Environ Monit Assess
DOI 10.1007/s10661-011-2448-1

Using GA-Ridge regression to select hydro-geological parameters influencing groundwater pollution vulnerability

Jae Joon Ahn · Young Min Kim · Keunje Yoo · Joonhong Park · Kyong Joo Oh

Received: 27 January 2011 / Accepted: 14 November 2011

Abstract For groundwater conservation and management, it is important to accurately assess groundwater pollution vulnerability. This study proposed an integrated model using ridge regression and a genetic algorithm (GA) to effectively select the major hydro-geological parameters influencing groundwater pollution vulnerability in an aquifer. The GA-Ridge regression method determined that depth to water, net recharge, topography, and the impact of vadose zone media were the hydro-geological parameters that influenced trichloroethene pollution vulnerability in a Korean aquifer. When using these selected hydro-geological parameters, the accuracy was improved for various statistical nonlinear and artificial intelligence (AI) techniques, such as multinomial logistic regression, decision trees, artificial neural networks, and case-based reasoning. These results provide a proof of concept that the GA-Ridge regression is effective at determining influential hydro-geological parameters for the pollution vulnerability of an aquifer, and in turn, improves the AI performance in assessing groundwater pollution vulnerability.

Keywords Ridge regression · Genetic algorithms · Groundwater pollution vulnerability · AI techniques · DRASTIC method

J. J. Ahn · Y. M. Kim · K. J. Oh (✉)
School of Information and Industrial Engineering,
Yonsei University,
262 Seongsanno, Seodaemun-gu,
Seoul 120-749, South Korea
e-mail: johanoh@yonsei.ac.kr

K. Yoo · J. Park
School of Civil and Environmental Engineering,
Yonsei University,
262 Seongsanno, Seodaemun-gu,
Seoul 120-749, South Korea

Introduction

Demand for water supply is increasing due to population growth and improvements in living standards (Goh 1995). Additionally, current climate change effects mean that there are more regions where surface water is scarce (Vano et al. 2010; Kundzewicz et al. 2008). However, the current water supply may not be enough to meet the rising water demand because of limited availability of clean freshwater resources (Ray and O'dell 1993; Al-Adamat et al. 2003; Dixon 2005; Gurdak et al. 2007). Groundwater is the major freshwater resource (Winter et al. 1999; Kundzewicz et al. 2008). Compared to the steady increase in usage of groundwater resources, the efforts to maintain groundwater conservation and contamination prevention have yet to be further made (Rupert 1999; Lee et al. 2007; Liu et al. 2008). Among such efforts, it is important to make decision on which aquifers have high priority in the policy of groundwater conservation and contamination prevention. Aquifers, which are sensitive to potential contamination, may have higher priority for groundwater conservation and

Published online: 29 November 2011

prevention against contamination. As an index of sensitivity of groundwater to contamination, "groundwater vulnerability" is generally used in policy making for groundwater conservation and contamination prevention (Aller et al. 1987; Rupert 1999).

The DRASTIC model is one of the most common methods to assess the vulnerability of groundwater (Kalinski et al. 1994; Rosen 1994; Melloul and Collin 1998; Secunda et al. 1998; Kim and Hamm 1999). This model is based on the concept of the hydrogeological setting that is defined as a composite description of all the major geologic and hydrologic factors that affect and control the groundwater movement (Aller et al. 1987). Once a DRASTIC index has been computed, it is possible to identify specific areas exposed to potential groundwater contamination. The higher the DRASTIC index is, the greater the groundwater vulnerability (Aller et al. 1987). Although the DRASTIC model is widely applied, there are some limitations reported to previous studies (Merchant 1994; Croskrey and Groves 2008). When calculating the DRASTIC index, the weights and parameters are selected by geological features. However, the model was developed based on certain type of aquifers in the USA, and it does not applied well to aquifers with dissimilar geological features. To apply the DRASTIC model in a different aquifer, weights of hydro-geological parameters have to be empirically determined. However, such empirical determination may not be cost-effective. To circumvent this limitation of the traditional groundwater vulnerability model, artificial intelligence (AI) algorithms may be good alternatives because of their capability of finding optimal correlation between input variables and target values without pre-existing information on weight values of input variables.

