PREDICTION OF PARTITION COEFFICIENTS FOR CADMIUM BY NEURAL NETWORKS

Nader Shariatmadari, Assistant Professor, Dept. of Civil Engineering of IUST
Mohammad Farrokhi, Assistant Professor, Dept. of Electrical Engineering of IUST
Amin Falamaki, Graduate student, Dept. of Civil Engineering of IUST

ABSTRACT
The partition or distribution coefficient (K_d); is an important parameter in estimating the potential for the adsorption of dissolved contaminants in soil pollution problems. It has been understood that K_d values can result in significant errors for predicting the impacts of contaminant migration or site - remediation options. The empirical predictor equations may be derived commonly by statistical analysis and take the form of a linear or nonlinear polynomial expression. In this study the artificial neural networks; ANNs; is used to predict the variation of the partition coefficient with variation of environmental components of soil. The objective is to investigate the feasibility of ANN technique for predicting of K_d variation with variation of environmental pH.To accomplish this object the database reported by EPA (1999) for cadmium adsorption were used. Results show that ANNs are more powerful tools than statistical analysis for prediction of partition coefficient variation with variation of environmental components of soil.

RÉSUMÉ
Le coefficient de partition ou de distribution (K_d), est un paramètre important pour l'estimation du potentiel d'adsorption des contaminants dissous en relation avec les problèmes de contamination des sols. Il est généralement compris que les valeurs de K_d utilisées peuvent engendrer des erreurs significatives dans la prédiction des impacts de la migration des contaminants ou pour le choix des options de réhabilitation. Les équations de prédiction empiriques peuvent être communément dérivées par des méthodes statistiques et prendre la forme d'expressions polynomiales linéaires ou non linéaires. Pour notre étude, les réseaux neuroniques artificiels (*artificial neural networks, ANNs*) sont utilisés pour prédire la variation du coefficient de partition avec la variation des conditions environnementales du sol. L'objectif est d'étudier la faisabilité de la technique par *ANN* pour la prédiction de la variation du K_d avec la variation du pH environnemental. Pour atteindre cet objectif, la base de données de l'EPA (1999) pour l'adsorption du cadmium a été utilisée. Les résultats montrent que les *ANNs* sont des outils plus puissants que l'analyse statistique pour la prédiction de la variation du coefficient de partition avec la variation des conditions environnementales du sol.

1. INTRODUCTION

The partition or distribution coefficient (K_d); is an important parameter in estimating the potential for the adsorption of dissolved contaminants in soil pollution problems. As typically used in fate and contaminant transport calculations, the K_d is defined as the ratio of the contaminant concentration associated with the solid to the contaminant concentration in the surrounding aqueous solution when the system is at equilibrium (EPA 1999). Soil chemists and geochemists have understood that K_d values can result in significant errors for predicting the impacts of contaminant migration or site - remediation options.

The constant K_d model and the parametric K_d model are the models that are used in predicting K_d. An important limitation of the constant K_d model is that it does not shows sensitivity to changing conditions. If some properties of groundwater (e.g., pH and solution ionic strength) change, a different K_d value should be applied in the model. In the parametric K_d model the K_d value varies as a function of empirically derived relationships with aqueous and solid phase independent parameters. Thus, it has the distinct advantage that considers new K_d values for each environmental condition. The empirical predictor equations may be derived commonly by statistical analysis and take the form of a linear and nonlinear polynomial expression. Table 1 shows some of the relations between K_d values and environmental condition in soils. Some of these statistical models are based on the limited database and show high errors and low correlation factors.

Some of the researchers tried to find relations for K_d for a certain soil in different environmental conditions. The relationship between equilibrium concentrations of lead and K_d values for a Hanford soil at a fixed pH was expressed by Eq.1 (Rhoads *et* al.1992) as:

$$K_d \text{ (ml/g)} = 9550\ C^{-0.335} \qquad [1]$$

where C is the equilibrium concentration of lead (µg/l).

