Estimation of Soil Distribution Coefficient Using Artificial Neural Networks Modelling for Chromium Contaminated Water

Mostafa Abolfazlzadehdoshsnbehbazari
Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada
Amin Falamaki
Department of Civil Engineering, Payam Noor University of Shiraz, Shiraz, Fars, Iran
Ania C. Ulrich
Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada

ABSTRACT
In this study, we determined the reliability of artificial neural networks (ANNs) in predicting soil distribution coefficients, K_d, in different soil types (Pacolet, Cloudland and Kenoma) and environmental conditions (pH, C and S content). The effect of ANNs geometry on the performance of the models was also assessed. The ANNs tested include the Multi Layer Perceptron (MLP), Radial Basis Function (RBF) and Hierarchical (HIER) Networks. In all cases, the correlation factors were greater than 0.984, demonstrating that ANNs are powerful tools for predicting K_d values and that MLP provided the best results.

RÉSUMÉ
Dans cette étude, nous avons déterminé la fiabilité des réseaux de neurones artificiels (RNA), pour prédire la variation du coefficient de la distribution des sols (K_d) pour les différents types des sols (Pacolet, Cloudland et Kenoma) et différentes conditions environnementales (pH et la teneur de C et de S). L'effet de la géométrie du RNA sur la performance des modèles est également évalué. Les réseaux neuronaux artificiels qui sont testés, incluent le Perceptron Multi Layer (PML), la Fonction de base Radiale (FBR) et les réseaux Hierarchial (HIER). En tous les cas, les facteurs de la corrélation sont supérieurs à 0,9841, ce qui montre que les RNA est un moyen puissant pour prédire des valeurs du K_d et que le PML mène aux meilleurs résultats.

1 INTRODUCTION

1.1 Sorption and Conventional Modeling Approaches

Sorption of a contaminant to soil is one of the most important processes in determining the fate of chemicals in the environment. Sorption is usually defined as the association of a dissolved contaminant with the surface of a solid material (Alley, 1993). Sorption may occur by adsorption to the surface of a solid, absorption within a solid, precipitation as a 3-dimensional molecular structure on the surface of the solid, or by partitioning into the organic matter (Sposito, 1989; Krupka et al. 1999).

The distribution coefficient, K_d is defined as the ratio of contaminant concentration in soil phase to the concentration in aqueous phase at equilibrium. K_d values depend on the type of contaminant and chemistry of aqueous and soil phases (Kaplan and Serne, 1995; EPA 2004). The most common adsorption models used in predicting K_d values are constant distribution coefficient, parametric sorption and isotherm adsorption models, including linear, Langmuir and Freundlich (Appelo and Postma, 1996; Rongbo Guo et al. 2002).

Contaminant properties and soil/aqueous environments exhibit multifaceted behaviour due to complex and imprecise geo-chemical processes associated with adsorption. In order to cope with the complexity of this process, artificial neural networks (ANNs) can be used and are well suited to model complex problems especially where the relationship between the variables is unknown (Hubrick, 1992).

1.2 Artificial Neural Networks Models

Artificial neural networks mimic the complicated behaviour of the central nervous system, based on how the human brain copes with problems as compared to conventional mathematical models and digital computers. Processing Elements (PE) which carry out the role of nodes in the human brain are arranged in three layers: an input layer, output layer and hidden layer. Complex and highly non-linear real world problems cannot be solved by traditional regression analysis, but ANNs can overcome these limitations by changing the function of transfer and in the case of highly non-linear phenomena by changing the number of hidden layers and nodes (Gardner and Dorling, 1998). ANN without any knowledge between the nature of input and output variables, tries to adjust their weights by using a training data set to find the input-output mapping with the smallest error. In fact, after summation of the input variables (X_0, X_1 … X_n) multiplied with the coefficients W_{ki} (weights), the result is passed through a transfer function. This function may be either a threshold logic, hard limiting function, sigmoid nonlinearity, or hyperbolic tangent. This process is summarized in Equations 1 and 2 and illustrated in Figure 1.

