Water Resour Manage (2007) 21:557–572
DOI 10.1007/s11269-006-9029-z

ORIGINAL ARTICLE

Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data

Raj Mohan Singh · Bithin Datta

Received: 30 June 2004 / Accepted: 4 April 2006

Abstract Groundwater pollution sources are characterized by spatially and temporally varying source locations, injection rates, and duration of activity. Concentration measurement data at specified observation locations are generally utilized to identify these sources characteristics. Identification of unknown groundwater pollution sources in terms of these source characteristics becomes more difficult in the absence of complete breakthrough curves of concentration history at all the time steps. If concentration observations are missing over a length of time after an unknown source has become active, it is even more difficult to correctly identify the unknown sources. An artificial neural network (ANN) based methodology is developed to identify these source characteristics for such a missing data scenario, when concentration measurement data over an initial length of time is not available. The source characteristics and the corresponding concentration measurements at time steps for which it is not missing, constitute a pattern for training the ANN. A groundwater flow and transport numerical simulation model is utilized to generate the necessary patterns for training the ANN. Performance evaluation results show that the back-propagation based ANN model is essentially capable of extracting hidden relationship between patterns of available concentration measurement values, and the corresponding sources characteristics, resulting in identification of unknown groundwater pollution sources. The performance of the methodology is also evaluated for different levels of noise (or measurement errors) in concentration measurement data at available time steps.

Keywords Groundwater pollution · Pollution source identification · Artificial neural networks · Concentration patterns · Measurement error · Missing observation · Inverse problem

R. M. Singh
Department of Civil Engineering, Motilal Nehru National Institute of Technology, Allahabad-211004, India

B. Datta (✉)
Department of Civil Engineering, Inddian Institute of Technology, Kanpur-208016, India
e-mail: bithin@iitk.ac.in

Introduction

In real world situations, for a polluted groundwater system, it is likely that temporal variation in concentration measurement values at an observation site may not be available at all time steps. Some of the initial concentration measurement values may not be recorded, resulting in only the partial breakthrough curves being available. Identification of unknown pollution sources is a difficult task even when concentration measurement data is available for all the time steps (Mahar and Datta, 2001; Singh *et al.*, 2003). The complexity of the problem increases when concentration measurement data are not available for all the time steps. In this study, the pattern matching capability of an ANN is exploited to identify unknown sources of pollution from these partial breakthrough curves. Data for training the ANN are simulated using a groundwater flow and contaminant transport numerical simulation model. A typical conservative pollutant is considered. The trained network is then utilized for the identification of pollution sources for specified concentration observation data at given locations.

Contamination observed at a location may result from a single source or combination of sources with varying injection rates and release periods. Often, these source characteristics are unknown. A necessary step in addressing the issues of groundwater contamination and its remediation is the accurate detection of the source characteristics that cause groundwater pollution. In their pioneer work, Gorelick *et al.* (1983) utilized classical optimization techniques to identify pollution sources in two hypothetical study areas. They used least squares regression and linear programming together with the response matrix approach to identify the sources. Wagner and Gorelick (1986) employed nonlinear multiple regression techniques to estimate aquifer parameters, and a linear source term for a one dimensional hypothetical column system. Estimation of the linear source term was found to be highly sensitive to the introduction of measurement error.

Datta *et al.* (1989) used statistical pattern recognition to develop an expert system for the identification of unknown groundwater pollution sources. The pattern learning and recognition capabilities of an optimal statistical classifier using dynamic programming, as well as an expert knowledge base, were combined to solve the identification problem. Bagtzoglou *et al.* (1992) presented the random walk method to identify sources of groundwater pollution. Wagner (1992) presented nonlinear maximum likelihood estimation for simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modeling.

Skaggs and Kabala (1994) recovered the release history of a groundwater contaminant using current spatial measurement of contaminant concentration. They considered effects of measurement errors, parameter estimation errors, and numerical instability on performance of the method. Mahar and Datta (1997) specifically considered optimal design of a dynamic monitoring scheme for improved identification of pollution sources in groundwater. Mahar and Datta (2000) included transient flow condition. Mahar and Datta (2001) considered simultaneous estimation of aquifer parameters and identification of unknown pollution sources. Aral *et al.* (2001) embedded a progressive genetic algorithm with groundwater simulation models for the identification of contaminant source locations and release history in aquifers. Some of the recent advances in this area are discussed in Atmadja and Bagtzoglou (2001a, b) and Bagtzoglou and Atmadja (2003).

Singh *et al.* (2002, 2003) presented results for identification of unknown pollution sources using an artificial neural network, when complete breakthrough curves are available. Singh *et al.* (2003) considered multiple scenarios incorporating measurement errors. The results were promising even with large measurement errors. Datta and Chakrabarty (2003) proposed the use of a linked optimization-simulation approach where the simulation model is externally linked to the optimization model to solve this source identification problem. This approach is potentially suitable for solving large-scale identification problems.

Each of these methodologies has both advantages and limitations. Methods based on the response matrix approach assume linearity of the groundwater system. Methods based on embedding techniques are more computationally intensive, and are not suitable for very large-scale problems. Gorelick (1983) concluded that numerical difficulties are likely to arise for large scale problems using the embedding technique. Also, these methods are sensitive to measurement errors and are affected by different starting points of solution.