In recent times, artificial neural networks (ANNs) have been applied to many geotechnical and environmental problems and showed some degree of success. The application of ANNs may overcome the limitations of traditional methods. In this study the ANNs is used to predict the variation of the partition coefficient, K_d, with variation of environmental components. The objective is to investigate the feasibility of ANN technique for predicting

Table 1. Relations Between Kd Values And Environmental Condition In Soils

RELATION SHIP	Species	1.1.1 Reference	Comments
K_d =-0.54 +0.45(pH).	Cadmium	EPA 1999	Different soils
log (K_d)=1.2 +1.0 log (CEC)	Cesium	Akiba and Hashimoto (1990)	A large number of soils, minerals, and rock materials
K_d(ml/g)=1639-902.4(pH)+150.4(pH)^2	Lead	Gerritse *et al.* (1982) Rhoads *et al.*(1992)	Different soils
K_d = 284.6 (DCARB)+27.8 (CLAY)-594.2	Plutonium	EPA 1999	Different soils
K_d=488.3(DCARB)+29.9(CLAY)-119.1(pH)-356.8 (EC)	Plutonium	EPA 1999	Different soils
K_d =25.7 (DCARB)+12.14 CLAY)+2.41 K_d<767.5 K_d=286.0(DCARB)+21.3(CLAY)-81.2 K_d>767.5	Plutonium	EPA 1999	Different soils
log (K_d)=-0.13 +0.69(pH).	Thorium	EPA 1999	The pH range of 4 to 8

DCARB = The Concentrations Of Dissolved Carbonate Of Soils

CLAY = The Clay Content Of Soils

EC = Electrical Conductivity

of K_d variation with variation of environmental pH.To accomplish this object the database reported by EPA (1999) for cadmium adsorption were used.

2. DATABASE

EPA (1999) reported cadmium K_d values and some important ancillary parameters that have been shown to influence cadmium sorption. Data set were from studies that reported K_d values and were conducted in systems consisting of

- Natural soils (as opposed to pure mineral phases)
- Low ionic strength solutions (<0.1 M)
- pH values between 4 and 10
- Solution cadmium concentration less than 10^{-5} M
- Low humic materials concentrations (<5 mg/l)
- No organic chelates (such as EDTA)

Totally, 170 cadmium K_d values were tabulated by EPA (1999). Of the 170 cadmium Kd values, 62 values had associated clay content data, 170 values had associated pH data,22 values had associated CEC data,63 values had total organic carbon data, 170 values had associated cadmium concentration data, and 16 had associated aluminum/iron-oxide data.

2.1 APPROACH AND REGRESSION MODELS

According to EPA report, linear regression analyses were conducted between the parameters and cadmium K_d values. The coefficients of correlation from these analyses are presented in Table 2. The largest correlation coefficient was between pH and log(K_d). This value is significant at the 0.001 level of probability (EPA 1999). Attempts at improving this correlation coefficient using additional variables, *i.e.*, using multiple-regression analysis, were not successful (EPA 1999).

Table 2 .Correlation coefficients (r^2) of the cadmium Kd data set for soils EPA (1999).

	K_d	log(K_d)	Clay Conc.	pH	CEC	TOC	Cd
K_d	1						
log(K_d)	0.69	1					
Clay Conc.	-0.04	0.03	1				
pH	0.5	0.75	0.06	1			
CEC	0.4	0.41	0.62	0.35	1		
TOC	0.2	0.06	0.13	-0.39	0.27	1	
Cd	-0.02	-0.1	-0.39	0.22	-0.03	-0.09	1
Conc.Fe Oxide	0.18	0.11	-0.06	0.16	0.19	0.18	0.01

Figure1 shows the cadmium K_d values as a function of pH. A large amount of scatter exists in these data (EPA 1999). At any given pH, the range of K_d values may vary by 2 orders of magnitude. This is not entirely satisfactory, but as explained above, using more than 1 variable to help categorize the cadmium K_d values was not fruitful (EPA 1999). The regression equation 2 presents the line in Figure 1:

$$\log K_d = -0.54 + 0.45(pH) \quad [2]$$

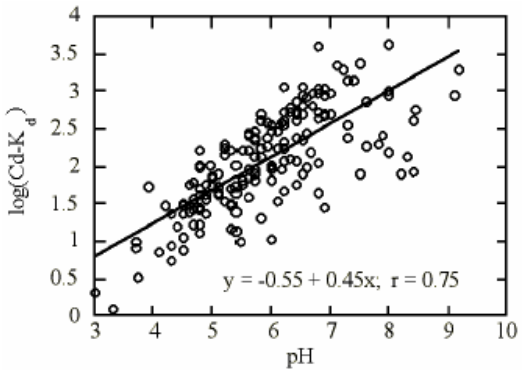

Figure 1. Relation between cadmium K_d values and pH in soils.

