Identification of clandestine groundwater pollution source locations and their release flux history

A Chakraborty¹² and O Prakash¹

¹Department of Civil and Environmental Engineering, Indian Institute of Technology Patna, Bihta, Bihar 801106, India
²E-mail: anirban.pce15@iitp.ac.in

Abstract. Large numbers of untreated toxic wastes are often buried underground, as a method of disposal. Such methods of clandestine disposal have led to widespread pollution of groundwater resources. These methods of wastes disposal are often adopted to avoid the cost incurred in proper treatment of the waste and untraceability of such clandestine disposals. However, the effect of such disposals manifests in form of groundwater pollution, which if left unchecked would potentially pollute the entire aquifer. Such polluted aquifers need to be reclaimed by implementing proper remediation techniques. However, the effectiveness of any remediation technique would depend on the precise knowledge of the unknown pollutant source/s characteristics, in terms of their Numbers, their Locations and their Release Flux history referred to as NLRF. This study presents a noble technique for simultaneously estimating the unknown number of clandestine pollutant sources, their locations along with their release flux history. Simulated Annealing (SA) is used in Linked Simulation Optimization (LSO) based framework. In this developed methodology, the number of sources and their respective locations and release flux histories are treated as unknown decision variables, and are estimated simultaneously along with the source release flux history.

1. Introduction
In many parts of the world, groundwater is the only source of fresh water. Global water consumption is estimated to have tripled during the last 50 years [1]. Sustainability of these resources are seriously threatened due to widespread pollution of groundwater aquifers from anthropogenic activities [2]. Several chemical and toxic pollutants which are byproducts or waste products of industrial activities are responsible for groundwater aquifer pollution. Industries producing such toxic chemicals often resort to improper methods of disposal of such wastes which eventually lead to groundwater pollution. Clandestine burial of untreated toxic wastes underground is often adopted by industries to offset the cost of proper treatment prior to disposal. Post dumping of the waste underground, the surface is dressed to cover the trace of such disposals. Over a period of time these toxic wastes leach into the groundwater with rainfall recharge and spreads along with groundwater flow, thus polluting entire groundwater aquifer [3].

A large number of polluted aquifers need to be reclaimed by adopting suitable remediation technique in order to ensure sustainable supply of drinking water for all. However, the effectiveness of any remediation technique will depend on the precise estimation of the pollution source characteristics [4]. When pollution is initially observed in groundwater, generally very little is known about the number of sources or their locations causing pollution. Groundwater pollution is usually first encountered in an arbitrarily located well or group of supply wells. Identification of unknown NLRF of sources from
randomly obtained pollution concentration measurement data is a challenging task. Characterization of unknown groundwater pollutant sources is an inverse problem [5] as location and activity initiation time have to be backtracked in space and time, respectively. Generally, this inverse problem is ill-posed and non-unique [5]. There are several techniques like least square based methodology [6], surrogate models [7-9] which are used for characterization of groundwater pollution sources. For estimating the unknown source characteristics, LSO method [10] holds good as it can efficiently perform with limited concentration measurement data availability scenario. Many researchers [11-14] gave methods for estimating the release flux histories of pollutant sources using optimal monitoring network assuming explicitly known potential pollution source locations. An SA based LSO technique [15] was given for determining activity initiation times for pollution sources. Release flux histories using LSO recovered for a distributed source [16]. LSO based techniques are effective in reconstructing the source flux release histories only when the number and locations of the pollutant sources are well known, or the potential source locations are known with fair degree of certainty [17].

This knowledge about the number of sources, and their actual/potential locations [18] are implicitly assumed to be known in these existing methodologies of pollutant source identification, which is seldom true in real life scenarios of groundwater pollution originating from clandestine sources. To address this major gap a new methodology is proposed that can simultaneously identify the number of clandestine sources of pollutant, their respective locations in the study area and source flux release history even when there is no prior information regarding the number of sources and their locations. The number of unknown sources and their locations are treated as explicit unknown decision variables such that the solution of the optimization problem optimally estimates the number and locations of the actual sources, and release flux histories of these identified sources simultaneously.

