Method of Traceability of Pollution Source by Machine Learning

Meng Sun¹, Xue-Yi You²*

¹ School of Math, Tianjin University, 300350 Tianjin, China
² School of Environmental Science and Engineering, Tianjin University, 300350 Tianjin, China
*Corresponding author’s e-mail: xyyou@tju.edu.cn

Abstract: In terms of traceability of pollutant sources, traditional methods are not only time-consuming and labour-intensive, but also difficult to accurately locate pollution sources. This paper applies the particle swarm optimization algorithm (PSO) and genetic algorithm (GA) to identifying a single pollution source and two pollution sources, respectively. The results show that the position of pollution sources can be located well by the method of PSO and GA.

1. Introduction
The identification of pollution source is a very important topic in the field of environmental and chemical field. Over the past half century, researchers at home and abroad have published a large number of articles on traceability of pollutants in the field of environmental science and mathematics [1-2]. Generally speaking, the traditional traceability methods have many kinds, but the implementation cost is high. With the development of computer technology, the methods of traceability of pollution sources have gradually changed to rely on artificial intelligence algorithms. The common artificial intelligence algorithms are simulated annealing algorithm, genetic algorithm, particle swarm optimization algorithm, etc. But unfortunately, former researchers have not made in-depth study of the algorithms involved, which leads to the potential of each algorithm has not been fully exploited [3]. In this approach, the particle swarm optimization (PSO) is introduced and the results show that the position of pollution source can be located well by the PSO method.

2. Traceability of pollution source by particle swarm algorithm
In order to achieve the maximum information sharing and improve the applicability of traceability method of pollution source, the comparative information between a monitoring point and other monitoring points and the relationship between a monitoring point and its own historical information are desirable. By the establishment of the above relationships, all mobile monitoring platforms can make independent decisions without human intervention and obtain their next move direction and distance.

2.1 Introduction of particle swarm algorithm
PSO is a global random search algorithm based on swarm intelligence proposed by Kennedy and Eberhart, inspired by simulating the migration and clustering behavior of birds in the process of foraging[4]. In the following, the particle swarm optimization (PSO) algorithm is briefly introduced in combination with the situation of traceability of a single pollution source.
PSO randomly sets up a group of initial monitoring points, and then find the optimal solution through iteration. At each iteration, the monitoring point updates its coordinates by tracking two extremes (local maximum concentration here). The first is the extreme value found by the monitoring point itself, that is, the maximum concentration recorded by a single monitoring point in the search history, which is called the individual extreme value; The other extreme value is the maximum concentration found by the whole set of monitoring points, which is called the global extremum. Taking a monitoring point in the set of monitoring points as an example, the iteration method of monitoring points is given. In a 2-D plane search space, it is assumed that there is a set of N monitoring points, and one of the monitoring points is represented as a two-dimensional vector:

\[
x_i = (x_{i1}, x_{i2})
\]

The moving speed of the above monitoring points in the \(i\)-th iteration search is also expressed as a two-dimensional vector, which is recorded as:

\[
v_i = (v_{i1}, v_{i2})
\]

The optimal location (the coordinate of local maximum concentration) determined by the monitoring point in the \(i\)-th iteration search is as follows:

\[
P_{best_i} = (P_{i1}, P_{i2})
\]

The optimal location (the coordinate of global maximum concentration) determined by the set of monitoring points in the \(i\)-th iteration is as follows:

\[
G_{best} = \max(P_{best_i})
\]

After finding these two optimal locations, each monitoring point obtains its own velocity and position in the \(i+1\) iteration according to the following formulas:

\[
v_{i+1} = w \times v_i + C_1 \times \text{rand}(1) \times (P_{best_i} - x_i) + C_2 \times \text{rand}(1) \times (G_{best} - x_i)
\]

\[
x_{i+1} = x_i + v_{i+1}
\]

Among them, \(C_1\) and \(C_2\) are learning factors, called acceleration constants, and \(w\) is inertia factors. The right side of the first formula has three parts. The first part reflects the movement habits of monitoring points, which represent the tendency of monitoring points to maintain their previous speed; The second part reflects the memory of monitoring points to their own experience, which represents the trend of monitoring points approaching to their best historical position; The third part reflects the experience of collaborative cooperation and knowledge sharing among the monitoring points, which represent the trend that each monitoring point has the best historical location to provide to the collection.