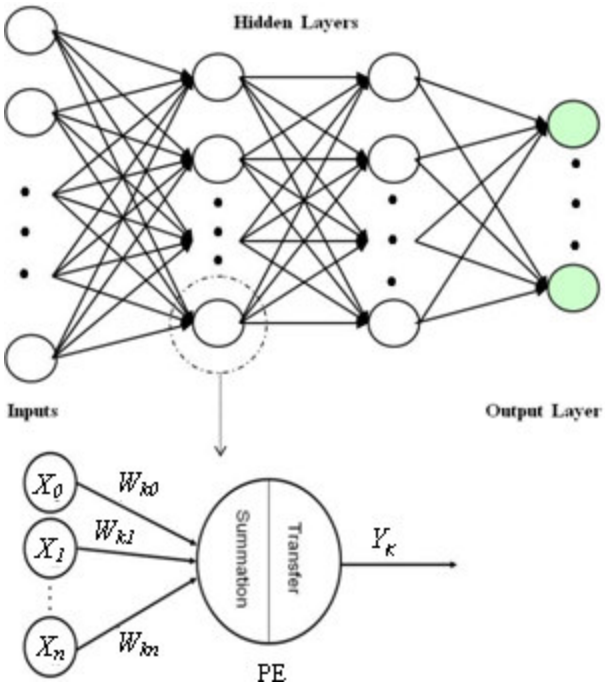

Figure 1. The ANN architecture

$$I = \sum_{i=0}^{n} W_{ki} X_i \quad [1]$$

$$Y_k = f(I) \quad [2]$$

In Equations 1 and 2, I is the result of summation, $f(.)$ is the transfer function and Y_k is the output of PE which provides the input of the PEs in the next layer.

Weight adjusting of variables in order to minimize errors is called training. After the training phase, the accomplished model performance is examined by using different sets of input-outputs to determine the reliability of the model (Shahin et al. 2001; Maier and Dandy, 2000).

Artificial neural networks have been successfully applied in various fields of geotechnical and environmental engineering such as settlement of foundations (Sivakugan, 1998), liquefaction (Ural and Saka, 1998; Najjar and Ali, 1998) and modelling of water treatment and quality (Baxter et al. 2001). ANN modeling of Pb(II) adsorption from aqueous solution by Antep pistachio shells showed the correlation coefficient of about 0.936 between ANN model outputs and experimentally measured model variables (Yetilmezsoy and Demirel, 2008). Fatemi (2007) investigated the ANN modeling of micelle-water distribution coefficient and obtained a root mean square error of 0.06 and 0.20 for training and test sets, respectively. In addition, ANN modeling of organic chemicals sorption on soil by Gao (1996) indicated a good fit between training and predicting distribution coefficient of organic carbon for the test set. Although there have been several studies on the successful use of ANN in modeling of contaminants distribution in soil, application of ANN in modeling of chromium adsorption into soils has not been previously reported. Sorption of every contaminant in a given environment would need a particular approach and may pose distinct challenges in terms of complexity of sorption media, variation in environmental conditions and so on. Moreover, using three different ANN models to compare the results is one of the advantages of the present study. In this study MLP, RBF and Hierarchical ANN models are used to predict the variation in the distribution coefficient, K_d, for chromium under different environmental conditions, variation of pH, carbon and sulphur content, on Pacolet, Cloudland and Kenoma soil; to determine if ANNs provide a reliable approach for predicting K_d variation in different soil types; the effect of the ANNs' geometry on the performance of ANN models; and if it is possible to use ANNs if environmental factors are varied.

2 MATERIAL AND METHODS

2.1 Data Base

The chromium adsorption data reported by Rai et. al. (1988) was used in this study. A review of this data indicated that a variety of factors influence the adsorption behaviour of chromium. These factors and their effects on chromium adsorption on soils and sediments were used as the basis for this study. Chromium adsorption on Pacolet, Cloudland and Kenoma soil is calculated as a function of pH, carbon and sulphur concentration of the soil. For Kenoma soil, measured K_d ranged from 1 to 28 (ml/g) for different pH, -log C (minus logarithm of carbon concentration) and -log S (minus logarithm of sulphur concentration). For this soil the number of tests is 15, nine of them used for training and six of them used for testing the ANN models. The number of reported K_d values used to test the Pacolet soil is 20 (14 of them used for training and six of them for testing ANN models). In this case measured K_d values range from 1 to 465 (ml/g). There are 23 K_d test results for Cloudland soil which ranges from 1 to 1443 (ml/g). Sixteen of these results were used for training and seven of them were used for testing the ANN models.

