM) (Bonton et al., 2011; Qin et al., 2013), pesticide root zone

model (PRZM-3) (Fontaine et al., 1992; Akbar at al., 2011), groundwater loading effects of agricultural management systems (GLEAMS) (Leone et al., 2009; Leonard et al., 1987). The main weaknesses associated with these models are (i) the need for large input data (Iqbal et al., 2012), and (ii) the limited regional scales applicability (Garnier et al., 1998; Anane et al., 2013).

Recently, machine learning (ML) and soft computing techniques such as artificial intelligence have been successfully applied for the prediction of hazard and risk in environmental sciences (Choubin et al., 2017a, 2017b; Ghorbani Nejad et al., 2017; Choubin et al. 2018b; Singh et al., 2018). However, the implementation of ML approaches for assessment of groundwater pollution risk are limited; and an integrated framework for groundwater risk assessment is still lacking. Hence, this study attempts to fill these gaps by proposing an integrated framework for groundwater risk assessment. Therefore, the main objectives of the current study are: (i) comparing the performance of three machine learning models (including two new algorithms for the first time, namely MDA and BRT, and a widely used algorithm, SVM) to map the groundwater pollution occurrence probability, (ii) using ensemble occurrence probability map to assess groundwater pollution risk, and (iii) proposing an integrated framework for groundwater risk assessment.

2. Materials and methods

2.1. Study area

The study area is Lenjanat plain in Isfahan province, in center of Iran, which covers about 1180 km^2. The plain is located between 51° 04′ to 51° 41′ E longitudes and 32° 04′ to 32° 31′ N latitudes (Figure 1). The plain is surrounded by calcareous mountains and elevations of the plain range between 1631 to 2337 m above sea level. The climate type in the study area is arid-cold. The mean annual precipitation is about 160 mm based on the rainfall data recorded during 1971 to 2017, which mostly falls in winter season. The

precipitation data during the sample collection period in April 2016 was about 26 mm. The concentration of precipitation is high in the western and northern part of the region and low in the eastern area. In contrary to the precipitation, the air temperature increases from west to east. The warmest month is August with a mean monthly temperature of 27 ° C and the coldest month of the year is January with a monthly mean temperature of 4 ° C.

The aquifer is isolated from the outside aquifers. Therefore, the main sources for discharging the aquifer are precipitation and the Zayandehrood River. The Zayandehrood River (Figure 1) drains the aquifer from the entrance into the plain to about 26 km, and approximately 17 km before exiting from the plain due to the pumping of groundwater it feeds parts of the aquifer (MOE, 1985). Average groundwater levels of the plain vary in depth from 4 to 120 m during 2000-2017 (available period). The highest depth is in some parts in the north and southeast of the region, and the smallest is near the Zayandehrood River.

The deposits in the study are related to the Permian to Quaternary periods. Bedrock of the aquifer is mostly Jurassic shale and in some areas is Cretaceous limestone. Alluvial deposits on bedrock include clay-marl sediments, which has a higher salt content than other sediments. This layer is without water or low water because of compression, and is considered as the bed rock of the upper alluvial layers. Alluvial deposits deposited on bedrock or clay-marl layer consists of clay, sand and gravel, with clay content varying in different regions and is at least 50%. The evaporated sediments like gypsum and salt crystals are found among alluvium deposits, and the percentage of clay and salt deposits in the lower layers increases (MOE, 1985).

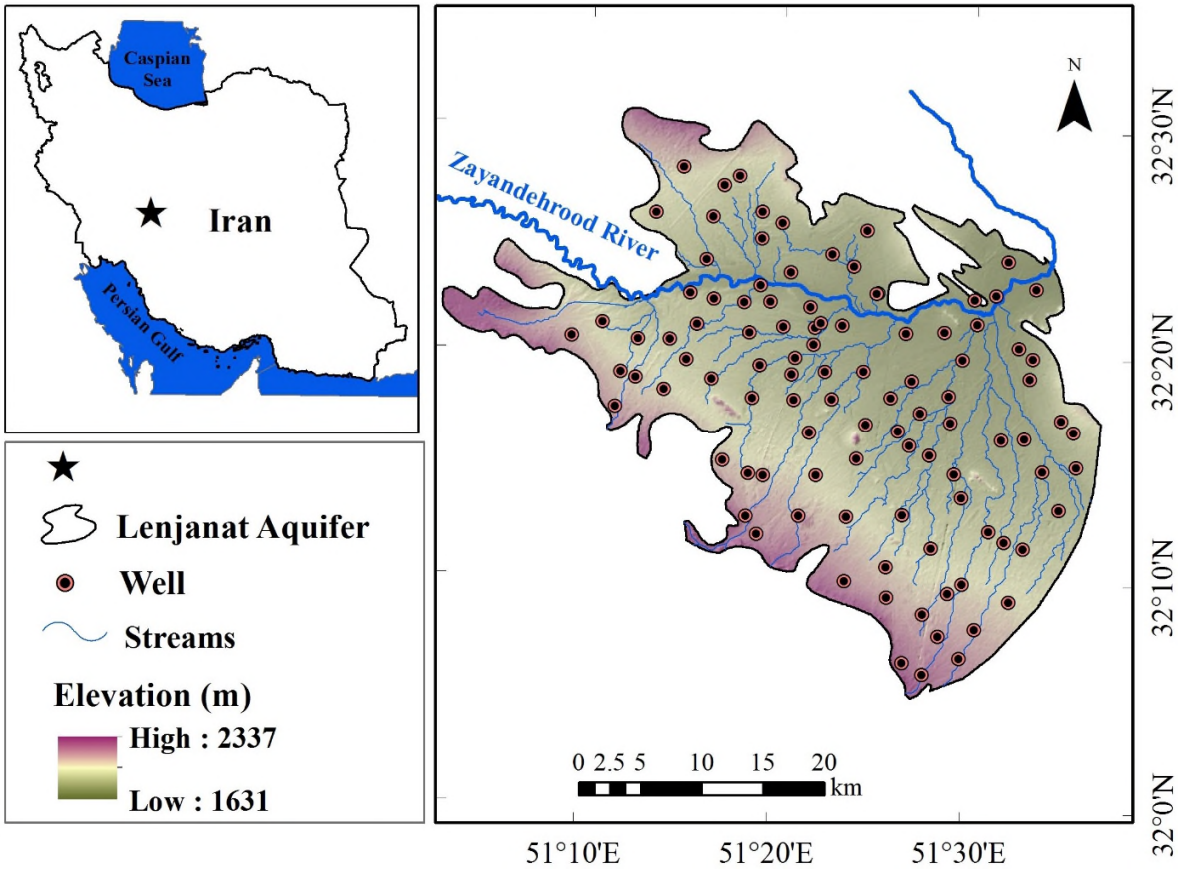

Figure 1: Location of the study area.

2.2. Methodology

The comprehensive methodological framework proposed in this study was constructed based on the following steps: (i) creation of groundwater vulnerability map by DRASTIC model, (ii) preparation of groundwater nitrate pollution map, (iii) creation of pollution occurrence probability map, and (iv) creation of groundwater pollution risk map. A step by step description of methodology is schematically given in Figure 2.