In most of these previously mentioned works the issue of missing data scenario is not addressed adequately. Therefore, an ANN based methodology is proposed to identify unknown groundwater pollution sources, when a portion of the concentration observation data is missing. The ANN based methodology proposed here uses patterns generated by a groundwater simulation model to train the ANN, using back-propagation algorithm. The partial breakthrough curves are the inputs, and corresponding source characteristics are the outputs, constituting a pattern for the training. An ANN model trained on a number of such generated patterns is finally utilized for the identification of unknown pollution sources in terms of source fluxes.

Simulation of groundwater flow and transport

The equation describing steady, two dimensional areal flow of groundwater through a nonhomogeneous, anisotropic, saturated aquifer can be written in Cartesian tensor notation (Pinder and Bredeoeft, 1968) as:

$$\frac{\partial}{\partial x_i}\left(T_{ij}\frac{\partial h}{\partial x_j}\right) = W; \quad i, j = 1, 2 \tag{1}$$

where T_{ij} = Transmissivity tensor (L^2 T^{-1}) = K_{ij} b; K_{ij} = hydraulic conductivity tensor (LT^{-1}) and b = saturated thickness of aquifer (L); h = hydraulic head (L); W = Volume flux per unit area (positive sign for outflow and negative sign for inflow) (L T^{-1}); and x_i, x_j = Cartesian coordinates (L)

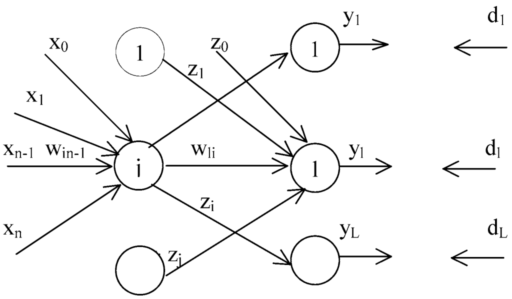
The equation describing transient, two-dimensional areal transport of a conservative solute through a saturated, rigid, and nondeformable aquifer, in Cartesian notation, can be written (Bear, 1972; Bredehoeft and Pinder,1973) as:

$$\frac{\partial(cb)}{\partial t} = \frac{\partial}{\partial x_i}\left(bD_{ij}\frac{\partial c}{\partial x_j}\right) - \frac{\partial}{\partial x_j}(bcv_i) - \frac{c'W}{\varepsilon}; \quad i, j = 1, 2 \tag{2}$$

where t = time (T); c = concentration of the dissolved chemical species (ML^{-3}); D_{ij} = coefficient of hydrodynamic dispersion (second-order tensor) (L^2 T^{-1}); c' = concentration of the dissolved chemical in a source or sink fluid (ML^{-3}); v_i = seepage velocity in the direction x_i (LT^{-1}); and ε = effective porosity of the aquifer (dimensionless). In this study, the product of the liquid volume disposal rate and the solute concentration of the source is treated is a single variable called source flux or disposal flux (MT^{-1}).

The ANN based methodology is utilized in this study for groundwater systems under steady state flow, and transient transport conditions. A numerical simulation model based on the method of characteristics (MOC) developed by the United States Geological Survey (USGS) (Konikow *et al.*, 1989) is utilized for simulating flow and transport processes in the

Fig. 1 A two layer fully interconnected MLP architecture



aquifer. These simulated solutions results are used as necessary inputs for training and testing the developed ANNs for source identification.

ANN based methodology

The ANN learns to solve a problem by developing a memory capable of associating a large number of example input patterns, with a resulting set of outputs or effects. ANN is discussed in ASCE Task Committee (2000), Rao *et al.* (2004), etc. An overview of artificial neural networks and neural computing, including details of basics and origins of ANN, biological neuron model etc. can be found in Hassoun (1999), Schalkoff (1997), and Zurada (1997).

Basic processing element of an ANN is the neuron or node. It performs work by two processes: – (i) internal activation; and (ii) transfer function. Internal activation adds up the values of incoming messages multiplied by a specific connection weight, also known as net. The variable net is defined as scalar product of the weight and input vector (Zurada, 1997). The transfer function calculates the activation level or threshold of the neuron from the internal activation. A typical transfer function is sigmoid or hyperbolic tangent function from which threshold is calculated as $y = f(\text{net}) = 1/(1 + \text{e}^{-\text{net}})$ for sigmoidal function, and $f(\text{net}) = \tanh(\text{net})$ for hyperbolic tangent function. When the processing elements or node are grouped together in a network of layers, they form the neural network architecture, also known as multilayer perceptron (MLP), as shown in Figure 1. For clarity, only selected connections are drawn. In Figure 1, $d_1, d_2, \ldots,$ represent the target values, $y_1, y_2, \ldots,$ the output values, and $x_1, x_2, \ldots,$ represent the input values, respectively. The vector $Z(z_0, z_1, \ldots, z_L)$ represents an intermediate layer of L neurons. The vector W with elements w_{ij} to denote the weight of jth neuron of a layer to ith neuron of previous layer. In this way, jth neuron may be from hidden layer or output layer.

Weight-space and back-propagation algorithm

An artificial neural network learns the approximation of the desired mapping (input vector to desired output vector) by repeatedly modifying network weights using an algorithm or learning rules. The entire process is called training.

Training an ANN is in fact a search in the so-called weight-space; that is, the space spanned by all possible weights in the network for a typical ANN topology (or architecture). The goal of the search is to find a point in weight-space which minimizes a certain error criterion. Training of an ANN is thus equivalent to performing a minimization procedure in weight-space with respect to the error criterion. One of the most prominent and widely used training algorithms is the back-propagation algorithm (Rumelhart *et al.*, 1986). It is based on the gradient descent minimization method.