Noisy input variables often hinder the power of AI techniques to find optimal correlation between input variables and target values (Kim et al. 2011). If influential input variables can be selected from noisy input variables, it would be beneficial for AI techniques to find optimal correlation of groundwater pollution vulnerability with hydro-geological characteristics of an aquifer. To address this issue, herein we proposed a novel method to pre-select influential hydro-geological parameters for AI techniques assessing groundwater pollution vulnerability. Ridge regression is very useful in adjusting for multicollinearity (Smith and Campbell 1980). A common problem in regression analysis is the significant correlation between input variables that causes coefficient estimates to be biased or unstable. However, ridge regression can reduce this problem by producing more stable estimators, which should result in a better overall estimation model. Genetic algorithm (GA) is a stochastic search technique that can explore large and complicated spaces on the ideas from natural genetics and evolutionary principle (Holland 1975; Goldberg 1989; Davis 1991). It has been demonstrated to be effective and robust in searching very large spaces in a wide range of applications (Klimasauskas 1992; Fogel 1993; Koza 1993; Han et al. 1997). GA is particularly suitable for multi-parameter optimization problems with an objective function subject to numerous hard and soft constraints. Based upon the advantages of GA and ridge regression methods, we developed a GA-Ridge regression method for selecting influential hydro-geological parameters in order to improve the performance of AI techniques to assess groundwater pollution vulnerability. To validate the beneficial effects by the GA-Ridge regression method, AI-estimated groundwater sensitivity to trichloroethene (TCE) contamination was compared with field-measured TCE sensitivity data from Woosan Industrial Complex, South Korea.

Methods

Architecture of the proposed model

Phase 1: variable selection

The proposed GA-Ridge model consists of two main phases. In phase 1, k is changed while conducting input variable selection under the ridge regression. For example, k and the input variable are selected using GA simultaneously. Technically, for variable selection, the parameter is assigned to 0 (discarding) or 1 (selecting), while the weight of k (a ridge constant) has a real number range from 0 to 1.

Phase 2: validation of the GA-Ridge model

Phase 2 focuses on proving the usefulness of the input variables selected by the GA-Ridge model. To guarantee the validity of the model, we evaluate groundwater vulnerability using AI techniques. The

overall structure of the integrated model is shown in Fig. 1.

Ridge regression

In linear multiple regression, a model is hypothesized in the following form:

$$Y = X\beta + \varepsilon, \tag{1}$$

where $E(\varepsilon)=0$; $E(\varepsilon'\varepsilon)=\sigma^2 I$; and X_{nxp} is a full rank matrix. To facilitate comparisons among models, variables are standardized so that the matrix $X'X$ is in the form of a correlation matrix, and the vector $X'Y$ is the vector of correlation coefficients between the criterion variable and all the input variables. The least square estimates and their variance–covariance matrix are then, respectively, $\widehat{\beta} = (X'X)^{-1}X'Y$ and $V\left(\widehat{\beta}\right) = \sigma^2 I(X'X)^{-1}$. The parameters obtained are unbiased, where $E\left(\widehat{\beta}\right) = \beta$. The difficulties in this standard estimation are a direct consequence of the average distance between $\widehat{\beta}$ and β. In particular, if L^2 is the distance between $\widehat{\beta}$ and β, then $L^2 = \left(\widehat{\beta} - \beta\right)'\left(\widehat{\beta} - \beta\right)$. This expression can be represented as $E(L^2) = \text{trace}[V(p)]$, which is equivalent to the following:

$$E\left(L^2\right) = \sigma^2 \text{trace}(X'X)^{-1} = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}, \tag{2}$$

where λ_i is the ith eigenvalue of the $X'X$ matrix. As the vector of X deviates from orthogonality, λ_i becomes smaller and $\widehat{\beta}$ can be expected to be farther from the true parameter β. The ridge regression is an estimation procedure based upon $\widehat{\beta}^* = \widehat{\beta}^*(k) = (X'X + kI)^{-1}X'Y$ for the standardized variables where k is a ridge constant. Therefore, $E\left(\widehat{\beta}^*\right) = (X'X + kI)^{-1}(X'X)\beta = Z_k\beta$, which is a biased estimate. Note that when $k=0$, $Z_k=I$ and $E\left(\widehat{\beta}^*\right) = \beta$ are the least squares unbiased estimates. Furthermore, $\widehat{\beta}$, for $k\neq 0$, is shorter than $\widehat{\beta}$ (i.e.), $\left(\widehat{\beta}^*\right)'\left(\widehat{\beta}^*\right) < \widehat{\beta}'\widehat{\beta}\Big)$. The variance–covariance matrix of the ridge regression estimate is $V\left(\widehat{\beta}^*\right) = \sigma^2 Z_k(X'X + kI)^{-1}$. The expected value of the squared distance between $\widehat{\beta}^*$ and β is

