Estimating Groundwater Pollution Source Location from Observed Breakthrough Curves Using Neural Networks

Jitendra Kumar[1], Ashu Jain[2] and Rajesh Srivastava[3]

[1] M. Tech. Student, [2] Assistant Professor, [3]Associate Professor
Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, India
{jitendra, ashujain, rajeshs}@iitk.ac.in

Abstract. This paper presents the results of a study aimed at estimating groundwater pollution source location from observed breakthrough curves using neural networks. Two different methods of presenting the breakthrough curves to the ANN are investigated. The feed-forward multi-layer perceptron (MLP) type artificial neural network (ANN) models are employed. The ANNs were trained using the back-propagation training algorithm on simulated data. A new approach for ANN training using back-propagation is employed that considers two different error statistics to prevent over-training or under-training of the ANNs. The preliminary results indicate that the ANNs are very efficient tools for estimating the distance of the potential pollution source from the observation well where breakthrough curve is measured.

1 Introduction

Rapid industrialization, increased use of pesticides, and tremendous increase in the number of underground petrol storage tanks in recent years have put the groundwater at a high risk of being contaminated by harmful chemicals. Once an aquifer is contaminated, it may take a long time and considerable expenditure to restore it to an acceptable state. Due to the huge financial implication of the clean-up, it becomes imperative to identify the source of the pollution so that suitable penalty could be imposed on the concerned industry/individual/agency. Complete identification of pollution source involves determination of source concentrations, duration, and location. The physical processes involved in the movement of water and contaminants in the aquifers are highly complex, non-linear, and dynamic processes affected by a wide range of physical variables. The identification of the pollution sources is much more complex in the sense that it requires an inverse modeling of the flow and contaminant transport. Over the past few decades, various investigators have looked at the problem of identification of groundwater pollution sources using a wide variety of techniques.

Bhanu Prasad (Editor): IICAI-05, pp. 1004-1017, 2005.

The simplest approach is to use forward simulations with assumed source location and release history and compare the solutions with the observed data. However it is not very efficient due to the infinite number of possible combinations and some type of optimization method has to be used to obtain the *best* solution. Probably the earliest such study was that of Gorelick et al. [1] who used forward-time simulations with an optimization model based on linear programming and multiple regressions. They incorporated the transport model as constraints in the form of a response matrix. Wagner [2] considered an inverse model as a non-linear maximum likelihood estimation problem for simultaneous model parameter estimation and source characterization. Mahar and Datta [3, 4] combined the identification of a pollutant source with the optimal design of a monitoring network for an efficient identification process. Mahar and Datta [4] used a classical nonlinear optimization technique to estimate the magnitude, location and duration of groundwater pollution sources under transient conditions. A different approach was proposed by Skaggs and Kabala [5]. They attempted to reconstruct the history of the plume using Tikhonov Regularization (TR). 1-D solute transport through a saturated homogeneous medium was studied with a complex contaminant release history and assuming no prior knowledge of the release function. Samarskaia [6] applied the TR with fast Fourier transforms to a groundwater contamination source reconstruction problem. Liu and Ball [7] used modified TR technique to study a contaminant release at Dover Air Force Base, Delaware. They used field measured concentration profiles in low-permeability porous media that underlie a contaminated aquifer at the Dover Air Force Base. Singh et al. [8] used Artificial neural networks (ANNs) for identification of unknown groundwater pollution sources. The ANN was trained to identify source characteristics based on simulated contaminant concentration measurement data at specified observation locations in the aquifer. The performance of ANN models was found to be very effective for source identification. Singh and Datta [9] utilized a trained ANN to simultaneously solve the problems of estimating unknown groundwater pollution sources and estimating unknown hydro-geologic parameters.