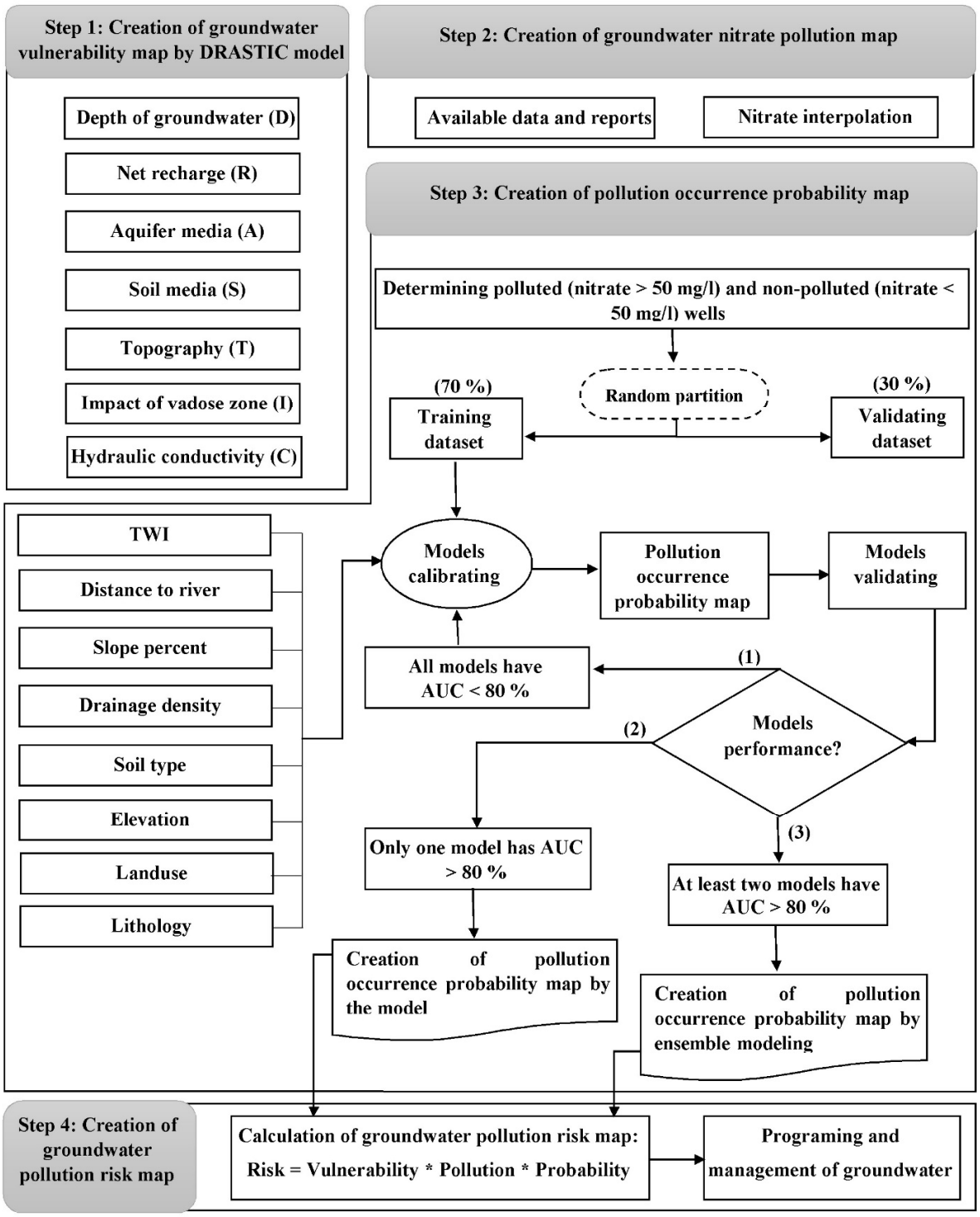

Figure 2: A step by step flowchart of the study. (TWI: topographic wetness index; AUC: area under curve).

2.3. Groundwater vulnerability mapping

In this study, DRASTIC procedure was used as a standard method to analysis the vulnerability of groundwater pollution. This method has been popularly used in many countries (such as in India and Iran; Rahman, 2008; Neshat et al., 2014), to assess the vulnerability of aquifers to pollution, due to the accessibility of input data from various government agencies.

The name DRASTIC corresponds with the first letter of input parameters for modelling vulnerability which includes: depth to water (D), net recharge (R), aquifer media (A), soil media (S), topography (T), impact of the vadose zone (I), and hydraulic conductivity (C) (Aller et al., 1987). More details of these parameters are fully described in Aller et al. (1987) and Rahman (2008).

Parameters are rated (r) from 1 to 10 and then for each parameter a weight (w) from 1 to 5 is assigned. The weight of each parameter is multiplied to its rate and summation of they indicate DRASTIC model as equation 1 (Rahman, 2008):

$$DRASTIC\ Index\ (DI) = Dr\,Dw + Rr\,Rw + Ar\,Aw + Sr\,Sw + Tr\,Tw + I\,rIw + Cr\,Cw \quad (1)$$

Where, w is weight of each parameter, and r is rate of each parameter in the DRASTIC model (Aller et al., 1987). Values of w and r for each parameter can be found in Aller et al. (1987).

2.4. Groundwater nitrate pollution mapping

In order to map groundwater pollution and risk, the nitrate concentration data from 102 wells were collected from the Iranian Water Resources Department (IWRD). Data points had a homogenous distribution (Figure 1) and were collected during April 2016. After obtaining the nitrate concentration in the samples, the Kriging interpolation method was applied for mapping the nitrate pollution in the

Lenjanat plain as represented in Figure 3. That the highest Nitrate concentration was recorded in the center of the (maximum value 172 mg/l), the lowest values were recorded is in the northwest of the plain (10 mg/l) (Figure 3).

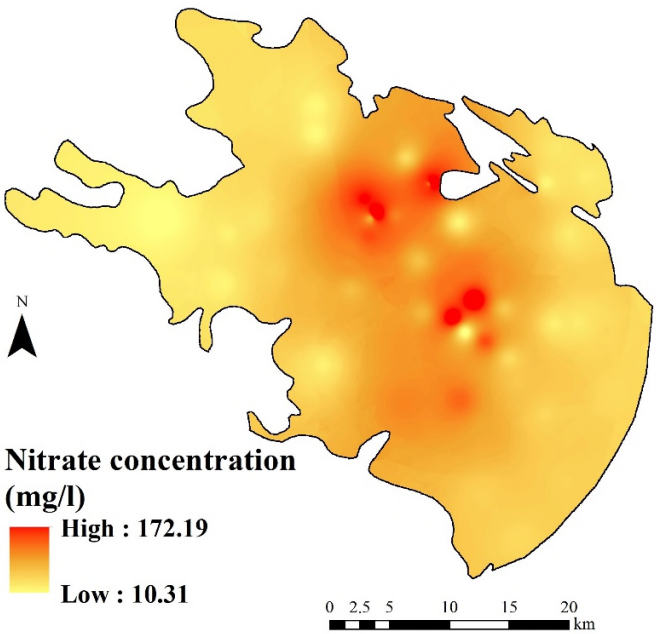

Figure 3: Groundwater pollution mapping of the Lenjanat plain.