2.2 ANN Models

Three types of ANN models, Multilayer Perceptron (MLP) network, Radial-Basis Function (RBF) network and Hierarchical network, were used to predict K_d values. The programming of these models was completed with MATLAB 6.5 and each network was trained for Kenoma, Pacolet and Cloudland soils, with nine, fourteen and sixteen sets of data, respectively (as described in section 2.1), which are the minimum number of records to achieve high accuracy in model training. These numbers were selected from trial and error runs during ANN model training. Each data set contained different pH, carbon and sulphur content values as they relate to measured K_d values.

2.2.1 Multilayer Perceptron Network

The structure of a MLP model consists of a number of hidden layers with several neurons in each hidden layer. The activation functions for hidden layer neurons are tangent hyperbolic and for the output layer is linear, in fact, linear function for transformation from the hidden space to the output space will increase the performance of the network (Vaziri et al, 2006). The MLP network specifications (the numbers of training samples, simulating samples, neurons in first hidden layer, neurons in second hidden layer and neurons in output layer), are summarized in Table 1.

2.2.2 Radial Basis Function Model

The basic form of the radial-basis function network includes three layers. Each layer has a completely different role. The input layer is made up of source nodes (sensory units) and connects the network to the environment. The role of the second layer is to apply a nonlinear transformation from the input space to the hidden space. This hidden layer in most cases has higher dimensions. The output layer is linear and supplies the response of the network to the activation pattern applied to the input layer. Table 1 demonstrates the summary of RBF network specifications used in this research.

2.2.3 Hierarchical Models

A modular approach involving two hierarchical levels of network architecture were adopted. This network includes two clusters containing two experts in each of them. Experts are networks with one neuron and each cluster contains a gating network with two neurons. The top level gating network beyond these clusters consists of two neurons. To obtain best results, different conditions were used (see Table 2 for a summary). The training of data for this model was carried out in a highly time consuming manner.

3 RESULTS & DISCUSSION

The available experimental data for chromium adsorption (Rai et al, 1988) were used to configure and evaluate the suitability of three ANN models in predicting the K_d values for three different soils. Statistical parameters such as correlation coefficient (Corr(x,y)), standard error (SE) and mean relative error (MRE) were calculated using Microsoft Excel for predicted and measured data. In fact, these statistical parameters are good indicators to check the prediction performance of the ANN networks. Equations 3 to 5 are the formula of these parameters and the calculated values are shown in Table 3. As it can be seen from the table, in all cases, correlation factors were no less than 0.984. The best correlation factor was obtained for Cloudland soil with a value of one which was obtained with the MLP model. The weakest correlation was 0.984 for Kenoma soil with the HIER 2 model.

$$Corr(K_{dm}, K_{dp}) = \frac{\sum (K_{dm} - \overline{K_{dm}})(K_{pm} - \overline{K_{pm}})}{\sqrt{\sum (K_{dm} - \overline{K_{dm}})^2 \sum (K_{dp} - \overline{K_{dp}})^2}} \quad [3]$$

$$SE = \frac{s}{\sqrt{n}} \quad [4]$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\left(\left.(K_{dpi} - K_{dmi})\right/K_{dmi}\right)\right| \times 100 \quad [5]$$

Where:

n is the number of samples,

K_{dm} and K_{dp} are the measured and predicted values of K_d respectively,

$\overline{K_{dm}}$ and $\overline{K_{pm}}$ are the average value of measured and predicted K_d respectively, and

s is the sample standard deviation.

Figures 2 to 4 show comparison of measured and predicted K_d values for all soils for training and testing data sets.