$$\begin{aligned} E\left(L^2(k)\right) &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2\beta'(X'X + kI)^{-1}\beta \\ &= tr\left[V\left(\widehat{\beta}^*\right)\right] + (\text{bias})^2. \end{aligned} \tag{3}$$

The first term on the right side of formula (3) is a continuous, monotonically decreasing function of k, and the second term is a continuous, monotonically increasing function of k. Therefore, if $\beta'\beta$ is bounded, there is a value of k such that $E\left(L^{2^*}\right) < E(L^2)$. To summarize, the ridge regression procedure produces estimates of $\widehat{\beta}^*$ that are biased and shorter than the least square estimates $\widehat{\beta}$, with a smaller variance, and they are closer to the true value of the coefficients. In other words, the entire objective of this procedure is to take a little bias in the parameter estimation and substantially improve the mean squares of the estimates and the prediction.

The ridge regression analysis has two interesting aspects. The first aspect is a ridge trace, or a two-dimensional plot of $\widehat{\beta}^*(k)$, and the residual

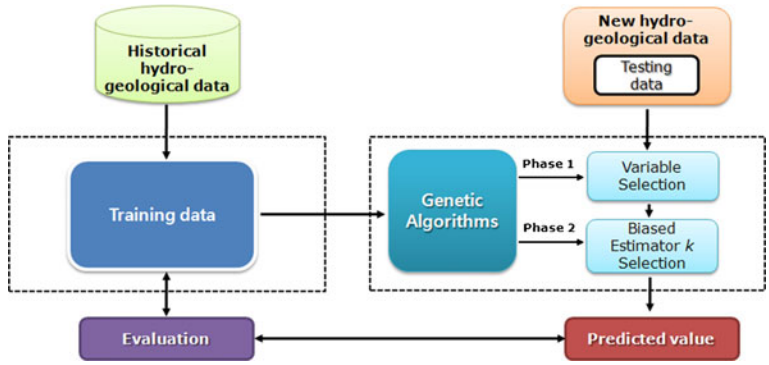

Fig. 1 Architecture of the proposed model

sums of squares [SSE(k)] for the values of k in the interval [0,1]. The trace portrays the complex interrelationships between non-orthogonal input variables and the effect of these interrelationships on the estimation of β. The second aspect is the determination of a value of k that provides a better estimate of β. It should be pointed out that, when the input variables are uncorrelated, the least squares estimates obtained are uniformly scaled by the quantity $1/(k+1)$, and the relative value of the regression coefficients is then independent of the choice of k.

Genetic algorithm

The GA performs the search process in four stages: (1) initialization, (2) selection, (3) crossover, and (4) mutation (Davis 1991). In the initialization stage, a population of genetic structures (known as chromosomes) that are randomly distributed in the solution space is selected as the starting point of the search. After the initialization stage, each chromosome is evaluated with a user-defined fitness function. The goal of the fitness function is to encode numerically the performance of the chromosome. For real-world applications of optimization methods, such as a GA, the choice of the fitness function is the most critical step. In this study, the GA plays a role in the optimization of variable selection and coefficients (or weights) for the input variables of the ridge regression to assess groundwater vulnerability.

Optimization scheme

The regression matrix formula (1) can be rewritten as follows:

$$Y_i = \beta_0 + D_1\beta_1 X_{i,1} + \cdots + D_p\beta_p X_{i,p} + \varepsilon_i,\ i = 1,\ 2,\ \ldots\ ,\ n. \tag{4}$$

Then, we assign the optimal coefficients $\left\{\beta_k : \sum_{k=1}^{p} \beta_k = 1,\ k = 1, 2, \cdots, p\right\}$ to each input variable, which minimizes the mean squared error (MSE) (5), while, simultaneously, D_1, D_2, D_p are changed to 0 or 1 denoting discard or selection of an input variable, respectively, and k is assigned to a real number [0, 1] through the GA process.

