Journal of Hydrology 499 (2013) 303–315

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

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

Identifying the origin of groundwater samples in a multi-layer aquifer system with Random Forest classification

Paul Baudron [a,d,e,*], Francisco Alonso-Sarría [b], José Luís García-Aróstegui [c], Fulgencio Cánovas-García [b], David Martínez-Vicente [a,b], Jesús Moreno-Brotóns [b]

[a] Fundación Instituto Euromediterráneo del Agua, Complejo Campus de Espinardo, Ctra. N301, 30100 Espinardo, Murcia, Spain
[b] University of Murcia, Institute for Water and Environment (INUAMA), Campus de Espinardo, 30100 Murcia, Murcia, Spain
[c] Geological Survey of Spain (IGME), Avda. Miguel de Cervantes, 45 – 5° A, 30009 Murcia, Murcia, Spain
[d] Institut pour la Recherche et le Développement, UMR G-EAU, Cemagref, 361 rue Jean-François Breton, BP 5095, 34196 Montpellier Cedex 5, France
[e] Univ Paris-Sud, Laboratoire IDES, UMR8148 IDES, Avenue du Belvédère, Bâtiment 504, Orsay F-91405, France

ARTICLE INFO

Article history:
Received 22 August 2012
Received in revised form 24 April 2013
Accepted 5 July 2013
Available online 15 July 2013
This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Chunmiao Zheng, Associate Editor

Keywords:
Multi-layer aquifer
Longscreen boreholes
Machine learning
Random Forest
Hydrogeochemistry
Hydrogeology

SUMMARY

Accurate identification of the origin of groundwater samples is not always possible in complex multi-layered aquifers. This poses a major difficulty for a reliable interpretation of geochemical results. The problem is especially severe when the information on the tubewells design is hard to obtain. This paper shows a supervised classification method based on the Random Forest (RF) machine learning technique to identify the layer from where groundwater samples were extracted. The classification rules were based on the major ion composition of the samples. We applied this method to the Campo de Cartagena multi-layer aquifer system, in southeastern Spain. A large amount of hydrogeochemical data was available, but only a limited fraction of the sampled tubewells included a reliable determination of the borehole design and, consequently, of the aquifer layer being exploited. Added difficulty was the very similar compositions of water samples extracted from different aquifer layers. Moreover, not all groundwater samples included the same geochemical variables. Despite of the difficulty of such a background, the Random Forest classification reached accuracies over 90%. These results were much better than the Linear Discriminant Analysis (LDA) and Decision Trees (CART) supervised classification methods. From a total of 1549 samples, 805 proceeded from one unique identified aquifer, 409 proceeded from a possible blend of waters from several aquifers and 335 were of unknown origin. Only 468 of the 805 unique-aquifer samples included all the chemical variables needed to calibrate and validate the models. Finally, 107 of the groundwater samples of unknown origin could be classified. Most unclassified samples did not feature a complete dataset. The uncertainty on the identification of training samples was taken in account to enhance the model. Most of the samples that could not be identified had an incomplete dataset.

1. Introduction

In complex multi-layer groundwater systems, the correct determination of the origin of a sample is the basic driving condition for a reliable interpretation of geochemical and hydrodynamic results. However, if there is no available information on the tubewell design, this driving condition can be hard to validate (Mayo, 2010). As a consequence, despite the large quantities of geochemical and piezometric data available, only those corresponding to fully documented tubewells should be used for investigation. Hence, there is a need for a tool that could provide an automatic and accurate estimation of the aquifer layer from which a water sample has been extracted. A possibility is to base this tool on geochemical criteria. Such a method must deal with additional difficulties as similar water types, temporal changes in the origin of groundwater, or having different ions analyzed in different samples. Moreover, it should be applicable with common major ion geochemistry. Such a tool could be helpful to all applications of geochemical data in Hydrogeology, as identifying anthropogenic transformation (e.g. Celle-Jeanton et al., 2009), understanding paleoclimates (e.g. Jiráková et al., 2009), determining mineralization processes (e.g. Gillon et al., 2012; Lorenzen et al., 2012), assessing groundwater flow patterns (e.g. Cronin et al., 2005) or calibrating groundwater flow models (e.g. Dahan et al., 2004), among other uses.

* Corresponding author at: Institut pour la Recherche et le Développement, UMR G-EAU, Cemagref, 361 rue Jean-François Breton, BP 5095, 34196 Montpellier Cedex 5, France. Tel.: +33 628 040 391.
E-mail addresses: paul.baudron@baudron.com (P. Baudron), alonsarp@um.es (F. Alonso-Sarría), j.arostegui@igme.es (J.L. García-Aróstegui), fulgencio.canovas@um.es (F. Cánovas-García), davidmv@um.es (D. Martínez-Vicente), jesusmorenobrotons@gmail.com (J. Moreno-Brotóns).

0022-1694/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2013.07.009

Statistical methods have been widely used in hydrology and Hydrogeology (e.g.; Adams et al., 2001; Lambrakis et al., 2004; Cloutier et al., 2008; Daughney et al., 2012). Generally, as a tool to subdivide and classify large hydrogeochemical datasets to facilitate interpretation. They might also be used to estimate mixing proportions (e.g. Valder et al., 2012). The techniques most applied are principal components analysis (PCA) and hierarchical cluster analysis (HCA). These techniques highlight tendencies inside groups of samples, allowing an easier representation of the results. However, these methods show several limitations, like the subjectivity of the criteria defining the classes, or its unsupervised nature. That is, they can be used to create a set of classes out of the whole dataset but they cannot assign samples to a set of a priori classes.

