ENVIRONMENTAL CONCERNS AND POLLUTION CONTROL IN THE CONTEXT OF DEVELOPING COUNTRIES

Real-time location algorithms of drinking water pollution sources based on domain knowledge

Xuesong Yan [1,2] · Zhengchen Zhou [1] · Chengyu Hu [1] · Wenyin Gong [1]

Abstract

The real-time location of pollution sources is the process of inverting pollution sources based on the dynamic optimization model constructed by the time-varying pollution concentration detected by the water quality sensor. Due to the vast quantities of the water supply networks, the water quality sensors will only be placed on critical nodes, resulting in multiple solutions. However, the increased monitoring data enhances the uniqueness of the solution. Combined with the real-time location of pollution sources, this work proposed a multi-strategy dynamic multi-mode optimization algorithm based on domain knowledge, which could guide the population search and avoid trapped into local optimal. The merging mechanism was used to keep the diversity of the population and prevent sub-population clustering on the same optimal solution. The simulation results showed that the algorithm could effectively solve the real-time location problem of pollution sources in different pipe networks and pollution scenarios.

Keywords Drinking water quality · Pollution sources · Real-time location algorithms · Dynamic multi-mode optimization · Multi-strategy · Domain knowledge

Introduction

With global industrialization and urbanization, drinking water pollution incidents occur frequently. Through real-time monitoring of the urban water supply network, relevant departments can take measures when water pollution occurs to avoid economic losses and social panic. The network system of drinking-water sensors can monitor the water quality in real-time. Once the pollution is found, the pollution source should be located in real time to reduce the pollution range, which is essential to ensure the safety of drinking water (Hu et al. 2018a, b, 2020; Gu et al. 2017, 2019a, b, c; Gu et al. 2017;Yan et al. 2020). Semih Kuter et al. use satellite data and snow classification for water resource management (Kuter et al. 2018, 2019).

Recently, the water supply network system detects the concentration of pollutants through the arranged water quality sensors and gives early warning. According to the monitored time-varying pollution concentration, the real-time location of the pollution sources reduces the scope of pollution sources in real-time, to locate the real pollution sources. The increase of pollution time and concentration data is more restrictive to the optimal solution, which makes it easier to find the pollution sources. Once the pollution is detected, the real-time location of the pollution sources can optimize the potential pollution sources and locate the real source. Besides, reducing the diffusion time of pollution in the water supply network, it ensures the safety of citizens' water use and reduces economic losses and negative impacts.

The algorithms of pollution sources location can be divided into particle backtracking, machine learning, and simulation-optimization. In early research, PBA was widely used to deduce pollution sources by tracking the state of the last time point (Shang et al. 2002; Laird et al., 2005, 2006). Then, machine learning and simulation-optimization appeared one after another. The machine learning algorithms calculate the pollution probability of each node in the water supply network

---

✉ Chengyu Hu
huchengyu@cug.edu.cn

[1] School of Computer Science, China University of Geosciences, Wuhan 430074, China

[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

to determine the highest probability is the pollution source. The simulation-optimization methods are transform the pollution sources lcocation problem into the optimization problem and use the intelligent optimization algorithms to solve it. The intelligent optimization algorithms have widely applications (Yang et al. 2018, Li et al. 2019, Wang et al. 2020a, b, 2021). Ayse Özmen, Fatma Yerlikaya-Zkurt, and their groups proposed the multivariate adaptive regression splines (MARS); it is a form of non-parametric regression analysis for building high-dimensional and nonlinear multivariate functions and applied in many fields of science, engineering, technology, finance, and control design in recent years. It is a modern methodology of statistical learning, data mining, and mathematical estimation theory which is important in both regression and classification (Özmen, 2016; Roche and Yalcinkaya, 2019, Yerlikaya-Zkurt and Taylan, 2020).

Huang et al. (Huang and Edward, 2009) considered the uncertainty of water demand, sensor measurement and modeling, so the algorithms of data mining combined with the maximum likelihood process are used to predict the location and time of pollution sources. Perelman et al. (Perelman and Avi, 2013 first used a clustering algorithm to simplify the water supply network system and then combined with a Bayesian network to predict the most likely pollution sources through probability.

Wang and Xin (Wang and Xin 2013) used the Markov Chain Monte Carlo Algorithm to predict the characteristics of pollution sources due to the uncertainty of the water supply network system. On this basis, Wang and Harrison (Wang and Harrison 2014) used a support vector regression algorithm to speed up the likelihood estimation to reduce the calculation time. Wagner et al. (Wagner et al. 2015) proposed a method based on ad-joint probability to identify the location, time, and quality of pollution sources. Such methods can not accurately locate the pollution source, but only give the possibility that the node is a source of pollution. However, with the increased pipe network scale, the calculation complexity of node probability rises dramatically.

Seth and co-workers (Seth et al., 2016) compared the algorithms for solving pollution source locations, which are the Bayesian probability method, pollution source state-based method, and simulation optimization method, where the last one can solve most pollution intrusion events. The simulation-optimization process continually searches for the optimal solution by comparing the similarity between the actual monitoring data of sensors and the simulation data set of EPANET. Avi and Elad (Avi and Elad 2005) used the genetic algorithm to maximize the random pollution matrix to search for pollution sources. Guan et al. (Guan et al. 2006) designed the optimization model as a corrector to identify the similarity between the simulation response data and the measured data of the monitoring point and then determine the pollution source. Zechman et al. (Zechman and Ranjithan, 2009) used

the global evolution strategy algorithm based on tree coding format to search for pollution sources. Lv et al. (Lv et al. 2010) used a simulation-optimization algorithm to compare the operation results of different input parameters, and the accuracy of successfully locating pollution sources reached 93.3%. Hu et al. (Hu et al. 2015) quantified the non-uniqueness of the localization solution of pollution sources. They proposed a parallel niche genetic algorithm based on MapReduce, which calculated the location, quality, and time vectors of pollution sources. Yan and other researchers (Yan et al. 2017a, b, 2018, 2019a; Gong et al. 2019) transformed the problem of pollution source location into different optimization problems and solved them with different intelligent optimization algorithms.

Based on the pollution source location, the real-time location is studied. De et al. (De et al. 2010) developed the alternative practical method to locate pollution sources in real-time by judging the consistency of the candidate source node's pollution possibility state and time interval with the sensor data. Liu et al. (Liu et al. 2011) proposed adaptive dynamic optimization algorithms, including strategies of adaptive mutation and multiple groups. Once the sensor detects the pollution, the algorithm can locate the source in real-time and narrow the scope. Costa et al. (Costa et al. 2013) first found the paths between pollution points and sources with PBA, then evaluated possible sources, and finally searched the source in real-time according to the continuous reading of the sensor. Wang and Zhou (Wang and Zhou 2017) proposed the time series Bayesian method. First, obtain the probability distribution of sensor observations of polluted nodes in offline mode. Then calculate the posterior probability in real-time and decompose it into a hierarchical tree structure. Finally, the nodes with the highest likelihood are polluted ones in real time. Yan et al. (Yan et al. 2019b) simulated uncertain water demand through the Gaussian models and proposed a real-time positioning algorithm, which can find real pollution events with fewer sensor data in a short time.

The main algorithms of pollution source location are machine learning and simulation optimization. However, the latter is only suitable for small- or medium-sized pipe network nodes because of the slow solution speed.

Materials and methods

Simulation software

The simulation software of this work is EPANET (Rossman 2000), which can simulate and analyze the water supply system. EPANET traces chemical concentration, pipeline water flow, and node pressures by simulating the characteristics of reservoirs, valves, pumps, pipelines, and nodes, which is beneficial to water supply enterprises and scientific research.

In the EPANET water supply network system, any node can be used as the source water quality. In the time mode of source water quality, pollutants can be dynamically injected into a node and diffuse. Different ways of injecting pollutants at the source cause the actual pumping of pollutants to be changed. Source injection methods include mass injection, flow step injection, set point injection, and concentration injection.

The mass injection is to inject equal quality pollutants according to the set mass curve. Flow step injection is to insert fixed concentration pollutants into the water inlet end of the nodes, while set point injection is to insert the pollutants into the outlet. The above three methods are applied to simulate the invasion of tracer, disinfectant, and pollutant without considering the water demand of the nodes. However, concentration injection is suitable for simulating the source of water supply, because it changes the pollutant concentration at the source node, which is related to water demand.

Problem modeling

In the drinking water supply network, sensors at key nodes monitor water quality in real time. Once pollution is detected, it will warn and continuously record real-time water quality. With the increase of information collected, it is essential to locate the characteristics of pollution sources in real-time, such as the location of pollution sources, injection time, duration, and quality vector. Sensors are only arranged to key nodes owing to the large scale of the water supply network. It makes the location of pollution sources "symmetrical," leading to the non-uniqueness of solutions. This problem belongs to multi-mode optimization (Hu et al. 2015; Yan et al. 2019c).

