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.

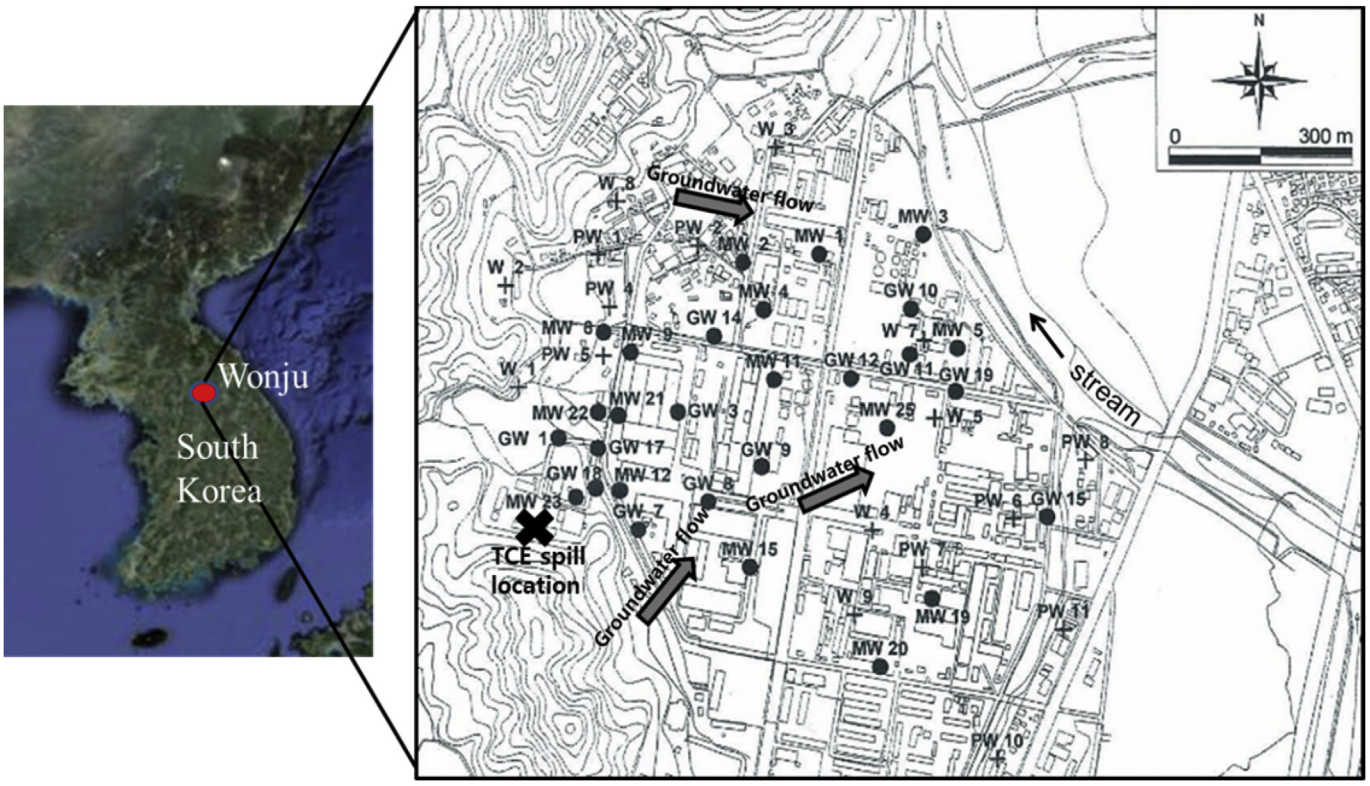


Fig. 1. Location of the Woosan Industrial Complex area. The groundwater wells at the study site are shown. The TCE spill location, which is regarded as the primary source of TCE contamination, is also shown by (**X**).

objective, this study attempted to determine the location(s) of the potential TCE NAPL residuals.

2. Materials and methods

2.1. Study site

The study site is located in an industrial complex in Wonju City, located approximately 75 km east of Seoul, the capital of the Republic of Korea (Fig. 1). The study site is approximately 0.65 km^2 and contains approximately 40 public, commercial, and residential buildings. The study site is surrounded by low-relief mountains to the west and a stream to the east that has a mean width of 40 m and flows to the north. Thirty groundwater pumping wells for industrial and domestic purposes were previously located on the study site, but 18 wells were closed because of TCE contamination, leaving only 12 pumps remaining in operation in 2004 (EMC, 2003; KECO, 2008). The aquifer at this site consists of weathered and fractured Jurassic biotite granite overlain by soil and alluvial deposits that are 10–15 m thick (EMC, 2003; KECO, 2008; Baek and Lee, 2011).

The alluvial deposits consist primarily of two types of sediments: silty sand and coarse sand. The subsurface geologic properties of the western and eastern regions of the study site are significantly different. Steep slopes exist in the western part of the study site, and gentle slopes exist in the eastern part. Most of the wells in the western part are located in a weathered granite aquifer at depths of greater than 20 m. An asphalt laboratory is situated on a hill in the western mountains and is regarded as the primary source of contamination (Fig. 1). The highest TCE levels were found at this location in the soil and groundwater survey (EMC, 2003; KECO, 2005; KECO, 2008). Groundwater generally flows from the west (the mountainous area) to the east (the stream). The hydraulic gradients are steeper in the west and gentler in the east near the stream (EMC, 2003; KECO, 2008). The water levels are higher in the wet season (June to early September) relative to the dry season (November to early May), but the flow directions and hydraulic gradients were largely unaffected by the season (EMC, 2003; KECO, 2008).

2.2. Data collection and preprocessing

TCE concentration data and hydrogeological properties were taken from previous groundwater contamination surveys conducted by the Wonju Environmental Management Corporation

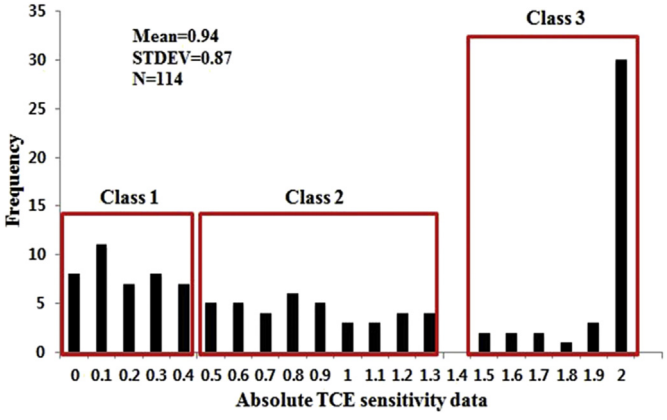


Fig. 2. Histogram of the classified absolute TCE sensitivity data. The values of the absolute TCE sensitivities were sorted from low to high (zero to two).