$$\mathrm{MSE}(Y) = \frac{1}{n-1}\sum_{t=1}^{n}\left(Y_t - \widehat{Y}_t\right)^2 \tag{5}$$

DRASTIC model

To compare the performance of the GA-Ridge regression method and AI technique with the traditional pollution vulnerability model, the DRASTIC model developed by Environmental Protection Agency was used in this work. The DRASTIC model has been used to identify areas that are more vulnerable to contamination than others, or to prioritize areas that need more groundwater monitoring. This model includes various hydro-geological settings whose physical characteristics affect groundwater qualities on a regional basis (Aller et al. 1987). The DRASTIC model has seven parameters representing several hydro-geological properties, including depth to water (D), net recharge (R), aquifer media (A), soil media (S), topography (T), impact of the vadose zone (I), and hydraulic conductivity (C). Their definitions are presented in Table 1. The DRASTIC INDEX (DI), which reflects groundwater pollution vulnerability, is calculated using the following formula (Aller et al. 1987):

$$\mathrm{DI} = D_R D_W + R_R R_W + A_R A_W + S_R S_W + T_R T_W + I_R I_W + C_R C_W, \tag{6}$$

where

D_R and D_W	Rating and weight assigned to the depth to water table
R_R and R_W	Rating and weight for range of aquifer recharge
A_R and A_W	Rating and weight assigned to aquifer media
S_R and S_W	Rating and weight for soil media
T_R and T_W	Rating and weight assigned to topography
I_R and I_W	Rating and weight assigned to vadose zone
C_R and C_W	Rating and weight given to hydraulic conductivity

Table 1 Descriptions and weights of the hydro-geological parameters used in the DRASTIC model (Aller et al. 1987)

Parameters	Description	Weight
Depth to water (*D*)	The depth from the ground surface to the water table, deeper water table levels imply lesser chance for contamination	5
Net recharge (*R*)	The amount of water which penetrates the ground surface and reaches the water table, recharge water represents the vehicle for transporting pollutants	4
Aquifer media (*A*)	The saturated zone material properties, which controls the pollutant attenuation process	3
Soil media (*S*)	Uppermost and weathered part of the ground, soil cover characteristics influence the surface and downward movement of contaminants	2
Topography (*T*)	The slope and slope variability of the land surface, steeper slopes signify higher groundwater velocity	1
Impact of vadose zone media (*I*)	The zone above the water table which is unsaturated, it controls the passage and attenuation of the contaminated material to the saturated zone	5
Hydraulic conductivity (*C*)	The ability of the aquifer materials to transmit water, which, in turn, controls the rate at which groundwater will flow under a given hydraulic gradient	3

According to the DRASTIC model guidelines, the rating and weight values are assigned to the hydro-geological parameters of each sampling point. The weight values of the hydro-geological parameters are presented in Table 1.

Study site and data collection

In this study, data and information on TCE contamination and hydro-geological properties were obtained from previous reports by the Korean Environmental Management Corporation for the Woosan Industrial Complex located at Wonju City, Kangwon Province, South Korea (Lee et al. 2003). For the following data-mining procedures, the TCE data and their corresponding hydro-geological properties were collected from February and August in 2003, 2004, and 2008, when TCE contamination and hydro-geological properties were well characterized at the site. We calculated the normalized change in TCE concentration between two TCE sampling points at a location for the target values (formula (7)), and used these values as a quantitative measure of TCE sensitivity in the studied site.

$$\text{TCE sensitivity} = \frac{C_{i+1} - C_i}{(C_{i+1} + C_i)/2} \tag{7}$$

Where C_{i+1} and C_i correspond to the TCE contamination concentrations at $t=i+1$ and $t=i$, respectively, at a location. Prior to the AI procedures, the TCE sensitivity data were divided into three classes called "classes 1, 2, and 3." The standard for the classification process was based on the TCE sensitivity data and on the number of wells that corresponds to the TCE sensitivity. In other words, the number of wells was mapped on a table according to the TCE sensitivity data, and then three classes were divided so that all the classes would contain an equal number of wells (Fig. 2). The classified TCE sensitivity data were used as the target values in this study.