2. Methodology

The proposed methodology uses LSO approach. The LSO approach incorporates the groundwater flow and pollutant transport simulation model as set of binding constraints in the optimization formulation which is solved iteratively. Every potential solution of the optimization model must fulfill the simulation of flow and transport models given by the Eqn.1. Eqn. 2 represents the constraints of search space for pollutant sources based on the hydrogeological parameters of the study area.

\[ C_{est}^k_{loc} = \sum_i^t \left( C_{sit}, L_{x_i}, L_{y_i}, L_{z_i}, t \right) \]  

Subjected to;

\[ L_{x_{min}} \leq L_x \leq L_{x_{max}}; \ L_{y_{min}} \leq L_y \leq L_{y_{max}}; \ L_{z_{min}} \leq L_z \leq L_{z_{max}} \]  

SA is used for solving the optimization problem. The basic concept of SA is derived from thermodynamics [19]. Each iteration of the optimization model decreases the residual error given in Eqn. (3), by suitably adjusting the values of the explicit decision variables based on the control parameters (initial temperature, temperature decreasing rate, etc.).

\[ \text{Minimize}, F = \sum_{k=1}^{nk} \sum_{loc=1}^{nloc} (C_{est}^k_{loc} - C_{obs}^k_{loc})^2 \]  

\( C_{est}^k_{loc} \) is the simulated concentration obtained by MT3D at any predetermined location loc and at the end of monitoring time step k, nk is the maximum number of monitoring time steps, nloc is the total number of measurement locations, \( C_{obs}^k_{loc} \) is the observed concentration at monitoring location loc and at the end of monitoring time step represents the grid location of the ith source of pollution such that the source is \( L_{x_i}, L_{y_i}, L_{z_i} \) confined within the actual study area by choosing the values of to be the boundary of the study area \( L_{x_{max}}, L_{y_{max}}, L_{z_{max}} \): The maximum number of potential sources is denoted by n.

In the LSO problem, decision variable is formed by unknown source characteristics i.e. NLRF such that the optimal values of these decision variables gives the best approximation of these unknown source characteristics. In this study, MATLAB (2017a) toolbox “simulannealbnd” for SA is used as optimization algorithm. The binding constraints in Eqn. (1) and Eqn. (2) essentially show the clubbing of Simulation and Optimization model.
3. Description of the study area
The study area adopted to test the efficiency of proposed method in estimating NLRF is given in figure 1. The study area comprises of a three dimensional homogeneous, anisotropic unconfined aquifer of area 1000m x 1000m. The hydro-geological parameters of the study area are given in table 1. The entire study area is discretised into 20 x 20 x 1 grids, having two clandestine sources of pollution denoted by red triangle. Observed pollutant concentration measurements at 100, 200, 300, 400 and 500 days are simulated using GMS 10.1 for three monitoring well locations denoted by blue circle. The entire activity duration of the sources is divided into five stress periods each of 100 days. In order to test the methodology, two different scenarios are solved in which the upper limit (n) on the number of potential sources (P) present in the study area is specified n= 2 and 3 respectively for identifying actual source characteristics (S). All the 20 x 20 x 1 grid locations are considered as possible location for the sources to be present for both the scenarios. It is to be noted that the methodology does not know the actual number of clandestine sources present in the area and relies on the optimal values of the decision variables to estimate the number of sources, their locations and source flux release for each stress period given by SIJ, where I represent the source number and J represents the stress period number of actual source. Similarly, PIJ represents potential source.

![Figure 1. Plan view of discretized study area.](image)

The concentration data is numerically simulated [20] using the given site condition using MT3Dms. These concentration data are perturbed with 10 percent random error as shown in Eqn. (4) and Eqn. (5). Both non-erroneous and erroneous concentration data are used in identification of NLRF of clandestine sources for testing the robustness of the developed method.

Where, \( C_{obs}^{k}_{loc} \) is the observed concentration at an arbitrary monitoring location at k monitoring time step; \( C_{obs}^{pert}_{loc} \) is the perturbed concentration using latin-hypercube distribution; \( \mu_{pert} \) is the standard deviation of 10 % of latin-hypercube distribution; The term, \( err \), is obtained from Eqn. (5).

\[
C_{obs}^{k}_{loc} (1 + err) = C_{obs}^{pert}_{loc}
\]
\[
err = \mu_{pert} \times rand
\]
Table 1. Hydro-geological properties of the study area.

| Name of the parameters          | Value     |
|----------------------------------|-----------|
| Length of the area               | 1000 m    |
| Width of the area                | 1000 m    |
| Length of each grid              | 50 m      |
| Width of each grid               | 50 m      |
| Hydraulic conductivity          | 20 m/day  |
| Horizontal anisotropy            | 1.29      |
| Longitudinal dispersivity        | 25 m      |
| Transverse dispersivity          | 10 m      |
| Porosity                         | 0.3       |
| Source’s grid location           | (7,4), (9,4) |
| Observation well location        | (7,6), (5,10), (4,8), (2,6) |

4. Result and Discussion
The performance evaluation results of the developed methodology for finding the unknown number of clandestine sources, their locations and flux history are presented in figure 2 and figure 3 for scenario 1, where maximum number of pollution sources is assumed to be n=2. It is to be noted that 'n' is the actual number of sources present in the study area and is not known to the methodology. Therefore, the methodology sets a maximum limit on the potential number of sources 'n'. The results show the identified source locations and estimated release flux history using error-free and erroneous concentration measurement data.