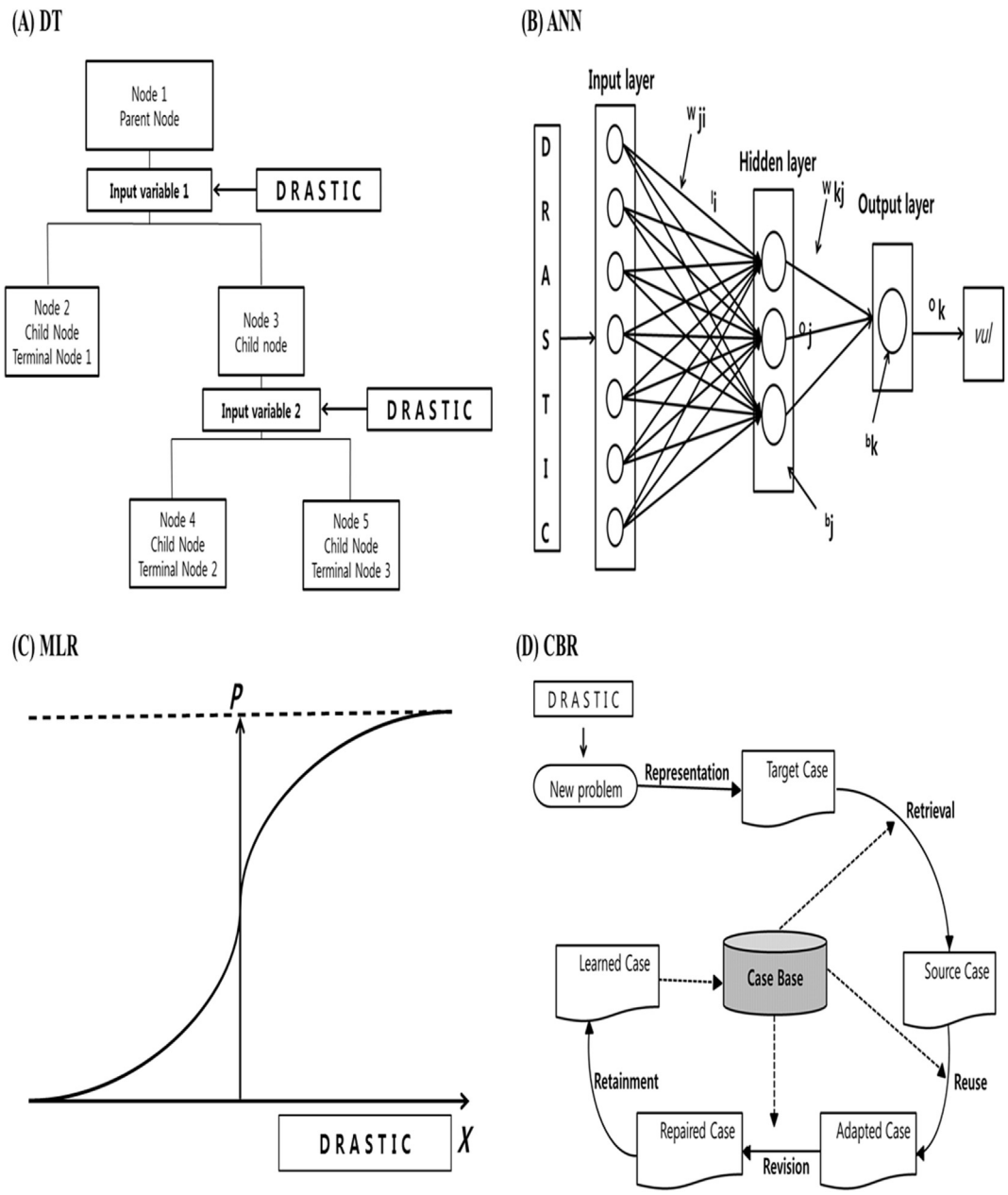


Fig. 3. Conceptual schemes of the data mining methods used in this study: the structure of CART process in DT (A), ANN process (B), MLR process (C), and CBR process (D).

(EMC) and the Korea Environment Corporation (KECO) (EMC, 2003; KECO, 2005; KECO, 2008). The TCE contamination data were collected from available wells three times in 2003, 2004, and 2008: 44 samples in February, 21 samples in April and 49 samples in August. The obtained data locations are shown in Fig. 1.

This study focused on seven hydrogeological properties (Depth to water [D], Net recharge [R], Aquifer media [A], Soil media [S], Topography [T], Impact of vadose zone media [I], and Hydraulic conductivity [C]) that are also considered in the DRASTIC method (Aller et al., 1987; Fijani et al., 2013; Mair and El-Kadi, 2013; Rodriguez-Galiano et al., 2014). These hydrogeological properties have been widely used by researchers in data mining models to assess groundwater vulnerability and sensitivity (Dixon, 2005; Fijani et al., 2013; Mair and El-Kadi, 2013; Rodriguez-Galiano et al., 2014). The D, A, I, and C data for the study site were obtained from EMC and KECO groundwater survey reports (EMC, 2003; KECO, 2005; KECO, 2008). The R data were calculated via the SCS-CN method (Mishra and Singh, 2003) using soil type data from the Korean Soil Information System [KSIS] (http://soil.rda.go.kr/) and the Korea Environmental Geographic Information System [KEGIS] (http://egis.me.go.kr/egis/) and precipitation data from the Wonju meteorological station. The S and T data were also obtained from the EMC and KECO reports and the KSIS and KEGIS web databases. Table 1 is a summary of the TCE concentrations and hydrogeological properties for all the data collected within the study site in 2003, 2004, and 2008.

For data normalization, the min–max normalization method was used in this study (Eq. (2.1)). This method is the simplest normalization technique and is widely used in statistical procedures (Huber, 1981; Breiman et al., 1984).

