18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Modeling large scale geogenic contamination of groundwater, combination of geochemical expertise and statistical techniques

Amini, M.[1], **A. Johnson**[1], **K.C. Abbaspour**[1] **and K. Mueller** [1]

[1] *Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Duebendorf, Switzerland*
Email: manouchehr.amini@eawag.ch

Abstract: There is an increasing interest in modeling groundwater contamination, particularly geogenic contaminant, on a large scale both from the researcher's as well as policy maker's point of view. However, modeling large scale groundwater contamination is very challenging due to the incomplete understanding of geochemical and hydrological processes in the aquifer. Despite the incomplete understanding, existing knowledge provides sufficient hints to develop predictive models of geogenic contamination. In this study we used a global database of fluoride measurements (>60,000 entities), as well as global-scale information relevant to soil, geology, elevation, climate, and hydrology to evaluate several hybrid methods. The hybrid methods were developed by combining two classification techniques including classification tree (CART) and knowledge based clustering (KBC) and three predictive techniques including multiple linear regression (MLR), adoptive neuro-fuzzy inference system (ANFIS) and logistic regression (LR). The results indicated that combination of classification techniques and nonlinear predictive method (ANFIS and LR) were more reliable than others and provided a better prediction capability. Among the different hybrid procedures, combination of KBC-ANFIS and also CART - ANFIS resulted in larger sensitivities and smaller false negative rates for both training and test data sets. However, as the CART classifier is very unstable and very sensitive to re-sampling, the combination of KBC and ANFIS or LR is preferred

Keywords: *Knowledge Based Clustering, Classification and Regression Tree, Adoptive Neuro Fuzzy Inference System, Logistic Regression, Hybrid methods*

1. INTRODUCTION

With the growing groundwater demand for drinking and irrigation, groundwater contamination, particularly geogenic contamination such as arsenic and fluoride, is of increasing concern (WHO, 2001; Zaporozec, 2002). With an incomplete knowledge of where such contamination may occur, modeling large scale groundwater contamination has received considerable attention in recent years (Feenstra et al., 2007; Amini et al., 2008a, 2008b). These models are generally based on statistical approaches, rather than numerical because, though more realistic and transferable, the application of numerical models is still limited to small scale studies as they need a detailed set of input parameters that generally are not available on a large scale. Statistical model of geogenic contamination needs to be able to deal with complex data (different types, continuous, categorical and extremely heterogeneous) while being flexible enough to account for different possible geohydrological settings. As different statistical approaches are inherently different and perform differently, choosing the optimum approach is a difficult task. In addition, a single modeling approach may not be able to handle different types of data (Nga et al., 2007). Generally, the available data come from different sources and are often complex, unbalanced, and contain missing values. Besides, the relation between the independent variable and dependent variables are often strongly nonlinear. Moreover, the processes may be different in different regions; hence, in each region a different set of independent variables may play a significant role, or the same variable may play opposing role in different regions. In this type of applications, a combination of classification and predictive techniques, or hybrid method, is needed to tackle the complexity of the data, and also to take into account the existing knowledge in the models, especially in delineating different regions.

In this study we use a global fluoride database to evaluate the performance of several hybrid methods by combining two classification techniques, (classification tree and knowledge based clustering) and three predictive techniques (multiple regressions, logistic regression and Adoptive neuro-fuzzy inference system). Classification tree (CART) is a statistical procedure for classifying the data according to the measured dependent variable, whereas “knowledge based clustering” (KBC) takes into account the existing knowledge about processes involved as well as the measurements.

2. MATERIAL AND METHOD

2.1. Database

We used a global database of fluoride to test the different modeling procedures in this study. We collected over 60,000 geo-referenced measured groundwater fluoride concentrations from 25 countries (Figure 1). A detail description of the database is given by Amini et al. (2008a). Globally available information related to climate, geology, hydrology, soil, landuse, elevation, and slope were also collected from different sources and a multi-layer global database was created in the ArcGIS (ver. 9.1) environment. An overview of the databases is given at the site (http://www.wrq.eawag.ch/index_EN).

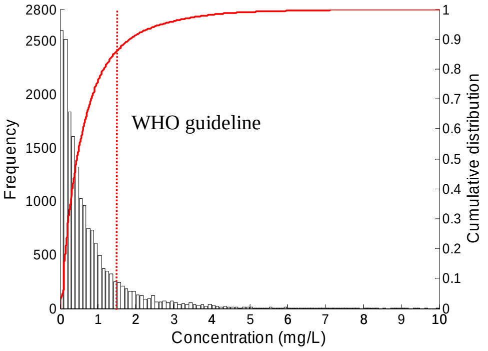

Figure 1. Frequency and cumulative distribution of fluoride concentrations.

