ESTIMATION OF DISTRIBUTION COEFFICIENTS FOR CHROMIUM BY NEURAL NETWORK
Nader Shariatmadari, Assistant Professor, Dept. of Civil Eng. – Iran University of Science and Technology, Tehran, Iran
Amin Falamaki, Graduate student, Dept. of Civil Eng. - IUST, and Faculty of Yasuj Azad University

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. 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 empirical predictor equations may be derived commonly by statistical analysis and take the form of a linear or nonlinear polynomial expression. In recent times, artificial neural net works (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 (or distribution) coefficient, $K_d$, with variation of environmental components. The objective is to investigate the feasibility of ANN technique for predicting of $K_d$ variation with variation of environmental components for a certain soil. To accomplish this object the database reported by Rai et. al. (1988) for chromium adsorption were used. Results show that ANNs are powerful tools for prediction of partition coefficient variation with variation of environmental components for a certain soil.

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 in the solid material to the contaminant concentration in the surrounding aqueous solution when the system is at equilibrium (EPA [2]). 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 empirical predictor equations may be derived commonly by statistical analysis and take the form of a linear or nonlinear polynomial expression. In recent times, artificial neural net works (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 (or distribution) coefficient, $K_d$, with variation of environmental components. The objective is to investigate the feasibility of ANN technique for predicting of $K_d$ variation with variation of environmental components for a certain soil. To accomplish this object the database reported by Rai et. al. (1988) for chromium adsorption were used. Results show that ANNs are powerful tools for prediction of partition coefficient variation with variation of environmental components for a certain soil.

2. CONSTANT PARTITION COEFFICIENT MODEL.
The constant partition coefficient, $K_d$, is defined as the ratio of the quantity of the adsorbate (i.e., metal or radionuclide) adsorbed per unit mass of solid to the quantity of the adsorbate remaining in solution at equilibrium. 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.

3. THE PARAMETRIC PARTITION COEFFICIENT MODEL.
Another practical conceptual model for adsorption is called the parametric $K_d$ model. in this 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. Most 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 Rhoads et al.[5] as:

\[
K_d (ml/g) = 9550 \times C^{-0.335}
\]  

(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. Shariatmadari et al. (2004) used ANNs to predict the variation of the partition (or distribution) coefficient, \( K_d \), of cadmium, with variation of environmental components of soil. The objective was to investigate the feasibility of ANN technique for predicting of \( K_d \) variation with variation of environmental \( pH \). Results showed that ANNs are more powerful tools than statistical analysis for prediction of partition coefficient variation with variation of environmental components for a certain soil. In this study the ANNs is used to predict the variation of the partition (or distribution) coefficient, \( K_d \), with variation of environmental components. The objective is to investigate the feasibility of ANN technique for predicting of \( K_d \) variation with variation of environmental components for different soils. The database reported by Rai et al.(1988), for chromium adsorption were used. The review of chromium \( K_d \) data obtained for a number of soils reported by Rai et al.(1988).indicated that a number of factors influence the adsorption behavior of chromium. These factors and their effects on chromium adsorption on soils and sediments were used as the basis for this study.