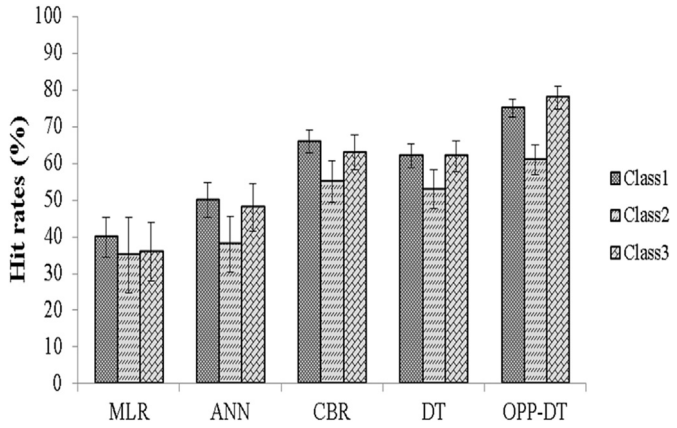


Fig. 4. Comparison of the hit rate results of different data mining and DRASTIC methods. The Y-error bar indicates the normalized standard deviation. OPP-DT denotes ordinal pair-wise partitioning applied with a DT. Class 1, Class 2, and Class 3 represent low, medium, and high TCE sensitivity, respectively.

$$V' = \left(\frac{V - min}{max + min}\right) \tag{2.1}$$

where v indicates the raw spectral value; *min* and *max* are the minimum and maximum values of the dataset, respectively; and V′ represents the normalized value.

Because the TCE concentrations in this study changed over time in 2003, 2004, 2008, the variations in the TCE concentrations were calculated according to the time recorded at the TCE sampling well point. This value was then used as a quantitative measure of the TCE sensitivity. This TCE sensitivity was calculated using Eq. (2.2). Additionally, to focus on the quantitative degree (low to high) of TCE variations according to the time interval, absolute TCE sensitivity data, as calculated by Eq. (2.3), were also used in this study.

$$\text{TCE sensitivity} = \left(\frac{C_{i+1} - C_i}{C_i}\right) \tag{2.2}$$

where Ci+1 and Ci correspond to the TCE concentrations at same sampling well points at different times.

The absolute TCE sensitivity was calculated using Eq. (2.3):

$$\text{Absolute TCE sensitivity} = |\text{TCE sensitivity}| \tag{2.3}$$

In this study, the calculated absolute TCE sensitivity values were converted into discrete values, and the TCE sensitivity data were grouped into three categories (Pal and Mather, 2003; Kim et al., 2011). The values of the absolute TCE sensitivity were sorted from low to high (from zero to two), with the data in the 0–33 percentiles (low) grouped into Class 1; the data in the 33–66 percentiles (moderate) grouped into Class 2; and the data in the 66–100 percentiles (high) grouped into Class 3 (Fig. 2). Class 3 represents a dramatic increase or decrease in TCE variations over time, Class 2 represents general variations in TCE concentrations over time, and Class 1 indicates relatively little variation in TCE concentrations over time at a study site.

2.3. Data mining model application

In this study, to estimate the groundwater pollution sensitivity to TCE contamination, four representative data mining methods – DT, ANN, MLR, and CBR – were applied. These methods have successfully predicted the specific vulnerability or pollution risk of an aquifer in many previous studies (Eisenberg and McKone, 1998; Pesch et al., 2008; Zhang et al., 2008; Vega et al., 2009; Cho et al., 2011; Ahn et al., 2012; Fijani et al., 2013).

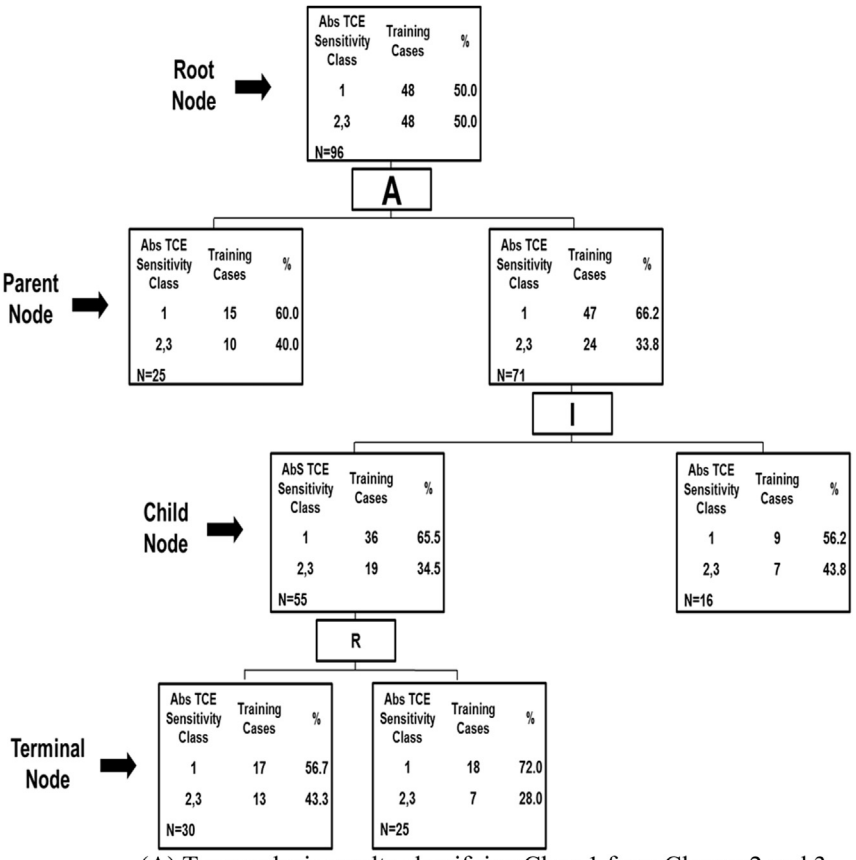
The four proposed data mining procedures used the seven hydrogeological properties as independent variables and discrete values of TCE sensitivity as the dependent variable. The rating and weight values from the DRASTIC model were not considered in this study, but the seven hydrogeological parameters from the DRASTIC model were used in the data mining. To apply the CART method in DT to this study, seven properties (*D*, *R*, *A*, *S*, *T*, *I*, *and C*) were used as independent variables, TCE sensitivity data were used as dependent variables, and the Gini index was used to determine the dataset. A generic illustration of the CART output is presented in Fig. 3(a). To apply the ANN method in this study, seven neurons (*D*, *R*, *A*, *S*, *T*, *I*, and *C*) were used in the independent layer, three neurons were used as the hidden layer, and one neuron (TCE sensitivity) was used as the dependent layer (Fijani et al., 2013). The traditional process involved in ANNs can be represented by a schematic learning process, as shown in Fig. 3(b). To apply the MLR method in this study, seven properties (*D*, *R*, *A*, *S*, *T*, *I*, and *C*) were used as the independent variables, and the TCE sensitivity data were used as the dependent variables. The traditional process involved in the MLR method can be represented by a schematic cycle, as shown in Fig. 3(c). To apply the CBR method in this study, seven properties (*D*, *R*, *A*, *S*, *T*, *I*, and *C*) were used as the independent variables, and the TCE sensitivity data were used as the dependent variables. The traditional process involved in the CBR method can be represented by a schematic cycle, as shown in Fig. 3(d).