ANN based methodology for pollution source identification

The groundwater pollution source identification problem is a complex inverse problem, and it is further complicated due to partially missing concentration observation measurement values. Universal function approximator capability of a feed forward multilayer perceptron with back-propagation algorithm is utilized to solve this difficult identification problem. Spatially and temporally varying observed concentration data are used to identify unknown source fluxes causing the contamination. In a given study area, application of the proposed methodology involves the procedures as outlined in the schematic representation of Figure 2. A numerical simulation model MOC (Konikow *et al.*, 1989) is utilized for generation of patterns for training and testing.

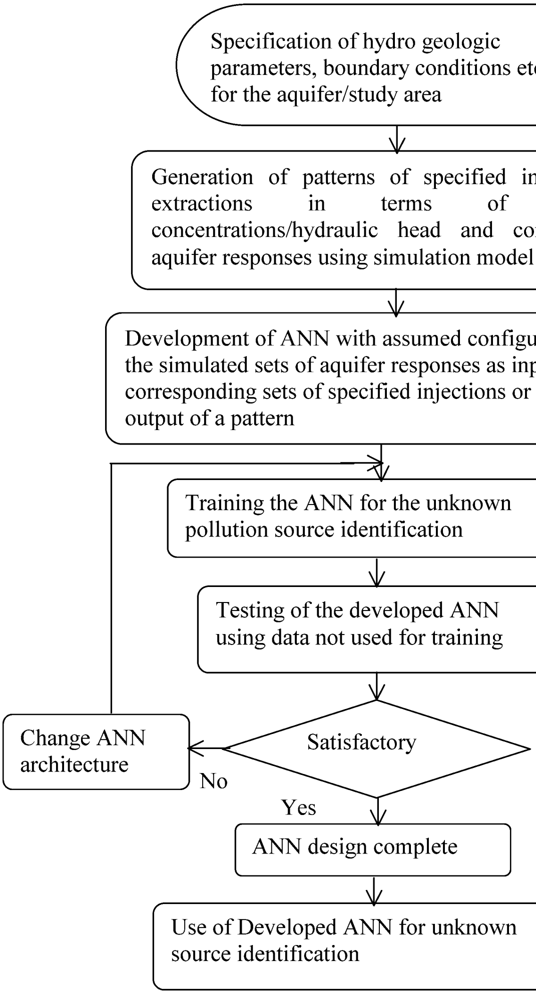

Specification of hydro geologic parameters, boundary conditions etc. for the aquifer/study area
Generation of patterns of specified injections or extractions in terms of observed concentrations/hydraulic head and corresponding aquifer responses using simulation model
Development of ANN with assumed configuration with the simulated sets of aquifer responses as inputs and corresponding sets of specified injections or extraction as output of a pattern
Training the ANN for the unknown pollution source identification
Testing of the developed ANN using data not used for training
Satisfactory
No
Change ANN architecture
Yes
ANN design complete
Use of Developed ANN for unknown source identification

Fig. 2 Schematic representation of the source identification methodology using ANN

ANN design is accepted to be satisfactorily completed when its performance during training and testing satisfies the stopping criteria based on some statistical parameters. Selection of best performing architecture of the ANN model among the various tested architectures, completes the training process. The trained ANN model can be subsequently used to identify unknown groundwater pollution sources.

Illustrative application of ANN methodology for identification of sources with missing observation data

The illustrative application of the methodology considers convective transport and hydrodynamic dispersion. It is assumed that the solute is conservative and that gradients of fluid density, viscosity, and temperature do not affect the velocity distribution. Also, steady state flow and transient transport condition is assumed. The performance of the methodology is evaluated for the study area shown in Figure 3.

In the illustrative example, two time varying pollution sources are considered at S1 and S2. Four observation wells O1, O2, O3 and O4 are considered. The concentration of the pollutant in the recharge from the pond, and the initial concentration of pollutant in the aquifer are assumed to be zero. The finite difference grid sizes, the aquifer parameter values, the recharge rates from the pond, and true value of source fluxes used to generate observed concentration are given in Table 1. It is assumed that a 10 year time domain can capture the entire concentration breakthrough curve at a specified observation location. This 10 year time domain is also divided into 40 equal time steps.

It is assumed that no concentration observation data are available for the first 3 years of the 10 years time domain of concentration measurement. Therefore, out of a total 10 year time domain, observation data for 7 years only are available, as first 3 years of observation data are missing. Now the available concentration observation data for the 7 year period contain concentration measurements corresponding to 28 equal time steps, each of 3 month duration. For 4 observation wells, a total of 112 concentration measurement values constitute the input for the ANN model. Source fluxes causing these concentration values are the output for training the ANN.

The performance of the methodology is evaluated for different levels of noise (or measurement errors) in concentration measurement data. The boundary and initial conditions were

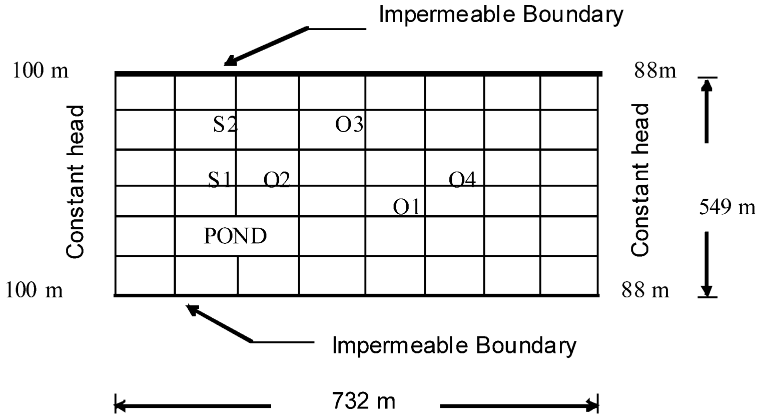


Fig. 3 Study area

Table 1 Flow and transport parameter values and source flux values used for simulating observed data for study area shown in Figure 1

