Modeling Earth Systems and Environment (2020) 6:1451–1466
https://doi.org/10.1007/s40808-020-00761-6

ORIGINAL ARTICLE

Mapping specific groundwater vulnerability to nitrate using random forest: case of Sais basin, Morocco

Abdelhakim Lahjouj[1,2] · Abdellah El Hmaidi[1] · Karima Bouhafa[2] · M'hamed Boufala[1]

Received: 24 December 2019 / Accepted: 21 March 2020 / Published online: 4 April 2020

Abstract

The objective of this study was to assess the groundwater vulnerability to nitrate (NO_3^-) pollution in the Sais basin, based on the drinking threshold (50 mg/L), using the random forest (RF) model. A spatial dataset consists of the nitrate concentrations observed in 154 water samples and 14 explanatory variables was considered in this research. These variables are rainfall, texture (sand, silt, and clay), lithology, organic matter, piezometric level, altitude, land use, calcium carbonate ($CaCO_3$), carbon/nitrogen ratio (C/N), slope, hydraulic gradient, and soil classification. 80% of the dataset was randomly selected for training and validation, and the remaining 20% for testing the RF model. The RF model was validated and tested using out-of-bag (OOB) error and receiver operating characteristic (ROC) curve. The error computed and the area under the curve for success rate were 0.11 and 82.2%, respectively. In addition, the RF result revealed that rainfall, sand content, clay content, piezometric level, organic matter, and lithology are the key factors determining groundwater vulnerability to NO_3^- in the Sais basin. However, using only these most important factors as RF inputs, the prediction accuracy was found to be slightly similar to that obtained using all variables. The groundwater vulnerability maps were created using the groundwater vulnerability indexes predicted. The most reliable groundwater vulnerability maps to NO_3^- showed that about 48 and 63% of the surface area of the basin are under high to very high vulnerability level, using all and most important explanatory variables, respectively. This study serves to determine the most vulnerable areas and to identify the factors affecting NO_3^- pollution in the Sais basin, to properly control and protect groundwater.

Keywords Groundwater vulnerability · Random forest · Machine learning · Nitrate pollution · Sais basin · Morocco

Introduction

Groundwater is an important natural resource. It constitutes the main source of water for industries and irrigated agriculture in the arid and semiarid areas (Nampak et al. 2014). The effective quality and quantity management of groundwater has become a major issue, since climate change, rapid population increase, and overuse of groundwater for irrigation can have major effects on groundwater. Therefore, to ensure the sustainable management of groundwater, the assessment of groundwater resources and associated pressure at the local scale are strongly required (Hasiniaina et al. 2010).

Nitrate (NO_3^-) is the most abundant pollutant in groundwater (Laftouhi et al. 2003; Moore et al. 2006). Indeed, NO_3^- concentrations increase with increasing and intensification of agricultural activities due to the overuse of nitrogen fertilizers (Nolan 2001; Puckett et al. 2011; Ki et al. 2015). Consequently, the consumption of water polluted by NO_3^- can be associated with health problems, such

✉ Abdelhakim Lahjouj
a.lahjouj@gmail.com

Abdellah El Hmaidi
elhmaidi@yahoo.fr

Karima Bouhafa
bouhafakarima@gmail.com

M'hamed Boufala
simoboufala@gmail.com

[1] Laboratory of Geo-Engineering and Environment, Department of Geology, Faculty of Sciences, Moulay Ismail University, P.O. Box 11202, Zitoune, Meknes, Morocco

[2] Laboratory of Water, Soil, and Plant, National Institute of Agricultural Research, Regional Center of Meknes, km 10, Haj Kaddour Road, P.O. Box 578 (VN), Meknes, Morocco

as methemoglobinemia and cancers for adults (Ward et al. 2005).

In the Sais basin, two aquifers are present: the lias and the plioquaternary aquifers. Their main uses are mainly for drinking and irrigation purposes. These aquifers have been the subject of several geomorphological, geological, hydrogeological, and geophysical studies (Taltasse 1953; Chamayou et al. 1975; Fassi 1999; Essahlaoui et al. 2001; Amraoui 2005). The plioquaternary aquifer is more heavily used for irrigation and drinking of the rural population, due to its shallow depth compared to the lias aquifer.

The geological and hydrogeological characteristics of the Sais basin may contribute positively to the NO_3^- pollution of the plioquaternary aquifer (Tabyaoui et al. 2003). Therefore, the assessment of its vulnerability degree can be an important tool for groundwater resource management, which allows determining the most affected area in the basin or presents a high risk of contamination by NO_3^-.

Groundwater vulnerability is defined as the degree of protection that the natural environment provides against groundwater pollution (National Research Council 1993). In fact, there are two types of groundwater vulnerability: The first type is the intrinsic vulnerability, which is assessed based on the characteristics of the natural environment, including aquifer, soil and climatic characteristics (Schnebelen et al. 2002). However, this type of vulnerability is considered static and invariable. Several methods have been proposed for assessing the intrinsic vulnerability, among others DRASTIC, GOD, SI, and SINTACS frameworks (Ghazavi and Ebrahimi 2015; Al-Shatnawi et al. 2015; Baghapour et al. 2016; El Himer et al. 2013). The second type is the specific vulnerability which concerns a specific pollutant or group of pollutants. This type is assessed using the intrinsic properties of the basin and the characteristics of the pollutant as well as anthropogenic factors related to the pollutant (Ribeiro et al. 2017). The specific vulnerability is assumed to be dynamic and closer to reality. Unlike the first type, the specific vulnerability can changes over time.

In recent years, machine learning techniques such as artificial neural network (ANN), support vector machine (SVM), random forest (RF), and decision tree (CART) have been applied in several fields. The RF model is robust and easy to apply compared to other machine learning techniques, it has the particularity to determine the importance of each explanatory variable in the prediction result. Besides, the RF model can provide good results compared to the multivariate statistics or other machine learning methods such as SVM and ANN (Breiman 2001; Liaw and Wiener 2002; Loosvelt et al. 2012; Ouedraogo et al. 2018).

In groundwater research, RF method has been used to predict NO_3^- and arsenic (As) concentrations in groundwater (Anning et al. 2012; Wheeler et al. 2015) and to assess the groundwater vulnerability (Rodriguez-Galiano et al. 2014; Mendes et al. 2016). These studies revealed that RF has a good prediction performance.

To the best of our knowledge, there are no previous studies that have assessed groundwater vulnerability using machine learning in Morocco. Furthermore, no study aimed to assess the specific groundwater vulnerability to NO_3^- in the Sais basin. However, Sadkaoui et al. (2013) have applied intrinsic methods in the Sais basin to assess groundwater vulnerability. Nevertheless, the rating proposed by some intrinsic frameworks such as DRASTIC (Aller et al. 1987) may differ depending on the study area specificities. Additionally, intrinsic vulnerability may ignore some important parameters which may affect the groundwater vulnerability. Consequently, the RF model may be a novel technique for the groundwater vulnerability assessment in Morocco.

The main objective of this study was to develop an accurate RF model to assess the specific groundwater vulnerability to NO_3^- of the plioquaternary aquifer of the Sais basin, using 14 parameters that may contribute to NO_3^- pollution.

The output of this research will contribute to:

1. Identify the most vulnerable areas to NO_3^- pollution in the Sais basin;
2. Determine the most important factors that control the groundwater vulnerability to NO_3^- pollution of the plioquaternary aquifer.

Materials and methods

Research area

The research area is the Sais basin, part of the Fez-Meknes region in Morocco (Fig. 1). The surface area of the basin is approximately 2100 km^2. The basin is located between the latitude 33°38′ to 34°4′N and longitude 5°49′ to 4°53′W. It is limited by the middle atlasic ranges in the south and the rife ranges in the north (Fig. 2). The geological setting is mainly dominated by the lacustrine limestone of the lias. The altitude of the study area varies between 185 m in the north and 1047 m in the south at the middle atlas ranges, with an average of 600 m. The study area is characterized by a Mediterranean climate (Amraoui 2005). The mean annual rainfall recorded by three stations located at Douyet (Northeastern of the basin), Meknes and Ain Taoujdate during the period 1981–2018 is 468 mm. The Sais basin is characterized by high agricultural activity due to good soil fertility. The agriculture is conducted under rainfed and irrigated conditions. Moreover, the Sais basin contains several lithological classes, including sandstone, siltstones, marlstone, alluvium and oncolite limestone, representing, respectively, 39, 18.4, 18, 11.6, and 10.6% of the total surface area of the basin.