Figure 2. Identified location of source grid for n=2.

The actual number of sources present in the study area are estimated based on the identified grid locations of the potential sources. Figure 2 shows the x and y co-ordinates of grid containing the sources. The result shows that the method is able to find the two actual sources and their respective grid locations, while using both erroneous and non-erroneous concentration observation data with 100 percent accuracy.
Figure 3. Identified result of source flux for n=2.

Figure 3 shows the unknown source-flux variables for every identified clandestine source for each stress period (S11, S12, S13, S14, S15, S21, S22, S23, S24, S25) marked on the x-axis. The estimated source flux for both the identified sources matches closely with the actual flux value when tested with error free data. There is significantly larger deviation of identified flux from the actual flux value in case of S13, S22 and S24 in case of erroneous observation data. Early convergence of objective function value due to unadjusted SA parameters could be the reason for the deviation in the estimated source flux values.

Figure 4. Comparison of breakthrough curve for n=2.
To test the accuracy of the results estimated by the developed methodology, a forward simulation is performed using the estimated source flux values. Concentration versus time breakthrough curve (figure 4) is plotted for this scenario (n=2) for both erroneous and non-erroneous data and compared with the actual breakthrough curve. The breakthrough curve while using non-erroneous data shows a closer fit as compared to the one generated while using erroneous data, thus, showing a good recovery of actual breakthrough curve. The goodness of fit for non-erroneous case gives a value of $R^2=1$ and for erroneous case gives a value of $R^2=0.99$ (refer figure 5).

![Figure 5. Comparison of goodness of fit, n=2.](image)

Results of source location identification for scenario 2, where maximum number of plausible sources are assumed to be 3, is presented in figure 6 and 7. It is to be noted that there are only two number of actual sources present in the study area, but this is unknown to the methodology and therefore assumes the actual number of sources to be three (n=3). The main strength of this developed methodology lies in the fact that even while making an incorrect assumption regarding the number of actual sources, it is...
still able to find the correct number of sources and their locations. The methodology identifies the source characteristics for both the conditions i.e. non-erroneous case which is denoted by NE and erroneous case which is denoted by ER represented on the X axis.

![Figure 7](image_url)

**Figure 7.** Identified source flux values for n=3.

The methodology is able to find the co-ordinates of all the three potential source locations for both erroneous and non-erroneous observation data. On careful examination of the grid co-ordinates it is seen that the identified potential sources P1 and P2 occupy the same grid location and therefore represent the same actual source S1. Potential source P3 correctly represents the actual source location S2. Thus, all the source locations are identified with 100 percent accuracy.

![Figure 8](image_url)

**Figure 8.** Comparison of breakthrough curve for n=3.
Since P1 and P2 represent one source together, hence the equivalent source flux will be the summation of the source fluxes for both the potential sources (P1 and P2) for every stress period. Figure 7 shows the unknown source-flux variables for every identified clandestine source for each stress period (P1J+P2J=S1J, J represent stress period) marked on the x-axis. The summation of the estimated source flux for P1 and P2 matches closely with the actual flux value when tested with error free data.

![Estimated vs Actual Concentration](image)

**Figure 9.** Comparison of goodness of fit, n=3.

The accuracy of the estimated source fluxes using this methodology is tested by comparing the breakthrough curves. Estimated source flux values are used in a forward simulation to generate the breakthrough curves. Comparison of the breakthrough curves (figure 8) for scenario 2 (n=3) using actual observed data and estimated source characteristics (both erroneous and non-erroneous) show a close fit. The goodness of fit (figure 9) for non-erroneous case gives a value of $R^2 = 0.99$ and for erroneous case gives a value of $R^2 = 0.98$. This implies that recovery of source characteristics by the developed methodology is accepted. The source characteristics using non erroneous concentration data gives better match than that while using erroneous concentration data. The maximum deviation of breakthrough curve for the erroneous data is 8 percent which is well within acceptable limits. It is to be noted that in real life scenarios of groundwater pollution there is seldom any information about the release flux history of the sources. Therefore, comparative study between actual breakthrough curve and breakthrough curve using estimated source characteristics is the only way to evaluate the accuracy of the result and performance of the methodology. In this study, a good match between these two breakthrough curves show superior performance of developed methodology.

