Journal of Cleaner Production 122 (2016) 277–286

Contents lists available at ScienceDirect

Journal of Cleaner Production

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

Decision tree-based data mining and rule induction for identifying hydrogeological parameters that influence groundwater pollution sensitivity

Keunje Yoo [a], Sudheer Kumar Shukla [a], Jae Joon Ahn [b], Kyungjoo Oh [b], Joonhong Park [a, *]

[a] School of Civil and Environmental Engineering, Yonsei University, 262 Seongsanno, Seodaemun-gu, Seoul, 120-749, South Korea
[b] School of Information and Industrial Engineering, Yonsei University, 262 Seongsanno, Seodaemun-gu, Seoul, 120-749, South Korea

ARTICLE INFO

Article history:
Received 9 September 2014
Received in revised form
22 September 2015
Accepted 25 January 2016
Available online 18 February 2016

Keywords:
Data mining
Groundwater pollution
Groundwater vulnerability
Trichloroethylene

ABSTRACT

This study aims to develop a new field-based approach that can estimate patterns of groundwater pollution sensitivity using data mining algorithms. Hydrogeological and pollution sensitivity data were collected from the Woosan Industrial Complex, Korea, which is a site contaminated by trichloroethylene (TCE). The proposed data mining algorithm procedure uses seven hydrogeological properties as input variables: depth to water, net recharge, aquifer media, soil media, topography, vadose zone media, and hydraulic conductivity. The observed TCE sensitivity was used as the target data. Initially, four data mining algorithms artificial neural network (ANN), decision tree (DT), case-based reasoning (CBR), and multinomial logistic regression (MLR) were tested. We found that the DT-based data mining and rule induction method shows better prediction accuracy and consistency than the other methods. We also used the ordinal pairwise partitioning (OPP) algorithm to improve the accuracy and consistency of the DT model. A classification and regression tree (CART) analysis of the OPP-DT model indicated that the net recharge (R), soil media (S), and aquifer media (A) were the major hydrogeological factors that influence groundwater sensitivity to TCE at the site. The results of this study demonstrate that the proposed model can provide more accurate and consistent estimates of groundwater vulnerability to TCE compared to the existing models.

1. Introduction

The intensive industrial use of chlorinated solvents, such as trichloroethylene (TCE), has caused these chemicals to be the most frequently detected type of groundwater contamination (USEPA, 2003; Rivett et al., 2014). TCE is highly carcinogenic to animals (USEPA, 2005), and its presence in groundwater is a substantial and considerable concern to human health (USEPA, 2003, 2005, 2006). Because TCE is a non-aqueous phase liquid (NAPL) with a density heavier than water, its introduction to the subsurface environment may result in the presence of persistent NAPL residuals in the unsaturated zone or weathered/fractured rocks and may cause vertical spreading of the groundwater contamination plume (Jackson, 1998; Chambers et al., 2004; Rivett et al., 2014). For long-term contaminated sites, the locations of NAPL residuals are generally difficult to determine. Because of the complex and problematic nature of TCE groundwater contamination, the remediation of long-term TCE-contaminated groundwater in a weathered/fractured rock environment is regarded as one of the most difficult remedial tasks.

Groundwater TCE contamination in the Woosan Industrial Complex in Wonju (Gangwon Province, South Korea) is a model case of long-term TCE-contaminated groundwater in a weather/fractured rock environment (EMC, 2003; Yang et al., 2003; KECO, 2008; Baek and Lee, 2011; Yang et al., 2012). In 1995, a significant amount of TCE NAPL was accidentally released into the subsurface environment at a location on the site. In the early phase, the groundwater was contaminated with high levels of TCE (undetected to 10 mg/L) that exceed the Korean Groundwater Quality Standard (TCE < 0.03 mg/L). After pump-and-treatment methods were used in 2003, the groundwater appeared to be clean. However, since 2006, TCE has re-appeared in the groundwater (KECO, 2008; Baek and Lee, 2011; Yang et al., 2012). Recent field studies suggest that the re-appearance of TCE in groundwater might be

* Corresponding author. Tel.: +82 2 2123 5798; fax: +82 2 312 5798.
E-mail address: parkj@yonsei.ac.kr (J. Park).

http://dx.doi.org/10.1016/j.jclepro.2016.01.075
0959-6526/

attributable to the presence of TCE NAPL residuals at the site (KECO, 2008; Baek and Lee, 2011; Yang et al., 2012; Rivett et al., 2014). Unidentified locations of multiple TCE NAPL residuals and hydrogeological complexity and heterogeneity in long-term TCE-contaminated sites may impair decision making with respect to planning groundwater protection and remediation (Rivett et al., 2012).