The real-time location of the pollution source is proposed based on the source location. From the perspective of simulation optimization, both are reverse-tracing problems, which is to solve the pollution source by comparing the difference between the cumulative concentration data of the potential pollution and the real. However, the simulation optimization process of pollution source location is static, while the other is dynamic. Figure 1 (a) shows that when the sensor detects pollution, only by accumulating enough concentration data can the optimization algorithm locate the pollution source. In previous studies, the cumulative concentrations calculated by the data model ranged from 24 to 48 h. When the pollution in the pipe network is detected, the source location cannot show the scope immediately, and the pollution will spread for 10–20 h. Therefore, this work proposes the real-time location of pollution sources. Figure 1 (b) shows that when pollution is detected, the real-time location of pollution sources will be optimized in a short time. The ever-increasing real observed concentration data makes the optimization algorithm more accurate. Unlike the cumulative concentration data of the data model of pollution source location, the real-time site is between the time when the pollution is first detected at the polluted time t_0 and the current time t_c. Moreover, the increased t_c
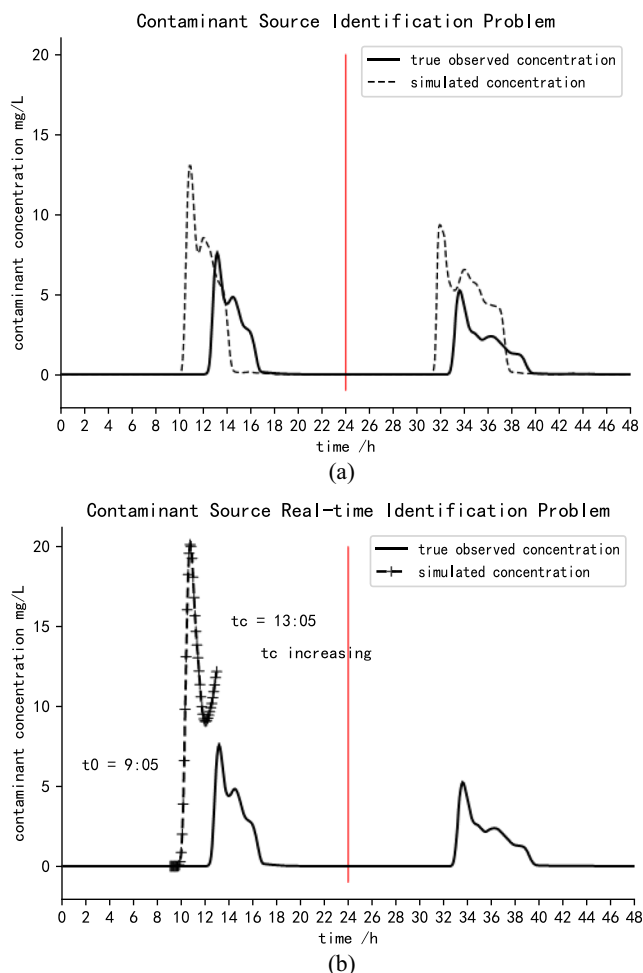

Fig. 1 Concentration comparison

increases concentration data and the accuracy of the optimization algorithm.

The non-uniqueness of the solution is more challenging in the real-time location of pollution sources rather than the location. The increased monitoring data will optimize the solution to reduce the quantity. As for optimization, it is necessary to compare the real sensor concentration accumulation data with the simulated data at a specific time and take this as the individual fitness value. With the spread of pollution sources, the pollution concentrations monitored by sensors vary at different times, resulting in the inability to compare individual fitness values. Therefore, this work expresses the optimization problem as Eq. (1),

$$\underset{\{L,\mathbf{M},T_0,D\}}{Minimize\ f} = \sqrt{\frac{\sum\limits_{t=t_0}^{t_c}\sum\limits_{i=1}^{Ns}\left(c_{it}^{obs}-c_{it}^{*}\right)^2}{\sum\limits_{t=t_0}^{t_c}\sum\limits_{i=1}^{Ns}\left(c_{it}^{obs}\right)^2}}\ Subject\ to\ \ \mathbf{M}$$

$$= \{m_1, m_2, \cdots, m_k\},\ \ m_i \geq 0\ \text{and}\ \ L = \{1, \cdots, N\}\ \ \ \ (1)$$

where f is the prediction error; t_0 the time when the sensor first detects the pollution in the real scene; t_c the current time; N_s the

Fig. 2 Simulation-optimization framework
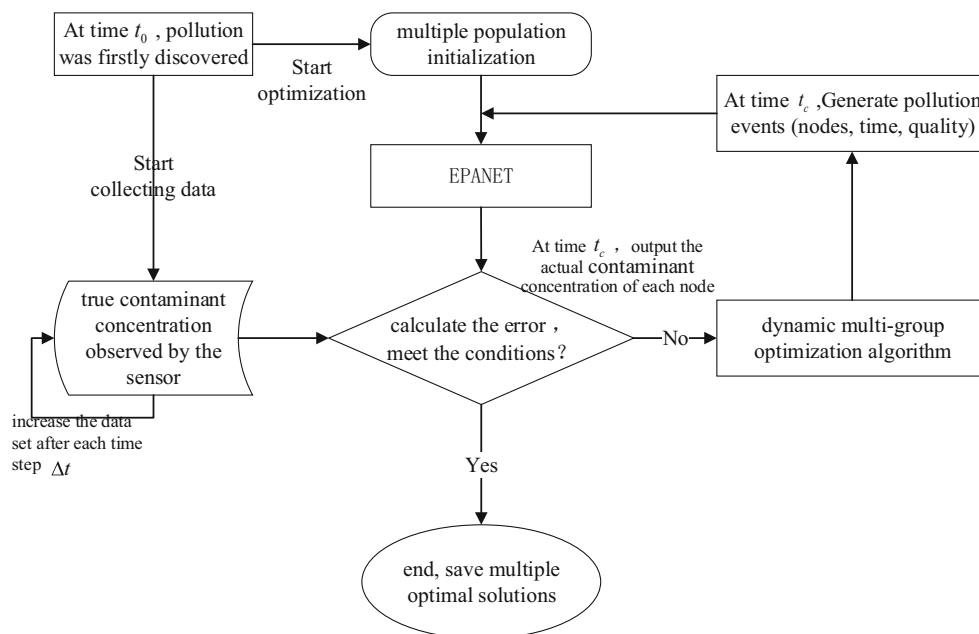
number of sensors; c_{it}^{obs} the pollutant concentration of the observation sensor i at time t; c_{it}^* the pollutant concentration of sensor i when EPANET simulates potential pollution event at time t, which can be obtained by function $\{L,\mathbf{M},T_0,D\}$; L the total number of nodes; $\mathbf{M}$ the injection vector of the pollution source; T_0 the initial injection time; and D the injection duration. The optimization goal is to obtain the optimal solution $\{L,\mathbf{M},T_0,D\}$ by finding a minimum value of f. This formula calculates the similarity between the simulated and the real concentration from t_0 to t_c. The data on increasing pollution concentration increases with the increase t_c, which leads to the optimization of the candidate

solutions. Also, the degree of non-uniqueness of the problem reduces, that is, the number of potential pollution events decreases.