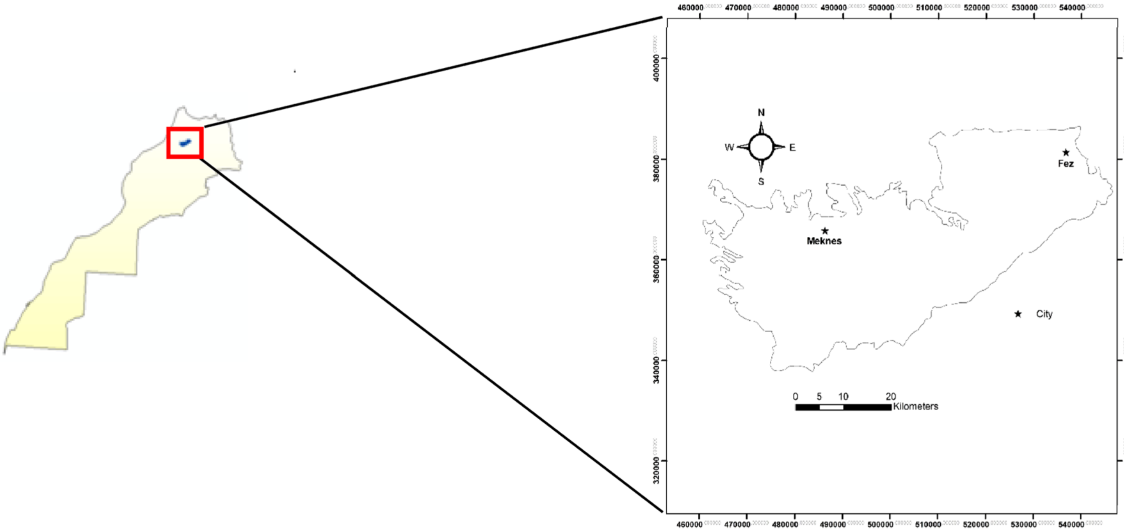
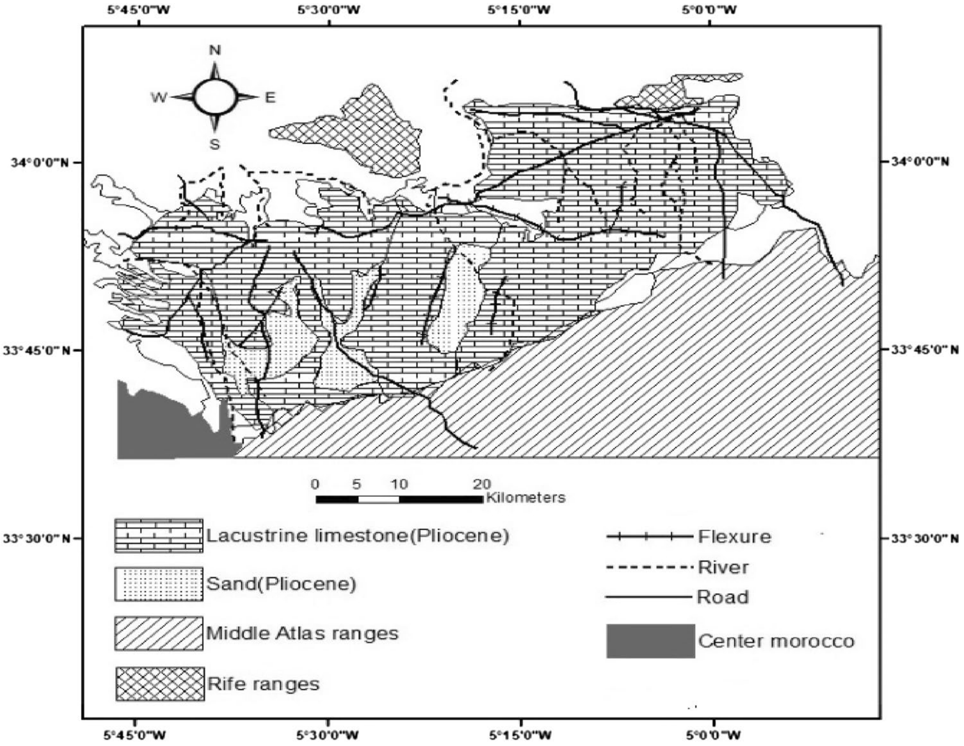


Fig. 1 Geographic location of the Sais basin

Fig. 2 Geological map of Sais basin



In the Sais plain, two aquifers are distinguished: lias aquifer which constituted by dolomitic limestone, and plioquaternary aquifer, composed of pliovillafranchien sandstone, conglomerate sand, and lake limestone (Essahlaoui et al. 2001; Tabyaoui et al. 2004; Amraoui 2005; Belhassan et al. 2010), the latter has a substratum from the upper Miocene with a depth exceeds 1000 m in some parts of the northeastern Sais basin. The recharge of the plioquaternary aquifer is done mainly by rainfall and irrigation water infiltration, as well as by the drainage of the lias aquifer from the southern part of the Sais basin (Sadkaoui et al. 2013).

Random forest

Random forest is a supervised nonparametric machine learning method, developed by Breiman (2001). It is based on multiple trees' decision algorithm (Rodriguez-Galiano et al. 2014; Catani et al. 2013; Micheletti et al. 2013). The method is used for data prediction and interpretation purposes. The RF model can be divided into a classification tree and a regression tree (Zabihi et al. 2016).

The RF can compute an unbiased error estimated by bootstrapping (Siroky 2009). The dataset used for RF is divided into two parts: training sub-dataset containing 2/3 of dataset randomly chosen with replacement, and validation sub-dataset containing the remaining 1/3. The validation sub-dataset is called out-of-bag (OOB) (Breiman 2001; Catani et al. 2013). The latter can be used to assess the prediction performance of RF and the input variables importance. In addition, RF presents some other interesting characteristics which justify its application in the groundwater vulnerability assessment:

- It can manage both categorical and numerical variables;
- It can determine the importance of each explanatory variable in the prediction result;
- It can learn complex patterns, without a linear relationship between the explanatory variables and dependent variable;
- It can handle outliers' data;
- It can handle a large dataset with high dimensionality;
- Its implementation is less complex compared to other machine learning techniques such as ANN and SVM.

The RF model uses two methods to assess the importance of explanatory variables used in the prediction. The first one is called the mean decrease accuracy (MDA), which is an indirect measure of the effect of each explanatory variable on the prediction accuracy (Calle and Urrea 2010). To compute MDA, RF uses the out-of-bag (OOB) dataset and permute each explanatory variable while others are fixed. Increasing the RF model error percentage indicates that the permuted variable is important (Naghibi et al. 2017). The RF model error is calculated from OOB sub-dataset based on the following formula (Grömping 2009):

$$\mathrm{OOB} - \mathrm{MSE} = \frac{1}{n\mathrm{OOB}, t} \sum_{\substack{i=1 \\ t \in \mathrm{OOBt}}}^{n} (yi - \hat{y}i)^2 \tag{1}$$

where yi and $\hat{y}i$ are, respectively, the observed and the mean of the predicted values from all trees; nOOB is the number of OOB observations in tree t and i is the OOB observation for the tree. Therefore, MDA can be an accurate tool for variable selection.

The second method is the mean decrease in the GINI, based on the heterogeneity decrease defined from the entropy. This tool determines the importance of explanatory variable j. It is the weighted sum of the decreases in the node heterogeneity, averaged over all trees using the GINI index. The GINI index can be used to explain the variable strength used as input in the RF model (Al-Abadi and Shahid 2016). The higher GINI value assigned to a variable indicates that it is more important in the prediction compared to other variables (Yang et al. 2019).

Observed nitrate concentrations

A total of 154 water samples of the plioquaternary aquifer in the rural area of the Sais basin were collected for NO_3^- analysis. Sampling campaigns were carried out in the spring and autumn seasons of 2013 (56 samples) and 2018 (98 samples). The samples were collected and stored at 2–4 °C and then analyzed within 24 h using the UV-Spectrophotometeric method. The distribution of observed NO_3^- concentrations in the different sampling campaigns is shown in Fig. 3. The mean NO_3^- concentrations were 60 and 64 mg/L in 2013 and 77 and 70 mg/L in 2018, respectively, in the Spring and Autumn season. Overall, the highest NO_3^- concentrations were observed in the north, northwestern and central parts of the basin.

Explanatory variables

In order to assess the groundwater vulnerability to NO_3^- in the Sais basin, a total of 14 explanatory variables related to the intrinsic and specific groundwater vulnerability to NO_3^- were used as RF model inputs (Fig. 4). All variables were mapped using geographic information system (GIS). Table 1 presents the 14 explanatory variables, their data sources, and their estimations methods. These variables are rainfall, texture (sand, silt, and clay), lithology, organic matter, piezometric level, altitude, land use, Calcium carbonate ($CaCO_3$), Carbon/nitrogen ratio (C/N), slope, hydraulic gradient, and soil classification. All variables were compiled within a 500-m-radius circular.

The explanatory variables were selected based on the following reasons:

- The slope is an important parameter that controls the runoff. A low slope contributes to water retention and therefore increases the probability of groundwater contamination (Tilahun and Merkel 2009).
- The altitude was selected based on the hydrogeology of the Sais basin. A part of the plioquaternary recharge is provided from the lias aquifer in the southern part of the basin, where the altitude is high. Which may contribute to the diminution of NO_3^- pollution by dilution.
- The piezometric level indicates whether the NO_3^- can rapidly reach the groundwater surface. The shallower water depth can increase the probability of NO_3^- contamination (Stigter et al. 2005).