In this study, the SAS Enterprise Miner (SAS Institute Inc., Cary, NC, USA, 2009) was used for the data mining learning (SAS, 2009). The cross-validation technique was also used in this research (Cawley and Talbot, 2003; Dixon, 2005; Chang et al., 2013). To examine the stability of the data mining models, 10-fold cross-validation was executed and the entire procedure was repeated 100 times. To evaluate the performance of the data mining models, the data were partitioned into a training set (70% of the dataset for each class) and a testing set (the remaining 30% of each dataset). The training set was used to determine an optimal value from one or more predictors during the learning phase of data mining. The testing set was used to evaluate the optimal value by verifying the prediction accuracy of the target data (TCE sensitivity data).

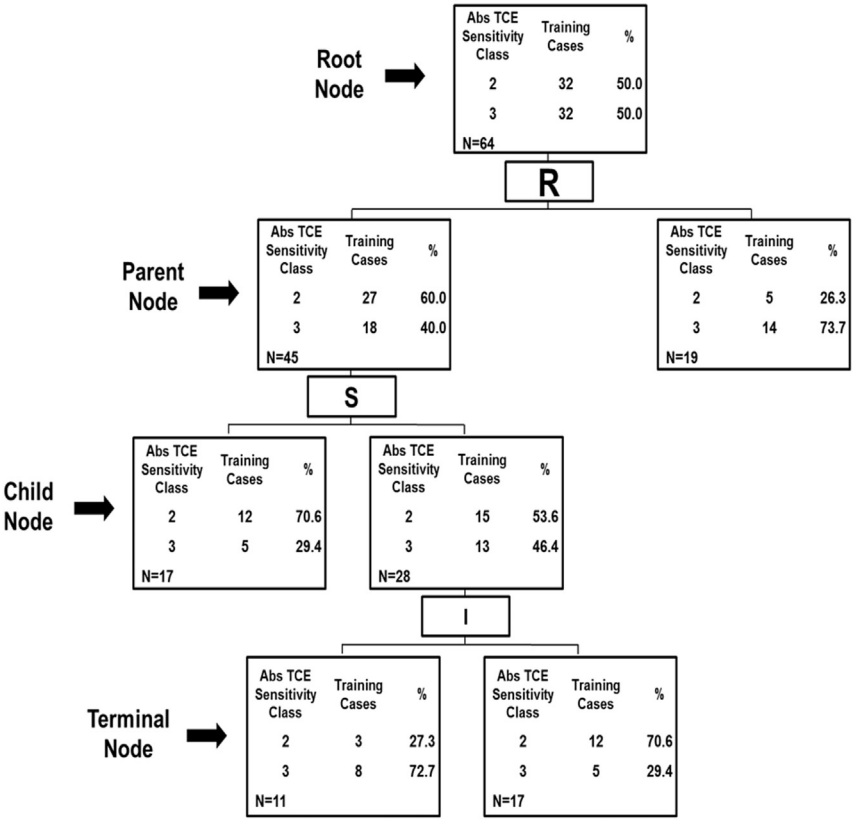
To improve the performance of the data mining results, the Ordinal Pairwise Partitioning (OPP) method was applied in this study (Kwon et al., 1997; Huang and Moraga, 2004; Kim et al., 2011; Park et al., 2011). The OPP method uses a partitioning approach with the format of (N) and (remaining classes). The goal of the OPP approach is to partition the dependent variable dataset into sub-datasets with reduced classes in an ordinal and pairwise manner according to the output classes (Kwon et al., 1997; Huang and Moraga, 2004; Kim et al., 2011; Park et al., 2011). In the case of a three-class scenario, the OPP method separates Class 1 from the remaining classes, and Classes 2 and 3 are then separated with the remaining data.

2.4. Performance criteria for classification

To evaluate the performance of a classification model, the performance measures based on the values of the confusion matrix generally include hit rate, accuracy, capture rate, and lift rate (Sokolova and Lapalme, 2009; Kim and Chang, 2011). The hit rate is an assessment criterion to compare the performance of the classification models (Sokolova and Lapalme, 2009; Kim and Chang, 2011; Kim and Moon, 2012). The hit rate is also widely



(A) Tree analysis results classifying Class 1 from Classes 2 and 3



(B) Tree analysis result classifying classes 2 and 3

Fig. 5. One example of a CART based tree analysis result. Each node is labeled with each class, training cases, and the number (N) of data samples in that group. The model is read from top down until terminal nodes appear. Influential variables are presented in parentheses at each root node to split. (A) indicates the result classifying class 1 from the others, and (B) indicates the result classifying classes 2 and class 3.

Table 2
Summary of the tree analysis from the OPP-DT method. Each ratio indicates the number of times each independent variable was regarded as the most influential variable in classifying the TCE sensitivity in the root node after 100 times runs.

		Most influential variable classifying TCE sensitivity data	
		Class 1 vs Class 2 and 3	Class 2 vs Class 3
Hydrogeology Properties	D	9%	11%
	R	22%	20%
	A	25%	20%
	S	19%	13%
	T	6%	7%
	I	11%	19%
	C	8%	10%

used to quantify the predictive power of models (Duman et al., 2012). In this study, the predictive accuracy of the proposed data mining models was assessed by dividing the number of correctly predicted cases by the total number of given cases in a given situation. The hit rate can be calculated as follows (Eq. (2.4)):

$$\text{Hit rate} = \frac{case\ of\ correct\ classification}{total\ of\ (case\ of\ correct\ classification + case\ of\ misclassification)} \quad (2.4)$$

