Science of the Total Environment 621 (2018) 697–712

Contents lists available at ScienceDirect

Science of the Total Environment

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

Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms

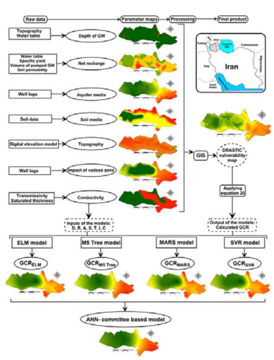
Rahim Barzegar [a,*], Asghar Asghari Moghaddam [a], Ravinesh Deo [b], Elham Fijani [c], Evangelos Tziritis [d]

[a] Department of Earth Sciences, Faculty of Natural Sciences, University of Tabriz, Tabriz, Iran
[b] School of Agricultural Computational and Environmental Sciences, Institute of Agriculture and Environment (IAg&E), University of Southern Queensland, Springfield, Australia
[c] School of Geology, College of Science, University of Tehran, Tehran, Iran
[d] Hellenic Agricultural Organization, Soil and Water Resources Institute, 57400 Sindos, Greece

HIGHLIGHTS

- A DRASTIC vulnerability map, with an r of 0.64, was developed for a multiple aquifer system (e.g. unconfined, semi-confined and confined)
- DRASTIC method was improved for mapping groundwater contamination risk (GCR) using ELM, SVR, M5 Tree and MARS algorithms
- An ANN committee based multi-model, with an r of 0.88, was constructed to exploit individual model advantages
- The subjectivity of the DRASTIC model was reduced based on the nature and hydrogeological features of the aquifers

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 24 August 2017
Received in revised form 15 November 2017
Accepted 16 November 2017
Available online xxxx

Editor: G. Ashantha Goonetilleke

Keywords:
DRASTIC
Groundwater contamination risk
Iran
Machine learning model
Multi-model

ABSTRACT

Constructing accurate and reliable groundwater risk maps provide scientifically prudent and strategic measures for the protection and management of groundwater. The objectives of this paper are to design and validate machine learning based-risk maps using ensemble-based modelling with an integrative approach. We employ the extreme learning machines (ELM), multivariate regression splines (MARS), M5 Tree and support vector regression (SVR) applied in multiple aquifer systems (e.g. unconfined, semi-confined and confined) in the Marand plain, North West Iran, to encapsulate the merits of individual learning algorithms in a final committee-based ANN model. The DRASTIC Vulnerability Index (VI) ranged from 56.7 to 128.1, categorized with no risk, low and moderate vulnerability thresholds. The correlation coefficient (r) and Willmott's Index (d) between NO_3 concentrations and VI were 0.64 and 0.314, respectively. To introduce improvements in the original DRASTIC method, the vulnerability indices were adjusted by NO_3 concentrations, termed as the groundwater contamination risk (GCR). Seven DRASTIC parameters utilized as the model inputs and GCR values utilized as the outputs of individual machine learning models were served in the fully optimized committee-based ANN-predictive model. The correlation indicators demonstrated that the ELM and SVR models outperformed the MARS and M5 Tree models, by virtue of a larger d and r value. Subsequently, the r and d metrics for the ANN-committee based multi-model in the testing phase were 0.8889 and 0.7913, respectively; revealing the superiority of the integrated (or ensemble) machine learning models when compared with the original DRASTIC approach. The newly designed multi-model ensemble-based approach can be considered as a pragmatic step for mapping groundwater contamination risks of multiple aquifer systems with multi-model techniques, yielding the high accuracy of the ANN committee-based model.

* Corresponding author.
E-mail address: rm.barzegar@tabrizu.ac.ir (R. Barzegar).

https://doi.org/10.1016/j.scitotenv.2017.11.185
0048-9697/

1. Introduction

Protection of groundwater from contamination is a major concern within in the context of total environmental research due to its prime significance as a vital natural resource and the multiple natural and anthropogenic pressures threatening its sustainability, especially in arid and semi-arid regions (Rezaei et al., 2017; Soltani et al., 2017). Researchers have attempted to evaluate the contamination risk of groundwater through assessments of its vulnerability, in order to facilitate preservation and protection efforts and provide robust expert system tools for decision-making in respect to unprecedented climate change groundwater management. Groundwater vulnerability can be defined as the ease of contamination to reach into the major groundwater system (Gogu and Dassargues, 2000; Panagopoulos et al., 2006; Vrba and Zoporozec, 1994). Numerous research methods e.g., COP (Daly et al., 2002), EPIK (Doerfliger and Zwahlen, 1997), ISIS (Civita and De Regibus, 1995), DRASTIC (Aller et al., 1987), AVI (Van Stempvoort et al., 1993), GOD (Foster, 1987), PI (Goldscheider et al., 2000), SINTACS (Civita, 1994), Time-Input (Kralik and Keimel, 2003), etc. have been designed and evaluated for their practicality in assessing groundwater vulnerability. A prominent approach is the DRASTIC method that constitutes an overlay constructed index-based method that combines seven important factors; each one is assigned a realistic numerical score and the respective weights of the attributes are related to groundwater contamination (Aller et al., 1987; Rahman, 2008). The weighted-attribute ratings are integrated to obtain an overall score for groundwater vulnerability, aiming to group the similar areas into distinct classes or categories (e.g., low, medium, and high vulnerability). In many previous works, the uncertainties, either in the method itself or in the data it utilized, are thoroughly discussed and a variety of methods have been developed to minimize and improve their reliability. As a result, the optimization of the DRASTIC method has been subjected to intensive research (Panagopoulos et al., 2006; Huan et al., 2012; Pacheco and Sanches Fernandes, 2013; Pacheco et al., 2015; Barzegar et al., 2016c; Kazakis and Voudouris, 2015; McLay et al., 2001), including a set of recent, similar usages in drought and heat wave map construction (Pandey et al., 2008; Jain et al., 2015; Kim et al., 2017).