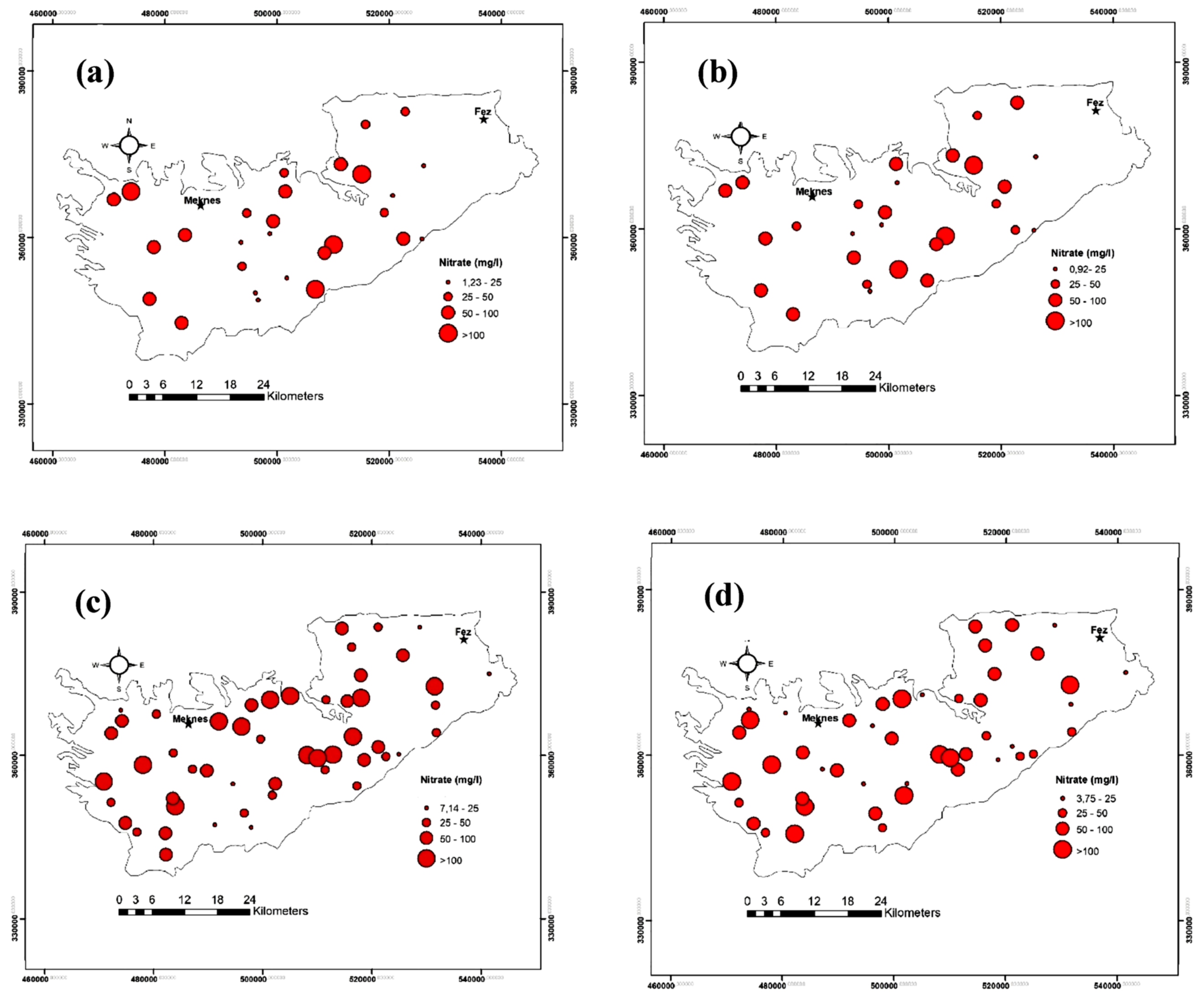


Fig. 3 Spatial distribution of NO_3^- concentrations in spring (**a**) and autumn (**b**) of 2013 and spring (**c**) and autumn (**d**) of 2018

- The rainfall contributes positively to groundwater recharge, which leads to the leaching of soil NO_3^- (Aslam et al. 2018).
- The hydraulic gradient is related to the groundwater flow direction (Rodriguez-Galiano et al. 2014). Which may contribute to the NO_3^- accumulation.
- Lithology can affect groundwater quality. It influences the facility of pollutant transfer to the aquifer (Chenini et al. 2015).
- Soil classification and texture can influence NO_3^- loss. NO_3^- leaching may be more important in sandy soils (Ahirwar and Shukla 2018). The texture components (sand, silt, and clay) were introduced in the RF model separately, to determine the most important component.
- Organic matter and C/N ratio are considered as parameters to be parameters related to the soil nitrogen cycle, which can contribute to NO_3^- losses. Moreover, Berdai et al. (2004) have considered these two parameters as important in the specific groundwater vulnerability to NO_3^-.
- Calcareous soils are characterized by high $CaCO_3$ content. The latter is considered as a factor dominating the ammonification and nitrification processes, which may increase NO_3^- leaching (Zarabi and Jalali 2012; Kutiel and Shaviv 1992).
- Land use is a parameter that represents a potential anthropogenic factor related to NO_3^- pollution. (Huang et al. 2017).

Modeling approach using RF

The groundwater vulnerability is generally understood as a contamination probability. Therefore, to obtain the groundwater vulnerability map to NO_3^-, the first step was rescaling

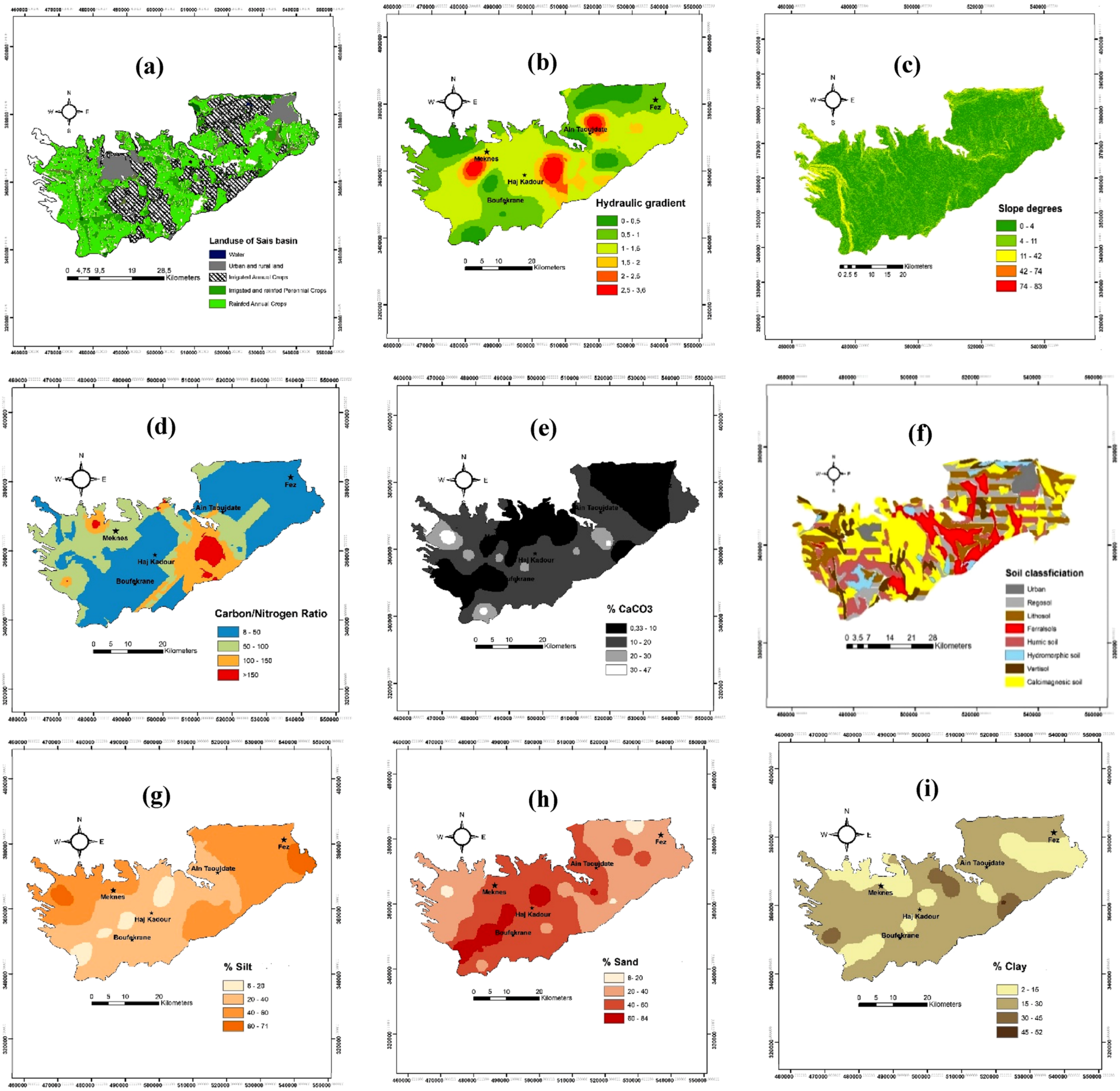


Fig. 4 Raster layers of explanatory variables used in Random Forest: **a** (Land use), **b** (Hydraulic gradient), **c** (Slope degrees), **d** (C/N ratio), **e** (%$CaCO_3$), **f** (Soil classification), **g** (% Silt), **h** (%Sand), **i** (%Clay), **j** (Altitude), **k** (Annual rainfall), **l** (%Organic matter), **m** (Piezometric level), **n** (Lithology)

the NO_3^- concentrations. The observed NO_3^- concentrations dataset observed in the 154 samples were divided into two groups, based on the threshold value of 50 mg/L. Concentrations that exceed the threshold were given a value equal to 1 (nitrate pollution) and concentrations lower or equal the threshold value equal to 0 (no nitrate pollution). The rescaled NO_3^- concentrations were used in the RF as output variable, while specific and intrinsic parameters as input variables. Secondly, the dataset (input and output) were split randomly into two sub-datasets. The first sub-dataset which contains 80% of dataset, was used for the training and validation and the remaining 20% was used for the testing of the RF model. It should be mentioned that RF model split the first sub-dataset (80% of dataset) into two groups, 2/3 for training and the remaining 1/3 for validation purposes. Figure 5 shows the methodology flowchart used for this study. The distribution of the training, validation, and testing samples are shown in Fig. 6. The RF implementation requires the number of trees