2.5. Groundwater pollution occurrence probability mapping

2.5.1. Proposed Framework

Nitrate concentration, obtained from the 102 wells, was used to provide a groundwater pollution occurrence probability map in the Lenjanat plain. According to the World Health Organization, a threshold (50 mg/l) was considered for separating the polluted and non-polluted wells (WHO, 2011). Wells from the study area were classified into: polluted when nitrate concentration were above 50 mg/l, and non-polluted when nitrate concentration were below 50 mg/l. Then, three machine learning models such as Boosted Regression Trees (BRT), Multivariate discriminant analysis (MDA), and Support Vector

Machine (SVM) were used to estimate the probability of pollution occurrence. The dataset was randomly partitioned into subsets where 70% of the data were used to calibrate the models and 30% were used for testing. For models' performance validation the receiver operating characteristic (ROC) curve method and area under curve (AUC) parameter were used (Ozdemir, 2011; Lee et al., 2012; Razandi et al., 2015). Performance threshold at AUC ≥ 80% was established based on Yesilnacar (2005); and the decision making cascade is described below (Figure 2). From the model performance, of the three ML models evaluated in this study, three different possible outcomes have been highlighted:

(i) all the models failed achieving a satisfactory performance (AUC was below threshold value) and were recalibrated until an accuracy of 80 % was achieved.

(ii) only one model achieved a satisfactory performance (over 80 % accuracy), no further recalibration was required and groundwater pollution occurrence probability map was produced by the model selected.

(iii) at least two models achieved a satisfactory performance (accuracy above 80 %), no further recalibration was required and groundwater pollution occurrence probability map was produced by Ensemble modeling which combine the single-models results into a synthesized model to improve the accuracy of modelling (Rokach, 2010). A weighted integration of specific models is mostly used in the ensemble methods (e.g., boosting and bagging) (Pourghasemi et al., 2017). In this study the ensemble modeling was conducted by using the equation (2):

$$EM = \frac{\sum_{i=1}^{n} (AUC_i * M_i)}{\sum_{i=1}^{n} AUC_i} \qquad (2)$$

Where, EM is the ensemble model, AUC_i is the AUC value of the ith single model (M_i).

2.5.2. Groundwater conditioning factors

In order to create the groundwater pollution occurrence probability map, eight conditioning factors (Figure 4) were applied based on the literature survey (Adiat et al., 2012; Park et al., 2014; Rahmati and Melesse, 2016; Golkarian et al., 2018). Selected conditioning factors include: topographic wetness index (TWI), distance to river, slope, drainage density, soil type, elevation, landuse, and lithology (Figure 2). TWI (Figure 4a) indicates situation of soil moisture, it is related to groundwater flow pattern (Rahmati and Melesse, 2016; Sajedi-Hosseini et al., 2018) and was calculated by 30 m pixel size digital elevation model (DEM) in SAGA GIS environment. Slope percent (Figure 4c) was another important factor affecting groundwater recharge and potentiality, and was obtained from DEM. Soil type has effects on infiltration, groundwater recharge, and subsurface flow (Rahmati and Melesse, 2016). The soil types investigated in this study area obtained from IWRD were the followings: Inceptisol, Entisol, and Aridisol (Figure 4e). Elevation was found to vary between 1631 to 2337 m in the study area (Figure 4f). The landuse map was obtained from IWRD, and includes 8 landuse classes: agriculture, rangeland, urban, rock, dry farming, orchard, woodland, and wetland (Figure 4g). Lithology and the geological formations have a key role on the groundwater hydrology. Lithology map was sourced from the Iranian Department of Geology Survey with the scale of 1:100,000. Most area of the plain is covered by Qft2 (Figure 4h). Table 1 describes the lithology of the study area.

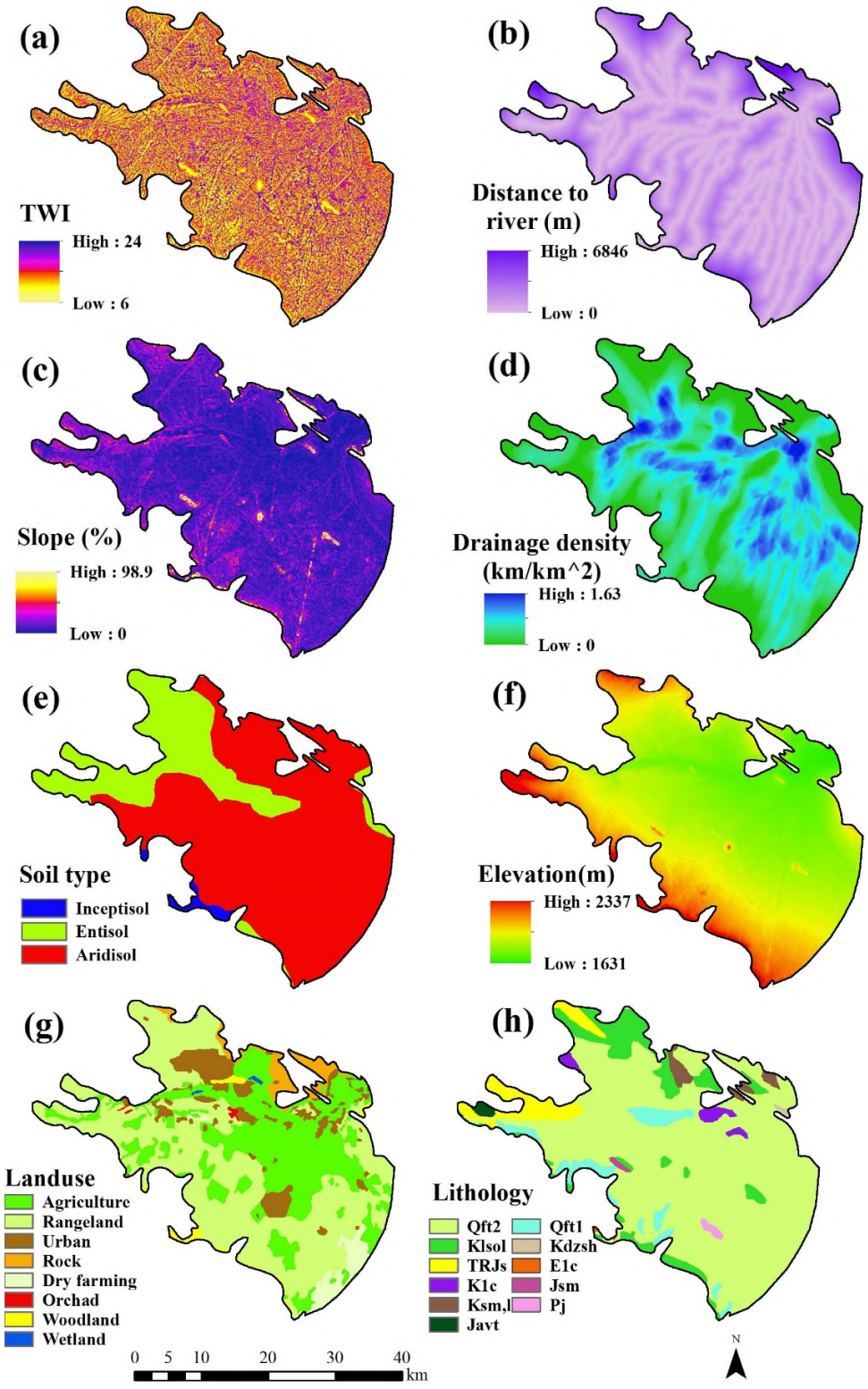

Figure 4: Groundwater conditioning factors: a) topographic wetness index (TWI), b) distance to river, c) slope, d) drainage density, e) soil type, f) elevation, g) landuse, and h) lithology.