3. NN MODEL

Work on artificial neural networks, commonly referred to as "neural network", has been motivated right from its inception by the recognition that the brain computes in an entirely different way from the conventional digital computer (Haykin, Simon 1994). To apply a NN to solving a real world problem, four basic steps are involved: (1) analyze the real world problem and select proper network architecture; (2) collect and pre-process data for training and testing; (3) design, train, and test the network model; and (4) deploy the network to the end user (Shi et al. 1998).

Among various network architectures, Multilayer perceptrons have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the error back-propagation algorithm. This algorithm is based on the error correction-learning rule. In the implementation of MLPs, data are categorized as input patterns and target patterns. The input patterns are fed to the network, which then performs feed-forward computations to calculate output patterns. The output patterns are compared with corresponding target patterns and the summation of the square of the error is calculated. The error is then back propagated through the network using the gradient-descent rule to modify the weights and minimize the summed squared error (Ellis et al. 1995). Thus, a good mapping between input patterns and target patterns can be achieved, resulting in a network capable of predicting the target pattern for a given input pattern.

In this research, two kinds of ANNs models were used to predict K_d values reported in literature, as follows: (1) Multilayer perceptrone network (MLP); (2) Generalized regression neural network (GRNN).

3.1 MLP MODEL

In back-propagation learning, we typically start with a training set and use the back-propagation algorithm to compute the synaptic weights of a multilayer perceptron by loading (encoding) as many of the training examples as possible into the network. The hope is that the neural network so designed will generalize. A network is said to generalize well when the input-output relationship computed by the network is correct (or nearly so) for input/output patterns (test data) never used in creating or training the network; the term "generalization" is borrowed from psychology. Here, of course, it is assumed that the test data are drawn from the same population used to generate the training data. Validation subset is typically 10 to 20 percent of the training set (Haykin, Simon 1994). Here 30 patterns were used for testing.

Designing BP network architecture includes determining the number of input and output variables (i.e., neurons in input and output layers) and selecting the number of hidden layers and neurons in each hidden layer. The number of hidden layers and number of neurons in each hidden layer in a BP network may affect the training efficiency and the precision of prediction. It is impossible to prove how many hidden layers and how many neurons in each hidden layer can result in the most effective training and the most accurate prediction, although the genetic algorithm can help us to some extent. The common practice is to experiment with different numbers of hidden layers and different numbers of neurons in each layer. The number of neurons in the input and output layers corresponds to the expected input and output variables of problem. Output variables are the expected answers to the problem, and the input variables are factors that affect the answers.
The NN program used was MATLAB 6.5. We have experimented with various BP networks with different hidden layers and different numbers of neurons in each hidden layer.

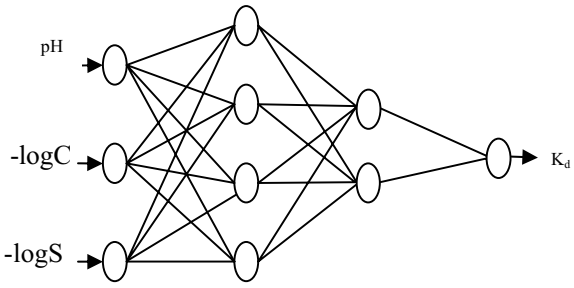

Figure 2. 3×4×2×1 MLP Network

3.2 GRNN MODEL

Generalized regression neural networks are a kind of radial basis network that is often used for function approximation. GRNNs can be designed very quickly.