Therefore, the number of potential solutions for the real-time location of pollution sources will decrease with time until finding the source event. In this work, through the limited sensor data set to search the whole water supply network, we can obtain multiple possible pollution sources. When the real time is considered, the data increases. The constraint on the solution is more significant, which causes the non-uniqueness degree of the solution to decrease. Thus, this work
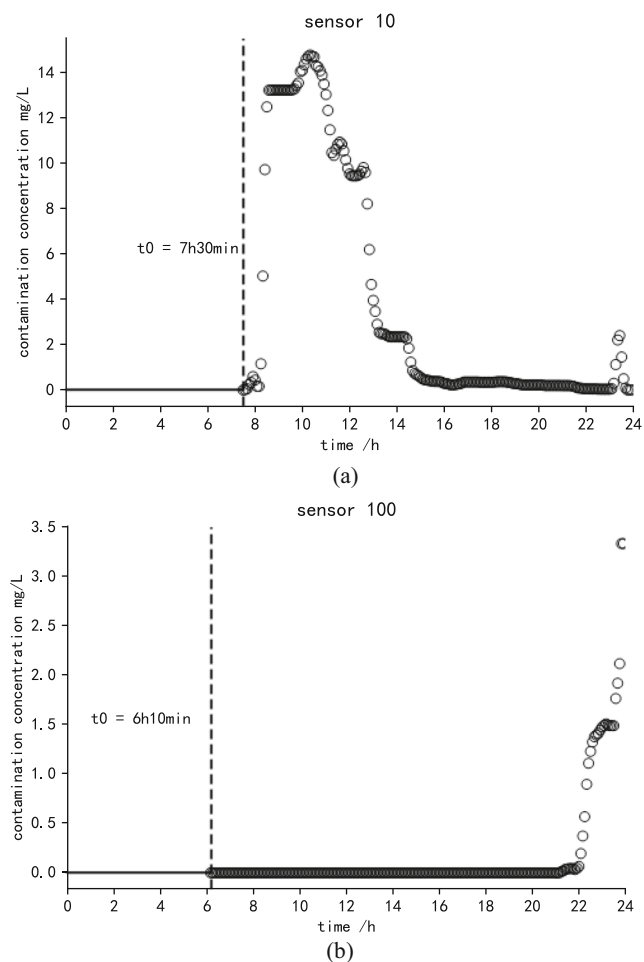
Fig. 3 Algorithm framework
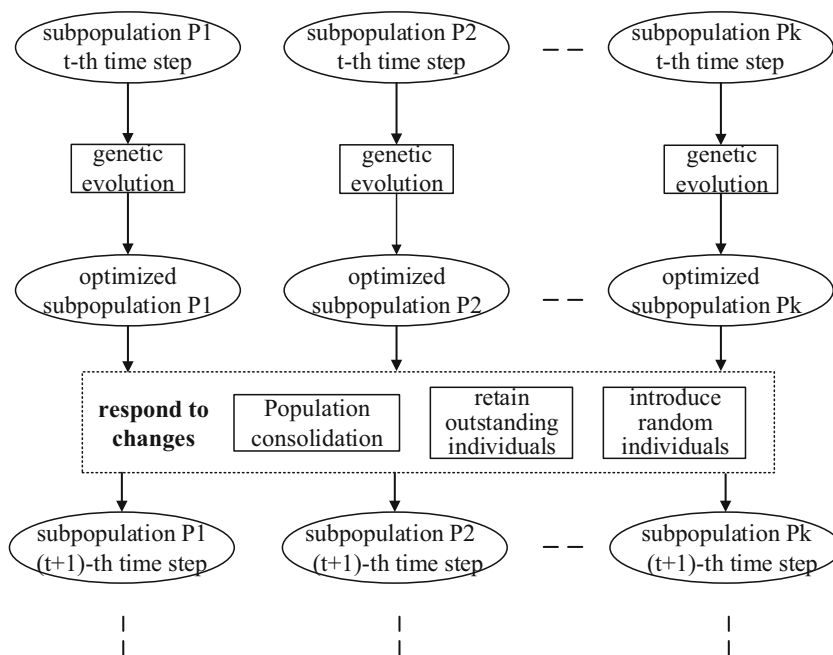
Fig. 4 Observed pollution concentrations

transforms the problem of the real-time location of pollution sources into a dynamic multi-mode optimization problem.

Real-time location algorithm based on domain knowledge

In this work, a simulation-optimization algorithm is used to solve the real-time location of pollution sources. In the water supply network system, the sensor detects the pollution for the first time at time t_c, which indicates that pollution intrusion exists. The pollution concentration information recorded by the sensor can be used as the real pollution source data after each time step Δt. In real life, when pollution is detected for the first time, relevant departments should launch a dynamic multi-mode optimization algorithm to search the pollution source information accurately. At a specific time t_c, a series of calculated pollution events are input to EPANET to generate the simulated concentration data.

Comparing the real concentration data with the simulation, we can screen out hidden pollution events to locate the pollution source information as soon as possible. Figure 2 shows a detailed flowchart for solving the real-time location of pollution sources. When the difference between the cumulative simulated concentration and the actual cumulative monitoring is 0 or less than a specific threshold value at the time t_c, the pollution event is the real pollution source (including pollution source location, injection time and duration, and quality vector).

Algorithm architecture

It is critical to detect environmental changes and the corresponding methods in dynamic optimization problems, which can be solved by dynamic multi-group optimization algorithms (Cruz et al. 2011; Nguyen et al. 2012; Jin and Branke 2005; Yang and Li 2010). The environmental change is that the collection of pollution information at each time step is increasing, which means that the real-time location is a static optimization problem at a specific time.

At time t, k sub-populations iterative evolve and optimize simultaneous. During iterative evolution, the population searches globally through selection, crossover, and mutation. If an individual of the population satisfies certain conditions, it enters the local optimizer for a local search to improve the accuracy of the algorithm. The premise is to assume that the optimal individual of each sub-population is a feasible solution to the problem. Before entering the next moment, the population adopts strategies of merging populations, retaining
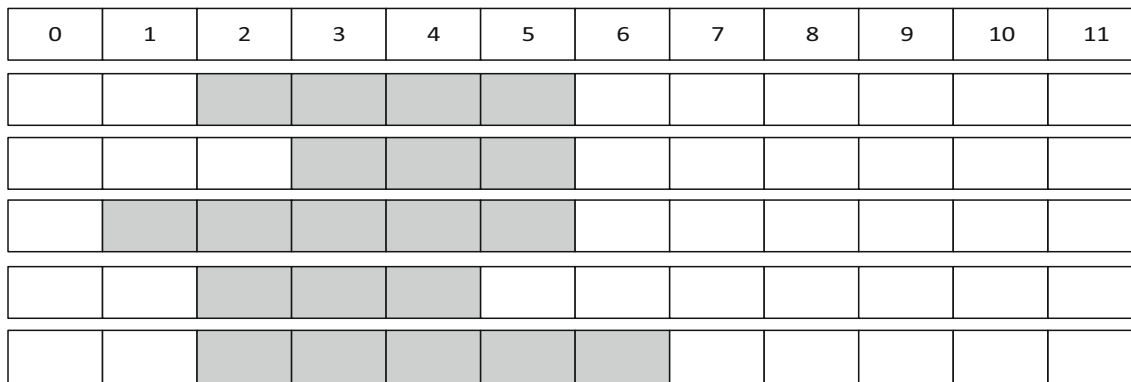

Fig. 5 Domain search process

Table 1 Pipe network parameters

Pipe Network Parameter	Net3_Rossman2000	BWSN1_Ostfeld2008	ky5_Jolly2013
Number of nodes	97	129	430
Number of pipes	119	178	507
Reservoir	2	4	4
Pool	3	3	3
Hydraulic step	1 h	1 h	1 h
Water quality step	5 min	5 min	5 min
Sensor position distribution	37, 61	10, 83, 100	6, 22, 30, 34, 40, 42, 43, 76, 80, 87

excellent individuals and introducing random individuals to respond to changes. When k populations are optimized, the node positions of the optimal individuals of some populations may be the same, which indicates that populations with the same search direction need to merge. Environmental change will lead to the premature convergence of k sub-populations and increase search costs. Sub-population needs to keep the optimal individuals of the same node while replacing the deleted individuals with new ones generated randomly. Figure 3 shows the specific optimization process.

Selection strategy based on domain knowledge

Previous real-time location studies of pollution sources had ignored the moment when the water quality sensor first detected pollution, which was essential for accurately locating pollution sources. For example, in the BWSN1_Ostfeld2008 Avi et al. 2008 pipe network, the 44th node was injected with a pollution event. Figure 4 shows the pollution concentrations detected by sensors 10 and 100 in a day. The left side of the dotted line indicates that the pollution concentration is zero, while the time when the pollution is first detected is 6:10. According to the experience of experts, if the time t_0^* for the first detection of potential pollution is farther from t_0, the less likely it is to be the real pollution.

Priority is given to the optimization of individuals with A close to B, which can find the real pollution events. The processing formula is as Eq.(2).

$$|t_0^* - t_0| < \varepsilon \tag{2}$$

Each individual in the population is a pollution event, and the first pollution monitoring time t0* can be obtained through EPANET. When the individual's satisfies Eq. (2), it can be regarded as a potential pollution event and will not be eliminated. Considering the individual fitness value, if the potential individuals selected by each sub-population meet the Eq. (2) dominate, the roulette wheel selection is improved by individual fitness value. Otherwise, it is improved by the individual's |t0*-t0| value. And the threshold ε=30 min.

Domain search strategy based on domain knowledge

At present, most studies have adopted single-point mutation for the injection start and duration time after the hybrid coding of pollution events. The two have no relevance, which makes the algorithm fall into local optimum. Experiments found that when the time series determined by the injection start and duration time highly coincides with the real pollution source, the mutation falls into local optimum. For example, the actual injection start and duration time are (2, 4), while the time sequence of (3,3) coincides with the real by 3, 4 and 5, which means the feasible solution containing the gene fragment may be a right solution. In the local optimum, the probability of (3,3) single point mutation to (2,4) is low, indicating that (3,3) is better than (5,4). Although the duration of (5,4) is consistent and (3,3) is inconsistent, individuals with (3,3) gene fragments have lower fitness values than (5,4). The above analysis shows that the higher the time series coincidence of pollution events, the smaller the prediction error. This work, from the time series of injection start and duration, proposed a domain search strategy. The start and duration time was taken as a whole to search for the time series with the highest degree of coincidence, keeping individual diversity and jumping out of the local optimum. The Harley coding principle adopted here made the time series obtained by the search different from itself by only one time period.