Table 1: Lithology of the study area.

Lithology unit	Lithology
Qft1	High level piedmont fan and valley terrace deposits
Qft2	Low level piedmont fan and valley terrace deposits
Klsol	Grey thick-bedded to massive orbitolina limestone
TRJs	Dark grey sandstone and shale
K1c	Red conglomerate and sandstone
Ksm,l	Calcareous shale and marl with intercalations of limestone
Javt	Andesitic volcanic tuff
Kdzsh	Marl, shale, sandstone, and limestone
E1c	Sandstone and polygenic conglomerate
Pj	Thick-bedded, dark-grey, partly reef type limestone
Jsm	Thin bedded argillaceous limestone and marl, thick-bedded to massive dolomitic limestone

2.5.3. Boosted Regression Trees (BRT)

Stochastic gradient boosting, or boosted regression trees (BRT), is a ML technique incorporated with a statistical method (Elith et al. 2006). BRT synthesizes boosting method with classification and regression trees (CART) (Elith et al., 2008). The BRT repeatedly fit many decision trees to improve model accuracy, which facilitate computing a mean from proper rules than to detection of a single prediction rule (Schapire, 2003). More details of BRT is described in the Schapire (2003) and Elith et al. (2008).

2.5.4. Support Vector Machines (SVM)

Another algorithm used in this study was SVM, which first introduced by Vladimir and Vapnik (1995). SVM is one of the most cogent prediction methods based on the structural risk minimization method and the statistical learning theory (Cortes and Vapnik, 1995; Choubin et al., 2018a). In this method, using the bands and an optimization algorithm, samples located in the boundaries of classes are identified and used

to compute an optimal decision boundary. These marginal samples are called support vectors (Cortes and Vapnik, 1995).

SVM algorithm have several standard conversions of kernel function (e.g., nu-svc: nu classification; C-svc: C classification; nu-svr: nu regression; eps-svr: epsilon regression; and etc.) (Choubin et al., 2018a). Selection of kernel type must be based on nature and characteristics of the phenomena (Arabgol et al., 2016). Many studies (e.g., Khalil et al., 2005; Arabgol et al., 2016; Choubin et al. 2018a) demonstrated that radial basis function (RBF) have favorable performance than other kernels in groundwater and hydrology predictions. So, in the current research, nu regression with a popular kernel function named radial basis function (RBF) was applied for production of pollution occurrence probability map.

2.5.5. Multivariate discriminant analysis (MDA)

Multivariate discriminant analysis (MDA) extracts a linear composition by including two or more variables, which are best, at discriminating among pre-determined independent groups (Hair et al., 1998). This process is conducted by maximizing the variance ratio between groups, to the variance within group. MDA derives the linear combinations as below (Hair et al., 1998):

$$Y = w_1 w x_1 + w_2 w x_2 + \dots w_n w x_n \qquad (3)$$

where, X_i (i =1,2,3,..., n) are independent variables, W_i (i =1,2,3,..., n) are discriminant weights, and Y is a discriminant score. More details of MDA model is described in Hair et al. (1998).

In this study, BRT, SVM, and MDA algorithms were implemented in R platform through SDM package (Naimi and Araújo, 2016).

2.6. Groundwater pollution risk mapping

Risk is described as a process to estimate the occurrence possibility in a particular event by a specific set of circumstances (Voudouris, 2009; Neshat et al., 2015). Groundwater pollution risk can be calculated by overlaying the vulnerability, pollution, and probability maps as equation 4 (Dewan, 2013):

$$Risk = Vulnerability * Pollution * Probability \quad (4)$$

The risk evaluation approach that we evaluated was based on previous studies (e.g., Neshat and Pradhan 2015; Kazakis and Voudouris, 2015; Neshat et al., 2015; Shrestha et al., 2016), where vulnerability, pollution, and probability maps were obtained by traditional statistical methods. In this work we propose a novel methodology which applies ML models (SVM, MDA, and BRT) to map the groundwater pollution occurrence probability, and we applied in parallel ensemble occurrence probability map for the assessment of groundwater pollution risk.

3. Results and Discussion

3.1. Groundwater vulnerability assessment

Groundwater vulnerability map (Figure 5) was produced by the DRASTIC model. DRASTIC index (DI) was obtained through equation 1. According to the Civita and De Regibus (1995) and Martínez-Bastida et al. (2010) the groundwater vulnerability map was classified into five classes of very low (DI<80), low (DI=80-120), moderate (DI=120-160), high (DI=160-200), and very high (DI>200). The east and west of study area indicate low and very low vulnerability, whereas the middle areas of the Lenjanat plain show the classes of high and very high. Variations of the nitrate concentration (Figure 3) match the vulnerability map produced by the DRASTIC method (Figure 5).

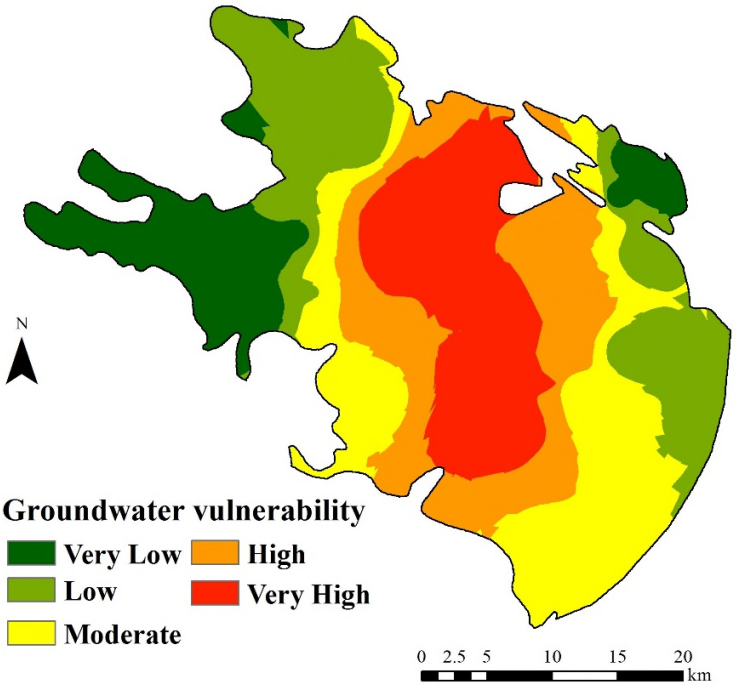

Figure 5: Groundwater vulnerability zones of Lenjanat plain.

3.2. Groundwater pollution occurrence probability results

All the three ML models investigated in this study, used eight conditioning parameters and nitrate for calibration and validation. Model calibration was repeated until a desired AUC value is achieved (>80%) which was also necessary condition to produce map of groundwater pollution. Table 2 indicates results of the models' performance in prediction of the groundwater pollution occurrence probability map. Where AUC ranged from 0.81 to 0.87 for models and corresponding value for ensemble model is equal to 0.89. Kappa statistic is a reliability metric, which ensures that agreement are not occurring by chance, and the models have good performance ($0.55 < K < 0.85$; Monserud and Leemans, 1992).