Table 1. Summary of MLP and RBF networks

No.	Name	NTS[1]	NSS[2]	NHN1[3]	NH2[4]	NON[5]
1	MLP Kenoma Network	9	6	4	2	1
2	MLP Pacolet Network	14	6	4	2	1
3	MLP Cloudland Network	16	7	4	2	1
4	RBF Kenoma Network	9	6	3	0	1
5	RBF Pacolet Network	14	6	9	0	1
6	RBF Cloudland Network	16	7	16	0	1

[1] NTS is number of training samples,
[2] NSS is number of simulating samples
[3] NHN1 is number of neurons in first hidden layer
[4] NHN2 is number of neurons in second hidden layer
[5] NON is number of neurons in output layer.

Table 2. Summary of hierarchical networks

No	Name	Type of the Soil	Description
1	HIER 1	Kenoma Soil	K_d was used as targets. Top level gating network
2	HIER 2	Kenoma Soil	K_d was used as targets. Top level gating network trained by output of each clusters.
3	HIER 3	Kenoma Soil	Natural logarithm of K_d was used as targets. Top level gating network trained by input data.
4	HIER 4	Kenoma Soil	The same as HIER 3 but the training time was doubled.
5	HIER 6	Pacolet Soil	Natural logarithm of K_d was used as targets. Top level gating network trained by input data.
6	HIER 7	Pacolet Soil	The same as network HIER 6 but more training time.
7	HIER 8	Cloudland Soil	(K_d/1600) were used as targets. Top level gating network trained by output of each clusters.
8	HIER 9	Cloudland Soil	Natural logarithm of K_d was used as targets. Top level

For all networks:
Number of clusters = 2
Number of experts in cluster = 2
Number of neurons in expert network = 1
Number of neurons in gating networks = 2

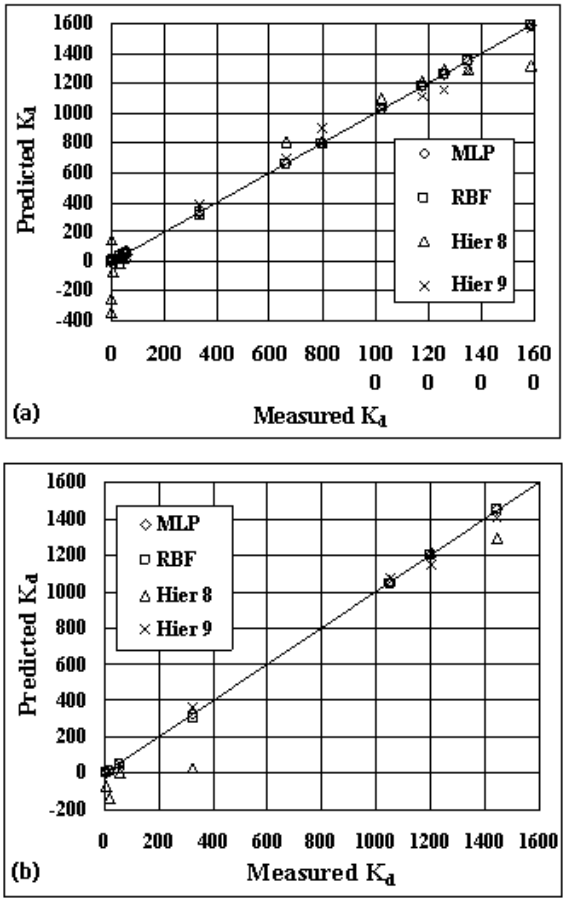

Figure 2. Comparison of measured and predicted K_d values for Cloudland soil. a) Training data, b) Testing data

3.1.1 Results for Cloudland Soil:

As shown in Table 3 and Figure 2, training the network using 16 training records was sufficient to achieve high accuracy. The highest coefficient of correlation (equal to 1) was obtained for training and testing sets of data by RBF and MLP networks. According to Figure 2, these two networks led to excellent predictions of K_d when compared to the measured values. Although RBF resulted in the lowest mean relative and standard errors (MRE = 0% and SE = 0) and the highest coefficient of correlation equal to 1) for the training data set, it was the MLP network which predicted the testing data set with the lowest error (SE = 0.71, MRE = 2.7) compared to the RBF network with SE = 7.39, MRE = 8.95 . HIER 8 predicted negative values for the testing data set, which means that the network failed. The HIER 9 network is the same as HIER 8, except that the natural logarithm of K_d was used as targets to train this network. For HIER 9, although good predictions were achieved for training the data set (Figure 2), error was much higher compared to the MLP and RBF networks. Better results may be obtained by increasing the number of epochs. It is clear that hierarchical networks in this case offer lower accuracy in prediction as is shown in Table 3.

