Science of the Total Environment 624 (2018) 661–672

Contents lists available at ScienceDirect

Science of the Total Environment

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

Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods

V.F. Rodriguez-Galiano [a,b], J.A. Luque-Espinar [c], M. Chica-Olmo [d], M.P. Mendes [e,*]

[a] Physical Geography and Regional Geographic Analysis, University of Seville, Seville 41004, Spain
[b] Geography and Environment, School of Geography, University of Southampton, Southampton SO17 1BJ, United Kingdom
[c] Unidad del IGME en Granada, Urbanización Alcazar del Genil, 4, 18006 Granada, Spain
[d] Departamento de Geodinámica, Universidad de Granada, Avenida Fuentenueva s/n, 18071 Granada, Spain
[e] CERIS, Civil Engineering Research and Innovation for Sustainability, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

HIGHLIGHTS

- Different Feature Selection approaches (FS) based on machine learning were evaluated.
- FS allowed to isolate and identify the main drivers of nitrate pollution in groundwater.
- Driving forces were more useful in predicting nitrates pollution in this case study.
- A novel feature, extracted from NDVI time series, was revealed as very promising.
- A Random Forest based wrapper outperformed the rest FS in predicting nitrates.

ARTICLE INFO

Article history:
Received 9 November 2017
Received in revised form 13 December 2017
Accepted 13 December 2017
Available online xxxx

Editor: D. Barcelo

Keywords:
Machine learning algorithms
Feature selection
Embedded methods
Wrapper methods
Groundwater
Nitrates

ABSTRACT

Recognising the various sources of nitrate pollution and understanding system dynamics are fundamental to tackle groundwater quality problems. A comprehensive GIS database of twenty parameters regarding hydrogeological and hydrological features and driving forces were used as inputs for predictive models of nitrate pollution. Additionally, key variables extracted from remotely sensed Normalised Difference Vegetation Index time-series (NDVI) were included in database to provide indications of agroecosystem dynamics.
Many approaches can be used to evaluate feature importance related to groundwater pollution caused by nitrates. Filters, wrappers and embedded methods are used to rank feature importance according to the probability of occurrence of nitrates above a threshold value in groundwater. Machine learning algorithms (MLA) such as Classification and Regression Trees (CART), Random Forest (RF) and Support Vector Machines (SVM) are used as wrappers considering four different sequential search approaches: the sequential backward selection (SBS), the sequential forward selection (SFS), the sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS). Feature importance obtained from RF and CART was used as an embedded approach. RF with SFFS had the best performance (mmce = 0.12 and AUC = 0.92) and good interpretability, where three features related to groundwater polluted areas were selected: i) industries and facilities rating according to their production capacity and total nitrogen emissions to water within a 3 km buffer, ii) livestock farms rating by manure production within a 5 km buffer and, iii) cumulated NDVI for the post-maximum month, being used as a proxy of vegetation productivity and crop yield.

1. Introduction

Nitrate in groundwater has been reported as a major problem all over the world. The Nitrates Directive (91/271/EEC, 1991) is an integral part of the water policy of the European Union (EU) and it was drawn up with the specific purposes of reducing water pollution caused by nitrates from agricultural sources and preventing further pollution.

* Corresponding author.
E-mail addresses: vrgaliano@us.es (V.F. Rodriguez-Galiano), ja.luque@igme.es (J.A. Luque-Espinar), mchica@us.es (M. Chica-Olmo), mpaulamendes@tecnico.ulisboa.pt (M.P. Mendes).

https://doi.org/10.1016/j.scitotenv.2017.12.152
0048-9697/

Different knowledge-driven and data-driven models can be used to recognise various sources of nitrate pollution and understand system dynamics. Knowledge-driven are models based on expert knowledge of processes that might have led to contamination in a given hydrogeological setting, but where no or very few data sample/pollution evidences are known to occur (Aller et al., 1987; Doerfliger and Zwahlen, 1997; Ribeiro, 2005). Data-driven models use objective evidence based on the associations between predictive variables and known occurrences of nitrate pollution (Solomatine et al., 2008). Within data-driven models, supervised machine learning algorithms (MLA) are normally applied from a set of training instances where each instance is described by a feature vector or attribute values (input variables) and a target feature expressed as a class label (classification) or a continuous value (regression) (Kohavi and John, 1998). In this case, the primary goal of predictive modelling is to maximise the accuracy (Motoda and Liu, 2002). Thus, the applicability of MLA on groundwater pollution issues is a consequence of their ability to recognise patterns of relationships among attributes and target feature, considering that there is some degree of uncertainty associated (Dixon, 2005). Indeed, MLA have been gradually used to predict nitrate concentration in groundwater, e.g., Random Forest (RF) (Rodriguez-Galiano et al., 2014; Tesoriero et al., 2017; Wheeler et al., 2015), Support Vector Machines (SVM) (Dixon, 2005; Khalil et al., 2005; Mohamad and Hassan, 2017), Artificial Neural Networks (Dixon, 2005; Khalil et al., 2005; Mohamad and Hassan, 2017; Nolan et al., 2015), Boosted Regression Trees and Bayesian Networks (Nolan et al., 2015), and Locally Weighted Projection Regression and Relevance Vector Machines (Khalil et al., 2005). Likewise, MLA have been applied to optimise subjective indexes methods for groundwater vulnerability assessment, e.g. (Fijani et al., 2013) and (Nadiri et al., 2017).