2.2 Numerical experimental results and analysis

The simulated concentration is obtained by using the two-dimensional in-plane river pollutant diffusion equation. We set up 10 initial monitoring points in the 100 x 100 plane and begin to use particle swarm optimization to complete the traceability experiment of a single source of pollution. The two-dimensional river pollutant diffusion equation is written as:

\[
c(x, y) = \frac{w}{h \sqrt{4 \pi v (x-x_0)}} \times e^{-\frac{v(y-y_0)^2}{4a(x-x_0)}}
\]

Table 1 gives the above model parameter settings.

| Symbol | Meaning                                      | Value | Unit |
|--------|----------------------------------------------|-------|------|
| W      | Pollutant discharged rate                    | 2000  | kg   |
| h      | Water depth                                  | 10    | m    |
| \(\alpha\) | Diffusion coefficient of pollutant in the y direction | 50    | m²/s |
| v      | Average flow rate of water flow in the x direction | 0.5   | m/s  |
| x      | Abscissa of the monitoring point             | x     | m    |
| y      | Longitudinal ordinate of the monitoring point | y     | m    |
| \(x_0\) | Abscissa of the pollution source              | 0     | m    |
| \(y_0\) | Longitudinal ordinate of the pollution source | 0     | m    |
Table 2 gives the results of parameter setting of PSO after debugging several times.

| Parameter                              | Name     | Value          |
|----------------------------------------|----------|----------------|
| Learning factor                        | $c_1$    | 1.5            |
|                                        | $c_2$    | 1.5            |
| Inertia factor                         | $w$      | 0.5            |
| The maximum number of iterations       | max_steps| 50             |
| Number of initial monitoring points    | population_size | 10            |
| Monitoring point throwing range        | area     | [-50,50]×[0,100]|
| Visualization range                    | all_area | [-80,80]×[-10,150]|
| Contaminant diffusion visualization    | image view|               |

Figure 1 gives the coordinates of monitoring points for 0, 10, 20, 30, 40 and 50 iterations, respectively. From the position map of monitoring points with different iterations, it can be seen intuitively that 9/10 monitoring points can converge to the vicinity of pollution sources successfully through 50 iterations on the premise that the initial monitoring points are 10. Naturally, the proportion of successful monitoring points can be obtained under different iterations.

When the coordinates of monitoring points satisfy $|x-30| < 3$ and $|y-30| < 3$, they are considered as successful monitoring points. Table 3 gives the proportion of successful monitoring points when the number of monitoring points is 10.
Obviously, it can be seen from the Table 3 that when the number of monitoring points is 10, the proportion of successful monitoring points increases steadily with the increase number of iterations, and it can be stabilized at a high success rate after about 30 iterations. It is worth noting that the optimized algorithm itself has the risk of reaching the only local optimum. Besides the reasons of the algorithm itself, the location of the initial monitoring points randomly set also increases the possibility of PSO falling into local optimum. In practice, the location of the monitoring points at the initial time is artificially arranged; the probability of local convergence can be reduced if PSO is used.

3. Traceability of pollution sources by genetic algorithms

In fact, the pollution source traceability experiment is actually a problem of "how to find the function extremum in a given range", and the corresponding anti-jamming problem can be understood as "how to avoid the extremum search method falling into the local" problem. Because the principle of PSO is relatively simple and there is little room for further optimization, it is difficult to enhance the anti-jamming ability of PSO by optimizing its parameters. If the anti-jamming problem cannot be solved, once there are multiple pollution sources in the plane, the PSO algorithm will face a great probability of falling into local optimum. Based on this analysis, we try to use genetic algorithm with double pollution source traceability[5].

3.1 Introduction of genetic algorithm

The idea of genetic algorithm (GA) comes from the idea of species selection in nature. It uses the steps of selection, crossover, recombination and mutation of gene fragments to select individuals that are more adaptable to the environment. After reciprocating, the excellent individuals can be obtained. In this experiment, the concentration of the monitoring point is set as a criterion to judge whether the monitoring point is "good". That is, the higher the corresponding concentration of the monitoring point is, the better the monitoring point is. The coordinates of the monitoring point can be set as the genes of the monitoring point. The steps of the algorithm cycle are finally determined as follows:

(1) Set a certain number of initial monitoring points, their location randomly and the number of cycles.

(2) Select two monitoring points according to roulette idea.

(3) Set the coordinates of the selected monitoring points as \((x_1, y_1)\) and \((x_2, y_2)\). According to the method of \((x_3, y_3) = (0.9x_1 + 0.1x_2, 0.9y_1 + 0.1y_2)\) and \((x_4, y_4) = (0.9x_2 + 0.1x_1, 0.9y_2 + 0.1y_1)\), then the geographic coordinates of the new monitoring points are obtained and the corresponding simulation concentration can be calculated.