As shown in Table 2 when the models achieved good performance (more than 80 % accuracy based on the Yesilnacar, 2005) the groundwater pollution occurrence probability maps produced (Figure 6a, b, c,

d). High values of pollution probability map are mostly related with low elevations, agriculture and urban landuses, and high drainage density.

According to the AUC and Kappa values, SVM indicates better performance (AUC= 0.87, Kappa= 0.85, and Mean Square Error, MSE= 0.16) rather than BRT and MDA. BRT (AUC= 0.84, Kappa= 0.84, MSE= 0.23) have higher performance than MDA (AUC= 0.81, Kappa= 0.80, MSE= 0.18) (Table 2). The main reason for poor performance of MDA model than BRT and SVM is related to the nature of the model itself. MDA model is a parametric method which requires that the data follows a normal distribution and it cannot handle the non-linear relationships between input and output variables (Xie et al., 2011). While SVM and BRT models are the non-parametric methods which are unlimited to the data distribution and can deal with the non-linear relationships.

The weighted integration of individual models was carried out based on the AUC values (Table 2) and equation 2. Then, the ensemble pollution occurrence probability map of groundwater was obtained (Figure 6d). The ensemble model indicates good performance as well (with AUC= 0.89, Kappa= 0.87, MSE= 0.16). Also, the mean and variance of the models prediction have been represented in the Table 3. The variance of the BRT and Ensemble models (is equal to 0.05) are lower than MDA and SVM (0.10 and 0.07 respectively) (Table 3).

The advantage of the SVM is good generalization. But capturing the critical variables by SVM is difficult. While MDA model can use variables which are best and critical (Xie et al., 2011) for modeling. Also, BRT combines boosting method with classification and regression trees (CART) (Elith et al., 2008). It fit many decision trees to improve model accuracy, which facilitate computing a mean from proper rules than to detection of a single prediction rule (Schapire, 2003). So, the advantages and disadvantages of models suggest that a combination of predictions from each individual model (i.e., Ensemble modeling)

can often result in better classification than individual models (Dietterich 2000; Lee et al. 2012b; Pourghasemi et al., 2017), as can be seen in this study (Table 2).

Table 2: Performance of the models in prediction of the groundwater pollution occurrence probability map during testing datasets

ML models	AUC	Kappa	MSE
BRT	0.84	0.84	0.23
MDA	0.81	0.80	0.18
SVM	0.87	0.85	0.16
Ensemble	0.89	0.87	0.16

Table 3: The mean and variance of the models in prediction of the groundwater pollution occurrence probability map

ML models	Mean	Variance
BRT	0.46	0.05
MDA	0.57	0.10
SVM	0.45	0.07
Ensemble	0.49	0.05

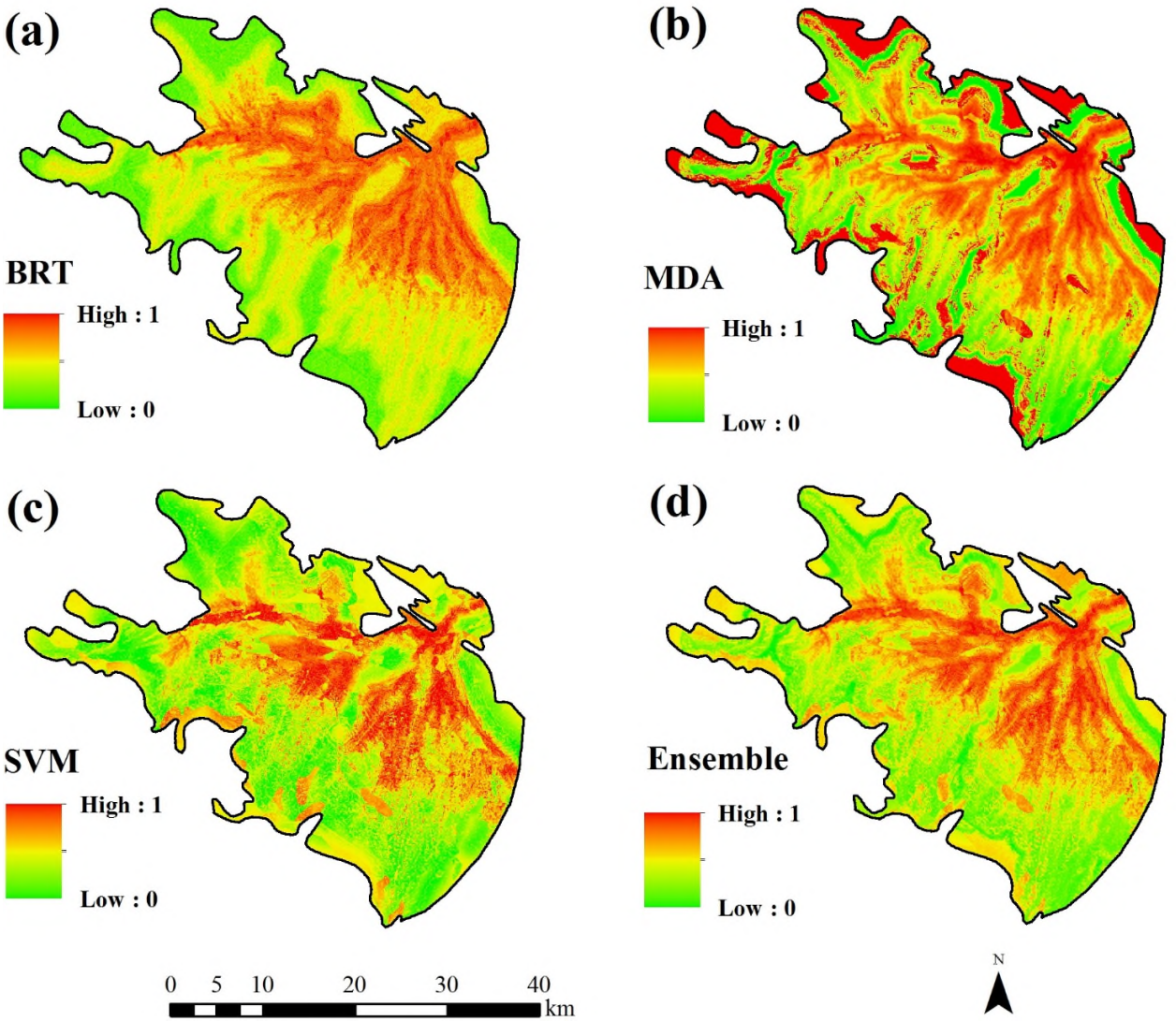

Figure 6: Groundwater pollution occurrence probability map: (a) BRT, (b) MDA, (c) SVM, and (d) Ensemble.

3.3. Groundwater pollution risk assessment

After producing the vulnerability, pollution, and probability maps using equation 4 the groundwater pollution risk map was created in ArcGIS environment. Then, risk map was classified by the equal interval

method into five categories of very low, low, moderate, high, and very high with areas of 171, 467, 232, 193, and 117 km^2, respectively (Figure 7).