The design of a supervised neural network may be pursued in a variety of different ways. The back-propagation algorithm for the design of a multilayer perceptron (under supervision) as described in the previous part may be viewed as an application of an optimization method known as stochastic approximation (Haykin, Simon 1994). In this part, we take a different

approach by viewing the design of a neural network as a curve-fitting (approximation) problem in a high-dimensional space. According to this viewpoint, learning is equivalent to finding a surface in a multidimensional space that provides a best fit to the training data, with the criterion for "best fit" being measured in some statistical sense. Correspondingly, generalization is equivalent to the use of this multidimensional surface to interpolate the test data. Such a viewpoint is indeed the motivation behind the method of radial-basis functions in the sense that it draws upon research work on traditional strict interpolation in a multidimensional space (Haykin, Simon 1994). In the context of a neural network, the hidden units provide a set of "function" that constitute an arbitrary "basis" for the input patterns (vectors) when they are expanded into the hidden-unit space; these functions are called *radial-basis functions.*

The construction of a *radial-basis function* (RBF) network in its most basic form involves three entirely different layers as illustrated in Figure 3. The input layer is made up of source nodes (sensory units). The second layer is a hidden layer of high enough dimension, which serves a different purpose from that in a multilayer perceptron. The output layer supplies the response of the network to the activation pattern applied to the input layer. The transformation from the input space to the hidden-unit space is nonlinear, whereas the transformation from the hidden-unit space to the output space is linear.

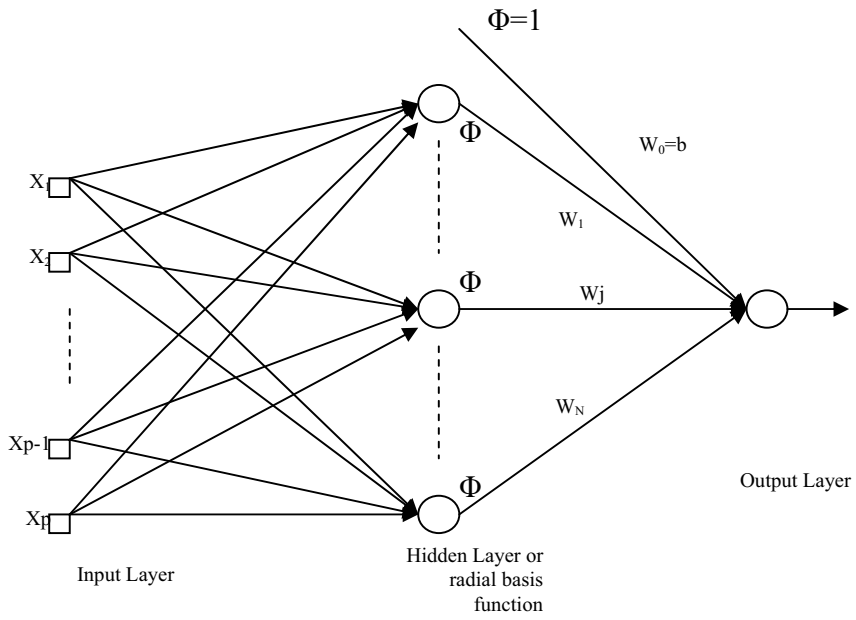

Figure 3. Radial Basis Function Network

RBF network creates as many radial-basis neurons as there are input vectors in X, and sets the first-layer weights to X. Thus, we have a layer of radial-basis neurons in which each neuron acts as a detector for a different input vector. If there are Q input vectors, then there will be Q neurons. Each bias in the first layer is set to 0.8326/SPREAD. This gives radial basis functions that cross 0.5 at weighted inputs of +/- SPREAD. This determines the width of an area in the input space to which each neuron responds. If SPREAD is 100, then each radial-basis neuron will respond with 0.5 or more to any input vectors within a vector distance of 100 from their weight vector. SPREAD should be large enough that neurons respond strongly to overlapping regions of the input space.

3.3 TRAINING AND TESTING OF MLP MODEL

All measured data were tabulated EPA (1999). 170 valid data patterns were extracted. As explained the largest correlation coefficient was between pH and $\log(K_d)$. Among them, 138 patterns were used for training. In back-propagation learning, we typically start with a training set and use the back-propagation algorithm to compute the synaptic weights of a multilayer perceptron by loading (encoding) as many of the training examples as possible into the network. The hope is that the neural network so designed will generalize. A network is said to generalize well when the input-output relationship computed by the network is correct (or nearly so) for input/output patterns (test data) never used in creating or training the network; the term "generalization" is borrowed from psychology. Here, of course, it is assumed that the test data are drawn from the same population used to generate the training data. Validation subset is typically 10 to 20 percent of the training set (Haykin, Simon 1994). Here 33 patterns were used for testing.