Table 2 Algorithm parameters

Parameters	Parameter description	Parameter size
k	Number of subpopulations	10
pop_size	Size of subpopulations	10
g_{Max}	Population iterations	20
P_c	Crossover probability	95%
P_m	Mutation probability	70%
Δt	Time step	10 min
t_{Max}	Positioning time	2 h
scale	Local search scale	2*10

Table 3 Pollution scenarios of Net3_Rossman2000 pipe network

Scenes Parameters	Node number	Pollution injection start time	Duration(h)	Injected pollution concentration(mg/L)
Pollution scene1	16	2	4	300, 180, 240, 180
Pollution scene 2	86	2	4	300, 180, 240, 180

In Fig. 5, it is assumed that the time interval is 0–11 for a total of 12 h, where the gray time quantum indicates that the pollution source has been injected. The second line represents the time series corresponding to (2,4). The adjacent search directions may be different, but only search in four-time series that differ by one time period, which is represented by (3,3), (1,5), (2,3), and (2,5) represented by 3, 4, 5, and 6 in the figure. Randomly selecting one of the searches results in a corresponding new start and duration time.

Local search strategy based on domain knowledge

Individual gene coding methods include integer coding and variable-length real-number coding. During selecting, crossing, and mutation, the dynamic algorithm is easier to find the integer coding genes composed of node position, injection start, and duration time, while the variable-length real-number coding quality vector is difficult to converge to the exact value. Experiments showed the fitness value of the dynamic optimization algorithm based on Darwin's idea dropped rapidly at the beginning and then slowly. It became invariant after iteration, which could not reach a minimal value. The algorithm accuracy was improved by searching for the quality

vector separately and increasing the optimization cost. Referring to the analysis of improved selection strategies, when the potential individual $|t_0{}^*-t_0|=0$, the quality of the individual is searched separately. In this work, the behavior of optimizing the quality vector of potential individuals satisfying $|t_0{}^*-t_0|=0$ is called the local search strategy of the whole multi-group dynamic algorithm.

Particle swarm optimization (PSO) algorithm is widely used in engineering. It has a simple algorithm, fast convergence, and few adjustment parameters, which is suitable for optimizing continuous problems. Therefore, the standard PSO algorithm is used in the local search of the quality vector for the real-time location of pollution sources. In this paper, let $\omega=0.8$, $c_1=c_2=2$.

Algorithm steps

This work proposed multiple strategies based on domain knowledge to analyze the characteristics of the real-time location of pollution sources. When the original Dynamic Multi-group Optimization algorithm (DMOA) was combined with the three improved strategies, the algorithm could quickly converge while improving accuracy. The steps of multi-

Fig. 6 Net3_Rossman2000 pipe network
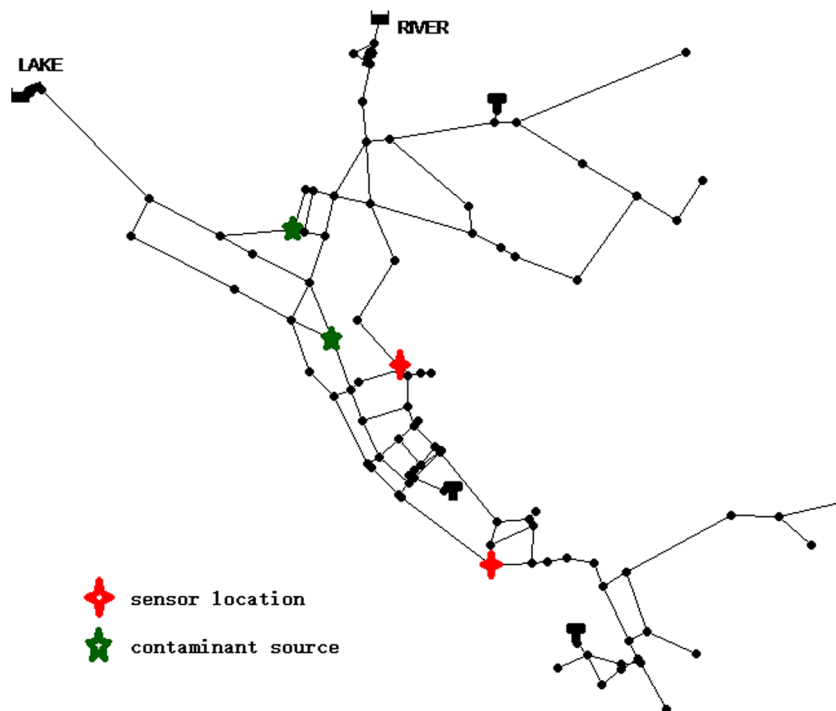
Table 4 Results of Net3_Rossman2000 pipe network

Pipe Network	Scale	Pollution incident	Algorithm	Positioning accuracy	Optimal fitness value	Average fitness value at successful positioning
Net3_Rossman2000	97	16	DMOA	36.67%	0.2324	0.3223
			MSDMOA	66.67%	0.0078	0.1353
		86	DMOA	46.67%	0.1001	0.3313
			MSDMOA	90%	0.0104	0.0979

strategy dynamic multi-mode optimization algorithm based on domain knowledge are as follows:

Step 1 Initialization: k sub-populations are initialized randomly, namely, $p_1, p_2,...p_k$, where the initial value of k depends on the size of the network.

Step 2 Population iteration: Let the current time $t_c=t_0+\Delta t$, and let the current iteration number $g=0$.

Step 2.1 Selection operator: Let $g=g+1$. Each sub-population adopts an improved selection strategy. That is when the potential individual of sub-population Pi dominates, the roulette wheel selection is improved by the prediction error. Otherwise, improved by the size of the individual $|t_0{}^*-t_0|$.

Step 2.2 Crossover operator: Discrete coding (position and time variables) adopts a two-point crossover, while continuous coding (mass vector) adopts real number recombination.

Step 2.3 Mutation operator: The position and quality vectors adopt a single point mutation, while the time variable takes the improved proximity search strategy in 3.3.

Step 2.4 Condition of stopping iteration: when $g \leq g_{Max}$, update the fitness values and elite individuals of each sub-
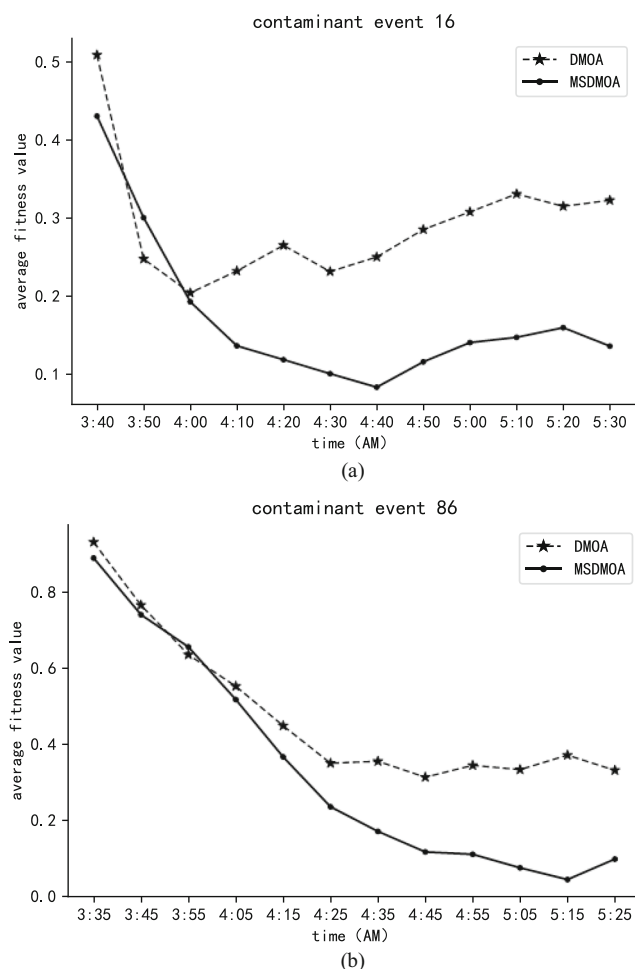
Fig. 7 Time-varying graph of the average fitness value of the Net3_Rossman2000 pipe network
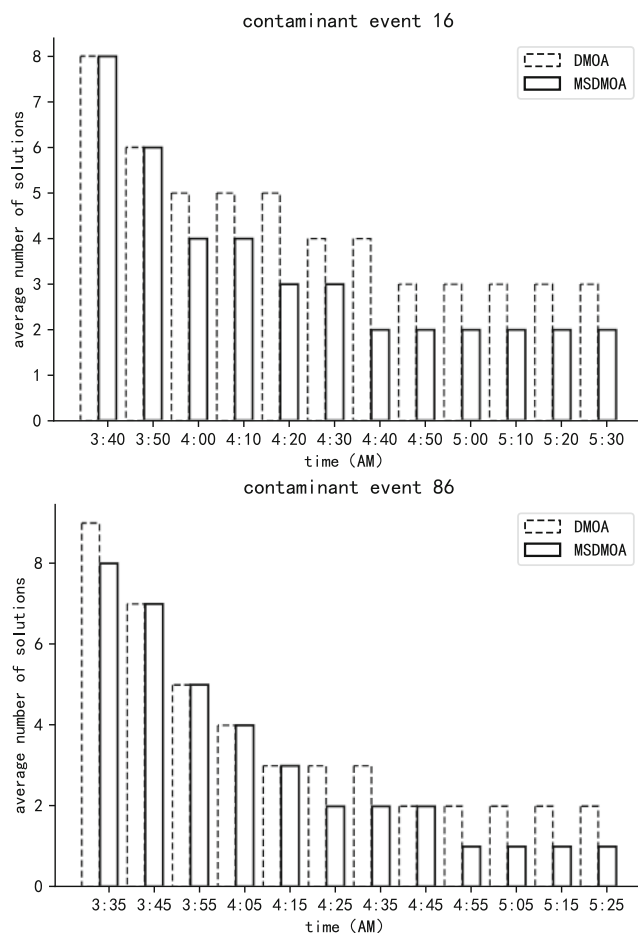
Fig. 8 Time-varying graph of the solutions to the Net3_Rossman2000 pipe network