It can been seen from Figure 7 that, the middle regions of the plain have high and very high risk of nitrate pollution. Main reason of this is existence of landuses of agriculture and urban in these regions. Main source of irrigation water in the plain is groundwater resources. Inorganic nitrate (like fertilizers) and organic nitrate sources (like human waste, wastewater) have higher nitrogen content causes nitrate pollution in groundwater (Dongol et al., 2005).

According to the landuse map (Figure 4g), it is found that high and very high risk of groundwater pollution mostly cover agriculture and urban areas. This result is in agreement with the Matzeu et al., 2017. Farmers' practices, overusing of chemical fertilizers contain nitrate compounds (like ammonium nitrate and ammonium sulfate), industrial wastewaters, and municipal sewage effluents can be main sources of nitrate (Amiri et al., 2014; Esmaeili et al., 2014). Therefore, according to previous literature and existence of urban area in the high and very high risk regions, industrial wastewaters, and municipal sewage effluents can cause increasing the risk of groundwater pollution in this area.

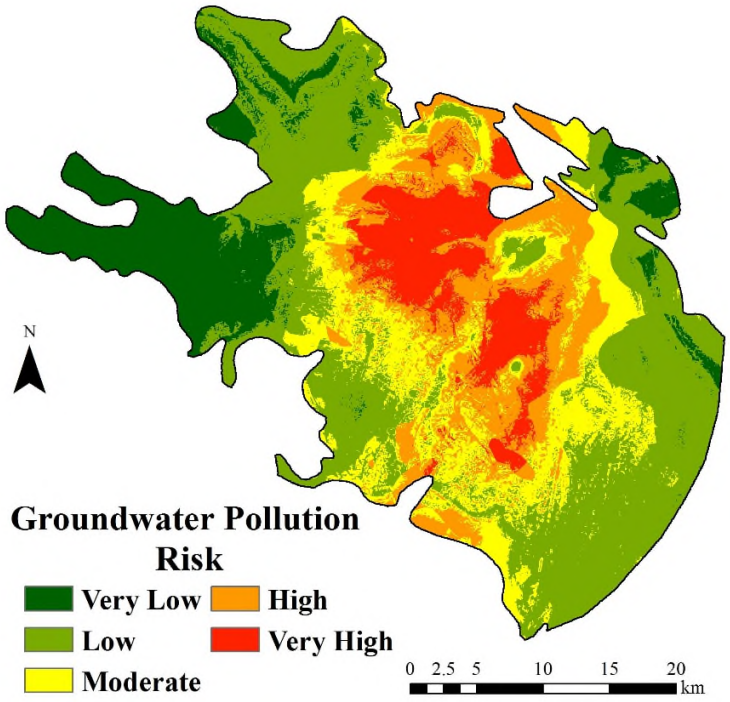

Figure 7: Groundwater pollution risk map of the Lenjanat plain.

This study has a few limitations that cab be addressed in the future studies. According to the available soil map only three soil types were considered. Future studies must be considered more detail soil map and use the soil hydrologic groups as indicating runoff and infiltration rates.

Although machine leaning models in this study indicated good performance but there is some uncertainty that affect results. Groundwater sampling strategy is also important which is another limitation. In this study, due to lack of financial support the sampling was conducted only once (April 2016), without considering the seasonal variations. The seasonal variations on the nitrate concentrations depends on (i) the input nitrate concentration, (ii) the groundwater flow scheme in the aquifer, and (iii) the mean transit time of the aquifer (Jódar et al., 2014). To avoid seasonal effects on the nitrate concentration value in groundwater it would have been a good idea for future studies to take groundwater samples with a given

frequency (monthly, bimonthly or quarterly) depending the hydrogeological setting, and then obtaining an averaged concentration for each sampling point.

The proposed risk contamination assessment method assumes the system as in a steady state, and does not consider contaminant migration through the aquifer that may "disconnect" the link existing between the contamination and the downgradient groundwater sampling points (Xu and Gómez-Hernández, 2016). This issue is very relevant because the estimated pollution occurrence probabilities may be affected.

4. Conclusion

Groundwater pollution risk assessment is a helpful implement for managing the groundwater resource, particularly in arid and semi-arid areas. This study developed a novel framework for assessing the groundwater pollution risk based on the ensemble modeling method. The proposed procedure highlighted that the risk is higher for central part of the plain due to, pollution, probability, and vulnerability maps. Based on the landuse map, it is verified that high and very high risk of groundwater pollution in the plain mostly are in accordance with the agriculture and urban land uses. To manage and control quality of groundwater in study area, it is important to reduce the use of nitrogenous fertilizers in irrigation. Furthermore, to avoid leaching of the soil nitrate, drip irrigation system should be replaced with flood irrigation practice.

Results of the study demonstrate that the development of risk assessment of groundwater pollution by the ensemble probability approach is possible. Our findings is reliable for the groundwater pollution risk assessment in regional scale and can create helpful information for support and understand groundwater pollution risk management decisions in semi-arid regions.

References

Adiat, K. A. N., Nawawi, M. N. M., & Abdullah, K. (2012). Assessing the accuracy of GIS-based elementary multi criteria decision analysis as a spatial prediction tool–a case of predicting potential zones of sustainable groundwater resources. *Journal of Hydrology*, *440*, 75-89.

Akbar, T. A., Lin, H., & DeGroote, J. (2011). Development and evaluation of GIS-based ArcPRZM-3 system for spatial modeling of groundwater vulnerability to pesticide contamination. Computers & geosciences, 37(7), 822-830.

Aller, L., Lehr, J. H., Petty, R., & Bennett, T. (1987). DRASTIC: a standardized system to evaluate groundwater pollution potential using hydrogeologic settings. *National Water Well Association, Worthington, Ohio, United States of America*.

Amiri, V., Rezaei, M., & Sohrabi, N. (2014). Groundwater quality assessment using entropy weighted water quality index (EWQI) in Lenjanat, Iran. *Environmental Earth Sciences*, *72*(9), 3479-3490.

Anane, M., Abidi, B., Lachaal, F., Limam, A., & Jellali, S. (2013). GIS-based DRASTIC, Pesticide DRASTIC and the Susceptibility Index (SI): comparative study for evaluation of pollution potential in the Nabeul-Hammamet shallow aquifer, Tunisia. *Hydrogeology Journal*, *21*(3), 715-731.

Arabgol, R., Sartaj, M., & Asghari, K. (2016). Predicting nitrate concentration and its spatial distribution in groundwater resources using support vector machines (SVMs) model. Environmental Modeling & Assessment, 21(1), 71-82.

Bonton, A., Rouleau, A., Bouchard, C., & Rodriguez, M. J. (2011). Nitrate transport modeling to evaluate source water protection scenarios for a municipal well in an agricultural area. Agricultural systems, 104(5), 429-439.

Choubin, B., & Malekian, A. (2017). Combined gamma and M-test-based ANN and ARIMA models for groundwater fluctuation forecasting in semiarid regions. *Environmental Earth Sciences*, *76*(15), 538. DOI: 10.1007/s12665-017-6870-8.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., & Kløve, B. (2018a). River suspended sediment modelling using the CART model: A comparative study of machine learning techniques. *Science of the Total Environment*, *615*, 272-281. DOI: 10.1016/j.scitotenv.2017.09.293.