It is clear that a variety of techniques have been investigated by researchers for pollution source identification. Recently, ANNs have also been employed for this purpose; however, most of the studies reported earlier have focused on identifying the source release history at potential locations. In real aquifers, identifying the location of the pollution source is extremely important for taking punitive measures. The objective of the present study is to investigate the use of ANN methodology to estimate the distance of the pollution source from an observed well where the measured concentration of the pollutant as a function of time (breakthrough curve) is available. The back-propagation training algorithm can be inefficient due to the use of sum square error (SSE) in determining the optimum level of training to ensure against under-training or over-training. In this study, we propose to use other statistics (e.g. average absolute relative error) during training of the ANN using back-propagation to determine the optimum level of training. The paper begins with a brief overview of the ANN technique followed by the model development before presenting the results and making concluding remarks.

2 Artificial Neural Networks

An Artificial Neural Network (ANN) is a densely interconnected network of several independent adaptive processing units called neurons. Similar neurons are arranged in one layer. Each neuron in a layer is connected with the neurons in the adjacent layer with what is known as the "connection strengths" or weights. The ANN models are specified by the network topology, neuron characteristics, and training/learning rules. A typical ANN employed in engineering applications consists of three layers, namely, input layer, hidden layer, and output layer (see Figure 1). The input data are presented to the ANN at the input layer, which are processed in a forward direction through the hidden layer(s), and the output from the ANN is computed at the output layer. This whole process is known as 'feed-forward mechanism'. The computed output at the output layer is compared with the known output and error is calculated. This error is then propagated backwards through the ANN and connection strengths are updated using generalized delta rule. This process of feed-forward mechanism and back propagation of errors is repeated until convergence is achieved. This whole process is known as the "training" of the network and this type of network is known as a "back propagation ANN". The most popular training method is called the generalized delta rule that is based on gradients descent [10]. A three layer feed-forward ANN with back-propagation training method [10] was employed in this study.

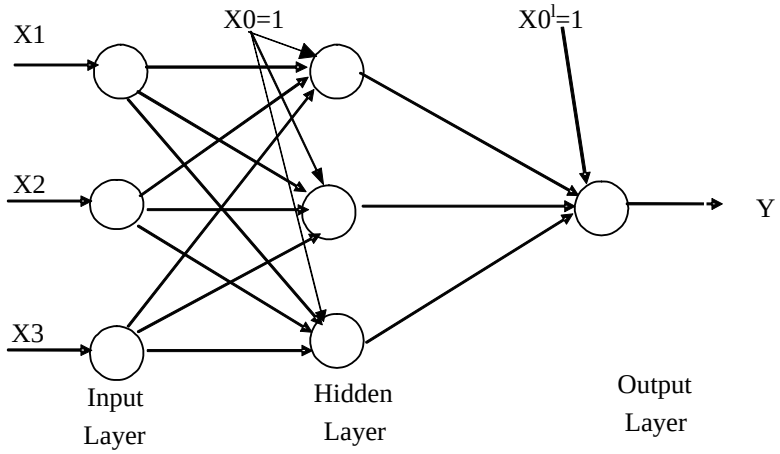

Figure 1: Structure of a 3- layer feed-forward ANN

3 Model Development

The ANN model development involves the following steps: (i) selection of data set for training and testing of the model, (ii) identification of the input vector, (iii) normalization (scaling) of the data, (iv) selection of the network architecture, (v) determining the optimum number of neurons in the hidden layer, (vi) training of the ANN models, and (vii) testing of ANN model using the selected performance evaluation statistics.

3.1 Data generation

The first step in the methodology would be to generate the training and testing patterns. The data for the problem of pollution source identification consist of observed breakthrough curve at observation wells and the source strengths, duration, and location. Since in the present study, the objective is to estimate the distance of the pollution source given a breakthrough curve, the pollution source is assumed to be active for constant duration injecting a conservative pollutant at a constant rate. The breakthrough curves at different distances were then calculated using the governing equations for the pollutant transport as follows:

The one-dimensional transport of conservative solutes through a homogeneous saturated semi-infinite porous media is represented by the advection-dispersion equation:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} - v\frac{\partial C}{\partial x} \qquad (1)$$

in which C is the concentration, t is time, D is the dispersion coefficient, x is the distance and v is the groundwater velocity. Its solution requires one initial and two boundary conditions which are dictated by the type of problem considered. In this study, we consider an initially uncontaminated aquifer with a pollutant source of constant strength releasing the pollutant till certain time and stopping after that. The initial and boundary conditions are then given by

$$\begin{aligned} &C(x,0) = 0 \\ &C(0,t) = C_0 \text{ for } t \le T_0 \text{ and 0 otherwise} \\ &C(\infty,t) = 0 \end{aligned} \qquad (2)$$

in which T_0 is the duration of release and C_0 is concentration at source. The solution of the above equations is obtained by utilizing the solution to a step input according to Ogata and Banks [11] as:

$$C = \frac{C_0}{2}\left[erfc\left(\frac{x - vt}{\sqrt{4Dt}}\right) + e^{\frac{vx}{D}} erfc\left(\frac{x + vt}{\sqrt{4Dt}}\right)\right] \quad \text{for } t < T_0 \qquad (3)$$

$$C = \frac{C_0}{2}\left[erfc\left(\frac{x-vt}{\sqrt{4Dt}}\right) + e^{\frac{vx}{D}} erfc\left(\frac{x+vt}{\sqrt{4Dt}}\right)\right] - \frac{C_0}{2}\left[erfc\left(\frac{x-v(t-T_0)}{\sqrt{4D(t-T_0)}}\right) + e^{\frac{vx}{D}} erfc\left(\frac{x+v(t-T_0)}{\sqrt{4D(t-T_0)}}\right)\right]$$

for $t > T_0$ (4)

Equations (3) and (4) were employed to generate the breakthrough curves using the following data: v = 0.1 m/day; D = 0.1 m^2/day; C_0 = 1000 mg/l; and T_0 = 120 days. The time interval of the breakthrough curves was taken as 30-days. A total of 2,000 input-output patterns were generated, of which 1,500 were used for training, 250 were used for validation, and 250 were used for testing. The data were scaled in the range of 0.1 and 0.9. A wide variety of standard performance statistics were employed to evaluate the performance of various ANN models.

3.2 Performance evaluation statistics

Six different standard performance statistics were employed for model development. These are normalized root mean square error (NRMSE), Nash-Sutcliffe efficiency (E), coefficient of correlation (R), average absolute relative error (AARE), threshold statistics (TS), and sum square error (SSE). The equations to calculate the first five are given below:

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum (XO - XE)^2}}{\frac{1}{N}\sum XO} \quad (5)$$

$$E = 1 - \frac{\sum (XE - XO)^2}{\sum (XO - \overline{XO})^2} \quad (6)$$

$$R = \frac{\sum (XO - \overline{XO}) \times (XE - \overline{XE})}{\sqrt{\sum (XO - \overline{XO})^2 \sum (XE - \overline{XE})^2}} \quad (7)$$

$$AARE = \frac{1}{N}\sum \left|\frac{XE - XO}{XO}\right| \times 100\% \quad (8)$$

$$TS_x = \frac{N_x}{N} \quad (9)$$

Where XO is the observed value of the variable, XE is the estimated value of the variable from a model, $\overline{XO}$ is the average observed value of the variable, $\overline{XE}$ is the average estimated value of the variable, N_x is the number of data points estimated for which the absolute relative error (ARE) is less than x%, N is the total number of data points predicted, and all the summations run from 1 to N. The value of x of 1%, 10%, and, 50% were considered in this study to compute threshold statistics.

3.3 ANN model development

The ANN models developed in this study consisted of three layers: an input layer, a hidden layer, and an output layer. The input vector represents the complete breakthrough curves. Two different ANN models were developed that differed in the manner of presenting the breakthrough curve to the input layer of the ANN. The first method consists of presenting concentration data at 30-day interval irrespective of the magnitude of the concentration. Each breakthrough curve consisted of a total of 73 ordinates, therefore the first model (called ANN-1 Model in this paper) had an architecture of 73-N-1. In the second method (called ANN-2 Model), only those measurements were included for which the concentration exceeded 0.001. The breakthrough curve thus obtained was divided into ten parts, and the pollutant concentrations at the eleven end points were computed. This time distribution of the pollutant concentrations was then modeled as 22-N-1 in ANN-2 model with the input layer including the 11 times and corresponding concentrations. The output neuron in both the ANN models represented the distance of the pollution source from the location where the breakthrough curve was observed.