Table 5 BWSN1_Ostfeld2008 pollution scenes of the pipe network

Scenes Parameters	Node number	Pollution injection start time	Duration(h)	Injected pollution concentration(mg/L)
Pollution scene1	23	2	4	300, 180, 240, and 180
Pollution scene 2	92	2	4	300, 180, 240, and 180

population and continue Step 2.1. Otherwise, continue Step 3.

Step 3 Local search: Considering the time cost of the algorithm, PSO search is performed separately on the quality vectors of the individuals satisfying $|t_0^*-t_0|=0$ after each sub-population iteration.

Step 4 Response to changes: Only the best individuals with the same position in each population are retained after merging sub-populations, while new individuals replace the original ones.

Step 5 Condition of stopping the algorithm: When $t_c-t_0<t_{Max}$, the algorithm ends. Otherwise, Step 2 continues.

Results and discussion

Setting of pipe network and algorithm parameters

This work compared the networks Net3_Rossman2000, BWSN1_Ostfeld2008, and KY5_Jolly2013 to verify the effectiveness of Multi-Strategy Dynamic Multi-Mode Optimization Algorithm (MSDMOA) based on domain knowledge. Table 1 shows the detailed settings of the three experimental pipe networks (Ostfeld et al. 2008). The number

of nodes, pipes, reservoirs, and pools in the innovative pipeline networks determines their scales and complexity, which in turn determines the location and number of sensors.

Assuming that all experiments are single-source injections, the algorithm proposed in this work is compared with DMOA to analyze the results. Table 2 is the parameters of the algorithm proposed in this work.

The parameters of experimental equipment are as follows CPU is Intel® Core™ i5-3230 2.6 GHz, RAM is 8G, Hard Disk is 1 TB, and the computer operating system is win7 32bit.

Algorithm performance indicators

There are few existing evaluation indicators about the real-time location of pollution sources, but this work gives useful performance indicators based on the characteristics of the problem. The purpose of real-time positioning research is to locate the pollution source through an optimization algorithm when the water quality sensor detects the pollution, which minimizes the harm. The following are performance indicators.

Positioning accuracy: Due to the non-uniqueness of the solution to the real-time location of the pollution sources, the best solution of the dynamic algorithm may be one or a group
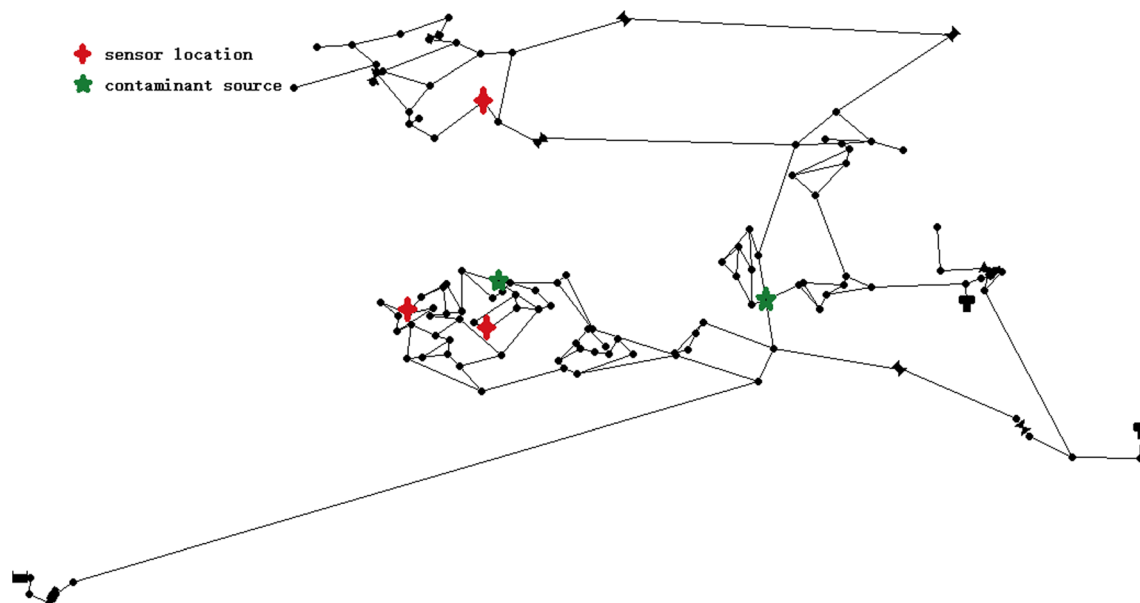
Fig. 9 BWSN1_Ostfeld2008 pipe network

Table 6 Results of BWSN1_Ostfeld2008 pipe network

Pipe network	Scale	Pollution incident	Algorithm	Positioning accuracy	Optimal fitness value	Average fitness value at successful positioning
BWSN1_Ostfeld2008	129	23	DMOA	60.00%	0.0303	0.1937
			MSDMOA	90.00%	0.0008	0.1004
		92	DMOA	53.33%	0.0516	0.3484
			MSDMOA	93.33%	0.0025	0.0744

of pollution events. The pollution position of the best solution is defined as $\{l_1,\ldots, l_N\}$, $N>0$, where N is the optimal solution set size. If the actual pollution source position is $l_{true}\in\{l_1,\ldots, l_N\}$, the optimization locates the pollution source. In the same pollution scene, R times of the same experiments are used to locate $R_{succeed}$ times, and the positioning accuracy rate is $p=R_{succeed}/R$.

Average fitness value for positioning success: The increase in observation time reduces the non-uniqueness of the problem, resulting in different observation time solutions. If the optimal solution set of the dynamic algorithm is N at the current time t_c, then the fitness value is $f_c=\Sigma N_i=l_{ture}f_i/N$. Since the result of inaccurate positioning is meaningless, the average

value f_c of successful positioning in multiple experiments is an important performance indicator.