2.2. Classification techniques

In this study we used two different classification techniques, classification tree (CART) and knowledge-based clustering (KBC). The classification tree (Breiman, 1996) sequentially grows a binary decision tree by splitting the predictor variables to reduce the conditional variation in the response variable. The best predictor can then be chosen using a variety of impurity or diversity measures. The goal is to produce subsets of the data that are as homogeneous as possible with respect to the target variable (see details in Breiman, 1996).

A detailed description of the knowledge based-clustering of fluoride concentrations is given in Amini et al, (2008a). This algorithm is an interactive procedure that requires the expertise of different disciplines. For the fluoride example we used the available statistical as well as geochemical and hydrological expertise. The goal was to delineate the regions that were as similar as possible in terms of their climatic, geochemical, and hydrological settings ensuring that similar fluoride-releasing processes were at work in each region. This procedure is similar to classification tree but includes supervision, meaning that the place and cutting value of variables in the tree is selected according to the existing knowledge. Here we briefly explain the procedure used to delineate regions for the case of fluoride, henceforth referred to as knowledge-based clustering or KBC. The same procedure could also be adapted for other contaminants.

1- Based on geochemical expertise and literature review, important geological, geochemical, climatic and soil conditions associated with the contaminant (in this study fluoride) in the groundwaters were identified.

2- The geochemical expertise was translated into IF-Then rules by finding appropriate cutoff values for continuous variables to split the data set. The cutoff values can be obtained either by experts or by statistical analysis of the data set.

3- The place of rules in the decision tree were defined by experts according to their influence on the target variable and statistically checked. A decision tree should be developed to delineate the regions with similar conditions.

4- A statistical test, here Kruskal-Wallis followed by least-significant difference test (LSD) for multiple comparisons, was used to compare the fluoride concentrations in the regions.

2.3. Prediction techniques

We used three different prediction techniques, namely logistic regression (LR), multiple liner regression (MLR) and adaptive neuro-fuzzy inference system (ANFIS). LR is a non-linear regression method. It uses a set of binary distributions, such as presence or absence of a characteristic, to derive coefficients for an equation that calculates the probability that a new case is of a certain class (Hosmer and Lemeshow, 2000). ANFIS is a particular type of fuzzy inference system (FIS) attached to a neural network with an adaptive learning procedure (Jang, 1993). The incorporation of fuzzy principles into the neural network provides more user flexibility and system robustness. For a given input/output data set, using ANFIS a FIS with specific membership functions and if-then rules can be constructed. The parameters of the membership functions can then be adjusted according to a learning procedure and the data being modeled.

2.4. Hybrid methods

The hybrid procedure is an integration of classifier, CART or KBC, and a predictive technique, MLR, LR or ANFIS. To briefly explain, the whole procedure consisted of the following steps:

1- Splitting the data set into two subsets for training (80%) and test (20%), using a stratified random sampling.

2- Classifying the training data set using either CART or KBC algorithms.

3- Filtering the significant variables for model development in each class using a stepwise regression.

4- Developing an ANFIS model for each class using the fluoride concentration and significant variables

5- Evaluating the models using both training and test data sets.

6- Propagating the uncertainty and calculating the probability of fluoride concentration exceeding the WHO guide value.

3. RESULTS AND DISCUSSION

3.1. Delineated clusters by KBC

Figure 2a illustrates the delineated region using geochemical expertise. A summary of statistical distribution of fluoride concentration in each region is given in Table 1. According to the origin of fluoride, geological information was regrouped into four categories including, “intrusive felsic rocks”, “volcanic felsic rocks and normal faults”, “sedimentary” and “rest”. The “intrusive felsic rocks” accounts for granitic rocks which known to be a major source of fluoride (Jacks et al., 2005). To capture the volcanic origin of fluoride (Ashley and Burley, 1994), combination of “volcanic felsic rocks” and “normal faults” were used. Our preliminary statistical analysis indicated a clear relationship between the median concentration of fluoride and distance to intrusive felsic rocks or extensional tectonic activities up to 1 decimal degree. Hence, a distance of 1 decimal degree was chosen in the subsequent analysis.