3. Results

3.1. Evaluation of different data mining methods

Fig. 4 summarizes the comparative performance results for the different data mining methods used in this study. According to the hit rate results, the CBR and DT approaches outperformed the ANN and MLR approaches. The CBR and DT methods provided a higher hit rate (overall value: 63% of CBR, 60% of DT) and consistency (3.0–5.2) than the MLR (overall value: 38%) and ANN (overall value: 45%). In addition, the standard deviations (Y-axis error bars) for the ANN and MLR approaches were larger than those for the other methods. Generally, the ANN is a robust nonlinear classifier method that provides satisfactory performance (Berry and Linoff, 2004; Zhang et al., 1998; Wong and Selvi, 1998; Maier and Dandy, 2000). In this study, however, our dataset was too limited to apply an ANN. For an ANN to provide satisfactory accuracy, at least 500–1000 training data samples are required (Anthony and Bartlett, 2002; Berry and Linoff, 2004; Sarangi and Bhattacharya, 2005; Liu and Jiang, 2013). However, only 114 data were used in this study. Furthermore, if only a small number of data are used for ANN training, an appropriate model is difficult to construct, resulting in over-fit problems in the predictions (Anthony and Bartlett, 2002; Sahiner et al., 2008). The MLR approach may also be infeasible or inappropriate for small sample sizes. Bull et al. (2002) noted that a limitation of the MLR method is its increased bias associated with small sample sizes. Our results suggest that the ANN and MLR approaches were not suitable for assessing the TCE vulnerability at the study site because of the small sample size.

The accuracies and coherencies of the DT and CBR approaches were equally higher than those for ANN and MLR. When OPP was applied to both approaches, their accuracies improved (Fig. 4). In DT, a rule between the input variables and target variable can be induced, whereas rule induction is not possible in CBR (Breiman et al., 1984; Eisenberg and McKone, 1998; Shin and Han, 1999; Berry and Linoff, 2004; Zhang et al., 2008; Kim et al., 2011). Because the OPP-DT approach outperformed the other tested data mining methods, it was selected as the optimal rule induction method for groundwater TCE sensitivity.

3.2. Identification of influential hydrogeological input variables

To induct a rule between the hydrogeological input variables and the target variable (TCE sensitivity), a CART-based tree analysis was conducted in this study. An example of the CART-based tree analysis results is shown in Fig. 5. The aquifer media (A) was identified as the most influential variable in the root node for

Table 3
One-way ANOVA results for the hydrogeological properties corresponding to the Classes 1 and 3 of the TCE sensitivity data in the training set.

Hydrogeological parameters		TCE sensitivity		*P* Value
		Class 1	Class 3	
D (m)	Mean	5.03	6.22	0.596
(Depth to water)	Std. Dev.	2.49	2.20	
R (%)	Mean	5.12	7.87	0.037
(Net recharge)	Std. Dev.	2.10	3.34	
T (%)	Mean	1.63	1.35	0.620
(Topography)	Std. Dev.	1.17	0.91	
C (cm/sec)	Mean	3.72×10^{-3}	5.19×10^{-3}	0.342
(Hydraulic conductivity)	Std. Dev.	0.01	0.01	
		Ratios (%)		
A (NA)[a]	Weathered rock	0.52	0.31	0.023
(Aquifer media)	Coarse sand and silt	0.30	0.41	
	Sandstone	0.18	0.28	
S (NA)[a]	Sand and concrete	0.61	0.39	0.031
(Soil media)	Sandy loam	0.30	0.58	
	Silty loam	0.09	0.13	
I (NA)[a]	Sand/Silt/Clay	0.28	0.12	0.077
(Impact of vadose zone)	Metamorphic/Igneous	0.67	0.70	
	Sand and gravel	0.05	0.18	

[a] Indicates not available.

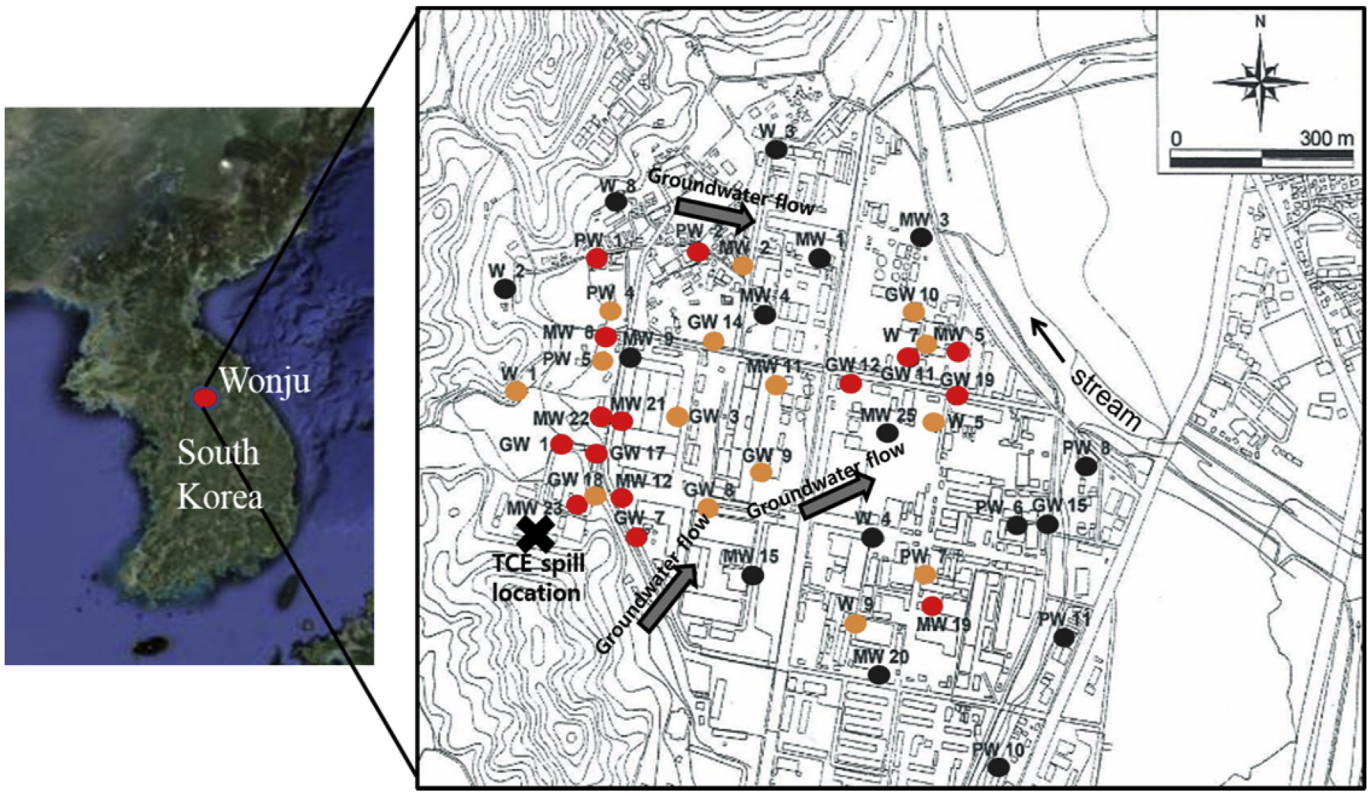


Fig. 6. The distribution of groundwater TCE sensitivity in Woosan Industrial Complex. Red points indicate high TCE sensitivity wells (TCE sensitivity 3), Orange points indicates medium TCE sensitivity wells (TCE sensitivity 2), and Black points indicate low TCE sensitivity wells (TCE sensitivity 1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

