Science of the Total Environment 476–477 (2014) 189–206

Contents lists available at ScienceDirect

Science of the Total Environment

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

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)

Victor Rodriguez-Galiano [a,*], Maria Paula Mendes [b], Maria Jose Garcia-Soldado [c], Mario Chica-Olmo [c], Luis Ribeiro [b]

[a] Geography and Environment, School of Geography, University of Southampton, Southampton, SO17 1BJ, United Kingdom
[b] CVRM-, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[c] Departamento de Geodinámica, Universidad de Granada, Avenida Fuentenueva s/n, 18071 Granada, Spain

HIGHLIGHTS

- Assessing of groundwater vulnerability to nitrate pollution using Random Forest algorithm
- Determination of the most significant predictors of nitrate pollution
- Application of a feature selection approach to reduce the number of explicative variables
- Predictive modeling of nitrate concentrations at or above the quality threshold of 50 mg/L

ARTICLE INFO

Article history:
Received 8 November 2013
Received in revised form 31 December 2013
Accepted 1 January 2014
Available online 24 January 2014

Keywords:
Random Forest
Groundwater
Vulnerability assessment
Machine learning techniques
Nitrates

ABSTRACT

Watershed management decisions need robust methods, which allow an accurate predictive modeling of pollutant occurrences. Random Forest (RF) is a powerful machine learning data driven method that is rarely used in water resources studies, and thus has not been evaluated thoroughly in this field, when compared to more conventional pattern recognition techniques key advantages of RF include: its non-parametric nature; high predictive accuracy; and capability to determine variable importance. This last characteristic can be used to better understand the individual role and the combined effect of explanatory variables in both protecting and exposing groundwater from and to a pollutant.

In this paper, the performance of the RF regression for predictive modeling of nitrate pollution is explored, based on intrinsic and specific vulnerability assessment of the Vega de Granada aquifer. The applicability of this new machine learning technique is demonstrated in an agriculture-dominated area where nitrate concentrations in groundwater can exceed the trigger value of 50 mg/L, at many locations. A comprehensive GIS database of twenty-four parameters related to intrinsic hydrogeologic proprieties, driving forces, remotely sensed variables and physical–chemical variables measured in "situ", were used as inputs to build different predictive models of nitrate pollution. RF measures of importance were also used to define the most significant predictors of nitrate pollution in groundwater, allowing the establishment of the pollution sources (pressures).

The potential of RF for generating a vulnerability map to nitrate pollution is assessed considering multiple criteria related to variations in the algorithm parameters and the accuracy of the maps. The performance of the RF is also evaluated in comparison to the logistic regression (LR) method using different efficiency measures to ensure their generalization ability. Prediction results show the ability of RF to build accurate models with strong predictive capabilities.

1. Introduction

Effective management of groundwater resources has become a global issue of concern since the rapid expansion of industrial and agricultural activities, the population increase and climate changes can have major effects on groundwater quality and quantity. Groundwater is frequently used for drinking water supply, by industry and agriculture. Hence, groundwater pollution can endanger human health and threaten those activities. Water quality issues are complex and diverse, and are deserving of urgent global attention and action (UN-Water, 2011).

The prevention, control and combat of groundwater pollution are addressed in various European Union (EU) and national legislative acts, since groundwater is considered a valuable natural source. The

* Corresponponding author at: Geography and Environment University of Southampton University Road Southampton SO17 1BJ.
E-mail address: vrgaliano@ugr.es (V. Rodriguez-Galiano).

0048-9697/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.scitotenv.2014.01.001

EU Water Framework Directive (2000/60/EC, 2000), WFD, and its daughter Directive on the Protection of Groundwater against Pollution (2006/118/EC, 2006), GWD, establish criteria for the definition of groundwater status (quality and quantity). Regarding nitrates, the GWD establishes the quality standard for assessing groundwater chemical status of 50 mg/L. Moreover, the Nitrates Directive (91/676/EEC, 1991) is an integral part of the WFD and it was drawn up with the specific purpose to reduce water pollution caused by nitrates from agricultural sources and prevent further such pollution. EU members are required to identify waters affected by nitrate pollution, designate nitrate vulnerable zones (NVZs). The NVZs are defined as areas where the groundwater contains or could contain (if no action is taken to reverse the trend) more than 50 mg/L of nitrates.

Groundwater quality monitoring and management is challenging. Groundwater pollution normally appears long-delayed in wells, springs and streams resulting in a very slow process of recovery of aquifer's quality, often during a few decades. Since groundwater moves slowly through the subsurface, the impact of anthropogenic activities may last for a relatively long time and for that reason, the environmental measures should be mainly focused on the prevention of the pollution (2006/118/EC, 2006; Alcalá and Custodio, 2014). Hence, the delineation of areas that are more vulnerable to groundwater contamination from anthropogenic sources has become an important management task for land use planning and for the establishing of agri-environment measures that can contribute for a good qualitative status of aquifers. Moreover, the vulnerability assessment to pollution by each pollutant is more consistent, or failing this by each class of pollutant (nutrients, pathogens, microorganics, heavy metals, etc.) individually, or by each group of polluting activities (unsewered sanitation, agricultural cultivation, industrial effluent disposal, etc.) separately (Foster et al., 2002).