Parameter or source fluxes duration	Value
Parameter values	
K_{xx} (m/s)	0.0001
K_{yy} (m/s)	0.001
ε	0.2
α_L (m)	30.5
α_T (m)	12.2
b (m)	30.5
Δx (m)	91.5
Δy (m)	91.5
Δt (month)	3
q_r (L/s)[a]	2.15
Source flux[b] at S1	
Year 1	48.8
Year 2	0.0
Year 3	10.0
Year 4	42.0
Year 5	36.0
Source flux at S2	
Year 1	0.0
Year 2	0.0
Year 3	0.0
Year 4	0.0
Year 5	0.0

[a] Clean water recharge from pond.
[b] Unit of source flux is in grams per second (g/s)

assumed known. Pollution sources were assumed to release a typical conservative pollutant in a uniform rate throughout each of the periods of activity considered.

Performance criteria

To evaluate the performance of the developed methodology, it is necessary to define the criteria by which its performance will be evaluated. Some statistical parameters are used to select the neural network architecture among various architectures, and to judge the predictive accuracy of selected architecture.

Performance evaluation criteria for network architectures

The following statistical parameters are used for quantifying the errors in training and testing for different ANN architectures.

Total error (E)

The total error of the network is defined as the normalized sum of the squared differences between the output of the network and the target values over all outputs and all patterns (Anguita *et al.*, 1994). Smaller the value of *E*, better is the performance of ANN in training.

$$E = \frac{1}{PN_l} \sum_{p=1}^{P} \sum_{l=1}^{N_l} \left(d_l^p - y_l^p\right)^2 \tag{3}$$

where, $P =$ total number of patterns; $N_l =$ total number of outputs $y_l^p = l^{\text{th}}$ output value for p^{th} pattern; and $d_l^p =$ target value at l^{th} output and p^{th} pattern.

Average absolute relative error (AARE)

The average absolute relative error represents normalized sum of the deviations between the output and target values of the ANN. It is defined as:

$$AARE = \frac{\sum_{p=1}^{P} \sum_{l=1}^{N_l} \left| d_l^p - y_l^p \right|}{\sum_{p=1}^{P} \sum_{l=1}^{N_l} d_l^p} \times 100 \tag{4}$$

Threshold statistics ($TS(K)$)

It measures the model performance at certain level of absolute relative error. Absolute relative error (k) represents the normalized deviation between the output values and target values. It is defined as:

$$k = \left| \frac{d_l^p - y_l^p}{d_l^p} \right| \times 100 \tag{5}$$

For the absolute relative error (k), Threshold Statistics, TS (k) is defined as:

$$TS(k) = \frac{n}{N} \times 100 \tag{6}$$

where $n =$ number of data points whose absolute relative error is less than k; and $N =$ total number of data points taken over all the outputs and all the patterns.

It is a measure of proportional number of data points for which the normalized errors are less than k. Larger this value, better is the performance of the ANN.

Correlation coefficient (R)

It is defined as:

$$R = \frac{\sum_{p=1}^{P} \sum_{l=1}^{N_l} \left(\Delta d_l^p\right)\left(\Delta y_l^p\right)}{\sqrt{\sum_{p=1}^{P} \sum_{l=1}^{N_l} \left(\Delta d_l\right)^2 \sum_{p=1}^{P} \sum_{l=1}^{N_l} \left(\Delta y_l^p\right)^2}} \tag{7}$$

where

$$\Delta d_l = d_l - \sum_{l=1}^{N_l} \frac{d_l}{N};$$

and

$$\Delta y_l = y_l - \sum_{l=1}^{N_l} \frac{y_l}{N}$$

Larger values of R represent better performance.

Performance evaluation criteria for developed ANN models

The ANN model is first selected on the basis of the above mentioned performance evaluation criteria used in the training phase. Normalized error is used as the evaluation criterion for evaluating the performance of the selected ANN models in identifying the unknown groundwater pollution sources.

Normalized error (NE)

The normalized error is used as a measure of over all source identification error. This is also a measure of the methodology performance, defined as (Mahar and Datta, 2001):

$$NE = \frac{\sum_{p=1}^{np} \sum_{l=1}^{ns} \left| ef_{l,p} - af_{l,p} \right|}{\sum_{p=1}^{np} \sum_{l=1}^{nl} af_{l,p}} \times 100 \tag{8}$$

where, *ns* = total number of potential source locations; *np* = total number of potential disposal periods at each locations; $ef_{l,p}$ and $af_{l,p}$ = model estimated average source flux and actual source flux respectively, at *l*th location and during *p*th disposal period.

Larger values of *NE* represent large absolute deviation between the estimated source fluxes and actual source fluxes.

Incorporating measurement errors

In some of the evaluation results presented, simulated values of the observed concentrations were perturbed to represent the effect of measurements errors. The perturbation of simulated concentrations is performed by adding randomly generated error terms to the numerically simulated concentrations. The normally distributed error terms represents the concentration measurement errors that generally occur in field measurements or laboratory tests. The perturbed concentration values are computed as follows (Mahar and Datta, 2001):

$$c_{\text{obs}}(x_n, t) = c_{ns}(x_n, t) + \varepsilon r \tag{9}$$

where, $c_{\text{obs}}(x_n, t)$ = measured or observed concentration at location x_n, at time t; $c_{ns}(x_n, t)$ = numerically simulated concentration at location x_n, at time t; and εr = random error term.