5. **Conclusions**

The developed LSO method for simultaneous identification of NLRF performs satisfactorily for both erroneous and non-erroneous observation data without deterministic information about the number of pollutant sources or their locations.

The method is able to ascertain the correct number of pollution sources even in scenarios where the initially presumed number of sources is more than the actual sources by superimposing multiple sources at the same location. In clandestine sources of groundwater pollution, it is impossible to correctly ascertain the actual number of sources prior. This method overcomes this critical limitation in the earlier
methodologies which require prior information about the actual number of sources present in the study area, or the precise potential source locations, for effective source characterization.

The method in its current form works only when the presumed number of potential sources in the study area is more than or equal to the actual number of sources. Limited evaluations conducted in this study shows the applicability of this method to real world problems of unknown groundwater pollution source identification where no prior information about the number of sources or their respective locations exists.

References

[1] Water U N 2012 Managing water under uncertainty and risk, The United Nations world water development report 4, UN Water Reports, World Water Assessment Programme (France: Paris)

[2] Alexander A C, Ndambuki J M, Salim R W and Manda A K 2018 Groundwater remediation optimization using solving constraint integer program (scip) Groundwater for Sustainable Development 7 176–84

[3] Barilari A, Londoño M Q, del Carmen Paris M, Lima M L and Massone H E 2020 Groundwater contamination from point sources. A hazard index to protect water supply wells in intermediate cities Groundwater for Sustainable Development 10 100363

[4] Datta B, Prakash O, Campbell S and Escalada G 2013 Efficient identification of unknown groundwater pollution sources using linked simulation-optimization incorporating monitoring location impact factor and frequency factor Water Resour. Manag. 27(14) 4959-76

[5] Tikhonov A N and Arsenin V Y 1977 Solutions of ill-posed problems Math. Comput. 32(144) 1-30

[6] Gorelick S M, Evans B and Remson I 1983 Identifying sources of groundwater pollution: An optimization approach Water Resour. Res. 19(3) 779-90

[7] Essouayd E, Verardo E, Pryet A, Chassagne R L and Atteia O 2020 An iterative strategy for contaminant source localisation using GLMA optimization and Data Worth on two synthetic 2D Aquifers J. Contam. Hydrol. 228 103554

[8] Ghafouri H R, Mosharaf-Dehkordi M and Afzalan B 2017 Identification of immiscible napl contaminant sources in aquifers by a modified two-level saturation based imperialist competitive algorithm J. Contam. Hydrol. 202 33–46

[9] Mahar P S and Datta B 1997 Optimal monitoring network and ground-water–pollution source identification J. Water Res. Plan. Man. 123(4) 199-207

[10] Guneshworo L, Eldho T I and Kumar A V 2018 Identification of groundwater contamination sources using meshfree RPCM simulation and particle swarm optimization Water Resour. Manag. 32(4) 1517-38

[11] Keshari A K and Datta B 2001 A combined use of direct search algorithms and exterior penalty function method for groundwater pollution management J. Porous Media 4(3) 1-21

[12] Prakash O and Datta B 2013 Sequential optimal monitoring network design and iterative spatial estimation of pollutant concentration for identification of unknown groundwater pollution source locations Environ. Monit. Assess. 185(7) 5611-26

[13] Li J, Lu W, Wang H, Fan Y and Chang Z 2020 Groundwater contamination source identification based on a hybrid particle swarm optimization-extreme learning machine J. Hydrol. 584 124657

[14] Barajas-Solano D A, Alexander F J, Anghel M and Tartakovskiy D M 2020 Efficient gHMC Reconstruction of Contaminant Release History Front. Environ. Sci. 7 1-13

[15] Jha M and Datta B 2013 Three-dimensional groundwater contamination source identification using adaptive simulated annealing J. Hydrol. Eng. 18(3) 307-17

[16] Xia X, Zhou N, Wang L, Li X and Jiang S 2019 Identification of transient contaminant sources in aquifers through a surrogate model based on a modified self-organizing-maps algorithm Hydrogeol. J. 27(7) 2535-50
[17] Bashi-Azghadi S N, Kerachian R, Bazargan-Lari M R and Solouki K 2010 Characterizing an unknown pollution source in groundwater resources systems using PSVM and PNN Expert. Syst. Appl. 37(10) 7154-61

[18] Kirkpatrick S, Gelatt C D and Vecchi M P 1983 Optimization by simulated annealing Science 220(4598) 671-80

[19] Ayvaz M T 2010 A linked simulation–optimization model for solving the unknown groundwater pollution source identification problems J. Contam. Hydrol. 117(1-4) 46-59