separating Class 1 from Classes 2 and 3, and the net recharge (R) was identified as the most influential variable in the root node for separating Class 2 from Class 3 (Fig. 5). To examine the coherency in the identification of the most influential input variables, 100 subsets of hydrogeological input variable data and target variable data (TCE sensitivity) were randomly selected, and their CART-based tree analyses were individually performed to identify the most influential input variables, similar to the example in Fig. 5. The net recharge (R), aquifer media (A), and soil media (S) highly influenced the TCE sensitivity in the separation of Class 1 from Classes 2 and 3, and net recharge (R), aquifer media (A), and impact of vadose zone (I) highly influenced TCE sensitivity in the separation of Class 2 from Class 3. Therefore, net recharge and aquifer media were the hydrogeological input variables that most influenced the TCE sensitivity in the site. Additionally, soil media and impact of vadose zone represent secondary influences.

To examine whether and how the hydrogeological characteristics are different between the Class 1 and Class 3 TCE sensitivity regions, a one-way analysis of variance (ANOVA) was conducted for the hydrogeological input variables (Table 3). The R, A, and S variables for the high-sensitivity (Class 3) regions were significantly different from those for the low-sensitivity (Class 1) regions (p-values <0.05). Net recharge values associated with the high-sensitivity regions were higher than those of the low-sensitivity regions. The aquifer media compositions in the high-sensitivity regions tended to feature less weathered rock and more course sand/silt and sandstone than for the low-sensitivity regions. The soil media compositions in the high-sensitivity regions tended to feature more sand/concrete and less sand/silty loam than the low-sensitivity regions. Unlike the results in Table 2, the values of the impact of vadose zone were not significantly different between the high and low TCE sensitivity regions. The other hydrogeological input variables (depth to groundwater table [D], topology [T], and hydraulic conductivity [C]) did not differ significantly between the high and low TCE sensitivity regions, a finding that is consistent with the conclusions based on Table 2.

4. Discussion

In this study, the groundwater sensitivity of the aquifer in the Woosan Industrial Complex to TCE contamination was assessed using a data mining approach. The DT and rule induction approaches exhibited a relatively high capacity to predict TCE sensitivity based on a limited dataset. These predictions were used to examine the temporal/spatial distributions of TCE contaminants in a real field site. The accuracy and consistency of the DT model were improved by using an additional OPP algorithm. In addition, the key hydrogeological parameters from the DRASTIC method, such as aquifer media, soil media, and recharge, which influence the groundwater TCE sensitivity, were successfully identified using the DT and rule induction approaches. The DT and rule induction approaches better reflect the characteristics of a heterogeneous aquifer, and the potential TCE NAPL residuals were successfully located in the Woosan Industrial Complex.

According to recent studies (Yang et al., 2012, 2014; Kaown et al., 2014), variations in the TCE concentrations at the Woosan Industrial Complex site appear to be influenced by leaching of residual DNAPL TCE trapped in the unsaturated zone. In addition, these studies suggest that DNAPL can migrate vertically through the unsaturated zone and be transported to the water table due to the presence of a good recharge area and concentrated summer precipitation events. During this transport, a portion of the DNAPL TCE may reach the saturated formation (Jo et al., 2010; Yang et al., 2012, 2014). Because the density of DNAPL is higher than water, it

penetrates into the saturated aquifer media until it meets an impermeable layer. TCE is also highly absorbent, obstructing groundwater flow in the aquifer and vadose media. Consequently, aquifer and vadose zone media are influential factors in the CART-based tree analysis of the study site.

Previous studies (Baek and Lee, 2011; Jo et al., 2010; Yang et al., 2012) have strongly suggested that this study site is heavily influenced by seasonal precipitation patterns and hydrogeological characteristics. In particular, because of the presence of a limited recharge area and concentrated summer precipitation events, groundwater recharge is an important hydrogeological factor in the groundwater TCE sensitivity at the Woosan Industrial Complex. Spatial and/or temporal variations in the recharge patterns due to variations in the surface conditions can be inferred from observed TCE concentration levels. Previous studies' results correspond to our CART-based tree analysis results, as shown Tables 2 and 3 This study site primarily consists of fractured granitic bedrock, which is overlain by weathered rock, coarse sand, and silt (Jo et al., 2010; Baek and Lee, 2011; Yang et al., 2012). As revealed in the CART-based tree analysis results (aquifer media results in Table 3), the aquifer and vadose zone areas associated with the Class 3 TCE sensitivity data are underlain by extremely coarse sand and silt and sandstone-bearing vertical fractures, unlike the areas associated with the Class 1 TCE sensitivity data. Clearly, these geologic features of the Class 3 areas can enhance the vertical TCE movement in the saturated zone. The soil media results in Table 3 also indicate that areal heterogeneities in the surface conditions might play an important role in classifying influential variables in our tree analyses with the exception of the nearby TCE spill location. The surface of the study site is largely covered by concrete pavement, which prevents precipitation from directly infiltrating the subsurface. In particular, the wells categorized as Class 3 TCE sensitivity (red points in Fig. 6) have relatively good conditions for surface infiltration, including sloped forests and grassy areas. However, the wells categorized as Class 1 TCE sensitivity were largely surrounded by impermeable concrete and paved surface conditions (black points in Fig. 6). The remaining Class 2 wells (orange points in Fig. 6) feature mixed surface media with both high- and low-infiltration conditions.