The effect of the climatic condition (Gupta et al., 2004) was captured by creating an ET-index, expressed as evapotranspiration over precipitation (ET/P). To find a relationship between ET/P and fluoride concentration, we plotted a series of ET/P thresholds, $(ET/P)_T$, with increments of 0.2 against the geometric mean of fluoride concentrations fulfilling the condition of $ET/P \geq (ET/P)_T$. We found a clear relationship between ET/P>2 and fluoride concentration. Hence, this ratio was used for a further delineation of the process regions as shown in Figure 2. In addition to geology and climate, the influence of sub-soil pH on fluoride concentration was considered. To this task two groups of soils with pH≥7.2 and pH<7.2 were differentiated because alkaline soils are known to have a positive correlation with fluoride concentration (Wang et al., 2002).

3.2. Delineated clusters by CART

The structure of classification tree obtained by CART is illustrated in Figure 2b. If the condition is fulfilled then the left branch is selected. In this tree, climatic parameters (ET/P and ET) are more significant than geological parameters. Among the different available geological parameters, distance to volcanic rocks, intrusive mafic rocks and intrusive felsic rocks are more significant and retained in the tree. These variables also correspond to the existing geochemical knowledge about the origin of fluoride. The presence of topsoil sand content (sand1) in the tree indicates, to some degree, the sedimentary depositions.

Table **1**. Summary statistic of fluoride concentration in regions delineated by knowledge based clustering (KBC) and CART (CTR).

Region	NO.*	Min.*	Mean†	Max.*	%>1.5
			Knowledge based clusters		
KBC1	3226	0.03	1.21^{a}	9.69	27.68
KBC2	2337	0.05	0.77^{b}	9.43	13.74
KBC3	2256	0.02	0.80^{b}	10	14.31
KBC4	6086	0.05	1.22^{c}	9.92	21.78
KBC5	1298	0.05	1.06^{a}	9.39	17.43
KBC6	720	0.05	0.97^{c}	8.38	17.06
KBC7	1822	0.02	0.41^{d}	9.01	3.23
KBC8	634	0.02	0.82^{b}	8.6	12.32
			Classification tree clusters		
CTR1	251	0.05	2.19^{a}	5.80	68.52
CTR2	472	0.10	1.29^{b}	8.96	25.31
CTR3	1155	0.05	2.23^{a}	9.69	64.42
CTR4	2144	0.03	0.99^{c}	8.20	21.54
CTR6	9745	0.00	0.54^{d}	9.67	5.93
CTR7	4612	0.02	0.78^{e}	9.92	11.60

* No. = number, Min. = Minimum, Max. =Maximum. †Same characters indicate the groups are not significantly different (p <0.05).

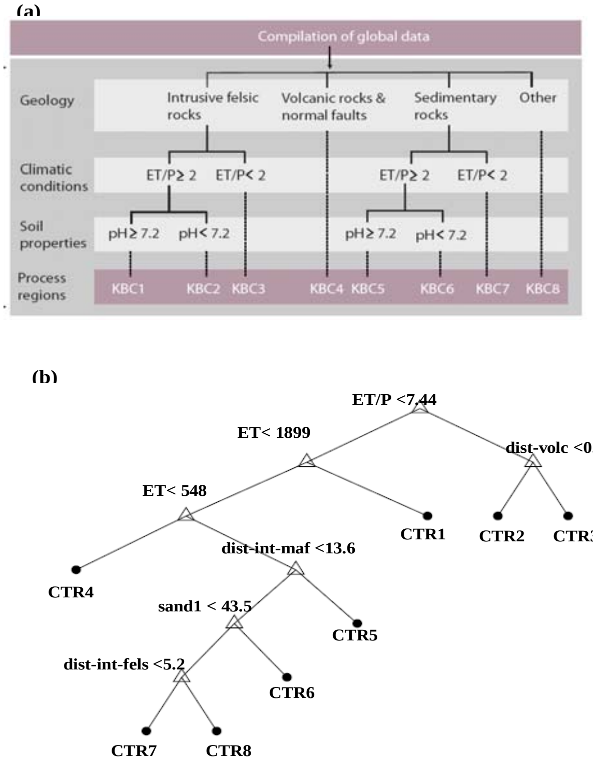

Figure 2. Delineated region by knowledge based clustering (a) and classification and regression tree (b). ET is evapotranspiration, P is precipitation, dist-volc is distance to volcanic rocks, dist-int-maf is distance to intrusive mafic rocks, dist-int-fels is distance to intrusive felsic rocks and sand1 is the topsoil sand content.