Here, the random variable εr is assumed to follow a normal distribution with mean = 0 and standard deviation = $a \cdot c_{ns}(x_n, t)$. Further the error term is defined as

$$\varepsilon r = e \times a \times c_{ns}(x_n, t) \tag{10}$$

where, a = a fraction ($0 \leq a \leq 1.0$); and e = normal deviate.

Standard normal deviate (e) are generated using MATLAB (Version 5.2.0.3084, 1998). The value of 'a' is varied from 0.05 to 0.3. Higher values of 'a' indicates higher level of noise in the concentration measurement data. In this study, it is assumed that values of $a < 0.10$ correspond to low noise level, $0.1 \leq a \leq 0.15$ correspond to moderate noise level, and $a > 0.15$ correspond to high noise level. Also, for performance evaluation purpose, a normal distribution of the errors is assumed. Any other suitable distribution function may be incorporated. The value of $c_{obs}(x_n, t)$ can be negative if 'e' is negative., 'a' is large and

$c_{ns}(x_n, t)$ is small. Generally, such a situation is less probable if '*a*' is small and '*e*' is also small. Otherwise a truncated normal distribution may be used.

Determination of network architecture

ANN architecture determines the number of connection weights (free parameters), and the way information flows through the network. Determination of appropriate network architecture (or topology) is most important and also the most difficult task in the ANN model building process.

In this study, the ANN model is implemented using a standard back-propagation algorithm (Rumelhart *et al.*, 1986; Anguita *et. al.*, 1994). Hyperbolic tangent transfer function, generalized delta learning rule, and the quadratic error function are chosen as internal parameters of the network. Hyperbolic tangent transfer function make back-propagation learning perform better (Haykin, 1994, 2000). The initial weights are randomly distributed between + 0.1 and – 0.1. Motivation for using small weights is to prevent premature saturation, while randomness is introduced as symmetry breaking mechanism so that the node may not become redundant. Keeping this in mind other authors, e.g. Maier and Dandy (1996) used this range of initial weights. Input and output values are scaled between – 0.9 and + 0.9 (close to the range of activation function i.e., -1 to $+1$).

During training, a large set of patterns are presented to the network. A set of input values together with the corresponding target value constitutes a pattern. The input values are simulated concentration measurements at the observation wells. As of concentration measurements for initial three years are assumed to be missing, the inputs comprise of the concentration measurement values for the remaining 7 years. The target values are corresponding source fluxes. These concentration measurements are simulated using a numerical simulation model (MOC) for the specified source fluxes and the study area. Source fluxes are randomly generated using a uniform distribution with specified upper and lower bounds. The upper and lower bounds may be determined in such a way that the ranges of possible values for source fluxes are adequate within these bounds. Out of total, 10,725 patterns, 5775 patterns were used for the training, and 4950 patterns were used for testing.

In the course of experimentation, it was observed that single hidden layer architectures lacked generalization ability when applied to the identification problem. In earlier work (Singh *et al.*, 2002) it was also observed that a network with two hidden layers performed better than networks with a single hidden layer. Better performance with two hidden layers may be due to the complexity and nonlinearity involved in mapping the input to the output for this problem.

The criteria used for terminating the training of the ANN is based on a total error (*E*) versus iterations graph, as well as the error criteria e.g. *E*, *R* etc. It is ensured that error surface graph has less fluctuations, total error (*E*) values are less than 10 percent, and correlation coefficients (*R*) are greater than 0.8. Although arbitrary, it is believed these are acceptable stopping criteria, while limiting the CPU time required.

Training and testing with missing data

In order to train the ANN for missing observation data scenario, the input values are the simulated concentration measurements at observation well locations shown in Figure 3. These data exclude the concentration values for the missing initial 3 years period. Total 112 observation data for all the four observation wells, 28 at each observation site, is used as input to ANN. The target values are 10 corresponding source fluxes (5 for each source

Table 2 Performance evaluation during the training and testing at 1000 iterations for missing data

ANN Model	AARE	TS(1)	TS(10)	TS(25)	TS(50)	TS(100)	R	E
				Training				
112-10-10-10	36.4	1.7	17.0	54.1	60.2	62.8	0.817	0.095
112-12-10-10	33.8	2.0	19.8	55.2	60.9	63.4	0.838	0.086
112-16-10-10	28.8	2.8	23.6	57.7	62.9	64.9	0.873	0.068
112-20-10-10	29.4	2.7	24.0	57.6	62.5	64.4	0.869	0.071
112-30-10-10	26.7	3.5	28.3	59.6	63.7	65.4	0.889	0.061
112-30-12-10	32.1	2.5	22.9	57.4	62.1	64.0	0.854	0.078
112-30-20-10	27.8	3.1	25.4	58.5	63.4	65.3	0.879	0.065
112-40-10-10	34.0	2.2	18.7	55.5	61.5	63.9	0.836	0.086
				Testing				
112-10-10-10	35.4	1.9	18.7	61.7	69.3	72.6	0.755	0.113
112-12-10-10	33.2	2.2	21.4	63.4	70.1	73.4	0.780	0.103
112-16-10-10	28.7	2.8	24.7	66.4	72.7	75.4	0.827	0.083
112-20-10-10	28.7	3.0	26.1	66.4	72.2	74.9	0.828	0.083
112-30-10-10	25.4	3.7	32.5	68.8	73.7	75.9	0.854	0.072
112-30-12-10	30.8	2.7	25.1	65.5	71.4	74.0	0.809	0.092
112-30-20-10	27.2	3.1	27.3	67.5	73.3	75.8	0.839	0.078
112-40-10-10	32.9	2.3	21.0	63.6	70.9	74.1	0.784	0.101