The next step in the development of the ANN model is the determination of the optimum number of neurons (N) in the hidden layer. The number of neurons in the hidden layer is, in fact, responsible for capturing (or mapping) the dynamic and complex relationship among various input and output variables considered. The sigmoid activation function was used as the transfer function at both hidden and output layers. This study employed the popular back-propagation training algorithm using step-wise learning with momentum factor. The value of learning coefficient of 0.075 and momentum correction factor of 0.075 was used while training. The value of N was varied from 1 to 20 and for each N, the back-propagation algorithm was used to minimize SSE at the output layer. Each of the ANN architectures was trained for a maximum of 50,000 iterations or when the SSE reached 0.0005. It was observed that none of the ANN architecture attained an SSE of 0.0005, which was probably due to the fact that the acceptable value of SSE chosen was very restrictive. Although the training data set is normally divided into two parts: training and validation subsets to prevent over-training or under-training of the networks, it may not be helpful always since the validation data set may not show the optimum level. Therefore, a combination of AARE and SSE was used to determine the optimal level of training. It was observed that the graph between SSE and the number of iterations showed a smooth decline in SSE as a function of the number of iterations during both training and validation data sets. However, the value of AARE fluctuated as a function of the number of iterations. A plot of the SSE and AARE during both training and validation data sets from the selected ANN architectures from ANN-1 Model are shown in Figure 2 through Figure 4.

