Estimation of Contaminant Transport Parameters using Genetic Algorithms

Anirudh Vemula[1], K.S. Hari Prasad[2]

Abstract

Groundwater is a valuable natural resource which has been contaminated by various anthropogenic activities. Many mathematical models have been used to describe the behaviour of groundwater systems under the various physical conditions encountered. The performance of these models is drastically affected by the accuracy with which the involved parameters can be determined. In the present paper, we investigate the feasibility of employing genetic algorithms to estimate the groundwater contaminant transport parameters; since the traditional methods of estimation are laborious, time consuming and expensive. For this purpose, the genetic algorithm is coupled with the Ogata and Banks model. The performance of this scheme is tested with pseudo-experimental data, obtained from the analytical solution with assumed values of parameters. The results indicate that this method can be used to estimate the parameters accurately.

1. Introduction

Groundwater is a valuable natural resource and is being used extensively for drinking, irrigation and industrial purposes. Due to increase in population and rapid industrialization this resource has been increasingly contaminated. The main sources of groundwater contamination are waste disposal sites, industrial effluents, agricultural and urban watersheds (Ward and Giger, 1985). Mathematical models have been effectively used to study the transport of contaminants in groundwater and to take appropriate measures for remediation (Charbeneau, 2000). However, with greater model sophistication comes a need for more intensive data requirements and precision in model prediction hinge on our ability to accurately determine the model parameters (Kool et al, 1987).

The transport parameters of soil medium are commonly determined by imposing restrictive initial and boundary conditions so that the governing equations can be inverted by analytical or semi-analytical methods. However, these direct inversion methods also have a number of limitations which restrict their practicality, in particular when used for calibrating field scale models. An alternative and more flexible approach to solving the inverse problem is to employ parameter estimation techniques. In such an approach, the model parameters are estimated by minimizing the deviations between the model predicted and field observed output. Contrary to direct inversion methods, the optimization approach does not put any inherent constraints on the form or complexity of the model, on stipulation of initial and boundary conditions. Thus a major advantage is that experimental conditions can be selected on the basis of convenience and expeditiousness.

2. Contaminant Transport in One Dimension

The one dimensional form of advection-dispersion equation for non reactive dissolved constituent in saturated, homogeneous, isotropic material under steady state uniform flow is described by (Charbeneau, 2000)

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} - v\frac{\partial c}{\partial x} \tag{1}$$

[1]*Undergraduate Student, Indian Institute of Technology, Roorkee, INDIA*
[2]*Assistant Professor, Department of Civil Engg., Indian Institute of Technology, Roorkee, INDIA*

where, x is the spatial coordinate (L), t is the time coordinate (T), c is concentration of contaminant (ML^{-3}). D is the hydrodynamic dispersion coefficient and v is fluid velocity (LT^{-1}). Usually D is expressed in terms of two components (Bean, 1972)

$$D = D_m + a_L v \tag{2}$$

D_m is the coefficient of molecular diffusion (L^2/T) and a_L is the longitudinal dispersivity (L).

Depending on the choice of applicable boundary conditions, analytical solutions are available for eq. (2). In the present work, we consider the Drichilet boundary condition – the inlet concentration is constant to be valid.

Initial Condition:

$$At\ t = 0,\ c = 0; \quad for\ 0 \le x \le \infty \tag{3}$$

Boundary Condition:

$$For\ t \ge 0,\ c = c_o; \quad x = 0 \tag{4}$$

$$For\ t \ge 0,\ c = 0; \quad x \to \infty \tag{5}$$

The solution for these conditions is given by Ogata and Banks (1961) as

$$\frac{c}{c_o} = \frac{1}{2}\left(erfc\left(\frac{x - vt}{\sqrt{4Dt}}\right) + \exp\left(\frac{xv}{D}\right) erfc\left(\frac{x + vt}{\sqrt{4Dt}}\right)\right) \tag{6}$$

when xv/D is sufficiently large, the second term of the above equation (3) may be neglected.

To solve the inverse problem, we require the observed experimental data – concentration of the contaminant at various locations at different time intervals. In the present work, we assume the values of the transport parameters (v, D) and then use the eq. (6) to determine the contaminant concentration. Genetic Algorithms are then employed along with the analytical model (see eq. (6)) to estimate these transport parameters and also the generated pseudo-experimental data.