Common to all aforementioned studies is an undeniable fact that for the induction of a MLA, the groundwater experts can use all available features, or select a smaller number of them. Nevertheless, if there is a large number of features, different negative effects might occur, i.e.: i) irrelevant features can result in overfitting training data (i.e. poor generalisation), thus, reducing the model accuracy; ii) models with high complexity may limit their interpretability and, therefore, hamper the decision making process and; iii) models with several features can be impractical and hard to replicate to other areas. To address this issue, it is possible to precede learning with a feature selection stage that strives to eliminate some noise and redundant data, establishing the most significant attributes (Reunanen, 2006; Witten and Tibshirani, 2010).

Feature selection (FS) is a process that selects a subset of original attributes, so that the feature space is optimally reduced according to a certain criterion (Blum and Langley, 1997; Dash and Liu, 1997; Zhang et al., 2006). The goal of FS is to reduce the amount of features, focusing on the relevant data and improving their quality and hence contribute to a better understanding of the processes (i.e. nitrate pollution of groundwater) that is driven by the selected features (Guyon and Elisseeff, 2003; Motoda and Liu, 2002). Several statistical methods can be employed in FS such as filters, wrapper and embedded methods (Fig. 1). The filter approach is a preprocessing step and use criteria not involving any learning machine and, by doing that, it does not consider the effects of a selected feature subset on the performance of the algorithm (Guyon and Elisseeff, 2006; Kohavi and John, 1998; Lal et al., 2006). Wrapper methods evaluate a subset of features according to accuracy of a given predictor (Guyon and Elisseeff, 2003; Kohavi and John, 1998). Search strategies are used within wrapper methods to yield nested subsets of variables, the variable selection being based on the performance of the learned model (Guyon and Elisseeff, 2003; Hilario and Kalousis, 2008). Embedded methods perform variable selection during the process of training and are generally specific to given learning machines (Guyon and Elisseeff, 2003). In this case, the learning step and the feature selection part cannot be separated (Lal et al., 2006).

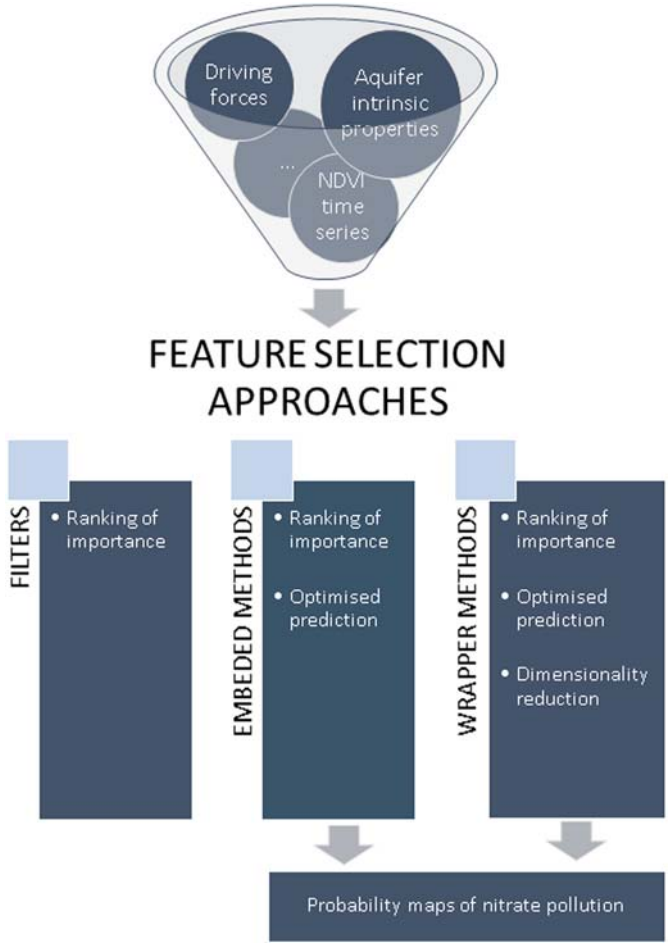

Fig. 1. Conceptual chart of feature selection for predictive modelling of groundwater nitrate pollution.

FS has been used to identify which variables are more relevant to predict nitrate concentration in groundwater, such as wrapper (Dixon, 2005; Khalil et al., 2005; Nolan et al., 2015; Wheeler et al., 2015) and embedded methods (Rodriguez-Galiano et al., 2014; Tesoriero et al., 2017). Wrappers or embedded methods include the use of non-parametric algorithms like decision trees, neural networks and support vector machines (Bazi and Melgani, 2006; Del Frate et al., 2005; Pal and Foody, 2010; Rodriguez-Galiano et al., 2012; Yu et al., 2002). Establishing features that are strongly related to nitrate pollution of groundwater can contribute to the establishment of better measures in the Action Programs (91/271/EEC, 1991), ensuring an effective reduction of groundwater pollution caused by nitrates and preventing further such pollution. In this study we aim to assess the performance of different FS methods (filters, wrapper and embedded) for defining which features can predict groundwater pollution by nitrates, using the following MLA: CART, Support Vector Machine and Random Forest. Furthermore, we intend to use a comprehensive database, where, as a novelty, new features are extracted from remotely-sensed time series of vegetation indices (weekly composites on an annual basis), allowing to infer the importance of agriculture in the prediction of groundwater nitrate pollution. The objectives of this study were: i) evaluation of the usefulness of different FS approaches; ii) recognition of the principal sources of nitrate contamination and understanding system dynamics and, iii) mapping of classifying probabilities of nitrate occurrence in groundwater above a threshold value.