The NN program used was MATLAB 6.5. We have experimented with various BP networks with one or two

hidden layers and different numbers of neurons in each hidden layer using the above collected data patterns. The networks with two or three hidden layers and different hidden neurons for hidden layers, respectively, have shown better agreement to the training patterns.

3.4 TRAINING AND TESTING OF GRNN MODEL

Similar to the MLP model, the GRNN model was trained and tested using the same data in the previous part. The NN program used was MATLAB 6.5. The network with one hidden layers and 138 hidden neurons for that hidden layer, has shown good agreement to the training patterns.

4. RESULTS AND DISCUSSION

Totally, four networks; GRNN, MLP 10x5x1, MLP 20x10x1 and MLP 10x5x3x1; are presented in this paper and compare to Mathematical model. The statistical accuracy parameters are tabulated in Table 2a, 2b and 2c for the training data, the testing data, and total (training and testing) data respectively. Some of this tabulated information is presented by next figures for convenience.

Table 2a. Training data

Model	Math. Model	GRNN	MLP 10x5x1	MLP 20x10x1	MLP 10x5x3x1
Mean Squared Error	428,661	180,696	239,415	253,168	242,584
Standard Error	519	304	235	206	210
Minimum Absolute Error	0.651	0.000	0.023	0.302	0.036
Maximum Absolute Error	3,668	2,839	3,608	3,689	3,656
Correlation Coefficient	0.366	0.712	0.628	0.613	0.636
Mean Relative Error (%)	199.4	139.5	96.0	101.1	95.5

Table 2b. Testing data

Model	Math. Model	GRNN	MLP 10x5x1	MLP 20x10x1	MLP 10x5x3x1
Mean Squared Error	78,473	92,255	53,066	56,281	53,636
Standard Error	212.5	296.3	131.2	140.5	127.5
Minimum Absolute Error	0.384	1.900	0.758	1.302	1.158
Maximum Absolute Error	768.4	1,051.3	610.6	645.4	625.5
Correlation Coefficient	0.410	0.562	0.604	0.575	0.609
Mean Relative Error (%)	130.3	152.1	108.4	102.8	105.8

Table 2c. Training and testing data

Model	Math. Model	GRNN	MLP 10x5x1	MLP 20x10x1	MLP 10x5x3x1
Mean Squared Error	361,081	163,628	203,453	215,173	206,121
Standard Error	475.3	302.5	218.5	195.2	197.1
Minimum Absolute Error	0.384	0.000	0.023	0.302	0.036
Maximum Absolute Error	3,668	2,839	3,608	3,689	3,656
Correlation Coefficient	0.370	0.691	0.627	0.610	0.634
Mean Relative Error (%)	186.1	141.9	98.4	101.4	97.5

Figure 4 presents the MRE (mean relative error) for predicted and measured for different models. It can be seen that for all sets of data the MLP networks show significant lower MRE. However, the GRNN shows higher MRE relative to the MLP. However, the number of hidden layers and neurons does not affect the MRE significantly. According to this figure, the MRE for total set of data is reduced from 186% for mathematical model to about 100% for MLP models.

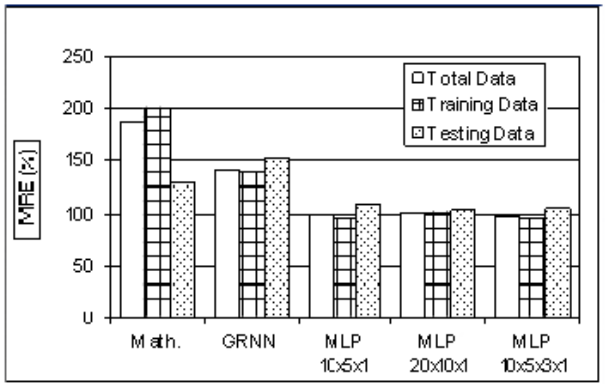

Figure 4. Mean relative error for different models

Figure 5 shows the correlation coefficient between predicted and measured K_d for different models. According to this figure, the correlation coefficient of mathematical model presented by EPA(1999); about 36 and 37%; are increased to more than 60% by ANNs models.