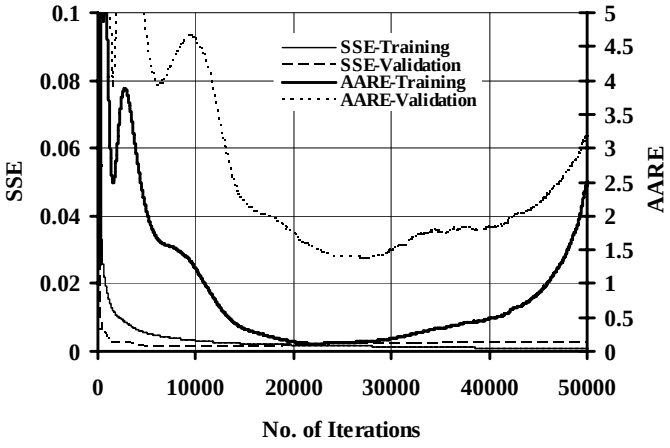

Figure 2: SSE & AARE v/s no. of iterations from 73-5-1 model

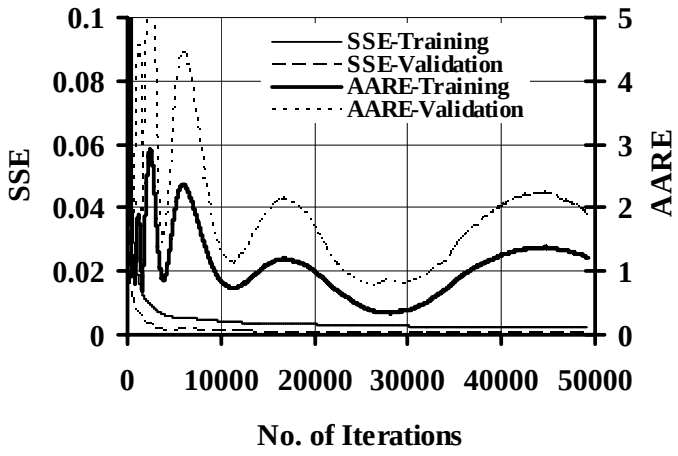

Figure 3: SSE & AARE v/s no. of iterations from 73-8-1 model

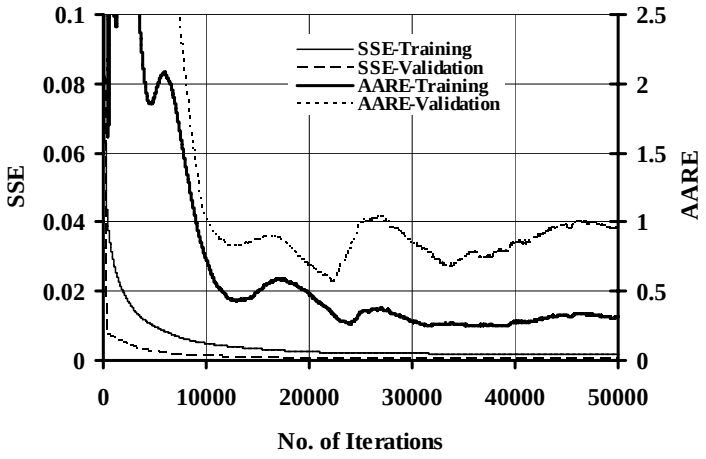

Figure 4: SSE & AARE v/s no. of iterations from 73-12-1 model

It is clear from these figures that it is difficult to determine the optimum level of training to ensure against over-training or under-training using SSE alone since the SSE v/s iterations curves during both training and validation continue to decrease with the number of iterations. However, this problem can be solved by examining the AARE that shows a definite dip in the AARE v/s iterations curve. Therefore, it is possible to stop the training when the AARE level is minimum during training and/or validation data sets. This procedure was employed in the present study to determine the optimal level of training for each of the ANN architectures investigated for both the methods.

Further, three error statistics, namely, R, sum square error (SSE), and AARE were used to determine the best ANN architecture (or optimal N). Figure 5 and Figure 6 show the graphs between the number of hidden neurons and different error statistics from ANN-1 and ANN-2 models, respectively.

For ANN-1 model, it can be seen from Figure 5 that the ANN architectures with 5, 9, 12, 15, and 18 hidden neurons have minimum SSE and AARE. The coefficient of correlation R is almost constant from all ANN architectures. Therefore, ANN architectures of 73-5-1, 73-9-1, 73-12-1, 73-15-1 and 73-18-1 were selected for further consideration. Similarly, for the ANN-2 model (see Figure 6), the ANN architectures having 2, 5, 7, and 11 hidden neurons have minimum AARE, and SSE is almost constant. Thus, ANN architectures of 22-2-1, 22-5-1, 22-7-1, and 22-11-1 were selected for further consideration. The results in terms of various performance statistics from the selected models during training, validation, and testing are presented in Table 1, and Table 2 from ANN-1 and ANN-2 models, respectively.

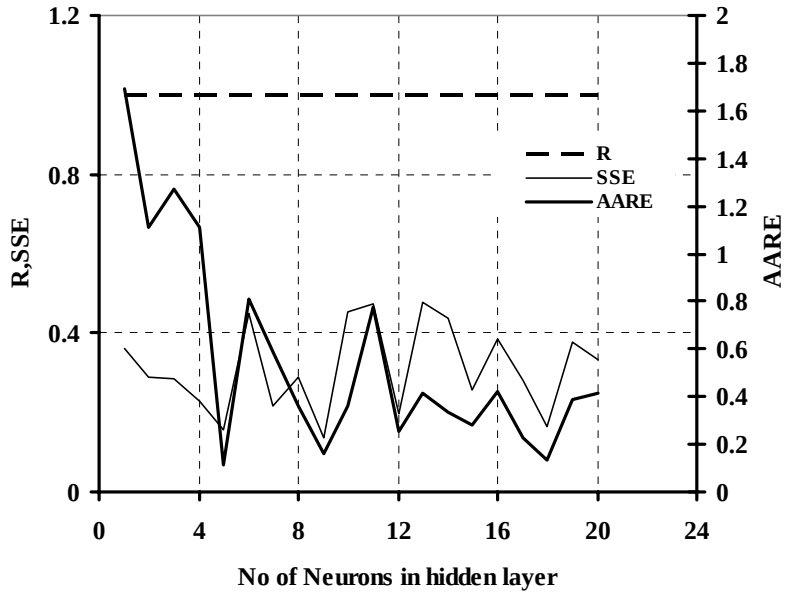

Figure 5: Error Statistics v/s no. of hidden neurons for ANN-1 model (73-N-1)

