Science of the Total Environment 644 (2018) 954–962

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

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

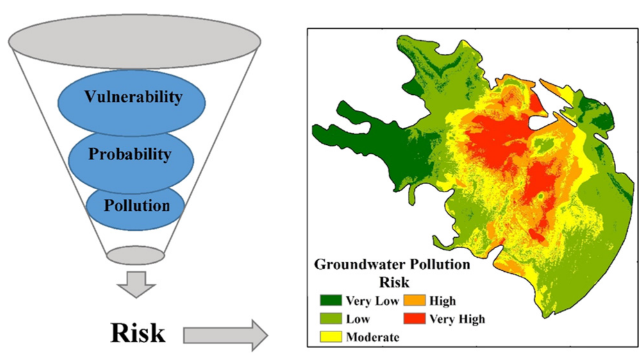
Farzaneh Sajedi-Hosseini [a], Arash Malekian [a,*], Bahram Choubin [b,*], Omid Rahmati [c], Sabrina Cipullo [d], Frederic Coulon [d], Biswajeet Pradhan [e,f]

[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] Young Researchers and Elites Club, Khorramabad Branch, Islamic, Azad University, Khorramabad, 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

HIGHLIGHTS

- Novel risk assessment framework for nitrate groundwater contamination in arid regions.
- Machine learning (ML) predicting vulnerability, pollution, and occurrence probability.
- ML allows quick regional evaluation of risk posed by nitrate in groundwater.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 2 June 2018
Received in revised form 4 July 2018
Accepted 4 July 2018
Available online xxxx

Editor: D. Barcelo

Keywords:
Groundwater pollution
Nitrate
Probability
Risk
Vulnerability
GIS

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% 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.

* Corresponding authors.
E-mail addresses: Malekian@ut.ac.ir (A. Malekian), Bahram.choubin@ut.ac.ir (B. Choubin).

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

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 groundwater 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 et 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 is 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 (Fig. 1). The plain is surrounded by calcareous mountains and elevations of the plain range between 1631 and 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

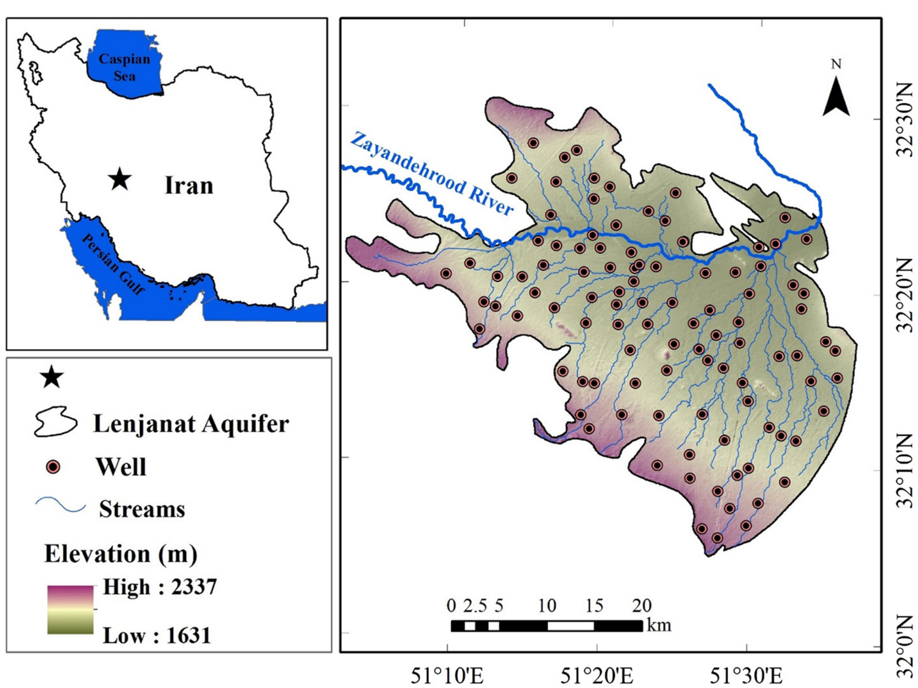


Fig. 1. Location of the study area.