Table 1. Relations Between \( K_d \) Values And Environmental Condition In Soils

<table>
<thead>
<tr>
<th>RELATION SHIP</th>
<th>Species</th>
<th>3.1.1 Reference</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K_d = -0.54 + 0.45(pH) ).</td>
<td>Cadmium</td>
<td>EPA 1999</td>
<td>Different soils</td>
</tr>
<tr>
<td>( \log (K_d) = 1.2 + 1.0 \log (CEC) )</td>
<td>Cesium</td>
<td>Akiba and Hashimoto (1990)</td>
<td>A large number of soils, minerals, and rock materials</td>
</tr>
<tr>
<td>( K_d (ml/g) = 1639 - 902.4(pH) + 150.4(pH)^2 )</td>
<td>Lead</td>
<td>Gerritse et al. (1982)</td>
<td>Different soils</td>
</tr>
<tr>
<td>( K_d = 284.6(DCARB) + 27.8(CLAY) - 594.2 )</td>
<td>Plutonium</td>
<td>EPA 1999</td>
<td>Different soils</td>
</tr>
<tr>
<td>( K_d = 488.3(DCARB) + 29.9(CLAY) - 119.1(pH) - 356.8(EC) )</td>
<td>Plutonium</td>
<td>EPA 1999</td>
<td>Different soils</td>
</tr>
<tr>
<td>( K_d = 25.7(DCARB) + 12.14 CLAY) + 2.41 )</td>
<td>Plutonium</td>
<td>EPA 1999</td>
<td>Different soils</td>
</tr>
<tr>
<td>( K_d &gt; 767.5 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( K_d = 286.0(DCARB) + 21.3(CLAY) - 81.2 )</td>
<td>Plutonium</td>
<td>EPA 1999</td>
<td>Different soils</td>
</tr>
<tr>
<td>( K_d &gt; 767.5 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \log (K_d) = -0.13 + 0.69(pH) ).</td>
<td>Thorium</td>
<td>EPA 1999</td>
<td>The pH range of 4 to 8</td>
</tr>
</tbody>
</table>

DCARB = The Concentrations Of Dissolved Carbonate Of Soils  
CLAY = The Clay Content Of Soils  
EC = Electrical Conductivity
4. 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 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 preprocess 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.

5. DATABASE

According to the US EPA [2], report the data set used to develop ANNs is from the adsorption data collected by Rai et al. (1988). They used four well-characterized soils and obtained soil samples from subsurface horizons to calculate the Kd values for Cr(VI) as a function of pH. All 4 soil samples were and characterized as to their pH, texture, CEC, organic and inorganic carbon contents, surface areas, extractable (hydroxylamine hydrochloride, and DCB) iron, manganese, aluminum, and silica, KOH extractable aluminum and silica, and clay. Table A1 and A2 as shown in Appendix I shows this database used for training and simulating networks respectively.

6. TRAINING AND TESTING OF 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 1994). Here 30 patterns were used for testing. Designing a 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. Various BP networks were experimented with one or two hidden layers and different numbers of neurons in each hidden layer. Table 2a shows the summery of MLP network specifications used in this research. We will show that a networks with four and two neurons in hidden layers and one neuron in output layer have shown good agreement to the training patterns (Fig.2). The activation functions for hidden layer neurons are tangent hyperbolic function and for output layer is linear function. The transformation from the input space to the hidden space is nonlinear. However, the transformation from the hidden space to the output space is linear.

![Figure 2. 3x4x2x1 MLP Network](image)

7. RESULTS AND DISCUSSION

Table 2 shows the summery of the MLP network specifications in this study.
6.1. RESULTS FOR KENOMA SOIL

For this soil, the network was trained by nine and tested for six input data. Results are shown in Table 3 for training and testing sets of data. A comparison between the measured partition factor and the predicted values for testing set of data are shown in Fig. 3.

The results show that training of network by nine training records is sufficient to predict with high accuracy. High coefficient of correlation (CF or $r^2$) is obtained, as shown in Table 3 for training and testing sets of data. So, MLP shows good results with low mean relative error (MRE) for training data set and equal to 1.42% and a high coefficient of correlation.

6.2. RESULTS FOR PACOLET SOIL

For this soil, the network was trained by 14 and tested for six input data. Results are shown in Table 4 for training and testing sets of data. A comparison between the measured partition factor and the predicted values for testing set of data are shown in Fig. 4.

The results show that training of network by 14 training records is sufficient to predict with high accuracy. High coefficient of correlation ($r^2$) is obtained, as shown in Table 4 for training and testing sets of data. It should be noted that although the MLP network has a high mean relative error (16.54%), but the prediction results for training and testing sets of data are acceptable. This high mean relative error is generating the errors of $K_d$ lower than 2 ml/g and according to Fig. 4 this network predicts excellent.

6.3. RESULTS FOR CLOUDLAND SOIL

For this soil, the network was trained by 16 and tested for seven input data. Results are shown in Table 5 for training and testing sets of data respectively. A comparison between the measured partition factor and the predicted values for testing set of data are shown in Fig. 5.