Choubin, B., Malekian, A., Samadi, S., Khalighi-Sigaroodi, S., & Sajedi-Hosseini, F. (2017a). An ensemble forecast of semi-arid rainfall using large-scale climate predictors. *Meteorological Applications*, *24*(3), 376-386. DOI: 10.1002/met.1635.

Choubin, B., Solaimani, K., Roshan, M. H., & Malekian, A. (2017b). Watershed classification by remote sensing indices: A fuzzy c-means clustering approach. *Journal of Mountain Science*, *14*(10), 2053-2063. DOI: 10.1007/s11629-017-4357-4.

Choubin, B., Zehtabian, G., Azareh, A., Rafiei-Sardooi, E., Sajedi-Hosseini, F., & Kişi, Ö. (2018b). Precipitation forecasting using classification and regression trees (CART) model: a comparative study of different approaches. *Environmental Earth Sciences*, *77*(8), 314.

Civita, M., & De Regibus, C. (1995). Sperimentazione di alcune metodologie per la valutazione della vulnerabilità degli acquiferi. Atti 2° Conv. Naz.

Collobert, R., & Bengio, S. (2001). SVMTorch: Support vector machines for large-scale regression problems. *Journal of machine learning research*, *1*(Feb), 143-160.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Dewan, A. (2013). *Floods in a megacity: geospatial techniques in assessing hazards, risk and vulnerability*. Springer Science & Business Media.

Dietterich, T.G. (2000). Ensemble methods in machine learning. In International workshop on multiple classifier systems. Springer, Berlin, Heidelberg. 1-15.

Dongol, B.S., Merz, J., Schaffner, M., Nakarmi, G., Shah, P.B., Shrestha, S.K., Dangol, P.M., Dhakal, M.P., 2005. Shallow groundwater in a middle mountain catchment of Nepal: quantity and quality issues. Environ. Geol. 49, 219–229.
Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., ... & Li, J. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 129-151.
Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802-813.
Esmaeili, A., Moore, F., & Keshavarzi, B. (2014). Nitrate contamination in irrigation groundwater, Isfahan, Iran. *Environmental earth sciences*, *72*(7), 2511-2522.
Feng, D., Zheng, Y., Mao, Y., Zhang, A., Wu, B., Li, J., ... & Wu, X. (2018). An integrated hydrological modeling approach for detection and attribution of climatic and human impacts on coastal water resources. *Journal of Hydrology*, *557*, 305-320.
Fontaine, D. D., Havens, P. L., Blau, G. E., & Tillotson, P. M. (1992). The role of sensitivity analysis in groundwater risk modeling for pesticides. *Weed Technology*, 716-724.
Foster SSD (1987) Fundamental concepts in aquifer vulnerability, pollution risk and protection strategy. In: Van Duijevenboden W, Van Waegeningh HG (eds) Vulnerability of soil and groundwater to pollutants, vol 38. TNO Committee on Hydrogeological Research, Proceedings and Information, The Hague, pp 69–86
Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
Garnier, M., Lo Porto, A., Marini, R., & Leone, A. (1998). Integrated use of GLEAMS and GIS to prevent groundwater pollution caused by agricultural disposal of animal waste. *Environmental management*, *22*(5), 747-756.
Ghorbani Nejad, S., Falah, F., Daneshfar, M., Haghizadeh, A. and Rahmati, O. (2017). Delineation of groundwater potential zones using remote sensing and GIS-based data-driven models. Geocarto International, 32(2), pp.167-187.
Golkarian, A., Naghibi, S. A., Kalantar, B., & Pradhan, B. (2018). Groundwater potential mapping using C5. 0, random forest, and multivariate adaptive regression spline models in GIS. *Environmental monitoring and assessment*, *190*(3), 149.
Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (Vol. 5, No. 3, pp. 207-219). Upper Saddle River, NJ: Prentice hall.
Hutchins, M. G., Abesser, C., Prudhomme, C., Elliott, J. A., Bloomfield, J. P., Mansour, M. M., & Hitt, O. E. (2018). Combined impacts of future land-use and climate stressors on water resources and quality in groundwater and surface waterbodies of the upper Thames river basin, UK. *Science of the Total Environment*, *631*, 962-986.
Iqbal, J., Gorai, A. K., Tirkey, P., & Pathak, G. (2012). Approaches to groundwater vulnerability to pollution: a literature review. *Asian Journal of Water, Environment and Pollution*, *9*(1), 105-115.
Iranian Ministry of Energy (IMOF). (2014). Rehabilitation and balance program for groundwater resources. 106pp.
Iranian Ministry of Energy (MOE), 1985. The report of geophysical studies in Mobarakeh (Esfahan). Tehran: Department of Groundwater. 586 pp.
Jódar, J., Lambán, L. J., Medina, A., & Custodio, E. (2014). Exact analytical solution of the convolution integral for classical hydrogeological lumped-parameter models and typical input tracer functions in natural gradient systems. Journal of hydrology, 519, 3275-3289.

Johnson, T. D., & Belitz, K. (2009). Assigning land use to supply wells for the statistical characterization of regional groundwater quality: Correlating urban land use and VOC occurrence. *Journal of Hydrology*, *370*(1-4), 100-108.
Kazakis, N. and Voudouris, K.S., 2015. Groundwater vulnerability and pollution risk assessment of porous aquifers to nitrate: modifying the DRASTIC method using quantitative parameters. Journal of Hydrology, 525, pp.13-25.
Khalil, A., Almasri, M. N., McKee, M., & Kaluarachchi, J. J. (2005). Applicability of statistical learning algorithms in groundwater quality modeling. Water Resources Research, 41, W05010. doi: 10.1029/2004WR003608.
Kumari, S., Jha, R., Singh, V., Baier, K., & Sinha, M. K. (2016). Groundwater Vulnerability Assessment Using SINTACS Model and GIS in Raipur and Naya Raipur, Chhattisgarh, India. *Indian Journal of Science and Technology*, *9*(41).
Lee, M.J., Choi, J.W., Oh, H.J., Won, J.S., Park, I., Lee, S. (2012). Ensemble-based landslide susceptibility maps in Jinbu area, Korea. Environ. Earth Sci. 67, 23–37.
Lee, S., Kim, Y.S. and Oh, H.J., 2012. Application of a weights-of-evidence method and GIS to regional groundwater productivity potential mapping. Journal of environmental management, 96(1), pp.91-105.
Leonard, R. A., Knisel, W. G., & Still, D. A. (1987). GLEAMS: Groundwater loading effects of agricultural management systems. *Transactions of the ASAE*, *30*(5), 1403-1418.
Leone, A., Ripa, M. N., Uricchio, V., Deak, J., & Vargay, Z. (2009). Vulnerability and risk evaluation of agricultural nitrogen pollution for Hungary's main aquifer using DRASTIC and GLEAMS models. *Journal of Environmental Management*, *90*(10), 2969-2978.
Majolagbe, A. O., Adeyi, A. A., & Osibanjo, O. (2016). Vulnerability assessment of groundwater pollution in the vicinity of an active dumpsite (Olusosun), Lagos, Nigeria. *Chem. Int*, *2*(4), 232-241.
Martínez-Bastida, J. J., Arauzo, M., & Valladolid, M. (2010). Intrinsic and specific vulnerability of groundwater in central Spain: the risk of nitrate pollution. *Hydrogeology Journal*, *18*(3), 681-698.
Matzeu, A., Secci, R., & Uras, G. (2017). Methodological approach to assessment of groundwater contamination risk in an agricultural area. *Agricultural water management*, *184*, 46-58.
McLay, C. D. A., Dragten, R., Sparling, G., & Selvarajah, N. (2001). Predicting groundwater nitrate concentrations in a region of mixed agricultural land use: a comparison of three approaches. *Environmental Pollution*, *115*(2), 191-204.
Monserud, R. A., & Leemans, R. (1992). Comparing global vegetation maps with the Kappa statistic. *Ecological modelling*, *62*(4), 275-293.
Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247-259.
Naghibi, S. A., & Pourghasemi, H. R. (2015). 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*, *29*(14), 5217-5236.
Naghibi, S. A., Moghaddam, D. D., Kalantar, B., Pradhan, B., & Kisi, O. (2017). A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. *Journal of hydrology*, *548*, 471-483.
Naimi, B., Araújo, M.B. (2016). sdm: a reproducible and extensible R platform for species distribution modelling. Ecography 39: 368–375.
Narany, T. S., Ramli, M. F., Aris, A. Z., Sulaiman, W. N. A., & Fakharian, K. (2014). Assessment of the Potential Contamination Risk of Nitrate in Groundwater Using Indicator Kriging (in Amol–Babol Plain, Iran). In *From Sources to Solution* (pp. 273-277). Springer, Singapore.