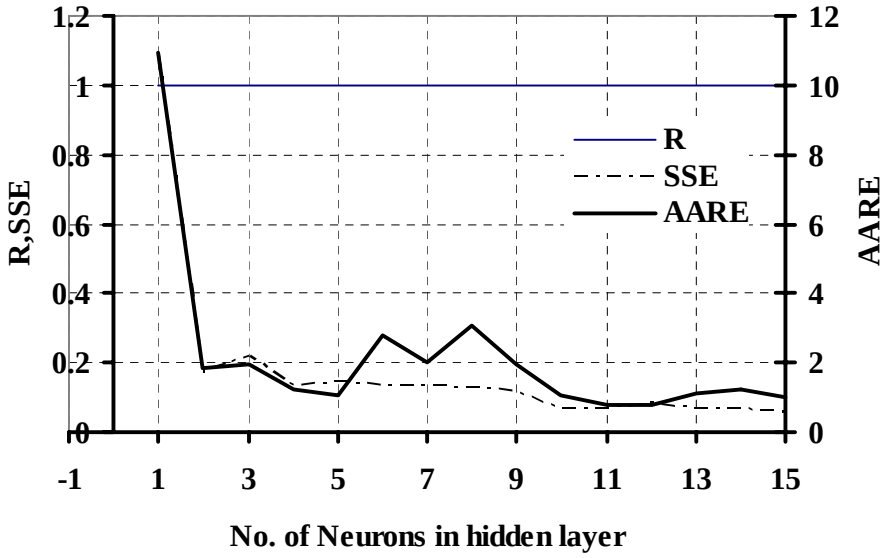

Figure 6: Error Statistics v/s no. of hidden neurons for ANN-2 model (22-N-1)

4 Results and Discussion

Analyzing the results from Table 1, it can be noted that the values of correlation coefficient and Nash-Sutcliffe efficiency in excess of 0.99 were obtained from all the models at the optimum number of iterations (mentioned in the Table 1 for each of the selected architectures). The values of R in excess of 0.97, and E in excess of 0.95 during both validation and testing from all the models represent excellent performance of ANN-1 model. The values of AARE of less than 0.3% from all models during training; less than 2% from all models (except 73-09-1) during validation; and less than 1.5% from all models (except 73-09-1) model during testing data set shows that the ANN-1 model is very efficient in predicting the distance of the pollution source from the location of the observed breakthrough curve very accurately. Further, analyzing the results in terms of various error statistics during training and validation, it can be seen that the 73-12-01 can be selected as the most suitable model for estimating the pollution source location.

Analyzing the performance of ANN-2 model from Table 2, it can be noted that the performance of all the models is excellent in terms of all the error statistics. All the selected models obtained R and E values in excess of 0.95 and AARE values in single digits, which can be characterized as very good. Based on all the results in Table 2, 22-11-1 model is found to be the best among this category. It obtained AARE values of 0.79%, 2.71%, and 2.38% during training, validation, and testing, respectively, which is the least as compared to all other models. The performance of the 22-11-1 model was very good in terms of NRMSE, TS, and SSE statistics also.

Comparing the performance of the two different methodologies from Table 1 and Table 2, it can be noted that although the performances of both the methodologies can be characterized as excellent but the first method performed slightly better, as expected. However, it must be emphasized that the second method involves far simpler ANN architecture. Comparing the performances of the best models based on the two methodologies (73-12-01 and 22-11-1), it can be observed that the performance of 73-12-1 ANN model is only marginally better than that of the 22-11-1 ANN model. Therefore, it can said that the 22-11-1 ANN model is the best model based on the principle of parsimony among all the models investigated in this study for the purpose of groundwater pollution source location estimation using breakthrough curves.