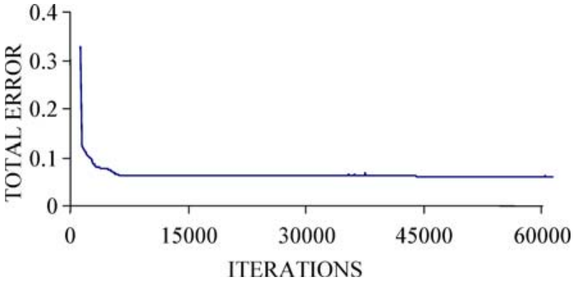


Fig. 4 Plot of error versus number of iterations for the ANN model 112-30-10-10

location at S1 and S2). The performances of some of the ANN architectures at 1000 iteration are shown in Table 2. The network with 112 inputs, 30 neurons in first hidden layer, 10 neurons in second hidden layer, and 10 outputs represented as 112–30–10–10, perform well both in training and testing mode. However, to reduce the total error value further, the training is performed with up to 60,000 iterations. From iterations versus total error graph (Figure 4), it is obvious that near 60,000 iterations the error surface value stabilizes. Also, *E,* and *R* values are well within the bounds set for stopping criteria. The values of *AARE* and *TS* considerably improved as the number of iterations is increased. The performance evaluation of the 112–30–10–10 network in training and testing at 60,000 iterations is presented in Table 3. Thus the ANN model represented by 112–30–10–10 architecture is selected for identification of the sources with missing concentration observation data.

Table 3 Performance evaluation during the training and testing at 60,000 iterations for missing data

ANN Model	AARE	TS(1)	TS(10)	TS(25)	TS(50)	TS(100)	R	E
				Training				
112-30-10-10	19.5	6.0	38.4	61.7	66.8	68.6	0.911	0.051
				Testing				
112-30-10-10	19.8	6.8	44.0	71.6	77.8	80.2	0.880	0.061

Table 4 Comparison of actual and predicted source fluxes using ANN model for error free concentration measurements for missing data

Potential source		Actual fluxes (g/sec)	Estimated source fluxes (g/sec)	
Duration (years)	Location		No Missing Data (Singh *et al.*, 2003)	Missing data
Year 1	S1	48.800	49.865	41.618
Year1	S2	0.000	2.707	4.466
Year 2	S1	0.000	1.257	0.000
Year 2	S2	0.000	0.000	0.000
Year 3	S1	10.000	11.708	11.884
Year 3	S2	0.000	3.717	0.000
Year 4	S1	42.200	41.391	44.582
Year 4	S2	0.000	0.0000	0.096
Year 5	S1	36.000	35.958	37.507
Year 5	S2	0.000	0.000	0.000
	NE		8.118	12.951

Identification results and discussion

Performances of the trained ANN models were evaluated for the two assumed scenarios: first without concentration measurements errors, and then with concentration measurements errors.

The identification results with missing concentration observation data are first compared with earlier obtained identification results without any missing data. The normalized error (*NE*) value increased from 8.118 with no missing data (Singh *et al.*, 2004) to 12.951 with missing data, assuming error-free concentration measurements. These results are shown in Table 4. Comparison of actual source fluxes with estimated source fluxes are shown in Figure 5. As less amount of information is utilized in the case of missing data, the deterioration in identification accuracy is expected.

Comparison with optimization approach

The problem of identification of unknown pollution sources with missing concentration measurement data is more complex than that with no missing data. The solution obtained by ANN based methodology is better than those of classical optimization based methodology for less complex design than that depicted in Figure 3 with missing data. Mahar (1995) developed non linear embedded optimization model to identify unknown pollution sources with missing concentration measurement data, considering only one potential source location at S1 with