3.1.2 Results for Kenoma Soil

The results shown in Table 3 and Figure 3 demonstrated that for Kenoma soil, training the network by nine records was sufficient to obtain high accuracy. As shown in Table 3, among the networks, MLP showed the best results with the lowest mean relative and standard errors (MRE = 1.42% and SE = 0.11) and highest coefficient of correlation (equal to 1) for the training data set. On the other hand, HIER 3 exhibited the lowest accuracy for the training data set. MLP and RBF networks resulted in the lowest SE (0.68 and 0.69) and the highest coefficient of correlation (0.997 and 0.997) for testing the data set. The RBF network had lower MRE values than MLP network. HIER 1 and HIER 2 networks achieved good predictions for the training data, but negative values were predicted for testing sets which means that the network fails.

Table 3. Statistical parameters for training and testing data

Soil Type	Network Type	Training Set of Data			Testing Set of Data		
		Correlation Factor	Standard Error	Mean Relative Error (%)	Correlation Factor	Standard Error	Mean Relative Error (%)
Kenoma	MLP	1	0.11	1.4	0.997	0.68	32.12
	RBF	0.997	0.83	6.31	0.997	0.69	22.7
	HIER 1	0.997	0.72	11.59	0.987	1.66	65.43
	HIER 2	0.997	0.78	13.74	0.984	1.88	71.6
	HIER 3	0.986	2.25	17.35	0.994	1.03	40.38
	HIER 4	0.995	1.21	8.94	0.997	0.74	22.9
Pacolet	MLP	1	1.22	16.55	1	0.76	2.03
	RBF	1	0.86	2.72	0.999	4.65	9.94
	HIER 6	0.939	96.1	37.16	0.984	39.89	17.5
	HIER 7	0.939	82.14	25.47	0.984	33.67	24.57
Cloudland	MLP	1	0.66	4.52	1	0.71	2.7
	RBF	1	0	0	0.999	7.39	8.95
	HIER 8	0.973	144.36	3765.6	0.987	115.62	419.93
	HIER 9	0.997	45.87	9.87	0.998	32.14	9.02

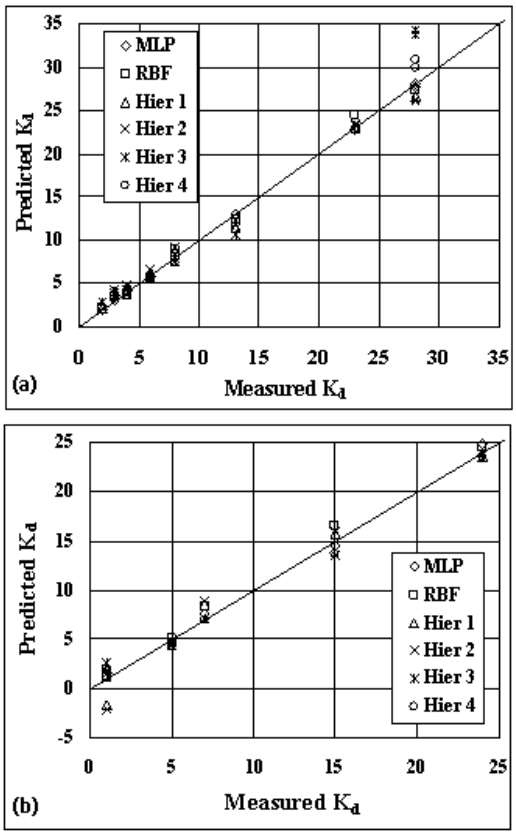

Figure 3. Comparison of Measured and Predicted K_d values for Kenoma Soil. a) Training set of data, b) Testing set of data