Table 1: Performance statistics from ANN-1 models

Model	Iterations	NRMSE	E	R	AARE	TS1	TS10	TS50	SSE
During Training									
73-05-1	26600	0.0022	0.999985	0.999992	0.11	99.27	99.93	100.0	0.002
73-09-1	23800	0.0019	0.999989	0.999995	0.16	99.20	99.93	99.93	0.001
73-12-1	31700	0.0023	0.999984	0.999992	0.25	96.60	99.67	100.0	0.002
73-15-1	12700	0.0026	0.999979	0.999991	0.28	94.47	99.80	100.0	0.003
73-18-1	38100	0.0021	0.999987	0.999994	0.13	98.67	99.87	100.0	0.002
During Validation									
73-05-1	----	0.1264	0.952106	0.976138	1.97	71.20	98.40	99.20	0.002
73-09-1	----	0.1266	0.951955	0.976066	3.81	56.40	96.40	99.20	0.005
73-12-1	----	0.1263	0.952171	0.976184	1.18	89.20	98.40	99.60	4E-04
73-15-1	----	0.1263	0.952155	0.976167	1.15	85.20	98.40	99.60	7E-04
73-18-1	----	0.1263	0.952165	0.976179	1.14	88.00	98.80	99.60	6E-04
During Testing									
73-05-1	----	0.1261	0.951907	0.976045	1.49	70.00	98.40	99.60	0.002
73-09-1	----	0.1263	0.951759	0.975970	2.72	54.80	96.40	99.20	0.005
73-12-1	----	0.1260	0.951962	0.976090	1.16	87.60	98.40	99.60	8E-04
73-15-1	----	0.1260	0.951959	0.976069	1.14	84.80	99.20	99.60	8E-04
73-18-1	----	0.1260	0.951971	0.976084	1.06	89.60	98.80	99.60	6E-04

Table 2: Performance statistics from ANN-2 models

Model	Iterations	NRMSE	E	R	AARE	TS1	TS10	TS50	SSE
During Training									
22-02-1	36800	0.0068	0.999863	0.999934	1.83	76.73	97.73	99.60	0.017
22-05-1	15000	0.0062	0.999883	0.999942	1.07	87.93	98.20	99.87	0.015
22-07-1	12300	0.0060	0.999893	0.999948	1.99	87.27	97.60	99.60	0.013
22-11-1	18500	0.0042	0.999947	0.999973	0.79	90.67	99.00	99.93	0.007
During Validation									
22-02-1	----	0.1348	0.951198	0.973506	5.31	12.40	96.40	99.20	0.140
22-05-1	----	0.1269	0.951687	0.975945	3.43	52.80	93.60	99.60	0.011
22-07-1	----	0.1267	0.951829	0.975935	4.87	54.00	96.40	98.80	0.008
22-11-1	----	0.1269	0.951679	0.975923	2.71	38.40	96.00	99.60	0.011
During Testing									
22-02-1	----	0.1348	0.945031	0.973350	4.53	12.80	96.00	99.20	0.145
22-05-1	----	0.1267	0.951488	0.975846	3.29	51.60	94.80	99.60	0.011
22-07-1	----	0.1265	0.951627	0.975813	3.81	54.80	96.00	98.00	0.008
22-11-1	----	0.1267	0.951488	0.975823	2.38	39.60	97.20	99.60	0.011

The results in the form of a scatter plot from 22-11-1 ANN model are presented in Figure 7. The narrow and uniform spread around the ideal line indicates that it was able to predict the distance of the pollution source very accurately for all magnitudes.