The results show that training of network by 16 training records is sufficient to predict with high accuracy. High coefficient of correlation ($r^2$) is obtained, as shown in Table 5 for training and testing sets of data. MLP network shows good results with a low mean relative error for training data set and equal to 4.53%, and a high coefficient of correlation. According to Fig. 8 this networks predict excellent.

6.4. RESULTS FOR OCALA SOIL

For this soil, the network was trained by 13 and tested for five input data. Results are shown in Tables 6a and 6b for training and testing sets of data respectively. A comparison between the measured partition factor and the predicted values for testing set of data is shown in Fig. 6. The results show that for MLP network, training of network by 13 training records is sufficient to predict with high accuracy. High coefficient of correlation ($r^2$) is obtained, as shown in Table 6 for testing set of data.
Predicted and Measured $K_d$ (ml/g) for Kenoma Soil

Measured data reported by Rai et al. 1988

Predicted and Measured $K_d$ (ml/g) for Pacolet Soil

Measured data reported by Rai et al. 1988

Predicted and Measured $K_d$ (ml/g) for Cloudland Soil

Measured data reported by Rai et al. 1988

Table 5. Measured and Predicted $K_d$ by MLP for Cloudland Soil

<table>
<thead>
<tr>
<th>Training Data</th>
<th>$K_d$</th>
<th>1</th>
<th>5</th>
<th>1</th>
<th>3</th>
<th>2</th>
<th>33</th>
<th>42</th>
<th>58</th>
<th>66</th>
<th>794.0</th>
<th>1023</th>
<th>1175</th>
<th>1349</th>
<th>1259</th>
<th>1585</th>
<th>CF</th>
<th>MRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP Prediction</td>
<td>1.3</td>
<td>4.2</td>
<td>1.1</td>
<td>3.1</td>
<td>2.2</td>
<td>33.3</td>
<td>42.3</td>
<td>58.7</td>
<td>331.6</td>
<td>659.9</td>
<td>794.8</td>
<td>1023.6</td>
<td>1175.9</td>
<td>1350.3</td>
<td>1258.5</td>
<td>1584.5</td>
<td>0.999</td>
<td>4.53</td>
</tr>
<tr>
<td>Testing Data</td>
<td>$K_d$</td>
<td>4</td>
<td>18</td>
<td>56</td>
<td>324</td>
<td>1047</td>
<td>1202</td>
<td>1445</td>
<td>CF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP Prediction</td>
<td>3.4</td>
<td>17.8</td>
<td>57.5</td>
<td>324.4</td>
<td>1048.4</td>
<td>1203.9</td>
<td>1447.2</td>
<td>0.999</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CF = Correlation Factor
MRE = Mean Relative Error

Table 6. Measured and Predicted $K_d$ by MLP for Ocala Soil

<table>
<thead>
<tr>
<th>Training Data</th>
<th>$K_d$</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>CF</th>
<th>MRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP Prediction</td>
<td>0.042</td>
<td>0.976</td>
<td>0.03</td>
<td>0.032</td>
<td>0.998</td>
<td>0.999</td>
<td>0.997</td>
<td>0.006</td>
<td>-0.002</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.023</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing Data</td>
<td>$K_d$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>CF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP Prediction</td>
<td>0.02</td>
<td>0.96</td>
<td>0.02</td>
<td>0.99</td>
<td>0.998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CF = Correlation Factor
MRE = Mean Relative Error

Figure 3. Testing Set of Data for Kenoma soil

Figure 4. Testing Set of Data for Pacolet soil

Figure 5. Testing Set of Data for Cloudland soil
8. CONCLUSIONS
ANNs is a powerful tool for prediction of partition coefficient variation with variation of environmental components for a certain soil. MLP Networks for prediction chromium partition coefficients of Kenoma, Pacolet, Cloudland and Ocala Soils are very accurate. The mean relative error for training data set of Kenoma, Pacolet, Cloudland and Ocala are 1.42, 16.54, 4.53 and 0.24 respectively. Although the MLP network has a high mean relative error (16.54%) for Pacolet Soil, but the prediction results for training and testing sets of data is acceptable. This high mean relative error is generating the errors of $K_d$ lower than 2 ml/g and according to Fig. 4 this network predicts excellent. High coefficient of correlation were obtained for both testing and training data sets of all four soils.

9. REFERENCES