Journal of Hydrology 519 (2014) 626–636

Contents lists available at ScienceDirect

Journal of Hydrology

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

Model-based clustering of hydrochemical data to demarcate natural versus human impacts on bedrock groundwater quality in rural areas, South Korea

Kyoung-Ho Kim [a], Seong-Taek Yun [a,*], Seong-Sook Park [a,b], Yongsung Joo [c], Tae-Seung Kim [d]

[a] Department of Earth and Environmental Sciences & KU-KIST Green School, Korea University, Seoul 136-701, South Korea
[b] Department of Natural Resources and Environmental Engineering, Hanyang University, Seoul, South Korea
[c] Department of Statistics, Dongguk University, Seoul, South Korea
[d] Division of Soil and Groundwater Research, National Institute of Environmental Research, Incheon, South Korea

ARTICLE INFO

Article history:
Received 9 March 2014
Received in revised form 21 July 2014
Accepted 29 July 2014
Available online 9 August 2014
This manuscript was handled by Peter K. Kitanidis, Editor-in-Chief, with the assistance of Martin Thullner, Associate Editor

Keywords:
Hydrochemistry
Bedrock groundwater quality
Model-based clustering
Normal (Gaussian) mixture model
Natural versus anthropogenic processes

SUMMARY

Improved evaluation of anthropogenic contamination is required to sustainably manage groundwater resources. In this study, we investigated the hydrochemical measurements of 18 parameters from a total of 102 bedrock groundwater samples from two representative rural areas in South Korea. We used model-based clustering with a normal (Gaussian) mixture model to differentiate the contributions of natural versus anthropogenic processes to the observed groundwater quality. Water samples varied in hydrochemistry from a Ca–Na–HCO_3 type to a Ca–HCO_3–Cl type. The former type reflected derivation of major ions largely from water–rock interactions, while the latter type recorded varying degrees of anthropogenic contamination. Among the major dissolved ions, fluoride and nitrate were shown to be good indicators of the two types, respectively. The results of model-based clustering showed that the bivariate normal mixture model, which was based on the covariance of nitrate and fluoride, was more robust than multivariate analysis, and provided better discrimination between the anthropogenic and natural groundwater groups. Model-based clustering to measure the degree of cluster membership for each sample also showed a gradual change in groundwater chemistry due to mixing between the two water groups. This study provided an example of the successful application of model-based clustering to evaluate regional groundwater quality and demonstrated that better selection of the dimensional structure (i.e., selection of optimal variables and number of clusters) based on hydrochemistry was crucial in obtaining reasonable clustering results.

1. Introduction

Groundwater chemistry is inherently controlled by water–rock interactions in the aquifer (Stumm and Morgan, 1996; Langmuir, 1997; Appelo and Postma, 1999). However, the natural quality of groundwater has progressively deteriorated in many countries due to diverse human impacts (Foster and Chilton, 2003; Morris et al., 2003). In South Korea, groundwater in fractured bedrock (gneisses and granitoids which occupy ca. 70% of the Korean Peninsula) forms the most important source of water supply, especially to rural communities where surface water reservoirs are generally absent. The deterioration in the quality of bedrock groundwater has partly resulted from a lack of effective management schemes, and thus the Korean government has recently made great efforts to sustainably manage bedrock aquifers (e.g., 'Master Plan for Managing Groundwater Resources'; MOCT and KOWACO, 2002). To properly manage available water resources, there is a need for a quantitative understanding of water quality deterioration from diffusive or non-point pollution sources, as well as the background chemistry associated with natural geochemical processes.

Through careful examination of the spatial variations in hydrochemical data, it is possible to estimate the relative contribution of natural versus anthropogenic effects on groundwater quality. Especially for large and/or complicated datasets, multivariate statistics (e.g., clustering) have been widely used for these determinations (Suk and Lee, 1999; Güler et al., 2002; Farnham et al., 2003; Güler and Thyne, 2004; Cloutier et al., 2008; Yidana, 2010). Diverse clustering methods such as fuzzy *c*-means (FCM) clustering, hierarchical clustering, and *k*-means clustering have been used to separate groundwater samples into homogeneous groups that

* Corresponding author. Tel.: +82 2 32903176; fax: +82 2 32903189.
E-mail address: styun@korea.ac.kr (S.-T. Yun).

http://dx.doi.org/10.1016/j.jhydrol.2014.07.055
0022-1694/

reflect the different source contributions to groundwater chemistry.

For successful application of clustering techniques, selection of suitable algorithms for the hydrochemical dataset is important (Templ et al., 2008). In hierarchical clustering and partitioning (or iterative relocation) methods, distance measurements (e.g., Euclidean distance) between observations are used. However, these methods each present their own challenges in determining the number of clusters and initial cluster centers. As an alternative technique, model-based clustering was recently adopted and can provide a principled statistical assessment of the practical problems that arise in applying clustering methods (McLachlan and Peel, 2000; Fraley and Raftery, 2002). The model-based clustering algorithms are not based on the distance measure, but use a probability mixture model to define each cluster as a subpopulation in the dataset, with the aim of optimizing the fit between the model and dataset. The implicit assumption of this method is that each cluster is represented by a parametric probability density, and the entire cluster structure can be modeled by a finite mixture.