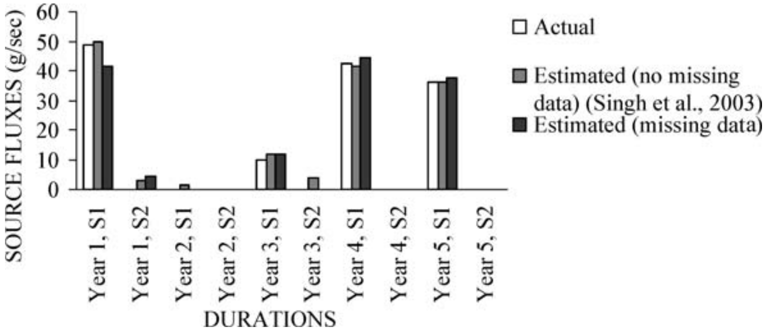


Fig. 5 Comparison of actual and estimated source fluxes

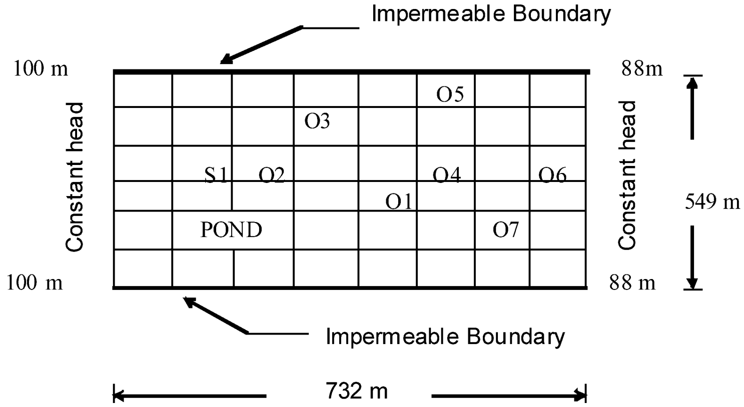


Fig. 6 Study area for optimization solution

seven observation wells. The study area is identical to that shown in Figure 3, except for the observation location. These details of the study area (Mahar, 1995) are shown in Figure 6. The identification problem as represented by Figure 6 and solved using the embedded optimization method incorporates seven observation wells compared to only four in the problem solved byusing ANN approach. Therefore, the problem solved by ANN approach is comparatively more complex. The identification error obtained by using the optimization approach for the missing data as represented by NE, is 11.6 percent. Identification error using ANN approach, for the more complex scenario with additional potential sources and only four observation locations is 12.9. It has been established (Singh, 2004) that the identification accuracy improved as the number of observation locations increased. The identification error is expected to further improve with seven observation wells instead of four. Also the error should decrease as the dimensionality of the identification decreases with less number of potential source locations. Therefore, it can be argued that the ANN approach is certainly comparable in efficiency in identification of sources with missing concentration observatiion data. Also, the ANN approach is much less complex computationally. This result is certainly encouraging.

The simulated concentration measurement values are perturbed using Equations (9) and (10). The performance of the ANN model with different levels of noise in concentration measurements is presented in Table 5. With no missing data case (Singh *et al.*, 2003), the

Table 5 Identification errors with different noise levels and sample sizes for missing data

	NE (sample size 20)		NE (sample size 30)	
A	No Missing data (Singh *et al.*, 2003)	Missing data	No Missing data (Singh *et al.*, 2003)	Missing data
0.05	8.700	12.992	8.613	13.165
0.10	9.836	16.299	9.696	16.139
0.15	11.297	21.206	11.076	20.545
0.20	12.838	26.750	12.534	25.417
0.30	16.119	38.347	15.604	35.714

identification error represented by *NE* values increased as the measurement noise level increased. Again, it is evident that due to lesser amount of information available in the missing data scenario, identification errors increased compared to the no missing data scenario. The trend of increasing *NE* value with increasing noise level is also evident. However, for high noise level ($a > 0.15$), the *NE* value for missing data scenario is almost two times that for the no missing data scenario.

Therefore, assuming an initial missing concentration measurement period of three years, in a ten years observation period since the start of activity of potential sources, the identification error increases in terms of *NE* value, compared to the case when no measurement data is missing. However, it is seen from Table 4 that even with high measurement noise ($a \geq 0.15$), the identification error are in the range of 25 percent or less. Therefore, the ANN model appears potentially applicable for identifying unknown groundwater pollution sources, with missing concentration measurement data.

No doubt, many limitations of the performance evaluation results presented here can be mentioned. These results are very much dependent on the proper training of the ANNs, as well as determination of the optimal architecture. This issue of optimal architecture has not been adequately addressed. The final architecture of the ANNs have been determined based on comparison of few specified architectures. Also, ANN performances would suffer if it is used in an extrapolation mode i.e. it has to identify patterns outside the ranges for which it has been trained and tested. Therefore, the training patterns need to be chosen especially based on potential source characteristics. Also, the features to be extracted from the patterns are important considerations (Datta *et al.*, 1989) contributing to the efficiency of the ANN. Another limitation is that a homogeneous aquifer study area was assumed for this performance evaluation.

The missing concentration measurement scenario evaluated here assumes that concentration measurements corresponding to the initial portion of the breakthrough curve is missing. There can be various other scenarios of missing concentration measurements. Such scenarios need to be incorporated in the evaluation process to fully establish the applicability of the proposed methodology.

Identification of unknown pollution sources in groundwater is an inverse problem. The issue of uniqueness of the obtained solutions and the question if the problem is ill-posed are vital (Datta, 2002). The identification problem becomes more complex and ill-posed as the measurement data become more sparse, or some data are missing, or uncertainties exist in specified parameter values or boundary conditions. Deterioration of identification results with missing concentration measurement data is also due to the increased complexity of the problem. Identification results based ANN architecture whose optimality can not be guaranteed may not be unique, even if a unique solution exists. Furthermore, the issue remains whether unique solutions exist, even if the ANN architecture is optimal, and identification problem is not ill-posed. Many of these issues remain to be addressed adequately.

Conclusions

The ANN based methodology is developed and evaluated for identification of unknown groundwater pollution sources, in terms of magnitude, location and timing. A realistic scenario of partially missing concentration measurement data is considered. The proposed ANN methodology has the potential to perform satisfactorily, even when the concentration measurement data are missing for a few initial time periods after the potential sources had become active. The ANN approach is straightforward, and does not require the formulation and solution of complex non-linear optimization models. The performance evaluations of ANN models under different levels of noise in concentration measurement are encouraging. However, more rigorous evaluation is necessary before the applicability and efficiency of the ANN approach is fully established, especially when a portion of concentration measurement data is missing.