The results of our CART-based tree analysis indicate that the high TCE sensitivity areas (Class 3 areas) have a finer aquifer media texture and better soil media infiltration conditions than the low TCE sensitivity areas. These geologic conditions are believed to strongly affect the groundwater recharge ability. Thus, the *p*-values of the net recharge results in Table 3 are significantly different between the areas associated with TCE sensitivity Classes 1 and 3. Based on the CART-based tree analysis, the DT and rule induction approaches better reflect the hydrogeological conditions at the study site and correlate well with the current available knowledge from previous studies (Jo et al., 2010; Baek and Lee, 2011; Yang et al., 2012).

In this study site, the main source and contaminant plumes were definitively identified from the aqueous phase concentration data and historical field-measured data (Jo et al., 2010; Baek and Lee, 2011). Our DT and rule induction results showed that certain wells located in the industrial complex, e.g., GW11, GW12, GW, 19, MW5, and MW 19, were determined to have a TCE sensitivity of Class 3 (Fig. 6). These Class 3 wells feature relatively unfavorable hydrogeological conditions due to limited recharge rates and paved surface conditions. This result suggests that TCE contaminant plumes may be transported by rainfall events in areas with relatively good surface infiltration conditions near the source area. Then, the TCE plumes migrate vertically through the unsaturated zone and are transported laterally from major source through the water table and saturated zone to down-gradient wells in the industrial complex. The groundwater flow direction is predominantly from west to east in the industrial complex area (Fig. 6). According to previous studies (Yang et al., 2012, 2014), high TCE concentrations were observed in the uppermost level. The observed dramatic decrease in TCE concentration with depth suggests that the source of the TCE is at or above the water table. This conclusion implies that the residual or free phase of the TCE contaminants preferentially exists near the water table and acts as a continuous source of aqueous-phase contaminants to the down-gradient wells. Yang et al. (2012) showed that the wells GW11, GW12, and GW19 were small TCE contaminant sources and that their source locations can be successfully identified with the combined use of seasonal impact analysis, a historical approach, and chemical fingerprinting tests.

Based on the results obtained from the tree analyses with statistical significance, this study successfully identified the key hydrogeological parameters – net recharge (R), aquifer media (A), and soil media (S) – that affect both large- and small-scale variations in the TCE sensitivity. Groundwater recharge plays an important role in TCE contamination with areal heterogeneity in the surface condition. In addition, highly weathered and fractured rocks hosting the aquifer can facilitate vertical TCE movement and form an extensive DNAPL plume. Our findings agree well with those of a recent study. Although the identification of these influential hydrogeological parameters may be relevant only for TCE-contaminated sites with hydrogeological properties similar to those of the Woosan Industrial Complex, the concept and framework used in this study can be applied to other types of aquifers with other contaminants. In particular, the DT and rule induction approaches are demonstrated to be more effective than other data mining algorithms when the available dataset is relatively small and features a high degree of nonlinear complexity. This finding has significant implications for environmental information research.

Acknowledgments

This work was supported by the Korea Ministry of Environment via a grant from The GAIA Project. In addition, this work was partially supported by the National Research Foundation (NRF) of Korea via a grant (No. 2011-0030040) funded by the Korea government (MSIP).

References

Aller, L., Bennett, T., Lehr, J.H., Petty, R.J., Hackett, G., 1987. DRASTIC: a Standardized System for Evaluating Groundwater Pollution Potential Using Hydrogeologic Settings. USEPA Report, 600/287/035.

Ahn, J., Kim, Y., Yoo, K., Park, J., Oh, K., 2012. Using GA-Ridge regression to select hydro-geological parameters influencing groundwater pollution vulnerability. Environ. Monit. Assess. 184, 6637–6645.

Anthony, M., Bartlett, P.L., 2002. Neural Network Learning: Theoretical Foundations. Cambridge University Press, New York.

Baek, W., Lee, J., 2011. Source apportionment of trichloroethylene in groundwater of the industrial complex in Wonju, Korea: a 15-year dispute and perspective. Water Environ. 25, 336–344.

Berry, M., Linoff, G., 2004. Data Mining Techniques. Indiana, USA, Wiley publishing, Inc, Indianapolis.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Tree. Champman and Hall, New York.

Bull, S.B., Mak, C., Greenwood, C.M.T., 2002. A modified score function estimator for multinomial logistic regression in small samples. Comput. Stat. Data. Anal 39, 57–74.

Cawley, G.C., Talbot, N.L.C., 2003. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern. Recogn. 36, 2585–2592.

Chambers, J.E., Loke, M.H., Ogilvy, R.D., Meldrum, P.I., 2004. Noninvasive monitoring of DNAPL migration through a saturated porous medium using electrical impedance tomography. J. Contam. Hydrol. 68, 1–22.

Chang, F.J., Tsai, W.P., Chen, H.K., Yam, R.S.W., Herricks, E.E., 2013. A self-organizing radial basis network for estimating riverine fish diversity. J. Hydrol. 476, 280–289.