Previous researchers have embraced different methods to optimize and modify the DRASTIC method, mainly to improve its practicality. For example, Neshat et al. (2014) applied the Wilcoxon rank-sum nonparametric statistical test and analytic hierarchy process (AHP) method to improve the overall DRASTIC method. Jafari and Nikoo (2016) proposed a new optimization-based methodology for determining groundwater risk using the DRASTIC model with genetic algorithm optimization model and Wilcoxon test. Neshat and Pradhan (2015) employed Dempster–Shafer theory (DST) for optimization of the DRASTIC method to demonstrate its ability for assessing groundwater risk in Kerman plain, Iran. Wang et al. (2012) evaluated groundwater contamination risk using hazard quantification, a modified DRASTIC method and groundwater value, in the Beijing plain, China. Nixdorf et al. (2017) developed a DRASTIC method based on public datasets at the highest available resolution in combination with numerical groundwater modelling for regional assessment of groundwater contamination risk in Songhua River Basin.

Among several statistical methods, the application of artificial intelligence (AI) has gained significant attention in the assessment of groundwater vulnerability and risk-based studies, due to their ability to cope with complicated environmental problems (Rodriguez-Galiano et al., 2014). The popularity of AI for groundwater vulnerability and risk studies nowadays has been increasing. For example, the study of Fijani et al. (2013) applied ANN, fuzzy logic, ANFIS and supervised committee machine artificial intelligence (SCMAI) models to optimize the original DRASTIC method applied in Maragheh-Bonab plain aquifer, Iran. They concluded that the SCMAI can potentially provide a better assessment of groundwater vulnerability by combining the advantages of individual data intelligent models (i.e., such studies have generated an ensemble of high-performance predictive models). As an example of their specific applications, such models have been proposed for the risk assessment of the Tabriz plain aquifer (Barzegar et al., 2016c), where the results revealed the capability of the integrated models in risk assessment. Other studies in this regard have also been prominent, for example, Dixon (2005a, 2005b) applied three fuzzy-DRASTIC models and then compared the results with those obtained from the ordinary DRASTIC. Afshar et al. (2007) applied the pseudo-trapezoidal membership function, Mamdani inference method, and central gravity defuzzification methods for the optimization of the DRASTIC method.

In last decades, groundwater contamination has been becoming a critical issue for water resources management in many nations, including Iran (Rahmati and Melesse, 2016; Mirzaei and Sakizadeh, 2016; Rahmati et al., 2015; Jalali, 2005). Therefore, developing a groundwater risk map can be considered as an effective way to protect groundwater from consequent contamination. The main objective of this study is to develop an optimized procedure that could be applied to the original DRASTIC method, in order to evaluate aquifer vulnerability and perform a risk assessment. This will be realised by a suite of four distinct AI models, namely the extreme learning machine (ELM); multivariate adaptive regression spline (MARS); support vector regression (SVR); and M5 Tree algorithms for spatial mapping of groundwater risk in Marand plain (North West of Iran). In addition, the research also focuses on the development of a practically-relevant method, where a multi-model, ensemble approach is applied using a committee of ANN-based models employed to integrate the results of individual AI-based models. The novelty of this research lies in the conceptualization, design and application of a new DRASTIC-based model for mapping the spatial vulnerability of Marand's plain aquifer system. The latter includes an interconnected, yet a complex system (Barzegar et al., 2017c), which consists of three different aquifer units (i.e., unconfined, semi-confined and confined). This research paper aspires to deliver and optimize a new DRASTIC-based method by using the committee of hybrid machine learning models. The contributions and implications of the research paper are far-reaching in terms of creating a proper understanding and assessment of the groundwater vulnerability in complex aquifer systems as an enduring challenge for hydrogeologists and related decision-makers. The combination of intelligent models, each presenting distinct learning abilities for optimal feature extraction process, provides a practical way to exploit their individual advantages, whilst eliminating their individual limitations to overcome the issue of the complex heterogeneity and thus provide a realistic assessment of contamination risks, especially in complex aquifer systems.

2. Materials and methods

2.1. DRASTIC vulnerability index

Groundwater vulnerability concept was first introduced by Margat (1968), based on the fundamental assumption that the physical environment provides to a degree natural protection to groundwater in regards to the contaminants entering the subsurface environment. The assessment of groundwater vulnerability to contamination has been subjected to intensive research during the previous years and a variety of methods have thus been developed. The simplest ones to apply, and for that reason are the most widely used, are the rating models. These methods classify each parameter that potentially influences the probability of the aquifer contamination, and leads to a weighted score that designates the vulnerability index of the groundwater (Foster, 1987; USEPA, 1996).

Rating methods assign numerical scores or ratings directly to the various physical attributes to develop a range of vulnerability classes. The index methods are one of the most commonly used and were among the earliest ones applied. The most widely employed index method is DRASTIC, named from the seven factors considered: Depth to groundwater, net Recharge, Aquifer media, Soil media, Topography,