In contrast, in the supervised classification approach, the prediction of the output class of any new sample is enabled by a set of decision rules (classification model) defined out of a set of labeled training samples. This approach enables the prediction of the correct output class for any new input case including the same predictor variables. Linear Discriminant Analysis (LDA) is a classical multivariate technique for supervised classification (Vaselli et al., 1997).

However, traditional statistical methods have been proven inadequate to identify complex patterns and relationships that could be revealed by more sophisticated procedures (De'ath and Fabricius, 2000). These new procedures include computer intensive machine learning techniques based on recursion, sampling and randomizations (Babovic, 2005; Prasad et al., 2006).

Approaches based on decision trees (Breiman et al., 1984) are among the most applied supervised classification methodologies. Random Forest (Breiman, 2001), is the one that have recently received most interest. It combines a large numbers of decision trees (usually 500–2000) to obtain a more accurate classification without overfitting the model to a specific dataset.

Studies using decision trees can be found in Remote Sensing (e.g. Guhimre et al., 2010), Medicine (e.g. Lempitsky et al., 2009), Genetics (e.g. Cutler and Stevens, 2006), Chemistry (e.g. Svetnik et al., 2004), Ecology (e.g. Cutler et al., 2007) or Soil Science (e.g. Schmidt et al., 2008). Only a few studies use supervised classification methods in Hydrogeology. Use of decision trees as a supervised classification method has been limited to the studies byLoos and Elsenbeer (2011) on the links between overland flow generation and topography, and by Peters et al. (2008) on groundwater-dependent vegetation patterns. LDA has been applied to classify groundwater samples only in rare occasions (e.g. Lambrakis et al., 2004). Other machine learning methods as Neural Networks can be found in Hydrogeology (e.g. Kurtulus and Razack, 2007), but they are more difficult to calibrate and were not used in the present study. Except the recent studies by Smith et al. (2010) on bacterial source tracking in lakes and Olson and Hawkins (2012) on stream base-flow water chemistry, we have not been able to find any studies using Random Forest neither in Hydrogeology nor for the analysis of hydrogeochemical datasets.

Our main goal was to test the Random Forest classification method to determine the origin of groundwater samples based on their geochemistry. The study was conducted in an intensively irrigated region with hundreds, many of them undocumented, tubewells. These tubewells provide a large geochemical dataset whose interpretation is hazardous due to the lack of design-information. Linear discriminant analysis and a simple classification tree were also used to compare results.

2. Study site

The Campo de Cartagena, in southeastern Spain (Fig. 1), is a 1440 km^2 coastal plain whose elevation ranges between 0 and 200 m a.m.s.l. The climate is semiarid with an average temperature of 18 °C and an average rainfall of about 300 mm per year. High variability is another characteristic of precipitation. Several years have registered values lower than 200 mm and, at the same time, more than 150 mm can be registered during a few days, mainly in spring and autumn. The main consequence is the lack of permanent watercourse, though several ephemeral streams drain the area. Groundwater and the Tagus–Segura water transfer, initiated in 1980 to derive water from the Tagus basin to the Segura basin, are the main sources of water supply (Baudron et al., in press; Rey et al., 2013).

The economy of the area relies on the agro-industrial sector with crops covering 1/3 of the total surface. Due to the low precipitation rate and a lack of permanent surface water, intensive irrigated agriculture has historically been mainly supported by groundwater extraction from the regional multi-layer aquifer system.

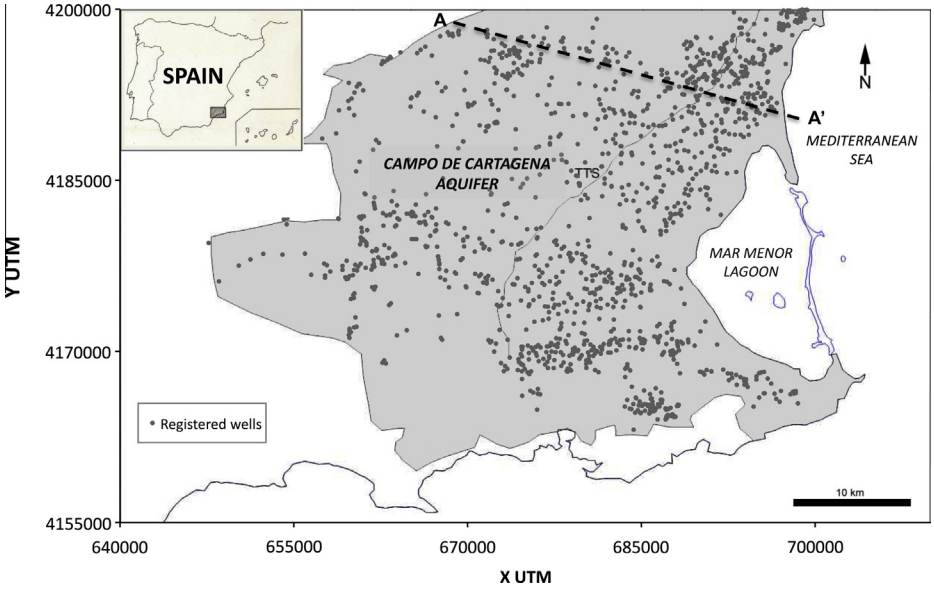

Fig. 1. Map of the study area, with the location of all registered wells and the geological cross-section of Fig. 2.