The estimation of groundwater pollution vulnerability (or sensitivity) is an important factor when prioritizing the planning of groundwater conservation and contaminant remediation actions (Gogu and Dassargues, 2000). Theoretically, groundwater vulnerability to a contaminant can be directly measured via the field observation of changes in contaminant concentration. However, such direct measurement is not practical due to its relatively high cost. Instead, index models and/or deterministic process-based models are generally used to estimate groundwater contamination sensitivity based on the available hydrogeological information of a site and/or the chemical and source characteristics of the groundwater contaminants (Aller et al., 1987; Dixon, 2005). In the Woosan Industrial Complex case study, the data on the temporal and spatial distributions of groundwater TCE concentrations were insufficient for deterministic process-based modeling (EMC, 2003; KECO, 2008; Baek and Lee, 2011), and model parameters for sorption, advection, and bulk density were not independently measured. These limitations make it difficult to use a process-based model to estimate the groundwater contamination sensitivity of the site. This difficulty is also generally true for other groundwater contamination field studies. In fact, only a very limited number of studies have completed detailed field characterizations to examine the temporal/spatial distribution of TCE contaminants near a localized source zone (Yang et al., 2012; Rivett et al., 2014). As an alternative, the DRASTIC model was developed by the U.S. Environmental Protection Agency (EPA) to evaluate the groundwater contamination potential for the entire United States (Aller et al., 1987). This model is based on the concept of the hydrogeological setting, which is defined as a composite description of all the major geologic and hydrologic factors that affect and control the groundwater movement into, through and out of an area (Aller et al., 1987). The acronym represents seven hydrogeological parameters taken into consideration in the evaluation procedure, as noted in Table 1. Each DRASTIC parameter is evaluated with respect to the others in order to determine the relative importance of each and is then assigned a relative weight, ranging from 1 to 5. The most significant parameters are given a weight of 5, while the least significant receive a weight of 1. The DRASTIC Index is then computed by applying a linear combination of all factors according to the following equation:

$$\text{DRASTIC Index} = \text{DrDw} + \text{RrRw} + \text{ArAw} + \text{SrSw} + \text{TrTw} + \text{IrIw} + \text{CrCw}$$

where the subscripts r and w are the corresponding ratings and weights, respectively.

The DRASTIC model (a representative index model [Aller et al., 1987]) was developed for hydrogeologically simple North American aquifers and may not be suitable for the complex and heterogeneous hydrogeological characteristics of the Woosan Industrial Complex site. Recently, data mining and rule induction approaches are frequently used to predict previously unknown events using already-available information. Data mining is potentially applicable in groundwater contaminant sensitivity analyses based on available hydrogeological information (Fijani et al., 2013; Pacheco et al., 2015). Decision Tree (DT)-based rule induction may be a suitable data mining option for predicting groundwater contamination sensitivity because it can be feasibly applied when only a small size of data are available, when sufficient knowledge of cause-and-effect relationships is lacking, and when complex nonlinear relationships exist in the available dataset (Singh and Datta, 2007; Kim et al., 2011; Ahn et al., 2012). In addition, rule induction that involves training with the relationships between measured independent and dependent variables can be used in identifying key independent variables influencing dependent variable values and in predicting previously unmeasured dependent variable values using their corresponding independent variable values (Breiman et al., 1984; Berry and Linoff, 2004). The suggested applicability of DT and rule induction in groundwater contamination sensitivity has yet to be evaluated.

In this study, the research objectives were (i) to evaluate the validity of use of DT and rule induction in predicting groundwater TCE sensitivity using hydrogeological input variables for a TCE-contaminated site and (ii) to develop a method for identifying key hydrogeological input variables influencing the groundwater TCE sensitivity of a study site. Using the results from the second

Table 1
Summary of TCE concentrations and their corresponding hydrogeological properties at the study site.

(a) Target variable (N = 114)

Variables	Unit	Mean	Max	Min	Std. Dev.
TCE	mg/L	0.14	3.50	0	0.81

(b) Input variables (N = 114)

Variables	Weight	Unit	Classified compositions	Mean	Max.	Min.	Std. Dev.
D(Depth to water)	5	m	Continuous numeric variables	6.54	13.65	2.04	2.43
R(Net recharge)	4	%		7.30	12.85	2.17	3.18
T(Topography)	1	%		1.55	5.00	0.50	1.52
C(Hydraulic conductivity)	3	cm/sec		3.90×10^{-3}	6.06×10^{-2}	5.30×10^{-4}	0.22
				Total composition (%)			
A(Aquifer media)	3	NA[a]	Weathered Metamorphic/Igneous	50.36			
			Coarse sand and silt	28.38			
			Sandstone	21.26			
S(Soil media)	2	NA[a]	Sand and Concrete	57.34			
			Sandy Loam	30.22			
			Silty Loam	12.44			
I(Impact of the vadose zone)	5	NA[a]	Sand and Gravel with significant Silt and Clay	30.00			
			Metamorphic/Igneous	47.54			
			Sand and Gravel	22.46			

[a] Indicates non-available.