Model training and testing

For proper training and testing performance, the number of training and testing data was evenly distributed between classes (Table 2). In this study, we evaluated the accuracy of the assessment models by comparing two different assessment models. One model used seven input variables and the other used the selected input model based on the GA-Ridge model. The accuracy was measured in hit rates of the testing set. The entire dataset was randomly partitioned into the training set (80% of the dataset for each class) and the testing set (remaining 20% of the whole dataset). The training dataset was used to find an optimal weight for the genetic learning. The testing dataset, on the other hand, was used to evaluate the weight and verify the forecasting accuracy of TCE sensitivity. As a result, the number of training and testing datasets was evenly distributed in each TCE sensitivity class (Table 2).

Compared AI techniques

In this study, we used nonlinear statistical and AI techniques to ensure the validity of the GA-Ridge

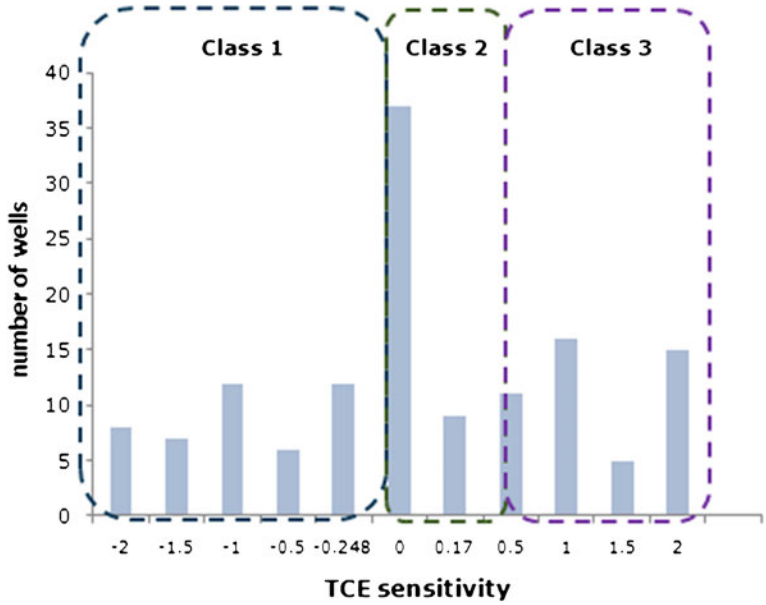

Fig. 2 Histogram of the TCE sensitivity data from the studied site

model for the groundwater vulnerability assessment model. These techniques included multinomial logistic regression (MLR), decision trees (DT), artificial neural networks (ANN), and case-based reasoning (CBR). These methods are popular classifiers currently being used.

In general, logistic regression is one of the nonlinear statistical methods used to estimate the probability of a binary outcome with upward or downward status (Hosmer and Lemeshow 1989). As a result, MLR models are well suited to forecasting problems with a categorical output variable. The MLR model assumes the following parametric regression model:

$$Y_{t+l} = 1/\left[1 + \exp\left\{-\left(\beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_p X_{pt}\right)\right\}\right], \quad (8)$$

where, β_0, β_1,…,β_p are the regression coefficients.

DT is broadly used for predictive modeling since it is easy to interpret and can model complex input–output relations with an automatic handling of missing values (Breiman et al. 1984; Cherkassky and Mulier 1998). We adopted the standard CART method by Breiman et al. (1984), which was developed based on an inductive learning approach. In addition, the Gini diversity index was employed to split tree nodes, and cross-validation was used to prune the trees.

CBR is a problem solving method that reuses cases and experience to find an appropriate solution to a given set of new cases (Shin and Han 1999). In this study, the nearest neighbor approach was used to retrieve the most similar cases. In addition, Euclidean distance and equally weighted voting methods were used for the distance and combination functions, respectively.

In this study, ANN was considered the most nonparametric method because of its data overfitting tendency (White 1989; Kaastra and Boyd 1996; Zhang et al. 1998). Among various ANN, a three-layer, fully connected back-propagation neural network (BPN) was used in this work. BPN training consisted of trial-and-error experiments to build an appropriate architecture for the BPN.