Cho, K., Sthiannopkao, S., Pachepsky, Y.A., Kim, K., Kim, J., 2011. Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network. Water. Res. 45, 5535–5544.
Dixon, B., 2005. Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: a GIS-based sensitivity analysis. J. Hydrol. 309, 17–38.
Duman, E., Ekinci, Y., Tanrıverdi, A., 2012. Comparing alternative classifiers for database marketing: the case of imbalanced datasets. Expert. Syst. Appl. 39, 48–53.
Eisenberg, J.N.S., McKone, T.E., 1998. Decision tree method for the classification of chemical pollutants: incorporation of across-chemical variability and within-chemical uncertainty. Environ. Sci. Technol. 32, 3396–3404.
EMC, 2003. Detailed Investigation Report on Contaminated Soil and Groundwater in the Woosan Industrial Complex and Joongang-dong Area in Wonju City. Environmental Management Corporation, South Korea.
Fijani, E., Nadiri, A.A., Moghaddam, A., Tsai, F.T.C., Dixon, B., 2013. Optimization of DRASTIC method by supervised committee machine artificial intelligence to assess groundwater vulnerability for Maragheh–Bonab plain aquifer. Iran. J. Hydrol. 503, 89–100.
Gogu, R.C., Dassargues, A., 2000. Current trends and future challenges in ground-water vulnerability assessment using overlay and index methods. Environ. Geol. 39, 549–559.
Huang, C., Moraga, C., 2004. A diffusion-neural-network for learning from small samples. Int. J. Approx. Reason 35, 137–161.
Huber, P.J., 1981. Robust Statistics. Wiley, New York.
Jackson, R.E., 1998. The migration, dissolution, and fate of chlorinated solvents in the urbanized alluvial valleys of the southwestern USA. Hydrogeol. J. 6, 144–155.
Jo, Y., Lee, J., Yi, M., Kim, H., Lee, K., 2010. Soil contamination with TCE in an industrial complex: contamination levels and implication for groundwater contamination. Geosci. J. 14, 313–320.
Kaown, D., Shouakar-Stash, O., Yang, J., Hyun, Y., Lee, K., 2014. Identification of multiple sources of groundwater contamination by dual isotopes. Groundwater 52, 875–885.
KECO, 2005. Detailed Investigation and Basic Remediation Design for Contaminated Soil and Groundwater in the Woosan Industrial Complex. Korea Environmental Corporation, Wonju City, Kangwon Province, Korea.
KECO, 2008. Long-term Monitoring of Groundwater Quality at TCE Contaminated Site in Woosan Industrial Complex. Korea Environment Corporation, Kang-wondo Roads & Bridges Maintenance Office.
Kim, C.h., Chang, W., 2011. Logistic regression in sealed-bid auctions with multiple rounds: application in Korean court auction. Expert. Syst. Appl. 38, 3098–3115.
Kim, K., Yoo, K., Ki, D., Son, I., Oh, K., Park, J., 2011. Decision-tree-based data mining and rule induction for predicting and mapping soil bacterial diversity. Environ. Monit. Assess. 178, 595–610.
Kim, Y.S., Moon, S., 2012. Measuring the success of retention management models built on churn probability, retention probability, and expected yearly revenues. Expert. Syst. Appl. 39, 11718–11727.
Kwon, Y., Han, I., Lee, K., 1997. Ordinal pairwise partitioning (OPP) approach to neural networks training in bond rating. Int. J. Intell. Syst. Acc. Finance. Manag. 6, 23–40.
Liu, B., Jiang, Y., 2013. A multitarget training method for artificial neural network with application to computer-aided diagnosis. Med. Phys. 40, 011908.
Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Modell. Softw. 15, 101–124.
Mair, A., El-Kadi, A.I., 2013. Logistic regression modeling to assess groundwater vulnerability to contamination in Hawaii, USA. J. Contam. Hydrol. 153, 1–23.
Mishra, S.K., Singh, V.P., 2003. Soil Consevation Sevice Curve Number (SCS-CN) Methodology. Kluwer Academic Publisher, Boston MA.
Pacheco, F.A.L., Pires, L.M.G.R., Santos, R.M.B., Sanches Fernandes, L.F., 2015. Factor weighting in DRASTIC modeling. Sci. Total. Environ. 505, 474–486.
Pal, M., Mather, P.M., 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote. Sens. Environ. 86, 554–565.
Park, J., Ki, D., Kim, K., Lee, S., Kim, D., Oh, K., 2011. Using decision tree to develop a soil ecological quality assessment system for planning sustainable construction. Expert. Syst. Appl. 38, 5463–5470.
Pesch, R., Schröder, W., Schmidt, G., Genssler, L., 2008. Monitoring nitrogen accumulation in mosses in central European forests. Environ. Pollut. 155, 528–536.
Rivett, M.O., Turner, R.J., Glibbery, P., Cuthbert, M.O., 2012. The legacy of chlorinated solvents in the Birminghamaquifer, UK: observations spanning three decades and the challenge of future urban groundwater development. J. Contam. Hydrol. 140–141, 107–123.
Rivett, M.O., Dearden, R.A., Wealthall, G.P., 2014. Architecture, persistence and dissolution of a 20 to 45 year old trichloroethylene DNAPL source zone. J. Contam. Hydrol. 170, 95–115.
Rodriguez-Galiano, V., Mendes, M.P., Garcia-Soldado, M.J., Chica-Olmo, M., Ribeiro, L., 2014. 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). Sci. Total. Environ. 476–477, 189–206.
Sahiner, B., Chan, H.P., Hadjiiski, L., 2008. Classifier performance estimation under the constraint of a finite sample size: resampling schemes applied to neural network classifiers. Neural. Netw. 21, 476–483.
Sarangi, A., Bhattacharya, A.K., 2005. Comparison of Artificial Neural Network and regression models for sediment loss prediction from Banha watershed in India. Agr. Water. Manag. 78, 195–208.
SAS, 2009. Web Report Studio 4.2: User's Guide. SAS Institute Inc, Cary, NC, USA.
Shin, K., Han, I., 1999. Case-based reasoning supported by genetic algorithms for corporate bond rating. Expert. Syst. Appl. 16, 85–95.
Singh, R., Datta, B., 2007. Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data. Water. Resour. Manag. 21, 557–572.
Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437.
USEPA, 2003. The DNAPL Remediation Challenge: is There a Case for Source Depletion? EPA/600/R-03/143.
USEPA, 2005. Trichloroethylene (TCE) Health Risk Assessment: Overview. http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=119268.
USEPA, 2006. Child-specific Exposure Factors Handbook 2006 (External Review Draft). http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=56747.
Vega, F.A., Matías, J.M., Andrade, M.L., Reigosa, M.J., Covelo, E.F., 2009. Classification and regression trees (CARTs) for modelling the sorption and retention of heavy metals by soil. J. Hazard. Mater 167, 615–624.
Wong, B.K., Selvi, Y., 1998. Neural network applications in finance: a review and analysis of literature (1990–1996). Inf. Manag. Amster 34, 129–139.
Yang, C., Prasher, S.O., Enright, P., Madramootoo, C., Burgess, M., Goel, P., Callum, I., 2003. Application of decision tree technology for image classification using remote sensing data. Agr. Syst. 76, 1101–1117.
Yang, J., Lee, K., Clement, T.P., 2012. Impact of seasonal variations in hydrological stresses and spatial variations in geologic conditions on a TCE plume at an industrial complex in Wonju, Korea. Hydrol. Process 26, 317–325.
Yang, J., Jun, S., Kwon, H., Lee, K., 2014. Tracing of residual multiple DNAPL sources in the subsurface using 222Rn as a natural tracer at an industrial complex in Wonju, Korea. Environ. Earth. Sci. 71, 407–417.
Zhang, G., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks: the state of the art. Int. J. Forecast. 14, 35–62.
Zhang, X., Lin, F., Jiang, Y., Wang, K., Wong, M.T.F., 2008. Assessing soil Cu content and anthropogenic influences using decision tree analysis. Environ. Pollut. 156, 1260–1267.