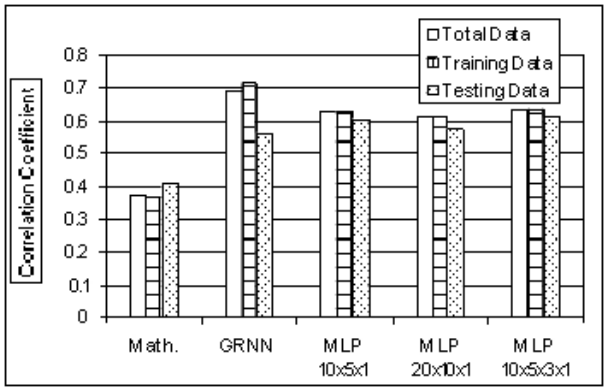

Figure 5. Correlation coefficient between predicted and measured K_d for different models

Such results are presented in Figure 6 for standard error. The standard error calculated by EPA mathematical model is about twice the MLP models.

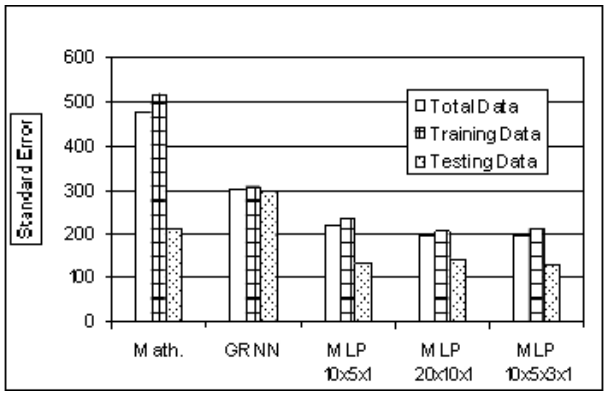

Figure 6. Standard error for different models

5. CONCLUSIONS

In this study the ANNs was used to predict the variation of the partition (or distribution) coefficient, K_d, with variation of environmental components of soil. The objective is to investigate the feasibility of ANN technique for predicting of K_d variation with variation of environmental pH.To accomplish this object the database reported by EPA (1999) for cadmium adsorption were used. Results show that ANNs are more powerful tools than statistical analysis for prediction of partition coefficient variation with variation of environmental components for a certain soil.

The MRE for total set of data is reduced from 186% for EPA mathematical model to about 100% for MLP models. Higher coefficients of correlation were obtained using the ANNs. The coefficient of correlation of mathematical model presented by EPA(1999); about 36 and 37%; are increased to more than 60% by ANNs models. Although results show the feasibility of pridiction of K_d, but more investigation is required.

6. REFERENCES

Akiba, D., and H.Hashimoto.1990."Distribution Coefficient of Strontium on Variety of Minerals and Rocks." Journal of Nuclear Science and Technology, 27:275-279.

Ellis, G. W., Yao, C., Zhao, R., and Penumadu, D. (1995). "Stress-strain modeling of sands using artificial neural networks." J. Geotechnical Engineering, ASCE, 121(5), 429-435.

EPA (1999) 402-R-99-004b Environmental Protection Agency, "Understanding Variation In. Partition Coefficient, Kd Values, Volume II. Review of Geochemistry and Available Kd Values for Cadmium, Cesium, Chromium, Lead, Plutonium, Radon, Strontium, Thorium, Tritium (3H), and Uranium".

Gerritse, R.G., R.Vriesema, J.W.Dalenberg, and H.P.De Roos.1982. "Effect of Sewage Sludge on Trace Element Mobility in Soils." Journal of Environmental Quality, 11:359-364.

Haykin, S. (1994). "Neural network" New York: Macmillan College Publishing Co.

Rhoads,K.,B.N.Bjornstad,R.E.Lewis,S.S.Teel,K.J.Cantrell, R.J.Serne,J.L.Smoot,C.T.Kincaid,and S.K.Wurstner.1992. "Estimation of the Release and Migration of Lead through Soils and Groundwater at the Hanford Site" 218-E-12B Burial Ground. Volume 1: Final Report. PNL-8356 Volume 1, Pacific Northwest Laboratory, Richland, Washington.

Shi, J., Ortigao, J. A. R., and Bai, j. (1998). "Modular neural networks for predicting settlements during tunneling" J. Geotechnical and Geoenvironmental Engineering, ASCE, 124(5), 389-395.