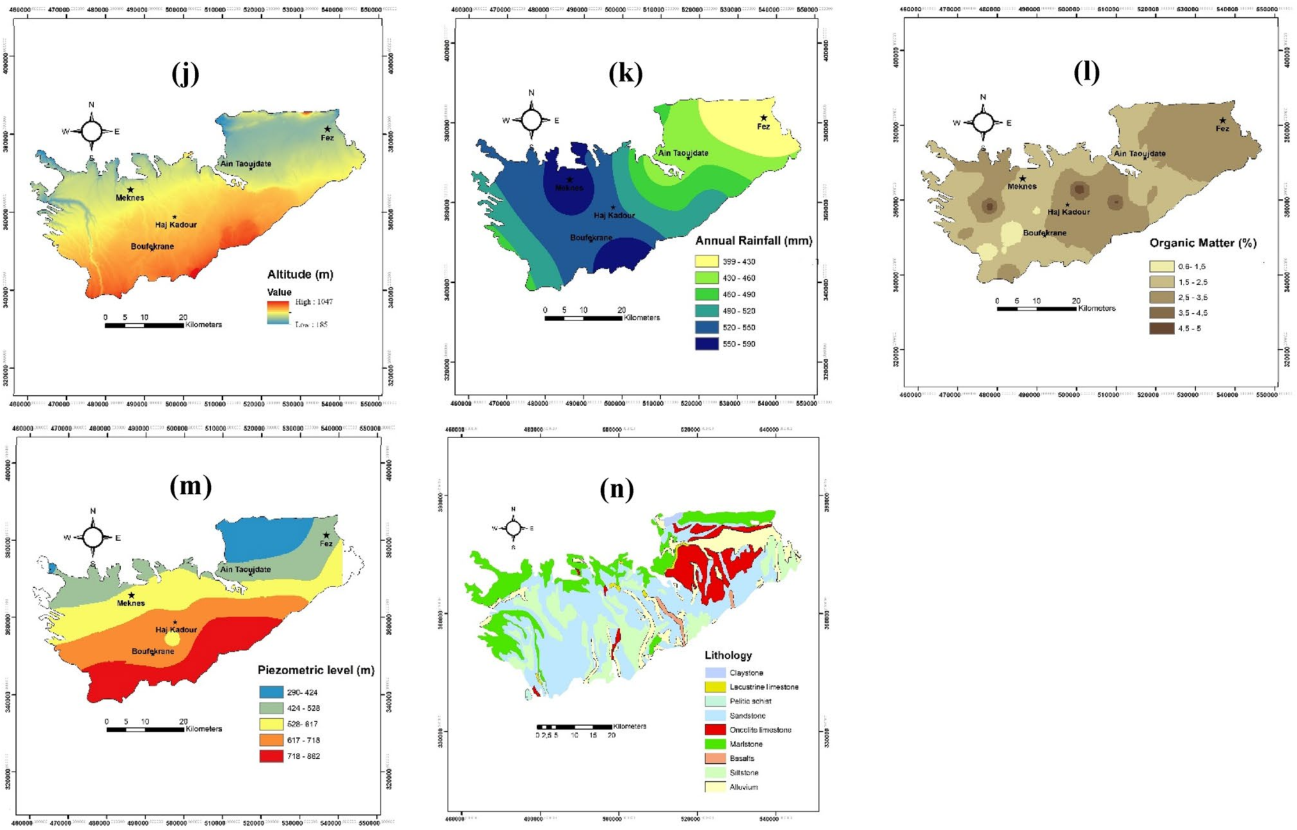


Fig. 4 (continued)

Table 1 Explanatory variables used in the RF model

Explanatory variable	Variable type	Source	Estimation method
Land use	Categorical	Provincial department of Agriculture (dated 2016)	
Hydraulic gradient	Numerical	Hydraulic basin agency (data collected in 2013 from 59 wells in the Sais basin)	For mapping the hydraulic gradient and piezometric level was used Kriging method in GIS
Piezometric level	Numerical		
Soil classification	Categorical	Hydraulic basin agency	
Slope	Numerical	Digital elevation model at scale 1:50,000 and 30 m of resolution	
Altitude	Numerical		
Lithology	Categorical	Ministry of mining, Energy an environment(data extracted from Geological map at scale 1/1,000,000 (dated 1985)	
Organic matter	Numerical	Data from results analysis of 82 soil samples of 1 m depth collected from 31 points in Sais basin in march 2018 and analyzed at the laboratory of National Institute of Agricultural research of Meknes. Only the results analysis of the 0–30-cm soil layer were used for this study	The analysis methods used are: Walkey and black method for organic matter; Kjeldahl distillation for total nitrogen; organic carbon was calculated from the % of organic matter; Bernard calcimeter method for $CaCO_3$; Robinson method for texture after elimination of organic matter using H_2O_2 (30%). For mapping the soil parameters,Kriging method in GIS was used
C/N ratio	Numerical		
$CaCO_3$	Numerical		
Texture (sand, clay, silt)	Numerical		
Annual rainfall	Numerical	Observed data by 9 stations located in the Fez-Meknes region during 1981–2018	For mapping the annual rainfall, IDW method in GIS was used

and the number of variables (m) used to determine the split at each node. Breiman (2001) recommends using m number close to 1/3 of all input variables. For this study, we used a maximum of 10,000 trees and we tested different numbers of random input variables (m_{try} = 1, 2, 3 and 4) at each node. The optimal m_{try} is one that computes the lowest error. The

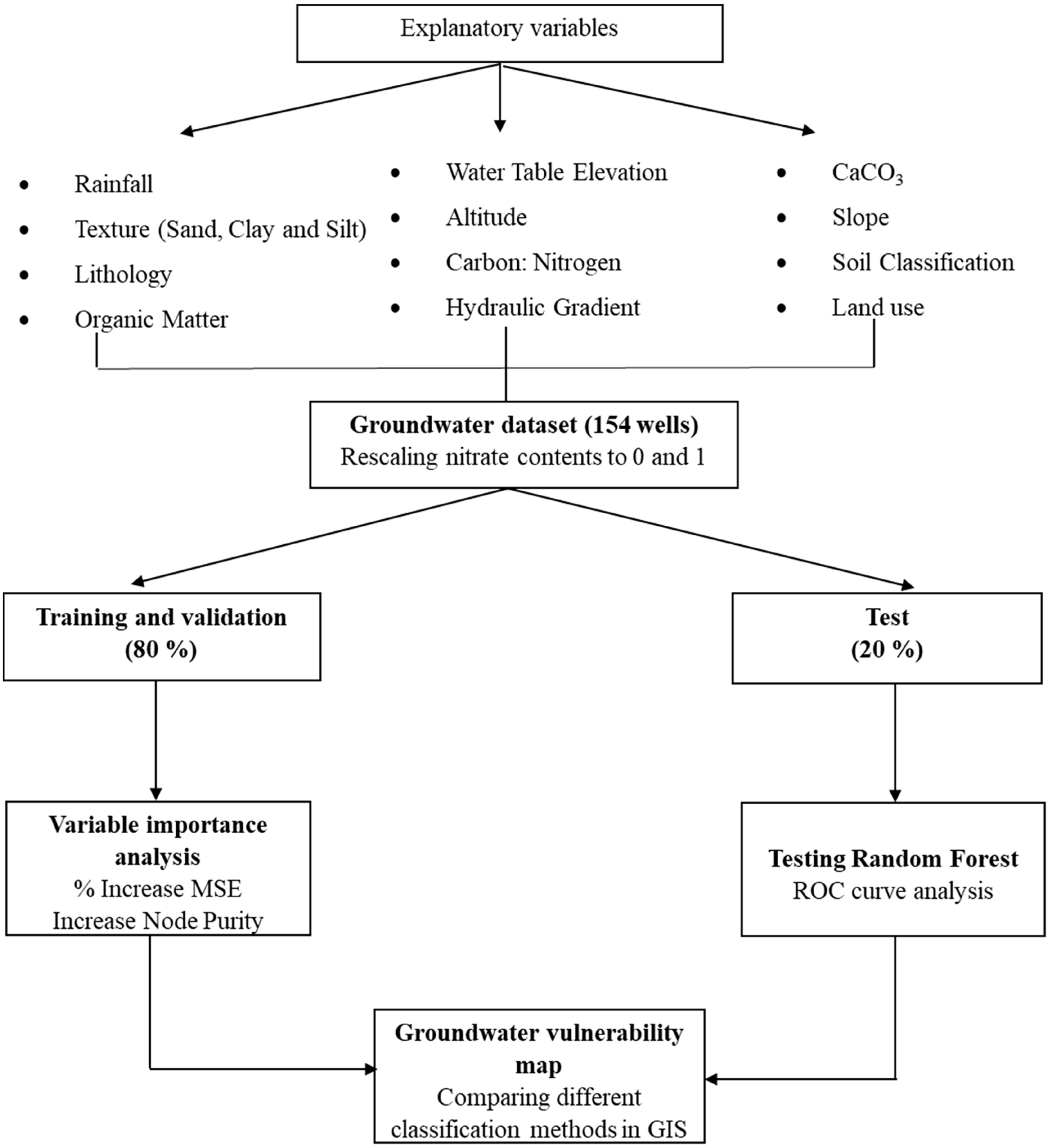


Fig. 5 Random forest flowchart used in this study

RandomForest package in R software (V 1.1.4) was used for the RF model. The optimal m_{try} was determined using the TuneRF function in R software.

Two modeling approaches based on variable importance were used for this study. The first approach (RF1), added all explanatory variables selected (14 variables) as model input. The second approach selected the most important variables in the RF1 model result and used them as input for a new RF implementation (RF2). The predicted values obtained by both RF models were considered as Groundwater Vulnerability Indexes (GVI).

The validation and the testing are essential steps in any study aimed at modeling using machine learning techniques. First, the GVI predicted by RF1 and RF2 were validated and compared based on the error computed by each m_{try} used, we retained the result with the lowest error. Second, the predictive accuracy of the RF model was tested using the receiver operating characteristic (ROC) analysis, through the ROCR package in R Software. The ROC curve allows calculation of the Area Under the Curve (AUC). The ROC plots the false-positive rate on the *X*-axis