Since the model-based clustering is based on a probability model, clustering results can be largely affected by the dimension (number of variables) of the input dataset. This suggests that the clustering structure of interest may be contained in a subset of the available variables and that some variables may be useless or even have a disadvantage in detecting a reasonable clustering result (Law et al., 2004; Tadesse et al., 2005; Raftery and Dean, 2006; Joo et al., 2009; Maugis et al., 2009). Thus, it is very important to assess the relevant variables. The model-based clustering can also provide a rigorous framework to assess the role of each variable in the clustering process (McLachlan and Peel, 2000). Therefore, model-based clustering has been shown to be a powerful tool for classification in many research fields (McLachlan and Basford, 1988; Banfield and Raftery, 1993). However, the practical applications have rarely been reported in hydrochemical or geochemical studies (Templ et al., 2008).

This study was based on a hydrochemical investigation of bedrock aquifers in representative rural areas of South Korea and applied model-based clustering with a normal (Gaussian) mixture model to separate the contributions of natural and anthropogenic processes on observed groundwater quality. To obtain reliable clustering performance, detailed hydrochemical knowledge of the aquifer was incorporated into the clustering process, and the clustering results were validated using statistical criteria. This case study demonstrated the significant influence of the dimensional structure and selection of optimal variables on the clustering results and also suggested the important role of hydrochemical interpretation in better predicting the cluster structure in advance of clustering analysis.

2. Materials and methods

2.1. Study areas

Hydrochemical surveys were conducted in selected rural areas on the outskirts of Boeun and Naju cities, South Korea (Fig. 1a) where bedrock groundwater forms an important source of the domestic and agricultural water supply. In both cities, about 70% of total households receive public main water supply, while in suburban areas with active agricultural activities, water use depends on groundwater from bedrock aquifers. The total populations and population densities of Boeun and Naju cities are ca. 15,000 and 240 per km^2 (Boeun), and ca. 95,000 and 150 per km^2 (Naju). Agricultural fields occupy about 30% (Boeun) and 39% (Naju) of the total land surface area and primarily consist of rice paddies (Boeun) or dry fields for vegetable and fruit production (Naju).

The topography of the Boeun area consists of a basin surrounded by steep mountains (ca. 400–900 m a.s.l.). The mean annual temperature, humidity, and precipitation in Boeun are 10.7 °C (−3.9 °C in January and 24.1 °C in July), 72.1% and 1260 mm/year, respectively. Naju is located on a wide plain with low elevations (<10 m to 20–50 m a.s.l.) and is surrounded by several mountains (ca. 300–450 m a.s.l.). The mean annual temperature and precipitation in the Naju area are 13.8 °C (1.1 °C in January and 28 °C in August) and ca. 1500 mm/year, respectively.

The geology of Boeun consists of Jurassic Boeun Granite (granodiorite and biotite granite), two-mica adamellite, and Quaternary alluvium (Fig. 1b). The geology of Naju is comprised of Jurassic biotite granite, Cretaceous volcanic-sedimentary rocks (andesite, andesitic tuff, and rhyolite) and quartz porphyry, and Quaternary alluvium (Fig. 1c). Geologic characteristics of the study areas can be available from the Geologic Information Search System (http://mgeo.kigam.re.kr/) of the Korea Institute of Geoscience and Mineral Resources (KIGAM). Wells for bedrock groundwater in both areas are predominantly situated in the Mesozoic granitoids.

2.2. Sampling and chemical analysis

A total of 102 bedrock groundwater samples (Boeun: 52, Naju: 50) were collected from pre-existing wells in January 2002 and February 2003 (Table 1; Fig. 1b and c). The average depth of the wells was 141 m (range: 54–300 m). The sampling and *in-situ* measurements of groundwater were performed based on the Standard Methods (APHA et al., 2001). Temperature, pH, Eh, electrical conductivity (EC) and dissolved oxygen (DO) were measured with portable meters (Orion Co.) after purging at least three well volumes (Appelo and Postma, 1999). Alkalinity was also measured in the field using the titration technique to determine the acid neutralizing capacity of the carbonate species (mainly HCO_3^-) in groundwater (Stumm and Morgan, 1996). Groundwater samples were filtered using 0.45 μm cellulose membranes and transferred into pre-washed HDPE bottles. Samples for the analysis of major cations were acidified to pH < 2 by adding a few drops of ultrapure nitric acid.

The chemical analyses were conducted at the Center for Mineral Resources Research (CMR) of Korea University, and also followed the Standard Methods (APHA et al., 2001). Dissolved cations (Na^+, K^+, Ca^{2+}, Mg^{2+}, Fe_{total}, Mn_{total}, and Sr^{2+}) and dissolved silica (SiO_2) were analyzed using Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES; Perkin Elmer OPTIMA 3000XL) and dissolved anions (Cl^-, SO_4^{2-}, NO_3^-, F^-) were determined using ion chromatography. The quality of the chemical analyses was carefully monitored by analyzing blanks and duplicate samples, as well as by calculating charge balances (Hounslow, 1995). Charge balance errors (C.B.) for the analyses were less than ±5%.

2.3. Model-based clustering

The model-based clustering (i.e., cluster analysis of a mixture model) was performed on the hydrochemical dataset to separate the samples into natural versus anthropogenic groups. Model-based clustering assumes that the multivariate dataset consists of a number of clusters and that each cluster can be described by a normal (Gaussian) distribution, that is, the dataset can be expressed as a mixture of multivariate normal distributions (McLachlan and Peel, 2000). With this assumption, the expectation maximization (EM) approach (Dempster et al., 1977) is used to calculate the maximum likelihood estimates of the parameters with normal distributions. In this study, the model-based clustering procedures were carried out by means of the Mixture Modeling (MIXMOD) program (Biernacki et al., 2006).