Algorithm performance discussion

Net3_Rossman2000 pipe network

In this experiment, the network of Net3_Rossman2000 was used. Table 1 shows the parameters of the pipe network, while Fig. 6 shows the topological structure. Table 3 shows the settings of two pollution scenarios, in which the pollution concentration continuously injected is in 1 hour.
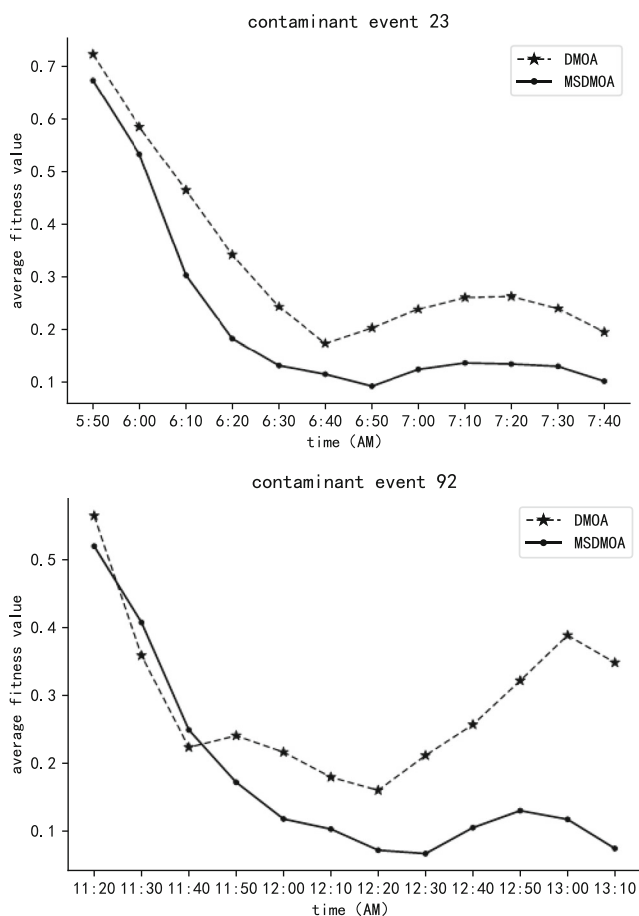
Fig. 10 Time-varying graph of the average fitness value of BWSN1_Ostfeld2008
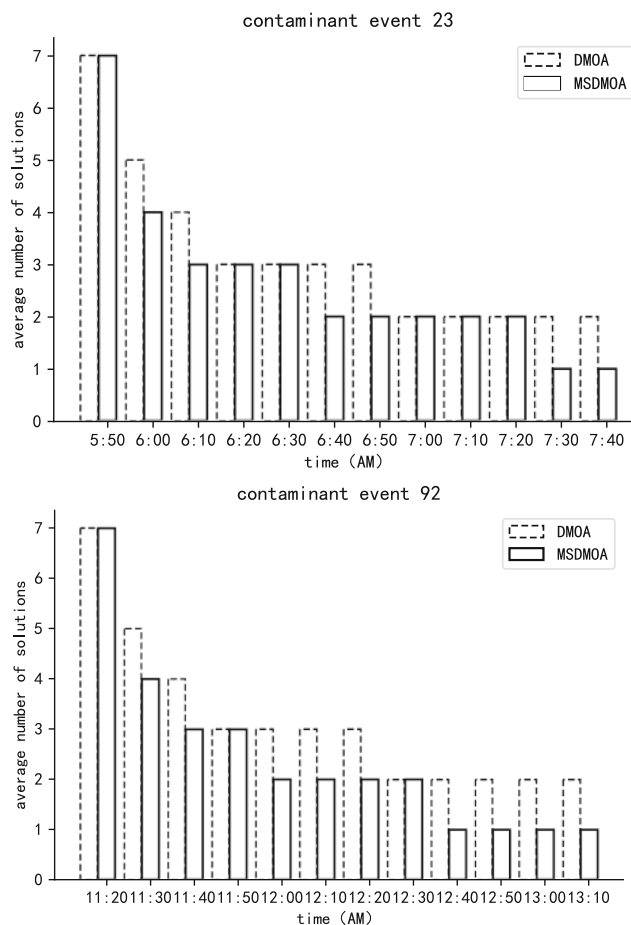
Fig. 11 Time-varying graph of solutions number to BWSN1_Ostfeld2008

**Table 7** Pollution scenarios of KY5_Jolly2013 pipeline network

Scenes Parameters	Node number	Pollution injection start time	Duration (h)	Injected pollution concentration(mg/L)
Pollution scene1	31	2	4	300, 180, 240, and 180
Pollution scene 2	155	2	4	300, 180, 240, and 180

Fig.6 shows the optimization results of the two pollution scenarios after the algorithms MSDMOA and DMOA are conducted 30 times, and 2 h after the sensor first detects the pollution. Table 4 shows the results of the algorithm performance and comparison indicators proposed above.

Table 4 shows that the location accuracy of DMOA based on multi-strategy is better than that of the general, meaning that the ability of MSDMOA to search and jump out of local optimization is better than that of DMOA. Besides, the optimal fitness value and the average fitness value of successful localization of MSDMOA are less than that of DMOA, but the convergence degree is reversed. The above indicates that the start time, duration, and concentration curve of MSDMOA to obtain the optimal solution. However, although the accuracy of MSDMOA is better than DMOA in the pollution incidents at the 16th node, there is still room for improvement. The accuracy of the algorithm improves with the increased positioning time or sensors; however, because of the damage and cost of pollution, it can only do more efficient positioning in limited time and sensors.

Pollution events 16 and 86 were first detected by water quality sensors at 3:30 a.m. and 3:25 a.m., respectively. Figure 7 shows the average real-time location fitness values of the two dynamic algorithms within 2 h after the pollution is first detected, where the first performance of MSDMOA is similar to DMOA. However, in the later stage, MSDMOA's fitness value decreases rapidly, and the searched pollution source is more accurate, which shows a better ability to jump out of local optimization and search. Besides, the pollution concentration detected by the sensor increases sharply at a specific time, which expands the environmental fluctuation. The average fitness curve risen slightly in the later period, indicating that the ability of MSDMOA to resist environmental changes needs to be improved. Figure 8 shows the changes in the number of optimal solutions to pollution sources overtime in 2 h, where MSDMOA performs better than DMOA, proving a strong search ability and fast convergence speed.
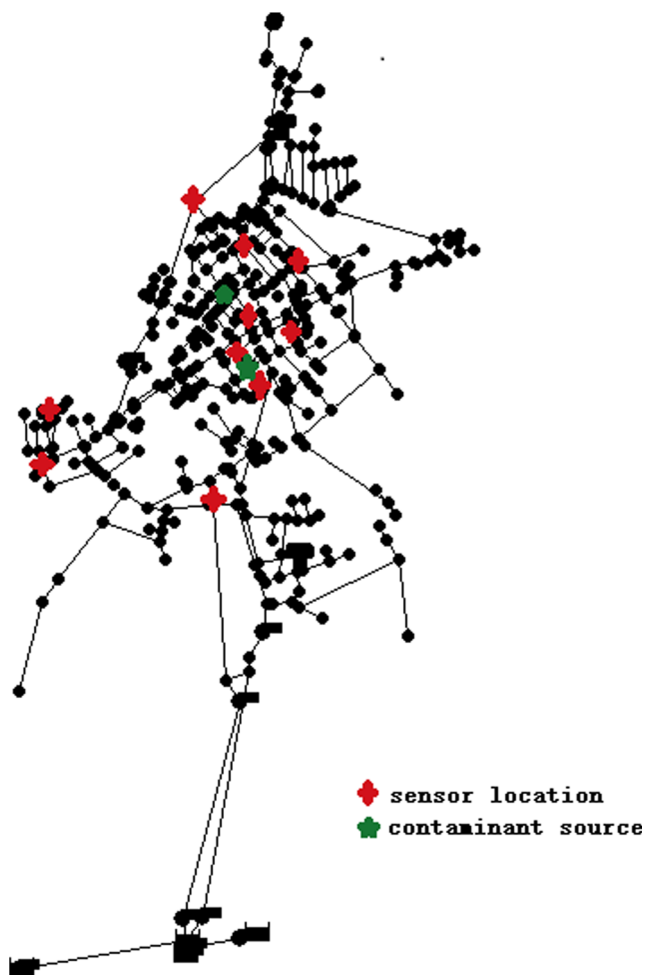
BWSN1_Ostfeld2008 pipe network

In this experiment, the BWSN1_Ostfeld2008 water supply pipe network was used. Table 1 shows the parameters of the pipe network, while Fig. 9 shows the topological structure. Table 5 shows the settings of the two pollution scenarios, where the pollution concentration continuously injected is in a unit of 1 h.

Figure 9 shows the optimization results of two pollution scenarios after MSDMOA and DMOA are conducted 30 times, 2 h after the sensor first detects the pollution. Table 6 shows the results of the algorithm performance and standard comparison indexes.
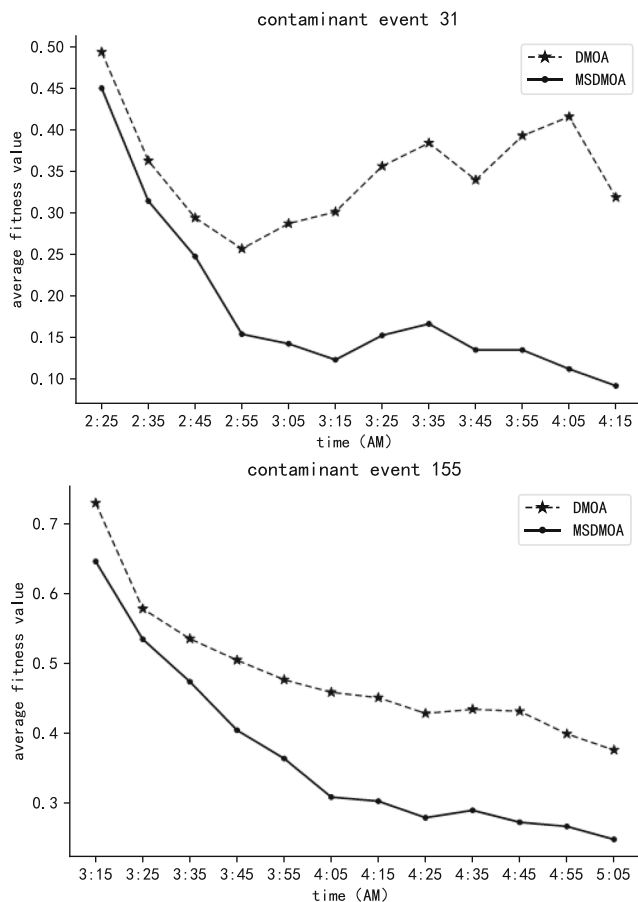
Table 6 shows that MSDMOA is superior to DMOA in positioning accuracy and optimal fitness values, indicating that MSDMOA has more potent abilities to jump out of local optimum and search. The average fitness value of MSDMOA at successful positioning is better than that of DMOA. However, it is opposite in pollution event 23, because jumping out of the local position at the later stage results in insufficient convergence of the individual. In general, MSDMOA has a

Fig. 12 KY5_Jolly2013 pipe network

Table 8   Results of KY5_Jolly2013 pipe network

Pipe network	Scale	Pollution incident	Algorithm	Positioning accuracy	Optimal fitness value	Average fitness value at successful positioning
ky5_Jolly2013	430	31	DMOA	43.33%	0.0967	0.3175
			MSDMOA	83.33%	0.0018	0.0906
		155	DMOA	50.00%	0.2247	0.3753
			MSDMOA	100.00%	0.0706	0.2475

robust search performance, but there are individuals with insufficient convergence.