Formulation of Objective Function

The objective function is defined as the sum of squares of difference between the model predicted and pseudo-experimental contaminant concentrations at different locations & times. If c_{obs} is the observed value and c_{pred} is the corresponding model predicted concentration for a given set of transport parameters, the sum of squares of difference (SSD) is calculated as

$$SSD = \sum_{i=1}^{N} \left(c_{obs} - c_{pred}\right)^2 \tag{7}$$

where, N is the total number of observations.

Eqn. (7) serves as the objective function in the present study and the target is to minimize SSD. Since, the Genetic Algorithm toolbox of MATLAB Development Environment (v7.0.1 R14) automatically determines the minima of a given objective function; we can directly define the fitness function as the SSD itself.

$$Fitness\ function = f(x) = \text{SSD} \tag{8}$$

Suitable termination criterion must be specified to stop genetic algorithm. We adopt the following criteria for the GA:

Max number of Generations: ∞
Max Stall Generations: 100

Max Stall Time Limit(s): 100

Maximum number of Generations is the number of generations at which the algorithm terminates. Maximum Stall Generations is the maximum number of generations within which there is no change in fitness value of the population. Similarly, Stall Time is the maximum time elapsed within which there is no change in fitness value.

Genetic Representation

Initially, a population of parameters proportional to the total string length is generated using a random generator. Generally, GAs have been developed by using binary coding in which a string (or chromosome) is represented by a string of binary bits that can encode integers, real numbers. In this study, the parameters are encoded as substrings of binary digits having a specific length. These substrings are joined together to form longer strings representing a solution. The length of the substring (i) is determined according to the desired solution accuracy and is dependent on the range of the parameter and the precision requirement of the parameter. McKinney and Lin (1994), gave the relationship between the substring length and the obtainable precision as

$$(D_k - C_k)10^{a_k} \leq 2^l - 1 \tag{9}$$

where, i_k = desired precision for the i^{th} parameter, i is the substring length, C_k and D_k = lower and upper bounds of the k^{th} parameter, k = index for the parameter:

The value of k^{th} parameter is decoded using the following equation.

$$Z_k = C_k + [(D_k - C_k)/(2^l - 1)]\sum_{j=0}^{l} I_j 2^j \tag{10}$$

Where Z_k = value of k^{th} parameter and l_j = binary digit at the j^{th} position in the string starting from right. Once the initial population is randomly generated, the fitness of each string is determined using eqn. (8).

Reproduction:

The strings generated in the initial population are chosen for participation in the reproduction process based on their fitness values. Many selection schemes such as deterministic sampling, stochastic sampling with or without replacement, stochastic tournament selection and fitness proportionate selection, are used for reproduction (Goldberg, 1989). In this work, the fitness proportionate selection is chosen, where a string is selected for the reproduction process with a probability proportional to its fitness. Thus the probability p_i of an individual member string i being selected is given by

$$p_i = s_i \Big/ \sum_{i=j}^{p} s_i \tag{11}$$

where s_i = fitness of individual string I and p is the population size. The scheme is implemented with the simulation of a roulette wheel with its circumference marked for each string proportionate to the string's fitness. The roulette wheel is rotated p times, each time selecting a copy of the string chosen by the roulette wheel printer. As the circumference of the wheel is marked according to a string's fitness, the roulette wheel mechanism makes s_i / s_{mean} copies of the i^{th} string in the re production. s_{mean} is the average fitness of the population given as

$$s_{mean} = \frac{1}{p} \sum_{i=1}^{p} s_i \tag{12}$$

The string with a higher fitness value represents a larger range in the cumulative probability values, and hence has a higher probability of being copied into the mating pool.

Mating (Cross over):

The general theory behind the cross over operation is that by exchanging important building blocks between two strings that perform well, the GA attempts to create new strings that preserve the best material from two parent strings. The cross over is a recombinant operator that selects two strings from the mating pool at random and cuts their bits at a randomly chosen position. This produces two "head" segments and two "tail" segments. The tail segments are then swapped over to produce two new full length strings. The number of strings participating in mating depends on cross over probability. If a cross over probability of p_c is used, only $100 \times p_c$ percent strings in the population are used in the cross over operation. Cross over has a wide range of possible types, i.e., one point, multipoint, uniform, intermediate arithmetical and entered arithmetical (Goldberg 1989). The effect of cross over depends on the site at which cross over takes place.