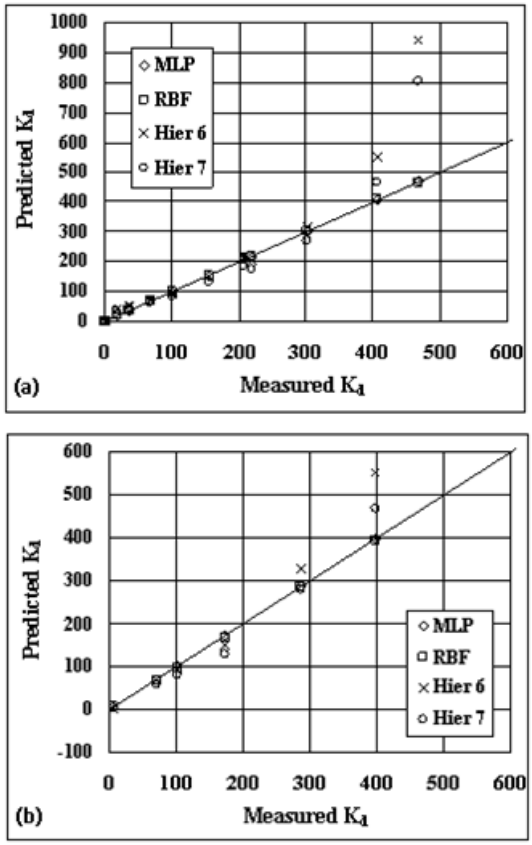

Figure 4. Comparison of Measured and Predicted K_d values for Pacolet Soil. a) Training set of data, b) Testing set of data

3.1.3 Results for Pacolet Soil

To obtain sufficiently accurate results (Table 1), 14 training records were used for training the networks for the Pacolet soil samples. Among the networks, MLP and RBF exhibited the best results with lower standard error (SE = 1.22 and 0.86) and a coefficient of correlation equal to 1 for the training data set. For the testing data these two methods showed coefficient of correlation values equal to 1 and 0.999 which indicated the accuracy of these methods in predicting K_d. It should be noted that although the MLP network has a high mean relative error (16.5%), the prediction results for training and testing sets of data is acceptable. This high mean relative error resulted from K_d=1 and 2 (ml/g) which have the highest relative difference between actual and training data. For HIER 6 and HIER 7 networks, good predictions were achieved for the training and testing data set, but were less precise relative to the results of MLP and RBF network models.

4 CONCLUSIONS AND FUTURE RECOMMENDATIONS

In this study we illustrated the use of artificial neural networks in predicting K_d values for chromium sorption in three different soils. The following conclusions can be drawn based on the results of this investigation:

1. As per statistical analysis, in all ANN models, correlation factors were not less than 0.984. This is an indicator that ANNs are powerful tools for predicting appropriate K_d values.
2. Among the networks tested in this study, MLP was selected as the best model with correlation factors of 0.997, 1 and 1 obtained for Kenoma, Pacolet and Cloudland soils respectively.
3. RBF Networks are very accurate for prediction of chromium partitioning coefficients of Kenoma, Pacolet and Cloudland soils.
4. It was found that HIER networks could not correctly predict small values of chromium distribution coefficients for Kenoma, Pacolet and Cloudland soils and application of this module sometimes resulted in negative values. Better results may be obtained by increasing the number of epochs, but it would require much longer training times (>10000 sec.).
5. Results obtained by ANN models indicated a better performance of networks for higher K_d values (generally for K_d values more than 10). Prediction of smaller K_d values were accompanied with higher MRE.

As this research has been conducted on just three kinds of soils, one specific metal and three ANN models, it is recommended to test more soil types, other metals in different environments and different ANN models, to find out if and how the results obtained in this study can be applied to those specific cases.

5 REFERENCES:

Alley, W. 1993. Regional Ground-Water Quality. Van Nostrand Reinhold, New York, NY, USA. 634 p.

Appelo, C. and Postma, D. 1996 C.A.J. Appelo and D. Postma, Geochemistry, Groundwater and Pollution, Balkema, Rotterdam. p. 536.