(4) Judge whether the concentration of the monitoring points \((x_3, y_3)\) and \((x_4, y_4)\) is meaningful (too large or too small can lead to the overflow of the algorithm, which affects the accuracy of the model). If the concentration of the two monitoring points is meaningful, choose two monitoring points having large concentration from the points of \((x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\) to replace the original monitoring point. If the concentration of at least one of the two new monitoring points is meaningless, then skip step (4) to step (5).

(5) In order to prevent the algorithm from convergence to a local area, a randomly selected monitoring point with meaningful concentration is replaced by a monitoring point with meaningful concentration generated randomly.

| Iteration | Numbers of successful monitoring points | Percentage of successful monitoring points |
|-----------|-----------------------------------------|-------------------------------------------|
| 10        | 1                                       | 10%                                       |
| 15        | 3                                       | 30%                                       |
| 20        | 7                                       | 70%                                       |
| 25        | 8                                       | 80%                                       |
| 30        | 10                                      | 100%                                      |
| 50        | 10                                      | 100%                                      |
| 75        | 10                                      | 100%                                      |

Table 3. The proportion of successful monitoring points.
(6) Back to step (2).

3.2 Numerical experimental results and analysis

The method of simple superposition of concentration is used to get the specific concentration of each point in the plane under the condition of double point source. Specifically, after calculating $c_1(x,y), c_2(x,y)$ according to the formula (7), they are summed. The other objective conditions, such as wind speed, are the same with that in the preceding section. The following two pollution sources are set up as:

(1) Secondary pollution source: coordinate (0,0), total amount of pollutant 2000g.

(2) Main pollution source: coordinate (30,30), total amount of pollutant 8000g.

Figure 2 gives the coordinates of monitoring points for 50, 100, 1000 and 10000 iterations. It can be found intuitively that when the number of iterations is 1000 and 10000, the GA is good at finding the source of the main pollution source. Because there are many overlapping points, Table 4 gives the final numerical coordinates of each monitoring point when the number of iterations is 1000. It can be found that most of the monitoring points (32/42) can find the coordinates of the main pollution source (30,30) successfully.

![Figure 2. The position of monitoring points with respect to iteration number](image)

Table 4 gives the proportion of successful monitoring points when the number of monitoring points is about 100. When the coordinates of monitoring points satisfy $|x-30| < 3$ and $|y-30| < 3$, they are considered as successful monitoring points.

| Iteration | Number of SMP | Percentage of SMP |
|-----------|---------------|-------------------|
| 50        | 0             | 0%                |
| 100       | 0             | 0%                |
| 500       | 0             | 0%                |
| 750       | 31            | 69.50%            |
| 1000      | 32            | 72.72%            |
| 3000      | 31            | 63.26%            |
| 5000      | 35            | 71.42%            |
| 10000     | 28            | 62.22%            |
4. Conclusions
The position of single pollution source and two pollution sources are identified by the method of the PSO and GA, respectively. The following conclusions can be drawn as:

1) When the initial monitoring point is 100, after the mutation point of iteration appears, then the traceability success rate can be stable (60% - 75%). The mutation point of iteration is closely related to the actual error range, and the number of iterations can be determined according to the user's accuracy requirements.

2) Although the traceability success rate is about 70% (in fact, it is unrealistic to hope that the traceability success rate is 100%), the final coordinates of the successful monitoring points are very accurate for the method of PSO and GA.

Acknowledgments
The research was financially supported by the National Key R&D Program of China through grant No. 2017YFC0211500.

References
[1] Sreedharan, P., Sohn, M., Gadgil, A., et al. (2006) Systems approach to evaluating sensor characteristics for real-time monitoring of high-risk indoor contaminant releases. Atmospheric Environment 40(19): 3490-3502.
[2] Wang, L.L., You, X.Y. (2015) Identification of indoor contaminant source location by a single concentration sensor. Air Quality Atmosphere Health, 8(1): 115–122.
[3] Bongers, S., Janssen, N.A., Reiss, B., et al. (2008) Challenges of exposure assessment for health studies in the aftermath of chemical incidents and disasters. Journal of Exposure Science and Environmental Epidemiology, 18(4): 341-359.
[4] Kennedy, J., Eberhart, R. (1995) Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks, 1942-1948.
[5] Datta, B., Chakrabarty, D., Dhar, A. (2011) Identification of unknown groundwater pollution sources using classical optimization with linked simulation. Journal of Hydro-environment Research, 5(1): 25-36.