Mutation:

Mutation is an important process that permits new genetic material to be introduced to a population. A mutation probability is specified that permits random mutations to be made to individual genes (e.g. changing *1* to *0* and vice versa for binary GAs). Mutation operator facilities the convergence towards an optimal solution even if initial population is far from the optimal solution. The binary GAs used mutation probability at very low rates ranging from 0.001 to 0.01.

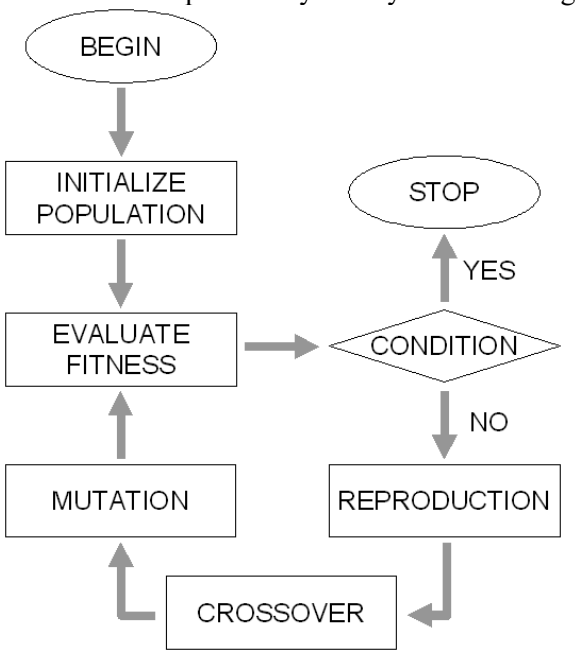


Fig. 1. Flow chart of Genetic Algorithm.

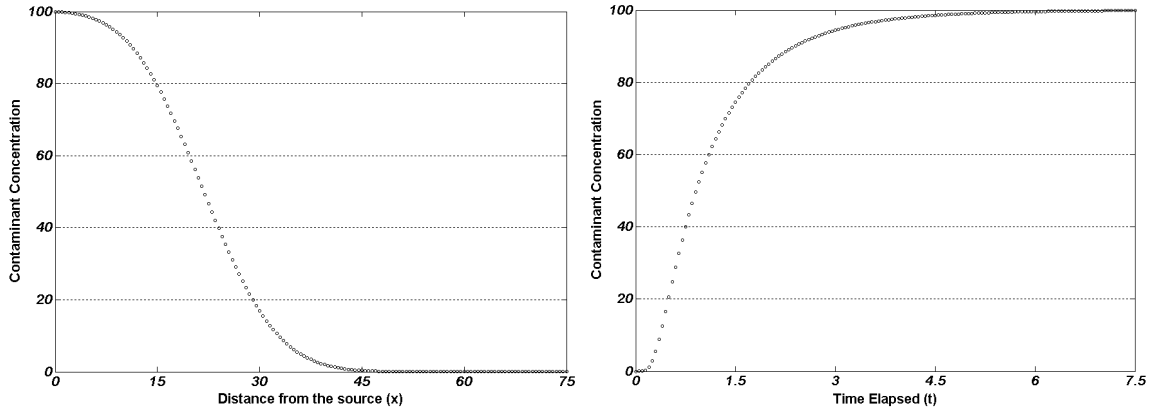
3. Experimental Setup and Results

In order to study the performance of the proposed schemes, two cases were considered and studied – variation of contaminant concentration with distance from source at given instant of time (t_o) and variation with time at given location (x_o). To generate the observed data, the following assumed parameter values (see Table 1) were used with eq. (6).