Groundwater vulnerability to contamination was defined by the National Research Council (1993) as “the tendency or likelihood for contaminants to reach a specified position in the groundwater system after introduction at some location above the uppermost aquifer”. There are two general types of vulnerability assessments. The first addresses intrinsic vulnerability, also named aquifer sensitivity and it is determined by characteristics of the aquifer and overlying material and, hydrogeological conditions. The second addresses, the specific vulnerability and it is determined by intrinsic characteristics of the aquifer as well as by anthropogenic factors such as land use and contaminant type (Vrba and Zaporozec, 1994). The specific vulnerability can be also defined as the sensitivity plus intensity, where ‘intensity’ is a measure of the source of contamination (Vowinckel et al., 1996). Therefore, specific vulnerability is fundamental factor in the assessment of pollution risk (Wang et al., 2012).

The assessment of groundwater vulnerability maps requires the application of diverse methods and techniques, based on the hydrogeological knowledge of the region under research and on the application of predictive models. With the aim of deciding which areas are vulnerable a large data volume can be collected which cannot be effectively analyzed without an adequate and efficient model. Information analysis and integration are paramount, as the final aim is to elaborate predictive spatial models which allow for the incorporation and combination of relevant variables related to the vulnerability to contamination. The vulnerability assessment of groundwater to contamination range in scope and complexity from simple, qualitative, and relatively inexpensive approaches to rigorous, quantitative, and costly assessments (Focazio et al., 2002). Several methods have been devised to vulnerability mapping. These can be categorized into knowledge-driven and data-driven types depending on the nature of the inference procedure used. Knowledge-driven are models which use subjective evidence based on expert knowledge of processes that might have led to contamination in a given hydrogeological setting, but where no or very few data samples/contamination evidences are known to occur. On the other hand, data-driven models use objective evidence based on the associations between evidential features (predictive variables) and known occurrences of nitrate contamination (Solomatine et al., 2008).

Within the context of knowledge driven models, subjective rating methods (index methods and hybrid methods) have become essential tools to support decision-making in vulnerability assessment. The groundwater vulnerability indexes such as DRASTIC (Aller et al., 1987), EPIK (Doerfliger and Zwahlen, 1997) and SI (Ribeiro, 2005) are often utilized to assess the groundwater vulnerability to non-point source nitrate pollution from agricultural areas.

Generally, subjective hybrid methods combine components of statistical, physically-based hydrogeological models or/and other objectives, incorporating subjective categorization and indexing of vulnerability (Focazio et al., 2002). For instances, numerous studies propose a combination of statistical and index methods with different approaches such as (1) to modified vulnerability indexes (Andrade and Stigter, 2009; Huan et al., 2012; Massone et al., 2010; Vías et al., 2010), (2) to incorporate some index variables using linear regression (McLay et al., 2001; Sonneveld et al., 2010) and, (3) to incorporate GIS and fuzzy rule-based model with rules specified from expert knowledge, to generate groundwater vulnerability maps (Dixon, 2005; Pathak and Hiratsuka, 2011).

Within data-driven models statistical multivariate methods exist such as logistic regression (Nolan et al., 2002; Ozdemir, 2011), Weights of Evidence (Antonakos and Lambrakis, 2007; Sorichetta et al., 2013; Sorichetta et al., 2012), and a set of methods known as artificial intelligence or machine learning such as Adaptive Neuro-Fuzzy Inference System (Shiri and Kişi, 2011; Talei et al., 2010), genetic algorithms (Azamathulla et al., 2008; Babbar-Sebens and Minsker, 2010; Katsifarakis et al., 1999), artificial neural networks (Banerjee et al., 2011; J. Huang et al., 2011a; Sahoo et al., 2006; Zare et al., 2011), support vector machines (Shiri et al., 2013; Tripathi et al., 2006; Yoon et al., 2011) and more recently Random Forest (Baudron et al., 2013). The multivariate statistical methods together with Artificial Neural Networks are the most commonly used in Hydrogeology studies. The main reason is its greater accessibility, as these techniques are included within different software packages. However, these techniques show a variety of problems such as their sensibility towards outlier values of logistic regression and the opacity of neural networks (Abrahart et al., 2008).

In recent years, machine learning has experienced significant development and new methods have been proposed to solve some of the problems described for widely used methods (Khalil et al., 2005). An emerging type of machine learning techniques which utilizes ensembles of regressions is receiving highlighted interest in other fields of knowledge (Friedl et al., 1999; Gislason et al., 2006; Hansen and Salamon, 1990; Krogh and Vedelsby, 1995; Sesnie et al., 2008; Steele, 2000). Ensemble learning algorithms use the same base algorithm to produce repeated multiple predictions, which are averaged in order to produce a unique model (Breiman, 2001; Friedl et al., 1999). An ensemble learning technique called Random Forest (RF) is increasingly being applied in land-cover classification from remotely sensed data (Pal, 2005; Sesnie et al., 2008) and other fields related to the environment and water resources (Booker and Snelder, 2012; Herrera et al., 2010; Z. Huang et al., 2011b; Loos and Elsenbeer, 2011; McGinnis and Kerans, 2012; Rodriguez-Galiano et al., 2012d; Vincenzi et al., 2011; Zhao et al., 2012).

RF offer a new approach to the problem of vulnerability mapping, as it is relatively robust to outliers and it can overcome the “black-box” limitations of artificial neural networks, assessing the relative importance of the variables and being able to select the most important variables (features) and reducing dimensionality. At the same time the parameterization of RF is very simple and it is computationally lighter than other machine learning methods (neural networks or support vector machines) (Rodriguez-Galiano and Chica-Rivas, 2012). Although RF is being currently used as a remote sensing data classifier (Rodriguez-Galiano et al., 2012b), its potential as a spatial modeling tool for vulnerability mapping is still underexplored due to its novelty.