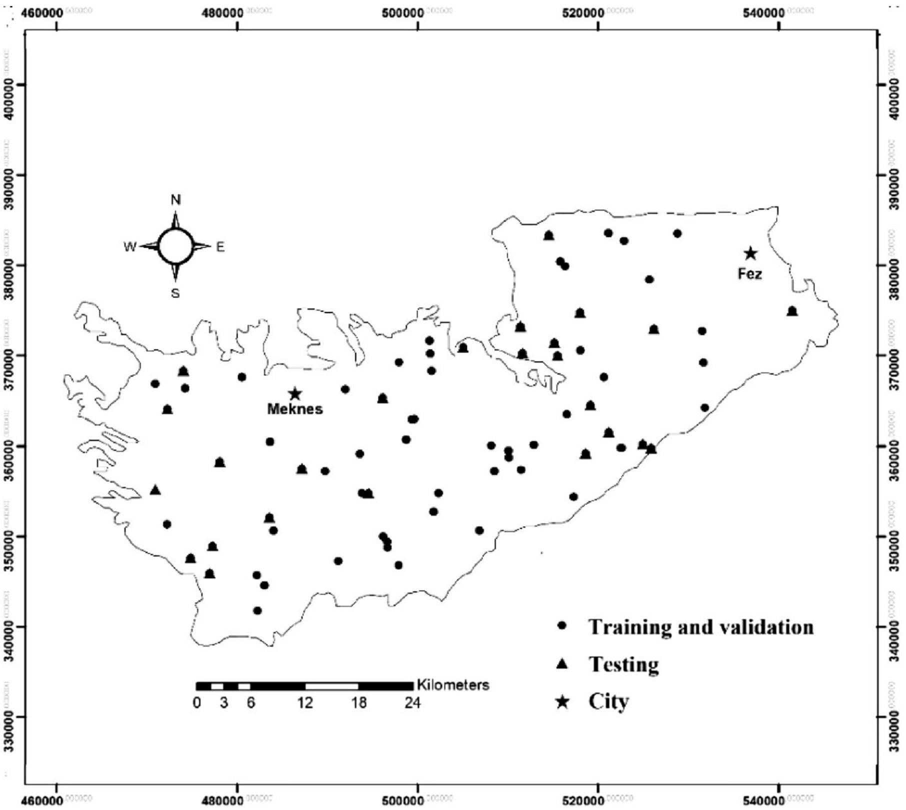


Fig. 6 Location of training, validation, and testing samples used in random forest

and the true positive rate on *Y*-Axis. It explains the trade-off between the two rates (Sezer et al. 2011; Ozdemir and Altural 2013; Akgun 2011). The classification of the prediction accuracy based on AUC can be described as follows: AUC > 0.9, excellent; 0.8 < AUC < 0.9, very good; 0.7 < AUC < 0.8, good; 0.6 < AUC < 0.7, average and 0.5 < AUC < 0.6, poor (Pourghasemi and Kerle 2016; Bradley 1997; Fawcett 2006).

Mapping groundwater vulnerability to nitrate

After the validation and testing, the vulnerability maps were created using all GVI predicted by RF1 and RF2 models, through the Kriging interpolation method in GIS. The most reliable interpolation retained, is the one that generated the lowest error. However, the average of some GVI coincident values was used in the interpolation.

The GVI were categorized into four vulnerability classes namely low, medium, high and very high. The most accurate map was obtained by comparing the different classification methods proposed by the GIS (quantile, natural breaks, geometrical interval, and equal interval), using the Spearman rank correlation (ρ) and one-way ANOVA, between the vulnerability classes and the observed NO_3^- concentrations. All the statistical tests were carried out by R Software.

Results and discussion

Random forest results

Accuracy of the random forest

Figure 7 shows the error computed as function of the number of trees, for each explanatory variable randomly

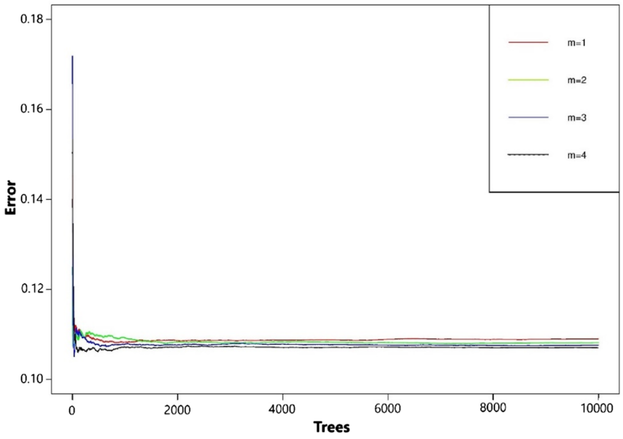


Fig. 7 Impact of the number of trees and random split variable (m_{try}) on the out-of-bag (OOB) error computed by Random Forest applied to all variables (RF1)

sampled (m_{try}) at each node, using the RF1 model. From this result, it can be observed that the error decreased when more trees are used. In fact, from 2000 trees, the error of each m_{try} was low and stable. The same result observed for the other m_{try} used. However, the m_{try} that computed the lowest error was 4, which is consistent with that recommended by Breiman (2001). Furthermore, the mean error value obtained was 0.1100, with a minimum and maximum values of 0.1091 and 0.1545, respectively.

Selection of the most important explanatory variables

The variable importance of the RF model is a particular output indicator of the relative contribution of each input variable in the prediction result. The comparison of variable importance was based on MDE (% increase in MSE) and the mean decrease in the GINI (% Increase in node purity). The importance of each explanatory variable is presented in Fig. 8. The high value indicates that the variable is more important.

As shown in Fig. 8a, the relative increase in MSE obtained was relatively high for all explanatory variables. It varies between 55 and 130.3%. This finding indicates that all

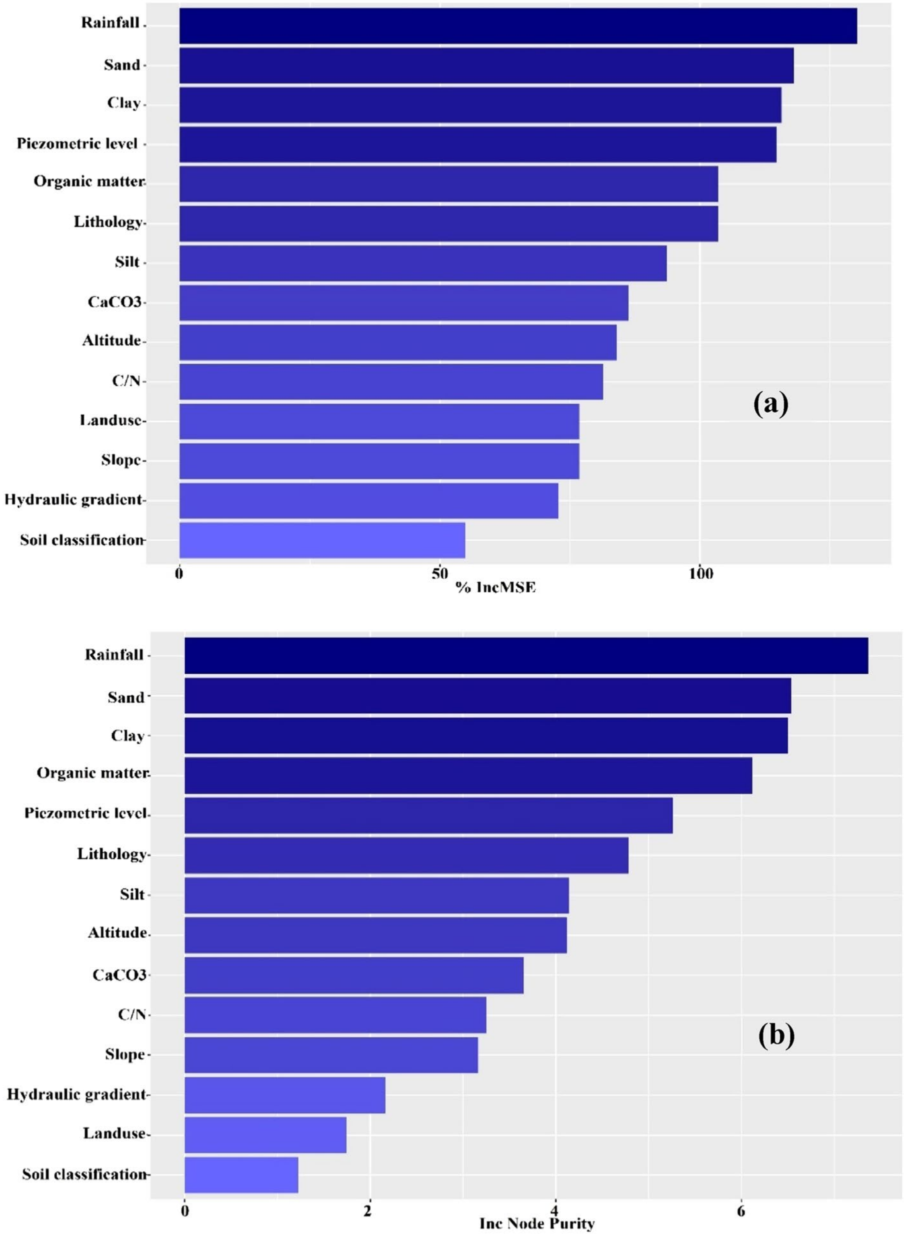


Fig. 8 Relative importance of variables using % increase MSE (**a**) and increase of node purity (**b**)

explanatory variables selected are considered to be controlling factors to groundwater NO_3^- pollution. Nevertheless, rainfall, sand, clay, piezometric level, organic matter, and lithology are the most important explanatory variables. The same result was obtained using the mean decrease in GINI (Fig. 8b) with different importance ranks.

According to the MDE results, the rainfall has the highest importance in GVI prediction, followed by sand and clay contents, with a value of 130, 118, and 116%, respectively. These results can be explained by the fact that rainfall contributes to groundwater recharge and therefore contributes to the NO_3^- leaching. Indeed, the areas where NO_3^- concentrations are high are located within areas containing high soil sand content, mainly in the central and western parts of the basin. Regarding clay importance, the result can be explained by its capability to protect groundwater against NO_3^- contamination due to its high retention capacity. The piezometric level was considered also as an important variable with a value of 115%. Concerning the importance of the organic matter, the result shows that the increase in MSE was 103.5%. This finding suggests that groundwater may receive high loads of organic nitrogen. NO_3^- leaching increases as a result of high mineralization in the case of high soil organic matter content (Hoffmann and Johnsson 1999; Kulabako et al. 2007). The same importance value was observed for lithology. However, the silt, $CaCO_3$, C/N and altitude have revealed medium importance. In contrast, land use, slope, hydraulic gradient, and soil classification are the less important parameters, with values of 54.91, 72.86, 76.81, and 76.82%, respectively.

According to the RF1 importance result, we selected the most important variables (Increase in MSE above 100%) as input for RF2, which are: rainfall, sand content, clay content, piezometric level, organic matter, and lithology.