Table 1. Parameter Values for the two cases

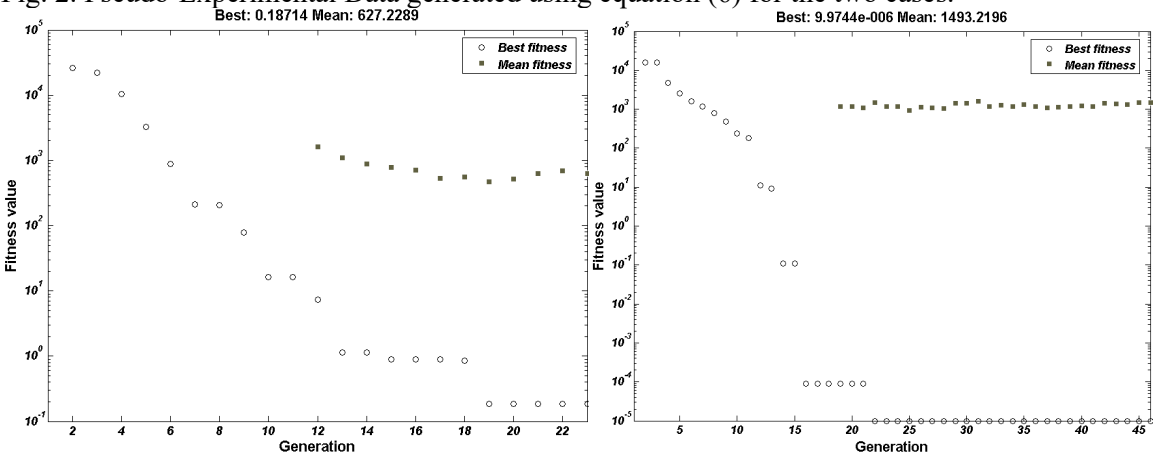
Parameter	Value
Source Concentration (C_o)	100
Velocity (v)	5
Dispersion Coefficient (D)	10
Time Elapsed (t_o)	4

Parameter	Value
Source Concentration (C_o)	100
Velocity (v)	5
Dispersion Coefficient (D)	10
Distance from Source (x_o)	6



Case (a) Case (b)

Fig. 2. Pseudo-Experimental Data generated using equation (6) for the two cases.



Case (a) Case (b)

Fig. 3. Simulation results for the two cases.

Table 2. Simulation Results for the two cases.

Parameter	Actual Value	Predicted Value	Relative Error
Velocity (v)	5	4.9792	0.42%
Dispersion Coefficient (D)	10	10.0736	0.74%

Parameter	Actual Value	Predicted Value	Relative Error
Velocity (v)	5	4.9999	0.002 %
Dispersion Coefficient (D)	10	10.0002	0.002 %

Table 2 presents the results of solving the inverse problem using GA for both the cases – at given instant of time and at a given location from the source. It can be observed that the relative error for both the cases is very less (< 1%), and hence the superior performance of the proposed scheme. It can be seen from the figure 3 that the mean fitness is not defined for initial

generations; which may be accounted to the large initial fitness values for the randomly generated population. The best fitness is, however, defined from the second generation and attains a limiting value.

4. Conclusion

Many remediation and protection activities require accurate models to simulate contaminant transport in groundwater. The performance of these models is heavily dependent on our ability to estimate the model parameters. The dispersion coefficient is known to vary spatially and its estimation is both laborious and expensive. Furthermore the estimated parameters may not accurately represent the field. Genetic algorithms (GA) have emerged as a powerful non-conventional optimization technique, which have been employed in the present work. The results suggest that the transport parameters estimated by solving the inverse problem using GA are in close agreement with the actual values.

References

1. Bear, J., 1972. Dynamics of Fluids in Porous Media. American Elsevier, New York.
2. Charbeneau, Randall J., "Groundwater Hydraulics and Pollutant Transport", by Prentice Hall, Upper saddle river, NJ 07458, 2000.
3. Dagan, G., 1989. Flow and Transport in Porous Formations. Springer-Verlag, New York.
4. de Marsily, G., 1986. Quantitative Hydrogeology. Academic Press.
5. Freeze, R.A., Cherry, J.A., 1979. Groundwater. Prentice-Hall, Englewood Cliffs, NJ.
6. Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley.
7. Neumann, S.P., 1982. Statistical Characterization of aquifer heterogeneities: an overview. In: Narasinuhan (Ed.), Recent Trends in Hydrogeology, Spec. Pap. Geol. Soc. Am. Vol. 189, Boulder, CO, pp. 81–102.
8. Ogata, A., and Banks R.B., (1961). A solution of the differential equation of longitudinal dispersion in porous media. U.S. Geological survey Prof. paper 411-A, Washington D.C., P.A-4.