References

American Society of Civil Engineers Task Committee on Application of Artificial Neural Networks in Hydrology (2000) Artificial neural networks in hydrology: preliminary concepts. J Hydrologic Engrg ASCE 5(2):115–123

Anguita D, Parodi G, Zunio R (1994) An efficient implementation of BP on RISC-based workstations. Neurocomputing 6:57–66

Aral MM, Guan J, Maslia ML (2001) Identification of contaminant source location and release history in aquifers. J Hydrologic Engrg ASCE 6(3):225–234

Atmadja J, Bagtzoglou AC (2001a) Pollution source identification in heterogeneous porous media. Water Resour Res 37(8):2113–2125

Atmadja J, Bagtzoglou AC (2001b) State of the art report on mathematical methods to reliable of groundwater pollution source identification. Environmental Forensics 2(3):205–214

Bagtzoglou AC, Atmadja J (2003) Marching-jury backward beam equation and quasi-reversibility methods for hydrologic inversion: application to contaminant plume spatial distribution recovery. Water Resour. Res 39(2):10:1–14

Bagtzoglou AC, Dougherty DE, Tompson AFB (1992) Application of particle methods to reliable identification of groundwater pollution sources. Water Resources Management 6(1):15–23

Bear J (1972) Dynamics of fluids in porous media, Elsevier, New York

Bredehoeft JD, Pinder GF (1973) Mass Transport in flowing water. Water Resour Res 9(1):194–210

Datta B (2002) Discussion of ‘Identification of contaminant source location and release history in aquifers’ by Mustafa M. Aral, Jiabao Guan, and Morris L. Masia. Journal of Hydrologic Engineering. ASCE 7(5):399–401

Datta B, Beegle JE, Kavvas ML, Orlob GT (1989) Development of an expert system embedding pattern recognition techniques for pollution source identification. Completion report for U.S.G.S. grant no. 14–08–0001–G1500, Dept. of Civ. Engrg., Univ. of California, Davis, Calif

Datta B, Chakrabarty D (2003) Optimal identification of unknown pollution sources using linked optimization simulation methodology. Proc. of Symposium on Advances in Geotechnical Engineering (SAGE 2003). I.I.T, Kanpur, India, pp. 368–379

Gorelick SM (1983) A review of distributed parameter groundwater management modeling methods. Water Resour Res 19(2):305–319

Gorelick SM, Evans B, Ramson I (1983) Identifying sources of groundwater pollution: an optimization approach. Water Resour Res 19(3):779–790

Haykin S (1994) Neural Networks: a Comprehensive Foundations, Mac Millan, New York

Haykin S (2000) Neural Networks: a Guided Tour. In: Singh and Gupta (eds) Soft Computing and Intelligent System, Academic Press, Sandiego, pp 71–80

Hassoun MH (1999) Fundamentals of Artificial Neural Networks, Prentice-Hall of India Private Limited, New Delhi

Konikow LF, Bredehoeft JD, Goode DJ (1989) Computer model of two-dimensional solute transport and dispersion in groundwater, U. S. Geol. Surv. Tech. Water Resources Invest. book 7, U.S. Geological Survey, Reston, Va

Mahar PS (1995) Optimal identification of groundwater pollution sources using embedding technique, PhD. Thesis, I.I.T., Kanpur

Mahar PS, Datta B (1997) Optimal monitoring network and ground-water-pollution source identification. Jl. Of Water Resources Plng and Mgmt ASCE 123(4):199–207

Mahar PS, Datta B (2000) Identification of pollution sources in transient groundwater system. Water Resource Management 14(6):209–227

Mahar PS, Datta B (2001) Optimal identification of ground-water pollution sources and parameter estimation. J Water Resources Plng and Mgmt ASCE 127(1):20–29

Maier RH, Dandy CG (1996) The use of artificial neural networks for the prediction of water quality parameters. Water Resour Res 32(4):1013–1022

MATLAB (1998) Using MATLAB. The Math Works, Inc

Pinder GF, Bredehoeft JD (1968) Application of the digital computer for aquifer evaluations. Water Resour Res 4(5):1069–1093

Rao SVN, Murty BS, Thandaveswara BS, Mishra GC (2004) Conjunctive use of surface and groundwater in coastal and deltaic syatems. J Water Resources Plng and Mgmt, ASCE 130(3):255–267

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representation by error propagation. Parallel Distributed Processin. MIT Press Cambridge Mass (1):318–362

Schalkoff RJ (1997) Artificial Neural Networks. The McGraw-Hill Companies, Inc., New York

Singh RM (2004) Identification of unknown pollution sources in groundwater using artificial neural networks and genetic algorithm, PhD. Thesis, I.I.T., Kanpur

Singh RM, Datta B, Jain A (2002) Identification Of Unknown Groundwater Pollution Sources Using Artificial Neural Network. Proc. of the International Conference on Advances in Civil Engineering (ACE-2002), I.I.T., Kharagpur, India. pp. 83–93

Singh RM, Datta B, Jain A (2004) Identification of unknown groundwater pollution sources using artificial neural networks. J Water Resources Plng and Mgmt ASCE 130(6):506–514

Skaggs TH, Kabala ZH (1994) Recovering the release history of a groundwater contaminant. Water Resour. Rese 30(1):71–79

Wagner BJ (1992) Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modeling. J Hydrol 135:275–303

Wagner BJ, Gorelick SM (1986) A statistical methodology for estimating transport parameters: theory and application to one dimensional advective dispersive systems. Water Resour Resear 22(8):1303–1315

Zurada, JM (1997) Introduction to Artificial Neural Systems, West Publishing Company