3.3. Statistical modeling of clusters

Although significant differences among delineated clusters indicate the clusters were successfully delineated in both methods, there is still a large heterogeneity within each cluster (Table 1), which needs to be captured by further statistical analysis. The influencing variables determined by stepwise regression for KBC classifier are given in Table 2. The coefficients in the table are standardized regression coefficients, indicating the influence of each variable on the predicted fluoride concentration. For example, in KBC1, P and ET/P have almost the same yet opposite effects on the prediction. These results support the necessity of classification prior to regression analysis for large data sets containing different types of variables. Not only do different variables influence the predictions in different regions, but the same variable may also have the opposite influence in different regions. For example, ET/P has a positive effect on fluoride predictions in KBC2 but negative effect in KBC3 (Table 2), which indicates the complexity of interactions between fluoride concentration in groundwater and environmental factors.

Table 2. Selected variables for regions delineated by knowledge based clustering (the numbers are standardized regression coefficients except for intercept).

Variables[†]	KBC1	KBC2	KBC3	KBC4	KBC5	KBC6	KBC7	KBC8
Intercept	-0.32	-1.40	0.04	0.73	-1.19	-1.18	-1.03	-0.68
Elevation	0.08	0.09	-0.16	-	0.06	-	0.17	0.27
Slope	-0.08	-0.07	-	-	-0.05	-0.10	-0.07	-
Evapotranspiration(ET)	-	-0.35	-	-0.26	0.36	0.65	-0.25	-
Precipitation(P)	-0.23	-	-	-	-0.24	0.21	-	-0.80
ET/P	0.24	0.31	-0.14	-	-	0.20	0.40	-
Temperature	-	0.56	0.11	0.37	-	-0.95	0.34	-
Runoff	-	-0.06	-0.12	-	0.14	-0.10	-	0.57
Irrigation	-	-	-0.06	-	-0.07	-	-	0.08
Topsoil sand	-	-	-0.07	-	-	-	-	-
Topsoil silt	-	0.25	-	-	0.14	-	-	-
Topsoil clay	-	-	-	-	-0.13	-	-	-0.20
Subsoil sand	-	-	-	-	-	-	-0.10	-
Subsoil clay	-	-	0.13	-	-	0.27	0.09	-
Topsoil C/N	-	-	0.06	-	-0.12	-	-	-
Subsoil C/N	-0.27	-	-	-	-	-0.15	-	-
Drainage_code	-	-0.23	-	-0.11	0.21	-	-	0.21
Subsoil pH	-	0.18	-	-	-0.18	-0.15	-	-
Subsoil OC	0.21	0.25	-	0.23	-	-	-	-
Subsoil N	-	-0.13	-	-	-	-	-	-
Subsoil CEC	-	-0.14	-0.18	-0.41	-	-0.16	-	-
Dist_V_rest	0.16	0.35	0.26	-0.19	-0.05	-0.34	-	-
Dist_V_fel	0.16	-	-0.17	-	0.18	0.24	-	0.37
Dist_int_maf	-	-	0.07	0.31	0.24	0.38	0.46	-
Dist_int_fel	-0.19	-0.08	-0.04	-	0.11	0.21	-	0.16
Dist_meta	-	-	-	-	0.18	-	-	-
Dist_fault	-	-0.39	-0.08	-0.30	-	0.83	0.19	-
Dist-Rivers	-	-	-	-0.19	-	-	-	-

[†] C = carbon, N=nitrogen, OC = organic carbon, CEC = cation exchange capacity, Dist = Distance (in decimal degree), V_fel=volcanic felsic rocks, V_rest= volcanic rest, int_maf=intrusive mafic rocks, int_fel= intrusive felsic rocks, meta= metamorphic rocks.

3.4. Comparison of different hybrid methods

A comparison of the sensitivity (SEN), specificity (SPE), positive predictive rate (FPR), and negative predictive rate (FNR) for the training set and the test set of classification techniques are shown in Table 3. The values in this table calculated for the probability cut off value of 0.5. In general, hybrid methods outperform the MLR and LR according to the calculated statistical measures. Among the different hybrid procedures, KBC-ANFIS and CART-ANFIS resulted in larger sensitivities and smaller false negative rates for both training and test data sets. All the models have large specificities, which indicate they perform well for the class of low fluoride concentrations. The area under curve (AUC) indicates the discriminatory power of the models. An AUC of 0.5 indicate the model is just a random model while an AUC values larger than 0.7 indicate the model is good. Based on AUC results the methods can be classified into two major categories, the first group with AUC less than 0.65 consist of LR and methods related to linear regression. The other group has AUC larger than 0.75 including combination of KBC and CART with LR and ANFIS (Fig. 6). These findings suggest that the relation between the model input parameters and fluoride concentration is strongly nonlinear even in the delineated regions. Although, CART-LR and KBC-LR result in a larger AUC than CART-ANFIS and KBC-ANFIS, their sensitivities are smaller. In other words, their performance for the class of high fluoride concentration, which is the target class, is not as good as those for class of low concentration. However, as the CART classifier is very unstable and very sensitive to resampling, the combination of KBC and ANFIS is preferred as it provides more robust predictions and also is flexible to account for geohydrological conditions.