Pollution events 23 and 92 were first monitored by water quality sensors at 5:40 and 11:10 am, respectively. Figure 10 shows the average fitness value of the two dynamic algorithms for real-time localization within 2 h of the first monitoring of pollution. The fitness value and the upward fluctuation of the MSDMOA curve are small in the later period, showing a stronger ability to adapt to environmental changes. Figure 11 shows that the non-uniqueness degree of the solution to the real-time location of pollution sources decreases with time (the increase of data). The performance of MSDMOA is better than that of DMOA, indicating the strong search ability and fast convergence speed.

ky5_Jolly2013 pipe network

This experiment uses the KY5_Jolly2013 water supply pipe network. Table 1 shows the parameters of the pipe network, while Fig. 12 shows the topological structure. Table 7 shows the settings of the two pollution scenarios tested in the water supply network, where the continuously injected pollution concentration is in 1 hour.

Figure 12 shows the optimization results of the two pollution scenarios after MSDMOA and DMOA are tested 30 times, while the pollution is first detected within 2 h. Table 8 shows the results according to the algorithm performance indicators and common comparison indicators.

Fig. 13 ky5_Jolly2013 average fitness values time-varying graph
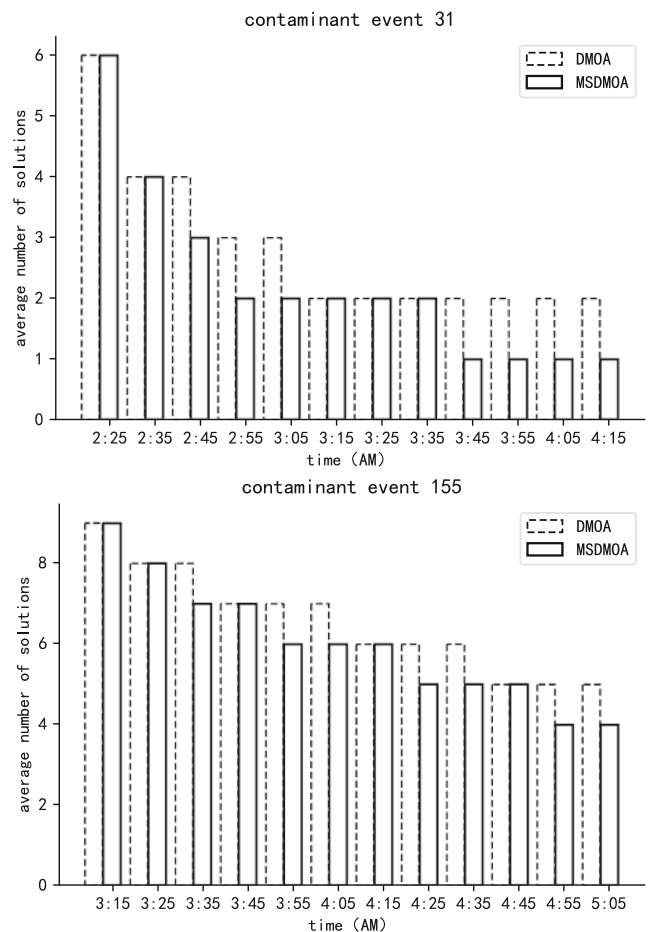
Fig. 14 KY5_Jolly2013 solution numbers time-varying graph

Table 8 shows that the positioning accuracy and optimal fitness value of MSDMOA are better than DMOA, indicating more potent abilities to jump out of local optimum and search. The average fitness value of MSDMOA at successful positioning is better than DMOA, except for pollution incident 155. The non-uniqueness degree of the solution to the pollution event 155 in the pipeline network is higher than that of the first two pipelines, which makes the average fitness value challenging to converge. Therefore, MSDMOA has excellent stability to be used on the big pipe network.

The first time that a water quality sensor detected pollution events 31 and 155 was at 2:15 a.m. and 3:05 a.m., respectively. Figure 13 shows the average fitness values of the two dynamic algorithms for real-time localization within 2 h of the first monitoring of pollution. MSDMOA has a fast convergence speed and a strong ability to adapt to environmental changes. Figure 14 shows the changes in the number of optimal solutions to pollution sources over time within 2 h, where MSDMOA performs better than DMOA. However, the number of the solutions with the two algorithms on the pollution event 155 fail to converge, indicating that the non-uniqueness of the solution increases along with the pipe network scale.

Conclusions

In this work, the simulation optimization method was used to solve the real-time location of drinking water pollution sources. Water quality sensors were only set at critical nodes of the water supply network, resulting in multiple solutions. The gradual increase of the monitoring data led to the optimization of the solution and the reduction of non-uniqueness. Therefore, the dynamic multi-mode optimization algorithm based on domain knowledge was designed to solve multiple solutions. The multi-group merging mechanism prevented the sub-populations from gathering at the same pollution location as the environment changed. Meanwhile, to improve accuracy and avoid falling into local optimality, this work used domain knowledge to enhance the evolution step by a multi-strategy hybrid search. The simulation results in three different scale pipe networks showed that the difficulties of locating real pollution events were various under different scales, topology, and pollution scenarios. The positioning accuracy of the algorithm in this work was higher than that of unmodified DMOA, especially for the excellent ability to jump out of the local optimal. Moreover, when it was converted to the same fitness value, the convergence speed of the proposed algorithm was higher than that of the unmodified DMOA. The algorithm proposed in this paper can locate the pollution source quickly and meet the real-time requirement of the location of drinking water pollution sources. The proposed algorithm can promote the method and technology in the field of pollution sources location, which has important scientific and practical significance.

Authors' contributions Conceptualization: Xuesong Yan, Chengyu Hu
Methodology: Xuesong Yan, Bin Wu
Formal analysis and investigation: Zhengchen Zhou
Writing—original draft preparation: Xuesong Yan, Zhengchen Zhou
Writing—review and editing: Bin Wu
Funding acquisition: Chengyu Hu
Resources: Chengyu Hu
Supervision: Chengyu Hu.

Data Availability All data generated or analysed during this study are included in this published article.

Declarations

Ethical approval and consent to participate Not Applicable

Consent for publication On behalf of all authors, I declare that this manuscript is the authors' original work and has not been published nor has it been submitted simultaneously elsewhere and all presentations of case reports must have consent for publication.

Competing interests The authors declared that they have no conflicts of interest to this work.

References

Avi O, Elad S (2005) Optimal early warning monitoring system layout for water networks security: inclusion of sensors sensitivities and response delays. Civ Eng Environ Syst 22(3):151–169. https://doi.org/10.1080/10286600500308144

Avi O, James GU, Elad S, Berry JW, Hart WE, Phillips CA, Watson JP, Dorini G, Jonkergouw P, Kapelan Z et al (2008) The battle of the

water sensor networks (bwsn): A design challenge for engineers and algorithms. J Water Resour Plan Manag 134(6):556–568. https://doi.org/10.1061/(ASCE)0733-9496(2008)134:6(556)

Costa DM, Melo LF, Martins FG (2013) Localization of contamination sources in drinking water distribution systems: a method based on successive positive readings of sensors. Water Resour Manag 27(13):4623–4635. https://doi.org/10.1007/s11269-013-0431-z

Cruz C, González JR, Pelta DA (2011) Optimization in dynamic environments: a survey on problems, methods and measures. Soft Comput 15(7):1427–1448. https://doi.org/10.1007/s00500-010-0681-0

De S, Annamaria E, Feng S, James GU (2010) Real-time identification of possible contamination sources using network backtracking methods. J Water Resour Plan Manag 136(4):444–453. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000050

Gong J, Yan X, Hu C, Wu Q (2019) Collaborative based pollution sources identification algorithm in water supply sensor networks. Desalin Water Treat 168:123–135. https://doi.org/10.5004/dwt.2019.24204

Gu X, Zhang Q, Li J, Vijay PS, Liu J, Sun P, He C, Wu J (2019a) Intensification and expansion of soil moisture drying in warm season over eurasia under global warming. J Geophys Res-Atmos 124: 3765–3782. https://doi.org/10.1029/2018JD029776