Baxter, C.W., Zhang, Q., Stanley, S.J., Shariff, R., Tupas, R-R.T and Shark, H.L. 2001. Drinking water quality and treatment: the use of artificial neural networks. Can. J. Civ. Eng. 28 (Suppl. 1): 26-35.

EPA, 2004. Understanding Variation in Partition Coefficient, K_d, Values; Volume III—Review of Geochemistry and Available K_d Values for Americium, Arsenic, Curium, Iodine, Neptunium, Radium, and Technetium [EPA 402-R-99-004C], http://www.epa.gov/radiation/cleanup/partition.htm).

Fatemi M. H., Karimian F. 2007. Prediction of micelle-water partition coefficient from the theoretical derived molecular descriptors. Journal of colloid and interface science pp:665-672

Gao, C. 1996. Predicting soil sorption coefficients of organic chemicals using a neural network model. Environ Toxicol Chem 15, pp. 1089–1096.

Gardner, M. W. and Dorling, S. R. 1998. “Artificial neural networks (The multilayer perceptron) - A review of applications in the atmospheric sciences.” Atmospheric Environment, 32(14/15), 2627-2636.

Hubick, K. T. 1992. Artificial neural networks in Australia. Department of Industry, Technology and Commerce, Commonwealth of Australia, Canberra.

Kaplan D.I. and Serne, R. J. 1995. Distribution Coefficient Values Describing Iodine, Neptunium, Selenium, Technetium, and Uranium Sorption to Hanford Sediments. PNL-10379, Pacific Northwest Laboratory, Richland, Washington.

Krupka, K.M., D.I. Kaplan, G. Whelan, R.J. Serne, and S.V. Mattigod. 1999. Understanding Variation in Partition Coefficient, K_d, Values -- Volume I: The K_d Model, Methods of Measurement, and Application of Chemical Reaction Codes. EPA 402-R-99-004A, U.S. Environmental Protection Agency, Office of Radiation and Indoor Air, Washington, DC.

Maier, H. R. and Dandy, G. C. 2000. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environmental Modelling & Software, 15(2000), 101-124.

Najjar, Y. M. and Ali, H. E. 1998. CPT-based liquefaction potential assessment: A neuronet approach. Geotechnical Special Publication, ASCE, 1, 542-553.

Rai, D., J. M. Zachara, L. E. Eary, C. C. Ainsworth, J. E. Amonette, C. E. Cowan, R. W. Szelmeczka, C. T. Resch, R. L. Schmidt, D. C. Girvin, and S. C. Smith. 1988. Chromium reactions in Geological Materials. EPRI-EA-5741. Electric Power Research Institute, Palo Alto, California.

Rongbo Guo, Xinmiao Liang, Jiping Chen, Qing Zhang, and Antonius Kettrup, 2002. Prediction of Soil Adsorption Coefficients from Retention Parameters on Three Reversed-Phase Liquid Chromatographic Columns, Analytical Chemistry, 74: 655-660.

Shahin, M.A., M.B. Jaksa, and Maier, H.R. 2001. Artificial neural network applications in geotechnical Engineering. Australian Geomechanics, Vol. 36, No. 1, pp. 49-62.

Sivakugan, N., Eckersley, J. D., and Li, H. 1998. Settlement predictions using neural networks. Australian Civil Engineering Transactions, CE40, 49-52.

Sposito, G. 1989. The Chemistry of Soils. Oxford University Press, New York, NY, USA. 277 p.

Ural, D. N., and Saka, H. 1998. Liquefaction assessment by neural networks. Electronic Journal of Geotechnical Engrg.

Vaziri, N. Hojabri, A. Erfani, A. Monsefi, M. Nilforooshan, B. 2007. Critical heat flux prediction by using radial basis function and multilayer perceptron neural network: a comparison study. Nucl. Eng. Des. 237, 377–385.

Yetilmezsoy, K. and Demirel S. 2008. Artificial neural network (ANN) approach for modeling of Pb(II) adsorption from aqueous solution by Antep pistachio (Pistacia Vera L.) shells J Hazard Mater 153:1288-300