Table 3. Comparison of the performance of models for training set and test set, the numbers except for AUC, are calculated for the probability cut off of 0.5.

Model	Training					Test				
	SEN	FNR	SPEC	FPR	AUC	SEN	FNR	SPEC	FPR	AUC
MLR	0.10	0.90	0.99	0.01	0.60	0.10	0.67	0.99	0.01	0.59
LR	0.13	0.87	0.98	0.02	0.65	0.13	0.87	0.98	0.02	0.65
CART-MLR	0.29	0.71	0.91	0.09	0.64	0.29	0.71	0.91	0.09	0.60
CART-ANFIS	0.43	0.57	0.98	0.02	0.80	0.35	0.65	0.98	0.02	0.80
CART-LR	0.36	0.64	0.98	0.02	0.82	0.32	0.68	0.98	0.02	0.82
KBC-MLR	0.33	0.67	0.91	0.09	0.69	0.33	0.67	0.90	0.10	0.68
KBC-ANFIS	0.45	0.55	0.90	0.10	0.76	0.43	0.57	0.89	0.11	0.66
KBC-LR	0.32	0.68	0.98	0.02	0.82	0.36	0.64	0.96	0.04	0.82

SEN = Sensitivity, SPEC= Specificity, FPR= False Positive Rate, FNR= False Negative Rate, AUC= Area under the Receiver Operating Characteristics (ROC) curve, KBC= Knowledge based clustering, ANFIS= Adoptive Neuro Fuzzy Inference System, CART= Classification and Regression Tree, MLR=Multiple linear regression, and LR= logistic regression.

ACKNOWLEDGMENTS

We are grateful to all the colleagues who have provided us with fluoride data.

REFERENCES

Amini, M., Mueller, K., et al. (2008a), Statistical Modeling of Global Geogenic fluoride Contamination in Groundwater. *Environ. Sci. Technol.* 42 (10): 3662–3668.

Amini, M., Abbaspour, K. C., et al. (2008b), Statistical Modeling of Global Geogenic Arsenic Contamination in Groundwater. *Environ. Sci. Technol.* 42 (10): 3669–3675.

Ashley, R. P. and Burley, M. J. (1994), Controls on the occurrence of fluoride in groundwater in the Rift Valley of Ethiopia. In: Nash at al. (Eds), Groundwater Quality. Chapman & Hall, London.

Breiman, L. (1996), The heuristics of instability and stabilization in model selection. *Ann. Statist.* 24, 2350-2383.

Feenstra, L., Vasak, L., et al. (2007), Fluoride in groundwater: overview and evaluation of removal methods." International Groundwater Resources Assessment Centre, Utrecht, The Netherland. www.igrac.nl.

Gupta S., Kumar, A., et al. (2004), Chemical analysis of ground water of Sanganer area, Jaipur in Rajasthan. *J. Environ. Sci. Eng.* 46(1): 74-78.

Hosmer, D. W. and Lemeshow, S. (2000), Applied Logistic Regression. 2nd ed. New York; Chichester, Wiley. ISBN 0-471-35632-8.

Jacks, G. B., Bhattacharya, P., Chaudhary, V., Singh, K. P. (2005), Controls on the genesis of some high-fluoride groundwaters in India. *Appl. Geochem.* 20: 221-228.

Jang, J. S. R. (1993), ANFIS: Adaptive-Network-based Fuzzy Inference Systems. *IEEE Transactions on Systems, Man, and Cybernetics* 23(3): 665-685.

Wang W. Y., Li, R. B., et al. (2002), Adsorption and leaching of fluoride in soils of China. *Fluoride* 35(2): 122-129.

WHO, (2001), United Nations Synthesis Report on Arsenic in Drinking Water. http://www.who.int/water_sanitation_health/dwq/arsenic3/en/.

Zaporozec, A., 2002. Groundwater contamination inventory: A methodological guide. IHP-VI, SERIES ON GROUNDWATER NO.2, UNESCO. http://unesdoc.unesco.org/images/0013/001325/132503e.pdf.