Table 2 Distribution of training and testing data

TCE sensitivity class	Training data	Testing data
Class 1	36	10
Class 2	36	10
Class 3	36	10
Total	108	30

Results and discussion

DRASTIC-estimated vs. field-observed TCE sensitivities

To examine the applicability of the traditional DRASTIC model for assessing the sensitivity of the

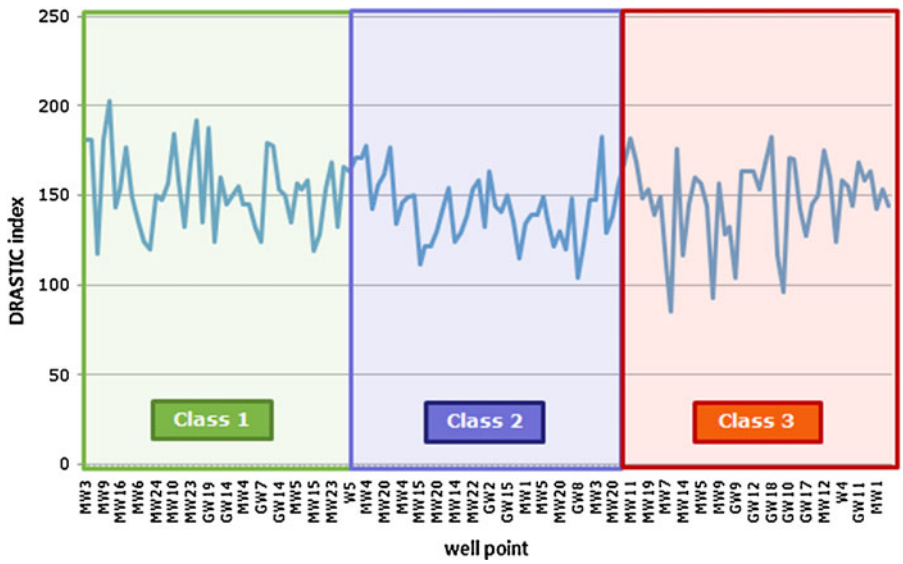

Fig. 3 DRASTIC index values and TCE sensitivity classes at the Woosan Industrial Complex

Korean aquifer to TCE contamination, we compared the DRASTIC-estimated vulnerability (DRASTIC index) values with the field-observed vulnerability (TCE sensitivity classes). No distinct trends were observed in response to the different TCE sensitivity classes (Fig. 3). This result indicates that the traditional DRASTIC model failed to describe the observed TCE sensitivity data in response to the hydro-geological characteristics of the aquifer site. Uncertainty in the TCE contamination source might be a possible explanation for the inconsistency between the DRASTIC model and the field-observed TCE sensitivity data. However, such uncertain effects may be compromised by the pre-treatment of TCE sensitivity data by normalization and classification. A plausible explanation may be that the weight values for the hydro-geological parameters guided by the traditional model (Table 1) are not suitable to reflect the hydro-geological characteristics of the studied site. This supports the need for an alternative approach to find the optimal correlation between the field-observed TCE sensitivity and hydro-geological parameters using AI techniques.

Table 3 The original (unselected) and GA-Ridge-selected (preselected) hydro-geological parameters

	Input variables
Unselected	*D* (depth to water), *R* (net recharge), *A* (aquifer media), *S* (soil media), *T* (topography), *I* (impact of vadose zone media), *C* (hydraulic conductivity)
Preselected	*D* (depth to water), *R* (net recharge), *T* (topography), *I* (impact of vadose zone media)

GA-Ridge-selected hydro-geological parameters

Among the seven hydro-geological parameters, GA-Ridge regression resulted in the selection of four influential parameters including depth to water, net recharge, topography, and impact of vadose zone media (Table 3).

In the traditional DRASTIC model, the weight values for the depth to water (*D*), impact of vadose zone media (*I*), and net recharge (*R*) were relatively large (i.e., D_w and I_w=5; R_w=1; Table 1). This is consistent with the finding from the GA-Ridge regression. However, the weight value for topography (*T*) was low in the DRASTIC model, while the GA-Ridge regression determined that topography was as an influential hydro-geological factor at the specific site. The DRASTIC model was developed based upon studies from North American regions with relatively flat and homogeneous hydro-geological settings. Meanwhile, the aquifer area of this study exhibited

Table 4 Hit rates for various AI techniques

	ANN	DT	MLR	CBR
Unselected	0.44	0.43	0.40	0.50
Preselected	0.47	0.61	0.47	0.59

more dynamic topography and heterogeneous hydro-geological settings. The hydro-geological characteristics of the Woosan Industrial Complex site may provide an explanation for the selection of topography by the GA-Ridge regression. In addition, this may also explain why the DRASTIC-estimated pollution vulnerability poorly correlated with the field-observed TCE sensitivity data (Fig. 3).