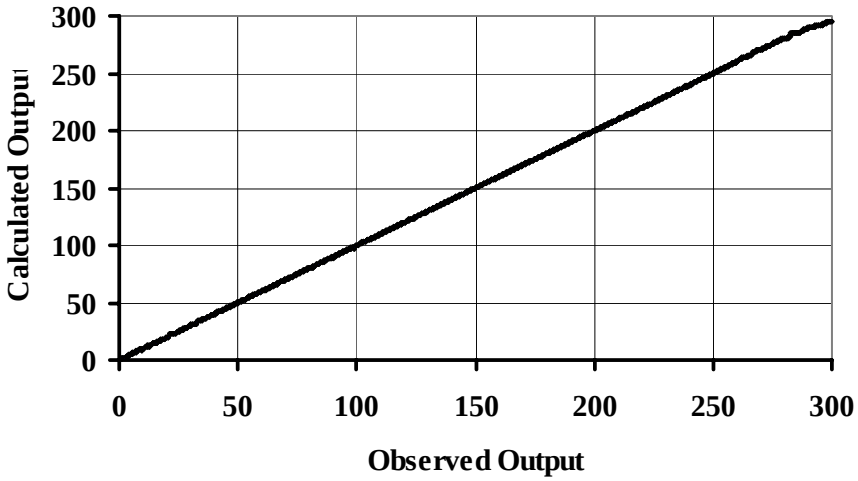

Figure 7: Observed and calculated distances from 22-11-1 ANN model

5 Summary and Conclusions

This paper presents the preliminary findings of a study aimed at estimating ground-water pollution source location using the pollution measurement data in the form of breakthrough curves. The feed-forward MLP type of ANN architecture was em-ployed to develop to develop various ANN models trained using back-propagation method. The data for ANN model development were generated using the analytical solution of the problem of one-dimensional steady flow and transient contaminant transport in homogeneous aquifer. The training data set was divided into training and validation to prevent over-training and/or under-training. In addition, two methods of presenting the breakthrough curves to the ANN input were investigated. A wide variety of standard performance statistics were used to evaluate the performance of various ANN models.

The preliminary results obtained in this study demonstrate that the ANNs can be very efficient tools of solving the complex problem of inverse modeling for pollution source identification. It has been found that the second method involving presenting only eleven concentration values and respective times to the ANN is sufficient to capture the complex inverse problem of source location identification. It has been found that using AARE in addition to SSE during ANN training can help determine optimum level of training in certain situations; however, it needs to be investigated

further. The limitation of the study presented has been that perfect data obtained from the analytical solution of the groundwater flow and transport problem were employed for ANN model development. In reality, the breakthrough curves obtained from aquifers contain many type of errors e.g. measurement errors etc. How the ANN models will be able to perform when presented with the noisy data remains to be investigated. It is hoped that further research efforts will focus in some of these directions.

References

1. Gorelick, S.M., Evans, B.E. and Remson, I. (1983). Identifying sources of groundwater pollution: an optimization approach. *Water Resour. Res.* , 19(3), 779–790.
2. Wagner, B.J. (1992). Simultaneously parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modeling. *J. Hydrol.*, 135, 275–303.
3. Mahar, P.S. and Datta, B. 1997. Optimal monitoring network and ground-water pollution source identification. J.Water Res. Pl. 123, 199–207.
4. Mahar, P.S. and Datta, B. 2000. Identification of pollution sources in transient groundwater systems. Water Res. Man. 14, 209–227. No. 3.
5. Skaggs, T.H. and Kabala, Z.J. 1994. Recovering the release history of a groundwater contaminant. Water Resour. Res. 30, 71–79. No. 1.
6. Samarskaia, E. 1995. Groundwater contamination modeling and inverse problems of source reconstruction. SAMS 18–19, 143–147.
7. Liu, C. and Ball, W.P. (1999). Application of inverse methods to contaminant source identification from aquitard diffusion profiles at Dover AFB, Delaware. *Water Resour. Res.*, 35(7), 1975–1985.
8. Singh, R. M., and Datta, B. (2004). Groundwater pollution source identification and simultaneous parameter estimation using pattern matching by artificial neural network. ENVIRONMENTAL FORENSICS, 5 (3): 143-153.
9. Singh, R.M., Datta, B., and Jain, A. (2004). Identification of unknown groundwater pollution sources using artificial neural networks, *J. Wat. Resour. Plng. & Mgmt.*, ASCE 130(6), 506-514.
10. Rumelhart, D.E., Hinton, G.E. and Williams, R. J. (1986a), “Learning representations by back-propagating errors,” *Nature*, 323, 533-536.
11. Ogata, A., and Banks, R.B. (1961). "A solution of the differential equation of longitudinal dispersion in porous media." U.S. Geol. Surv., Prof. Pap. No. 411-A.