The result revealed that the error tendency is relatively similar to the RF1 result. The lowest errors were computed from 2000 trees (Fig. 9). However, the best m_{try} for RF2 was 2, which computed the lowest error compared to other m_{try}. The mean error value obtained was 0.1099 with a minimum and maximum values of 0.1083 and 0.1750, respectively. Therefore, using the most important parameters can decrease slightly the OOB error.

Relative operating characteristics (ROC) curve

The ROC curve plots for both RF models are shown in Fig. 10. The AUC results are quite similar for both RF

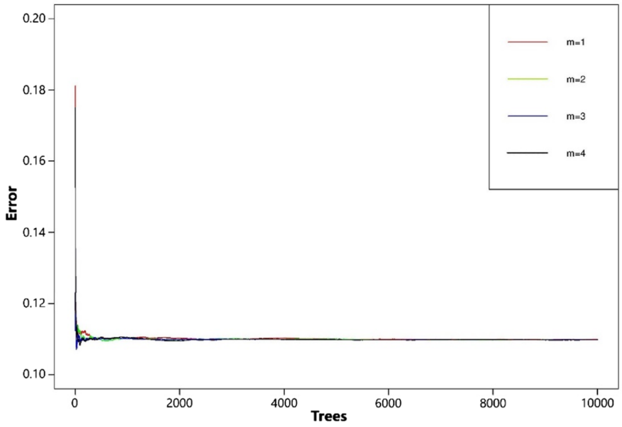


Fig. 9 Impact of the number of trees and random split variable (m_{try}) on the out-of-bag (OOB) error computed by Random Forest applied to the most important variables (RF2)

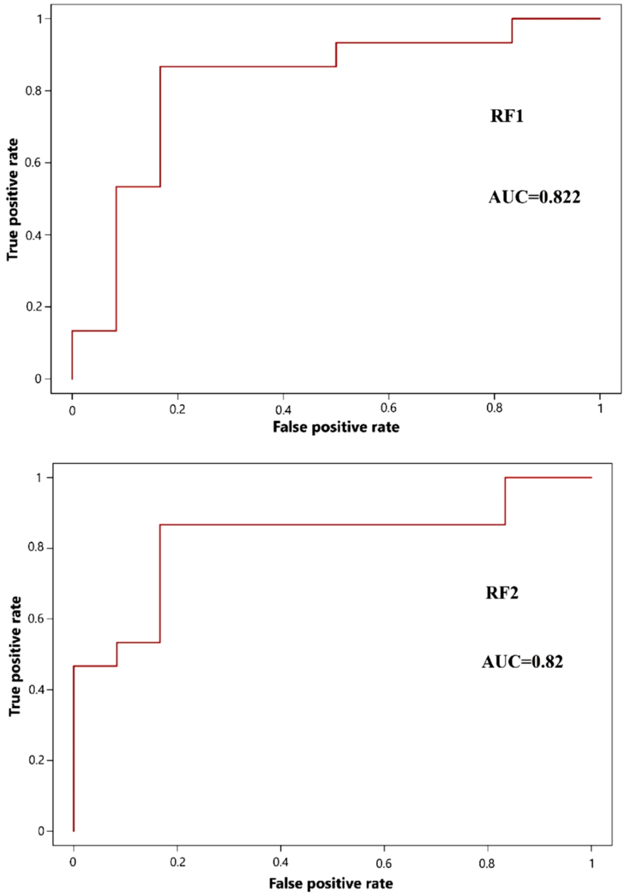


Fig. 10 ROC curve computed using RF1 and RF2 models

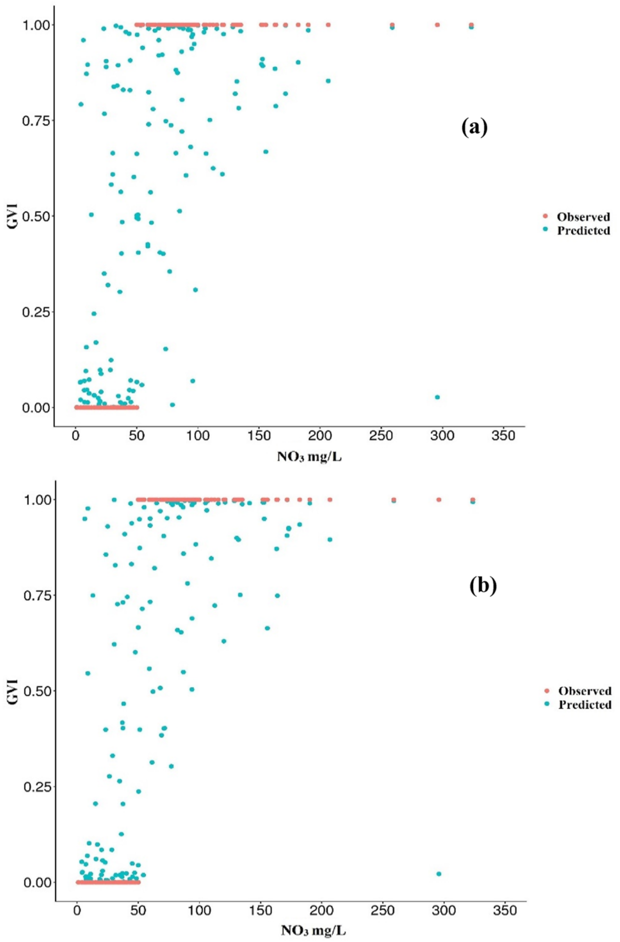


Fig. 11 predicted GVI through the RF1 (**a**) and RF2 (**b**), the red points represent the observed values (0 and 1)

models. The AUC were 0.822 and 0.82, which correspond to the prediction accuracy of 82.2 and 82% for RF1 and RF2 models, respectively. Therefore, both RF models produce very good prediction performance.

Mapping groundwater vulnerability to nitrate

As seen in Fig. 11, the predicted GVI increase significantly as a function of observed NO_3^- concentrations, these findings were similar for both models (RF1 and RF2). However, the predicted GVI obtained showed that RF2 predicts more accurately GVI compared to RF1. The predicted values range from 0.003 to 0.998 and 0.0019–0.999 for RF1 and RF2, respectively. Therefore, the removal of the less important explanatory variables caused a slight increase in the GVI prediction accuracy. This finding was consistent with the OOB errors computed.

The GVI predicted using both RF models were classified according to four vulnerability classes (low, medium, high and very high). The comparison between the classification methods based on the Spearman rank correlation (ρ) and Eta coefficient (η), showed that geometric interval and equal interval are considered as the most appropriates methods in RF1 and RF2, respectively (Table 2 and Table 3). These classification methods were used to create vulnerability classes.

The observed NO_3^- concentrations according to the vulnerability classes obtained are presented as a boxplot in Fig. 12. These plots summarize the observed NO_3^- concentrations by a central point which indicates the median, a box to indicate the variability around the median (25th and 75th

Table 2 Comparison between classification methods applied to RF1

Classification method	Which vulnerability class corresponds to the lowest concentration of NO_3?	Which vulnerability class corresponds to the highest concentration of NO_3?	Spearman rank correlation (ρ)	Eta (η)	ANOVA F-Statistics
Equal interval	Low	Very high	0.6500	0.2678	15.86
Quantile	Low	Very high	0.6534	0.2778	16.67
Geometric interval	Low	Very high	0.6547	0.2800	16.85
Natural breaks	Low	Very high	0.6460	0.2696	16

Table 3 Comparison between classification methods applied to RF2

Classification method	Which vulnerability class corresponds to the lowest concentration of NO_3?	Which vulnerability class corresponds to the highest concentration of NO_3?	Spearman rank correlation (ρ)	Eta (η)	ANOVA F-Statistics
Equal interval	Low	Very high	0.6645	0.2837	17.7
Quantile	Low	Very high	0.6464	0.2703	16.05
Geometric interval	Low	Very high	0.6514	0.2803	16.88
Natural breaks	Low	Very high	0.6441	0.2677	15.84

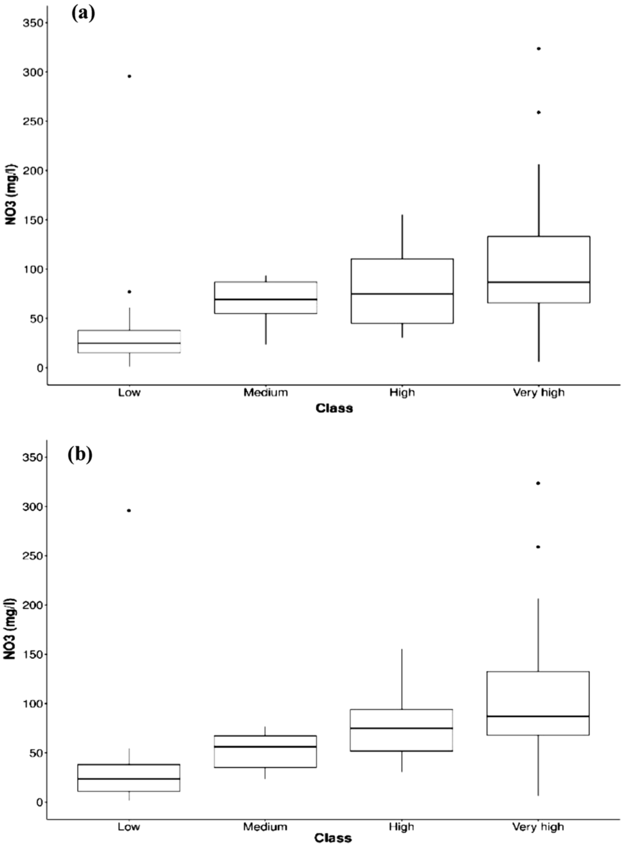


Fig. 12 Box plot of observed nitrate concentrations and vulnerability classes obtained through the RF1 (**a**) and RF2 (**b**)

percentiles), whiskers around the box to indicate the range of variables and the points to indicate the outliers' values. It can be observed that the vulnerability classes show their suitability for observed NO_3^- concentrations. The low class presents the lowest concentrations, while very high class contains the highest NO_3^- concentrations. However, the comparison between two RF models (Table 2 and Table 3), showed that the RF2 model was more reliable in GVI prediction, the Spearman rank correlation (ρ) and Eta coefficient (η) between vulnerability classes and observed NO_3^- concentration were up to 0.6645 and 0.2837, respectively, which are relatively greater than those obtained by the RF1 model (0.6547 and 0.2800, respectively).