Neshat, A., & Pradhan, B. (2015). An integrated DRASTIC model using frequency ratio and two new hybrid methods for groundwater vulnerability assessment. Natural Hazards, 76(1), 543-563.
Neshat, A., Pradhan, B., & Javadi, S. (2015). Risk assessment of groundwater pollution using Monte Carlo approach in an agricultural region: an example from Kerman Plain, Iran. *Computers, Environment and Urban Systems*, *50*, 66-73.
Neshat, A., Pradhan, B., Pirasteh, S., & Shafri, H. Z. M. (2014). Estimating groundwater vulnerability to pollution using a modified DRASTIC model in the Kerman agricultural area, Iran. *Environmental Earth Sciences*, *71*(7), 3119-3131.
Nobre, R. C. M., Rotunno Filho, O. C., Mansur, W. J., Nobre, M. M. M., & Cosenza, C. A. N. (2007). Groundwater vulnerability and risk mapping using GIS, modeling and a fuzzy logic tool. *Journal of Contaminant Hydrology*, *94*(3-4), 277-292.
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). Journal of Hydrology, 405(1-2), pp.123-136.
Park, I., Kim, Y., & Lee, S. (2014). Groundwater productivity potential mapping using evidential belief function. *Groundwater*, *52*(S1), 201-207.
Pourghasemi, H. R., Yousefi, S., Kornejady, A., & Cerdà, A. (2017). Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. Science of the Total Environment, 609, 764-775.
Qin, R., Wu, Y., Xu, Z., Xie, D., & Zhang, C. (2013). Assessing the impact of natural and anthropogenic activities on groundwater quality in coastal alluvial aquifers of the lower Liaohe River Plain, NE China. *Applied geochemistry*, *31*, 142-158.
Rahman, A. (2008). A GIS based DRASTIC model for assessing groundwater vulnerability in shallow aquifer in Aligarh, India. *Applied geography*, *28*(1), 32-53.
Rahmati, O., & Melesse, A. M. (2016). Application of Dempster–Shafer theory, spatial analysis and remote sensing for groundwater potentiality and nitrate pollution analysis in the semi-arid region of Khuzestan, Iran. *Science of the Total Environment*, *568*, 1110-1123.
Razandi, Y., Pourghasemi, H. R., Neisani, N. S., & Rahmati, O. (2015). Application of analytical hierarchy process, frequency ratio, and certainty factor models for groundwater potential mapping using GIS. *Earth Science Informatics*, *8*(4), 867-883.
Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33, 1–39.
Sajedi-Hosseini, F., Choubin, B., Solaimani, K., Cerdà, A., & Kavian, A. (2018). Spatial prediction of soil erosion susceptibility using FANP: Application of the Fuzzy DEMATEL approach. *Land Degradation & Development*. DOI: 10.1002/ldr.3058.
Schapire, R. E. (2003). The boosting approach to machine learning: an overview. Nonlinear Estimation and Classification, 171,149–171.
Shrestha, S., Semkuyu, D.J. and Pandey, V.P., 2016. Assessment of groundwater vulnerability and risk to pollution in Kathmandu Valley, Nepal. Science of the Total Environment, 556, pp.23-35.
Singh, S. K., Taylor, R. W., Rahman, M. M., & Pradhan, B. (2018). Developing robust arsenic awareness prediction models using machine learning algorithms. *Journal of environmental management*, *211*, 125-137.
Stigter, T. Y., Ribeiro, L., & Dill, A. C. (2006). Evaluation of an intrinsic and a specific vulnerability assessment method in comparison with groundwater salinisation and nitrate contamination levels in two agricultural regions in the south of Portugal. *Hydrogeology Journal*, *14*(1-2), 79-99.

Van Beynen, P. E., Niedzielski, M. A., Bialkowska-Jelinska, E., Alsharif, K., & Matusick, J. (2012). Comparative study of specific groundwater vulnerability of a karst aquifer in central Florida. *Applied Geography*, *32*(2), 868-877.
Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). Support vector method for function approximation, regression estimation and signal processing. In *Advances in neural information processing systems* (pp. 281-287).
Vladimir, V.N., Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York (314 pp).
Voudouris, K., Kazakis, N., Polemio, M., & Kareklas, K. (2010). Assessment of intrinsic vulnerability using DRASTIC model and GIS in Kiti aquifer, Cyprus. *European water*.
World Health Organization, 2011. Guidelines for Drinking-water Quality. Fourth ed. http://www.who.int/water_sanitation_health/publications/2011/dwq_guidelines/en/.
Xie, C., Luo, C., & Yu, X. (2011). Financial distress prediction based on SVM and MDA methods: the case of Chinese listed companies. Quality & Quantity, 45(3), 671-686.
Xu, T., & Gómez-Hernández, J. J. (2016). Joint identification of contaminant source location, initial release time, and initial solute concentration in an aquifer via ensemble Kalman filtering. Water Resources Research, 52(8), 6587-6595.
Yesilnacar, E.K., 2005. The Application of Computational Intelligence to Landslide Susceptibility
Zhou, J., Li, G., Liu, F., Wang, Y., & Guo, X. (2010). DRAV model and its application in assessing groundwater vulnerability in arid area: a case study of pore phreatic water in Tarim Basin, Xinjiang, Northwest China. *Environmental Earth Sciences*, *60*(5), 1055-1063.

A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination

Sajedi-Hosseini, Farzaneh

2018-07-11

Sajedi-Hosseini F, Malekian A, Choubin B, et al., (2018) A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Science of The Total Environment, Volume 644, December 2018, pp. 954-962

https://doi.org/10.1016/j.scitotenv.2018.07.054