GA-Ridge regression improved AI performances

Using the selected hydro-geological parameters by the GA-Ridge regression, our groundwater vulnerability assessment models with various nonlinear statistical and AI techniques (MLR, DT, CBR, and ANN in this study) were constructed as described in the methodology, and their hit rates were compared with the original assessment model without any selection among the seven hydro-geological parameters.

As shown in Table 4, the GA-Ridge regression improved the accuracy of all the tested AI methods, providing a proof of concept that the GA-Ridge regression is able to select hydro-geological parameters influencing TCE contamination sensitivity in the site. However, the magnitudes in improved accuracy differ among the tested AI techniques. Compared to the model using all seven hydro-geological parameters, the ANN showed slight improvement (3% for ANN), while the CBR, MLR, and DT models showed significant improvements (7% for MLR, 9% for CBR, and 18% for DT). These results suggest that the use of a GA-Ridge regression is effective at improving CBR, MLR, and especially DT model performances, but not effective for ANN. Possible explanations for these results are that ANN is not sensitive to the selection of input variables (Jacobs et al. 1991) and/or that the number of data points (N=138) was not great enough for an accurate ANN performance (Ryan et al. 1998).

Conclusion

This study proposes the GA-Ridge model to select proper hydro-geological parameters influencing groundwater pollution vulnerability. The traditional DRASTIC-estimated groundwater TCE vulnerability data showed no correlation with the field-observed TCE sensitivity data from the Woosan Industrial Complex area, indicating an inapplicability of the traditional groundwater vulnerability model to reflect the hydro-geological characteristics in the site. The GA-Ridge regression determined four hydro-geological parameters (depth to water, impact of vadose media, net recharge, and topography) influencing TCE contamination vulnerability in the aquifer. When using the selected input variables, the accuracy of AI models was improved when compared to those without the pre-selection steps. The GA-Ridge regression resulted in a significant improvement in the DT, CBR, and MLR performances. These finding provide a proof of concept that the GA-Ridge regression is capable of effectively selecting influential hydro-geological parameters in a contaminated aquifer, and such pre-selection of input variables can improve the accuracy of AI techniques for assessing groundwater pollution vulnerability, which is a crucial factor in planning and policy making for groundwater conservation and protection.

Acknowledgments This research was supported by the Korea Ministry of Environment via the GAIA project (grant number: 141-081-034). In addition, this research was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology (R33-10076).

References

Al-Adamat, R. A. N., Foster, I. D. L., & Baban, S. M. J. (2003). Groundwater vulnerability and risk mapping for the basaltic aquifer of the Azraq basin of Jordan using GIS, remote sensing and DRASTIC. *Applied Geography, 23*, 303–324.

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. National Water Well Association, EPA-600/2-87-035.

Breiman, L., Friedman, J. H., Olshen, R. A., & Charles, J. S. (1984). *Classification and Regression Trees*. New York: Wadsworth, Inc.

Cherkassky, V., & Mulier, F. (1998). *Learning from Data*. New York: Wiley.

Croskrey, A., & Groves, C. (2008). Groundwater sensitivity mapping in Kentucky using GIS and digitally vectorized geologic quadrangles. *Environmental Geology, 54*, 913–920.

Davis, L. (1991). *Handbook of Genetic Algorithms*. New York: Van Nostrand Reinhold.

Dixon, B. (2005). Groundwater vulnerability mapping: A GIS and fuzzy rule based integrated tool. *Applied Geography, 25*, 327–347.

Fogel, D. B. (1993). Applying evolutionary programming to selected traveling salesman problems. *Cybernetics and Systems, 24*, 27–36.

Goh, A. T. C. (1995). Back propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering, 9*, 145–151.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. New York: Addison-Wesley.

Gurdak, J., Mccray, J., Thyne, G., & Qi, S. (2007). Latin hypercube approach to estimate uncertainty in ground water vulnerability. *Groundwater, 45*, 348–361.

Han, I., Jo, H., & Shin, K. S. (1997). The hybrid systems for credit rating. *Journal of the Korean Operations Research and Management Science Society, 22*, 163–173.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. Michigan: University of Michigan Press.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hilton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*, 79–87.

Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing, 10*, 215–236.

Kalinski, R., Kelly, W., Bogardi, I., Ehrman, R., & Yaniamoto, P. (1994). Correlation between DRASTIC of VOC contamination of municipal wells in Nebraska. *Ground Water, 32*, 31–34.

Kim, Y., & Hamm, S. (1999). Assessment of potential for groundwater contamination using the DRASTIC/EGIS technique, Cheongju area, South Korea. *Hydrogeology Journal, 7*, 227–235.

Kim, K., Yoo, K., Ki, D., Son, I. S., Oh, K. J., & Park, J. (2011). Decision-tree-based data mining and rule induction for predicting and mapping soil bacterial diversity. *Environmental Monitoring and Assessment*. doi:10.1007/s10661-010-1763-2.

Klimasauskas, C. C. (1992). Hybrid neuro-genetic approach to trading algorithms. *Advanced Technology for Developers, 1*, 18–19.

Koza, J. (1993). *Genetic Programming*. Cambridge: MIT.

Kundzewicz, Z., Mata, L., Arnell, N., Doll, P., Jimenez, B., Miller, K., Oki, T., Sen, Z., & Shiklomanov, I. (2008). The implications of projected climate change for freshwater resources and their management. *Hydrological Sciences Journal, 53*, 3–10.

Lee, J., Sohn, Y., Seo, C., Jeon, K., Yoo, S., Jeong, J., Jo, J., & Jeong, H. (2003). Report of Soil and Groundwater examination to Woosan industrial complex and Area of Jungang-dong in Wonju. *Korea Environmental Management Corporation*, 7–18.

Lee, J., Yi, M., Yoo, Y., Ahn, K., Kim, G., & Won, J. (2007). A review of the National Groundwater Monitoring Network in Korea. *Hydrological Processes, 21*, 907–919.

Liu, J., Zheng, C., Zheng, L., & Lei, Y. (2008). Ground water sustainability; Methodology and application to the North China Plain. *Groundwater, 46*, 897–909.

Melloul, A., & Collin, M. (1998). A proposed index of water quality assessment: The case of Israel's Sharon region. *Journal of Environmental Management, 54*, 131–142.

Merchant, J. (1994). GIS-Based groundwater pollution hazard assessment: Critical review of the DRASTIC model. *Photogrammetric Engineering and Remote Sensing, 60*, 1117–1127.

Ray, J., & O'dell, P. (1993). Diversity: A new method for evaluating sensitivity of groundwater to contamination. *Environmental Geology, 22*, 345–352.

Rosen, L. (1994). A study of the DRASTIC methodology with emphasis on Swedish conditions. *Ground Water, 32*, 278–285.

Rupert, M. (1999). Improvements to the DRASTIC Ground-Water Vulnerability Mapping Method. National Water-Quality Assessment Program-NAWQA. USGS Fact Sheet Fs-066-99. U.S. Enred: http://id.water.usgs.gov/pdf/factsheet/DRASTIC.pdf.

Ryan, J., Lin, M. J., & Miikkulainen, R. (1998). Intrusion detection with neural network. *Neural Information Processing Systems, 48*, 72–77.

Secunda, S., Collin, M., & Melloul, A. (1998). Groundwater vulnerability assessment using a composite model combining DRASTIC with extensive agricultural land use in Israel's Sharon region. *Journal of Environmental Management, 54*, 39–57.

Shin, K. S., & Han, I. (1999). Case-based reasoning supported by genetic algorithms for corporate bond rating. *Expert Systems with Applications, 16*, 85–95.

Smith, G., & Campbell, F. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association, 75*, 74–81.

Vano, J., Scott, M., Voisin, N., Stöckle, C., Hamlet, A., Mickelson, K., Elsner, M., & Lettenmaier, D. (2010). Climate change impacts on water management and irrigated agriculture in the Yakima River Basin, Washington, USA. *Climatic Change, 102*, 287–317.

White, H. (1989). Learning in neural networks: A statistical perspective. *Neural Computation, 4*, 425–464.

Winter, T., Harvey, J., Franke, O., & Alley, W. (1999). Ground water and surface water: A single resource. U.S. Geological Survey Circular 1139

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*, 35–62.