The vulnerability maps obtained using both RF models are shown in Fig. 13. It shows that the northern, central, northeastern and western parts of the basin are the areas where the groundwater vulnerability to NO_3^- is classified as high to very high. These two classes cover, respectively, 25.04 and 22.9% of the total area, for RF1 and 36.38 and 26.5% for RF2 (Table 4). In these areas, the annual rainfall varies between 430 and 550 mm, while the sand content varies between 40 and 84%. Regarding clay content, it varies between 2 and 30%. As for the organic matter, the content

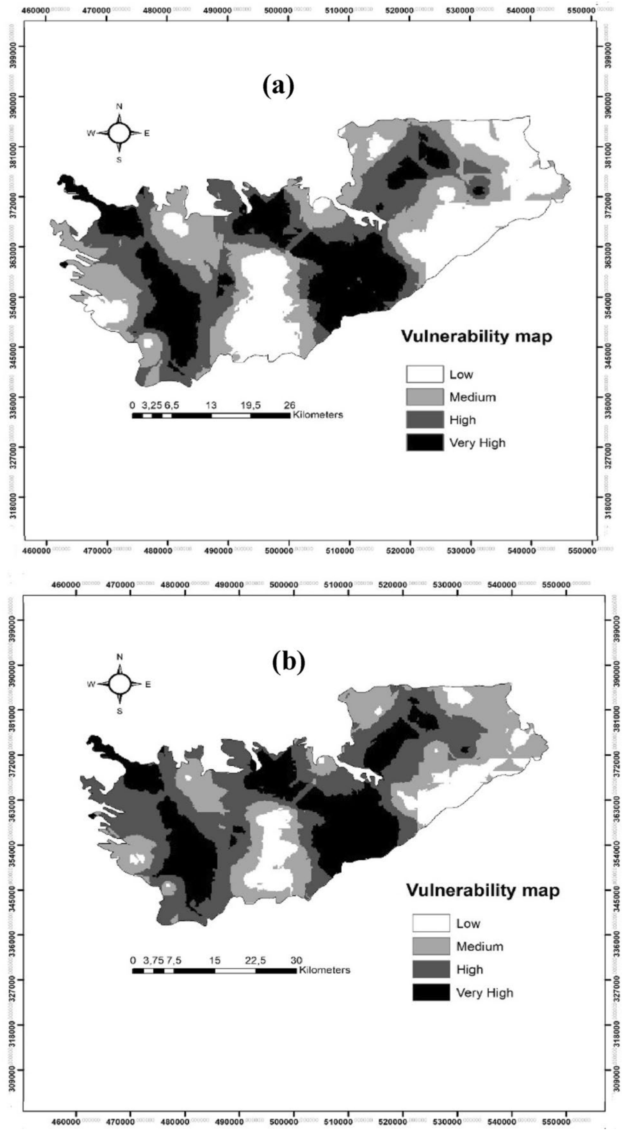


Fig. 13 Vulnerability maps obtained using RF1 (**a**) and RF2 (**b**)

varies between 1.5 and 5%. Moreover, three lithological classes are dominants in these areas, namely sandstone, marlstone, and oncolite limestone.

Concerning the medium vulnerability class, it occupies 27.71 and 26.14% of the total area, respectively for RF1 and

Table 4 Statistics of the groundwater vulnerability surface area

Classes	RF1 (% surface area)	Area (km^2)	RF2 (% surface area)	Area (km^2)
Low	24.80	510	11	232
Medium	27.71	582	26.14	549
High	25.04	526	36.38	764
Very high	22.90	482	26.5	546

RF2. This class is located mainly in some eastern, northern and western parts of the basin. However, the area that presents a low vulnerability does not exceed 24.80 and 11% of the total area, for RF1 and RF2, respectively, and located mainly in some southern (middle atlas limits) and eastern parts of the basin. However, these areas are characterized by high clay content. The latter varies between 30 and 52%. Moreover, two lithological classes are dominants, which are marlstone and siltstone.

Based on these results, the RF model provides good performance in the determination of groundwater vulnerability, this is due to its ability for learning non-linear relationships between NO_3^- concentrations and explanatory variables used in this study. However, the groundwater vulnerability maps to NO_3^- obtained can be improved continuously over time, when new input variables are considered, such as groundwater recharge and nitrogen fertilizer application.

Conclusion

Improving water management strategies need a robust method to assess groundwater vulnerability. The present study aimed to develop an accurate RF model for the prediction of groundwater vulnerability to NO_3^-. The observed NO_3^- concentrations in the Sais basin were rescaled to 0 and 1, based on the drinking threshold of NO_3^- (50 mg/L). The predicted values were considered as GVI. Fourteen explanatory variables related to the intrinsic and specific groundwater vulnerability were used as inputs in the RF model. These variables were rainfall, organic matter, soil texture (sand, clay, and silt), altitude, lithology, land use, C/N ratio, piezometric level, $CaCO_3$, slope, hydraulic gradient, and soil classification. The OOB-error and AUC were 0.1100 and 82.2%, respectively. Moreover, the study revealed that all explanatory variables used are considered to be controlling factors to groundwater NO_3^- pollution, with differing importance degrees. In fact, the rainfall, sand content, clay content, organic matter, piezometric level, and lithology were the most important predictors of GVI. Moreover, using only these important parameters as RF input showed that the OOB-error and AUC were of 0.1099 and 82%, respectively. The comparison between the observed NO_3^- concentrations and the vulnerability classes obtained showed that the RF2 model can produce slightly more accurate groundwater vulnerability map.

The results revealed that about 48 and 63% of the total surface area are under high to very high vulnerability to NO_3^-, using RF1 and RF2, respectively. While about 27.7 and 26.1% of the surface area are in medium vulnerability, and 24.8 and 11% of the surface area are in low vulnerability, using RF1 and RF2, respectively.

Base on the RF results, the most important factors in the prediction result should be taken into consideration when recommending nitrogen fertilization since the agricultural activity is intense in the Sais basin.

Nevertheless, NO_3^- pollution can be affected by other variables related to the biogeochemical process, overuse of nitrogen fertilizers and the land use change. Consequently, including these factors in the RF model may also improve the groundwater vulnerability map to NO_3^- in the Sais basin.

The current study is a novel application of machine learning technique in groundwater vulnerability assessment in Morocco. In the future, the RF model performance can be compared with other machine learning methods. This study will provide valuable information for groundwater management in the study area.

References

Ahirwar S, Shukla JP (2018) Assessment of groundwater vulnerability in upper Betwa river watershed using GIS based DRASTIC model. J Geol Soc India 91(3):334–340. https://doi.org/10.1007/s12594-018-0859-0

Akgun A (2011) A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at İzmir, Turkey. Landslides 9(1):93–106. https://doi.org/10.1007/s10346-011-0283-7

Al-Abadi AM, Shahid S (2016) Spatial mapping of artesian zone at Iraqi southern desert using a GIS-based random forest machine learning model. Model Earth Syst Environ. https://doi.org/10.1007/s40808-016-0150-6

Aller L, Bennett T, Lehr JH, Petty RH, Hackett G (1987) DRASTIC: a standardized system for evaluating groundwater pollution potential using hydrogeologic settings. USEPA Report 600/2-87/035, Robert S. Kerr Environmental Research Laboratory, Ada, Oklahoma

Al-Shatnawi AM, El-Bashir MS, Khalaf RMB, Gazzaz NM (2015) Vulnerability mapping of groundwater aquifer using SINTACS in Wadi Al-Waleh Catchment, Jordan. Arab J Geosci. https://doi.org/10.1007/s12517-015-2080-4

Amraoui F (2005) Contribution à la connaissance des aquifères Karstiques cas du Lias da la plaine du Sais et du causse moyen atlasique tabulaire. Université Hassan II Ain Chock, Faculté des Sciences, Casablanca, Maroc, Thèse de Doctorat d'Etat, p 249p

Anning DW, Paul AP, McKinney TS, Huntington JM, Bexfield LM, Thiros SA (2012) Predicted Nitrate and arsenic concentrations in basin-fill aquifers of the southwestern United States. US Geological Survey Scientific Investigations Report 2012–5065

Aslam RA, Shrestha S, Pandey VP (2018) Groundwater vulnerability to climate change: a review of the assessment methodology. Sci Total Environ 612:853–875. https://doi.org/10.1016/j.scitotenv.2017.08.237

Baghapour MA, Nobandegani AF, Talebbeydokhti N, Bagherzadeh S, Nadiri AA, Gharekhani M, Chitsazan N (2016) Optimization of DRASTIC method by artificial neural network, nitrate vulnerability index, and composite DRASTIC models to assess groundwater vulnerability for unconfined aquifer of Shiraz Plain, Iran. J Environ Health Sci Eng. https://doi.org/10.1186/s40201-016-0254-y

Belhassan K, Hessane MA, Essahlaoui A (2010) Interactions eaux de surface–eaux souterraines: bassin versant de l'Oued

Mikkes (Maroc). Hydrol Sci J 55(8):1371–1384. https://doi.org/10.1080/02626667.2010.528763

Berdai H, Soudi B, Bellouti A (2004) Contribution à l'étude de la pollution nitrique des eaux souterraines en zones irriguées: Cas du Tadla. Revue H.T.E. N° 128-Mars

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30(7):1145–1159. https://doi.org/10.1016/s0031-3203(96)00142-2

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

Calle ML, Urrea V (2010) Letter to the editor: stability of random forest importance measures. Brief Bioinf 12(1):86–89. https://doi.org/10.1093/bib/bbq011

Catani F, Lagomarsino D, Segoni S, Tofani V (2013) Landslide susceptibility estimation by random forest technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci 13:2815–2831. https://doi.org/10.5194/nhess-13-2815-2013

Chamayou J, Combe M, Genetier B, Leclercn C (1975) Le bassin de Fès-Meknès, ressource en eau du Maroc. Notes et mémoire Service Géologique, Maroc, Rabat

Chenini I, Zghibi A, Kouzana L (2015) Hydrogeological investigations and groundwater vulnerability assessment and mapping for groundwater resource protection and management: state of the art and a case study. J Afr Earth Sc 109:11–26. https://doi.org/10.1016/j.jafrearsci.2015.05.008

El Himer H, Fakir Y, Stigter TY, Lepage M, El Mandour A, Ribeiro L (2013) Assessment of groundwater vulnerability to pollution of a wetland watershed: the case study of the Oualidia-Sidi Moussa wetland, Morocco. Aquat Ecosyst Health Manag 16(2):205–215. https://doi.org/10.1080/14634988.2013.788427

Essahlaoui A, Sahbi H, Bahi L, El-Yamine N (2001) Reconnaissance de la structure géologique du bassin de Saiss occidental, Maroc, par sondages électriques. J Afr Earth Sci 32(4):777–789. https://doi.org/10.1016/s0899-5362(02)00054-4

Fassi O (1999) Les formations superficielles du Saiss de Fès et de Meknès des temps géologique à l'utilisation actuelle des sols. Notes et mémoire Services Géologique, Maroc, Rabat, n°389.

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Ghazavi R, Ebrahimi Z (2015) Assessing groundwater vulnerability to contamination in an arid environment using DRASTIC and GOD models. Int J Environ Sci Technol 12(9):2909–2918. https://doi.org/10.1007/s13762-015-0813-2

Grömping U (2009) Variable importance assessment in regression: linear regression versus random forest. Am Stat 63(4):308–319. https://doi.org/10.1198/tast.2009.08199

Hasiniaina F, Zhou J, Guoyi L (2010) Regional assessment of groundwater vulnerability in Tamtsag basin, Mongolia using drastic model. J Am Sci 6(11):65–78

Hoffmann M, Johnsson H (1999) Environ Model Assess 4(1):35–44. https://doi.org/10.1023/a:1019087511708

Huang L, Zeng G, Liang J, Hua S, Yuan Y, Li X, Dong H, Liu J, Nie S, Liu J (2017) Combined impacts of land use and climate change in the modeling of future groundwater vulnerability. J Hydrol Eng 22(7):05017007. https://doi.org/10.1061/(asce)he.1943-5584.0001493

Ki MG, Koh DC, Yoon H, Kim H (2015) Temporal variability of nitrate concentration in groundwater affected by intensive agricultural activities in a rural area of Hongseong, South Korea. Environ Earth Sci 74(7):6147–6161. https://doi.org/10.1007/s12665-015-4637-7

Kulabako N, Nalubega M, Thunvik R (2007) Study of the impact of landuse and hydrogeological settings on the shallow groundwater quality in a peri-urban area of Kampala, Uganda. Sci Total Environ 381(1):180–199

Kutiel P, Shaviv A (1992) Effects of soil type, plant composition and leaching on soil nutrients following a simulated forest fire. For Ecol Manag 53(1–4):329–343. https://doi.org/10.1016/0378-1127(92)90051-a

Laftouhi NE, Vanclooster M, Jalal M, Witam O, Aboufirassi M, Bahir M, Persoons E (2003) Groundwater nitrate pollution in the Essaouira Basin (Morocco). C R Geosci 335(3):307–317. https://doi.org/10.1016/s1631-0713(03)00025-7

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2(3):18–22

Loosvelt L, Petersb J, Skriverc H, Lievensa H, Van Coillied FMB, De Baets B, Verhoesta NEC (2012) Random forests as a tool for estimating uncertainty at pixel-level in SAR image classification. Int J Appl Earth Obs Geoinf 19:173–184. https://doi.org/10.1016/j.jag.2012.05.011

Mendes MP, Rodriguez-Galiano V, Luque-Espinar JA, Ribeiro L, Chica-Olmo M (2016) Applying random forest to assess the vulnerability of groundwater to pollution by nitrate. Geo ENV 2016. In: The 11th international conference on geostatistics for environmental applications. Lisbon, Portugal. geoENV2016BookofAbstractsMPM

Micheletti N, Foresti L, Robert S, Leuenberger M, Pedrazzini A, Jaboyedoff M, Kanevski M (2013) Machine learning feature selection methods for landslide susceptibility mapping. Math Geosci 46(1):33–57. https://doi.org/10.1007/s11004-013-9511-0

Moore KB, Ekwurzel B, Esser BK, Hudson GB, Moran JE (2006) Sources of groundwater nitrate revealed using residence time and isotope methods. Appl Geochem 21(6):1016–1029. https://doi.org/10.1016/j.apgeochem.2006.03.008

Naghibi SA, Ahmadi K, Daneshi A (2017) Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resour Manag 31(9):2761–2775. https://doi.org/10.1007/s11269-017-1660-3

Walsh ES, Kreakie BJ, Cantwell MG, Nacci D (2017) A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system. PLoS ONE 12(7):e0179473. https://doi.org/10.1371/journal.pone.0179473

Nampak H, Pradhan B, Manap MA (2014) Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. J Hydrol 513:283–300. https://doi.org/10.1016/j.jhydrol.2014.02.053

National Research Council (1993) Ground water vulnerability assessment: predicting relative contamination potential under conditions of uncertainty. The National Academies Press, Washington, D.C.

Nolan BT (2001) Relating nitrogen sources and aquifer susceptibility to nitrate in shallow ground waters of the United States. Ground Water 39(2):290–299. https://doi.org/10.1111/j.1745-6584.2001.tb02311.x

Ouedraogo I, Defourny P, Vanclooster M (2018) Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. Hydrogeol J. https://doi.org/10.1007/s10040-018-1900-5

Ozdemir A, Altural T (2013) A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan Mountains, SW Turkey. J Asian Earth Sci 64:180–197. https://doi.org/10.1016/j.jseaes.2012.12.014

Pourghasemi HR, Kerle N (2016) Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. Environ Earth Sci. https://doi.org/10.1007/s12665-015-4950-1

Puckett LJ, Tesoriero AJ, Dubrovsky NM (2011) Nitrogen contamination of surficial aquifer-a growing legacy. Environ Sci Technol 45:839–844. https://doi.org/10.1021/es1038358

Ribeiro L, Pindo JC, Dominguez-Granda L (2017) Assessment of groundwater vulnerability in the Daule aquifer, Ecuador, using the susceptibility index method. Sci Total Environ 574:1674–1683. https://doi.org/10.1016/j.scitotenv.2016.09.004

Rodriguez-Galiano V, Mendes MP, Garcia-Soldado MJ, Chica-Olmo M, Ribeiro L (2014) Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (southern Spain). Sci Total Environ 476–477:189–206. https://doi.org/10.1016/j.scitotenv.2014.01.001

Sadkaoui N, Boukrim S, Bourak A, Lakhili F, Mesrar L, Chaouni A, Lahrach A, Jabrane R, Akdim B (2013) Groundwater pollution of Sais basin (Morocco), vulnerability mapping by DRASTIC, GOD and PRK methods, involving Geographic Information System(GIS). Present Environ Sustain Dev 7:296–309

Schnebelen N, Platel JP, Nindre Y, Baudry D (2002) Gestion des eaux souterraines en Aquitaine Année 5. Opération sectorielle. Protection de la nappe de l'Oligocène en région bordelaise, Rapport, BRGM, Orléans, France

Sezer EA, Pradhan B, Gokceoglu C (2011) Manifestation of an adaptive neuro-fuzzy model on landslide susceptibility mapping: Klang valley, Malaysia. Expert Syst Appl 38(7):8208–8219. https://doi.org/10.1016/j.eswa.2010.12.167

Siroky DS (2009) Navigating random forests and related advances in algorithmic modeling. Stat Surv 3:147–163. https://doi.org/10.1214/07-ss033

Stigter TY, Ribeiro L, Dill AMMC (2005) 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. Hydrogeol J 14(1–2):79–99. https://doi.org/10.1007/s10040-004-0396-3

Tabyaoui FZ, Sahbi H, Elouazzani A, Chadli K, Essahlaoui A, Elouali A, Rouai M (2004) Etat de la pollution par les nitrates dans des eaux de la nappe plio-quaternaire du plateau de Meknès (Maroc). Geomaghreb, n°2, 63-75

Taltasse P (1953) Recherche géologique et hydrogéologique dans le bassin de Fès-Meknès. Notes et mémoires Service Géologique, Maroc, n°115, p 300

Tilahun K, Merkel BJ (2009) Assessment of groundwater vulnerability to pollution in Dire Dawa, Ethiopia using DRASTIC. Environ Earth Sci 59(7):1485–1496. https://doi.org/10.1007/s12665-009-0134-1

Ward MH, Dekok TM, Levallois P, Brender J, Gulis G, Nolan BT, VanDerslice J (2005) Workgroup report: drinking-water nitrate and health-recent findings and research needs. Environ Health Perspect 113(11):1607–1614. https://doi.org/10.1289/ehp.8043

Wheeler DC, Nolan BT, Flory AR, DellaValle CT, Ward MH (2015) Modeling groundwater nitrate concentrations in private wells in Iowa. Sci Total Environ 536:481–488. https://doi.org/10.1016/j.scitotenv.2015.07.080

Yang J, Griffiths J, Zammit C (2019) National classification of surface–groundwater interaction using random forest machine learning technique. River Res Appl. https://doi.org/10.1002/rra.3449

Zabihi M, Pourghasemi HR, Pourtaghi ZS, Behzadfar M (2016) GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. Environ Earth Sci. https://doi.org/10.1007/s12665-016-5424-9

Zarabi M, Jalali M (2012) Leaching of nitrogen from calcareous soils in western Iran: a soil leaching column study. Environ Monit Assess 184(12):7607–7622. https://doi.org/10.1007/s10661-012-2522-3

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.