Gu X, Zhang Q, Li J, Vijay PS, Liu J, Sun P, Cheng C (2019b) Attribution of global soil moisture drying to human activities: A quantitative viewpoint. Geophys Res Lett 46:2573–2582. https://doi.org/10.1029/2018GL080768

Gu X, Zhang Q, Li J, Vijay PS, Sun P (2019c) Impact of urbanization on nonstationarity of annual and seasonal precipitation extremes in China. J Hydrol 575:638–655. https://doi.org/10.1016/j.jhydrol.2019.05.070

Gu X, Zhang Q, Vijay PS, Zheng Y (2017) Changes in magnitude and frequency of heavy precipitation across China and its potential links to summer temperature. J Hydrol 547:718–731. https://doi.org/10.1016/j.jhydrol.2017.02.041

Guan J, Mustafa MA et al (2006) Identification of contaminant sources in water distribution systems using simulation – optimization method: case study. J Water Resour Plan Manag 132(4):252–262. https://doi.org/10.1061/(ASCE)0733-9496(2006)132:4(252)

Hu C, Dai L, Yan X, Gong W, Liu X, Wang L (2020) Modified NSGA-III for sensor placement in water distribution system. Inf Sci 509: 488–500. https://doi.org/10.1016/j.ins.2018.06.055

Hu C, Li M, Zeng D, Guo S (2018a) A survey on sensor placement for contamination detection in water distribution systems. Wirel Netw 24(2):647–661. https://doi.org/10.1007/s11276-016-1358-0

Hu C, Shu X, Yan X, Zeng D, Gong W, Wang L (2018b) Inline wireless mobile sensors and fog nodes placement for leakage detection in water distribution systems. Software Pract Exper 39:1–16. https://doi.org/10.1002/spe.2631

Hu C, Zhao J, Yan X (2015) A map reduce based parallel niche genetic algorithm for contaminant source identification in water distribution network. Ad Hoc Netw 35:116–126. https://doi.org/10.1016/j.adhoc.2015.07.011

Huang J, Edward AM (2009) Data mining to identify contaminant event locations in water distribution systems. J Water Resour Plan Manag 135(6):466–474. https://doi.org/10.1061/(ASCE)0733-9496(2009)135:6(466)

Jin Y, Branke J (2005) Evolutionary optimization in uncertain environments– a survey. IEEE Trans Evol Comput 9(3):303–317. https://doi.org/10.1109/TEVC.2005.846356

Kuter S, Akyürek Z, Weber GW (2018) Recent contributions to climate change and water resource management by applying novel analytics on satellite data. EWG-ORD 2018 Workshop OR for Sustainable Development: Establishing Policy and Measuring Goal Attainment, Complutense University of Madrid, Spain, July 5-7, 2018. pp. 1-11

Kuter S, Akyürek Z, Weber GW, Gütmen S (2019) Advancing Water-Resource Management: Application of Novel OR-Analytics - Snow classification on Sentinel-2 imagery by MARS. EWG-ORD 2019 Workshop Renewable: Energy, Health & Sustainability. Dublin, Ireland, June, 2019, pp1-10.

Laird CD, Biegler LT, van Bloemen Waanders BG, Bartlett RA (2005) Contamination Source Determination for Water Networks. J Water Resour Plan Manag 131(2):125–134. https://doi.org/10.1061/(ASCE)0733-9496(2005)131:2(125)

Laird CD, Lorenz TB, Bart GV (2006) Mixed-integer approach for obtaining unique solutions in source inversion of water networks. J Water Resour Plan Manag 132(4):242–251. https://doi.org/10.1061/(ASCE)0733-9496(2006)132:4(242)

Li S, Gong W, Yan X, Hu C, Bai D, Wang L (2019) Parameter estimation of photovoltaic models with memetic adaptive differential evolution. Sol Energy 2019(190):465–474. https://doi.org/10.1016/j.solener.2019.08.022

Liu L, Ranjithan SR, Mahinthakumar G (2011) Contamination source identification in water distribution systems using an adaptive dynamic optimization procedure. J Water Resour Plan Manag 137(2):183–192. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000104

Lv M, et al (2010) Notice of retraction investigation on backward tracking of contamination sources in water supply systems-case study. In 2010 The 2nd Conference on Environmental Science and Information Application Technology 484-487. doi:https://doi.org/10.1109/ESIAT.2010.5568348

Nguyen TT, Yang S, Branke J (2012) Evolutionary dynamic optimization: a survey of the state of the art. Swarm Evolut Comput 6:1–24. https://doi.org/10.1016/j.swevo.2012.05.001

Özmen A (2016) Introduction. In: robust optimization of spline models and complex regulatory networks. Contributions to Management Science. Springer, Cham. https://doi.org/10.1007/978-3-319-30800-5_1

Perelman L, Avi O (2013) Bayesian Networks for Source Intrusion Detection. J Water Resour Plan Manag 139(4):426–432. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000288

Roche R, Yalcinkaya F (2019) Electrospun polyacrylonitrile nanofibrous membranes for point-of-use water and air cleaning. Chemistry Open 8(1):97–103. https://doi.org/10.1002/open.201800267

Rossman LA (2000) Epanet 2 users manual, US environmental protection agency. Water Supply and Water Resources Division , National Risk Management Research Laboratory, Cincinnati, p 45268

Seth A, Klise KA, Siirola JD, Haxton T, Laird CD (2016) Testing Contamination source identification methods for water distribution networks. J Water Resour Plan Manag 142(4):04016001. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000619

Shang F, James GU, Marios MP (2002) Particle backtracking algorithm for water distribution system analysis. J Environ Eng 128(5):441–450. https://doi.org/10.1061/(ASCE)0733-9372(2002)128:5(441)

Wang C, Zhou S (2017) Contamination source identification based on sequential Bayesian approach for water distribution network with stochastic demands. IISE Trans 49(9):899–910. https://doi.org/10.1080/24725854.2017.1315782

Wagner DE, Roseanna MN, Cody C (2015) Adjoint-based probabilistic source characterization in water-distribution systems with transient flows and imperfect sensors. Water Res Plan Man 141(9):04015003. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000508

Wang F, Li Y, Zhou A, Tang K (2020a) An Estimation of Distribution Algorithm for Mixed-Variable Newsvendor Problems. IEEE Trans Evol Comput 24(3):479–493. https://doi.org/10.1109/TEVC.2019.2932624

Wang F, Li Y, Liao F, Yan H (2020b) An ensemble learning based prediction strategy for dynamic multi-objective optimization. Appl Soft Comput 96:106592. https://doi.org/10.1016/j.asoc.2020.106592

Wang F, Zhang H, Zhou A (2021) A particle swarm optimization algorithm for mixed-variable optimization problems. Swarm Evol Comput 60:100808. https://doi.org/10.1016/j.swevo.2020.100808

Wang H, Harrison KW (2014) Improving Efficiency of the Bayesian Approach to Water Distribution Contaminant Source Characterization with Support Vector Regression. J Water Resour Plan Manag 140(1):3–11. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000323

Wang H, Xin J (2013) Characterization of groundwater contaminant source using Bayesian method. Stoch Env Res Risk A 27(4):867–876. https://doi.org/10.1007/s00477-012-0622-9

Yan X, Gong W, Wu Q (2017a) Contaminant source identification of water distribution networks using cultural algorithm. Concurr Comp-Pract E 29(24):1–11. https://doi.org/10.1002/cpe.4230

Yan X, Hu C, Sheng VS (2020) Data-driven pollution source location algorithm in water quality monitoring sensor networks. Int J Bio-Inspir Com 15(3):171–180. https://doi.org/10.1504/IJBIC.2020.107474

Yan X, Li T, Hu C, Wu Q (2019b) Real-time localization of pollution source for urban water supply network in emergencies. Clust Comput 22:5941–5954. https://doi.org/10.1007/s10586-018-1725-y

Yan X, Sun J, Hu C (2017b) Research on contaminant sources identification of uncertainty water demand using genetic algorithm. Clust Comput 20(2):1007–1016. https://doi.org/10.1007/s10586-017-0787-6

Yan X, Yang K, Hu C, Gong W (2018) Pollution source positioning in a water supply network based on expensive optimization. Desalin Water Treat 110:308–318. https://doi.org/10.5004/dwt.2018.22330

Yan X, Zhao J, Hu C, Zeng D (2019c) Multimodal optimization problem in contamination source determination of water supply networks. Swarm Evol Comput 47:66–71. https://doi.org/10.1016/j.swevo.2017.05.010

Yan X, Zhu Z, Li T (2019a) Pollution source localization in an urban water supply network based on dynamic water demand. Environ Sci Pollut Res 26(18):17901–17910. https://doi.org/10.1007/s11356-017-0516-y

Yang P, Tang K, Yao X (2018) Turning High-dimensional Optimization into Computationally Expensive Optimization. IEEE Trans Evol Comput 22(1):143–156. https://doi.org/10.1109/TEVC.2017.2672689

Yang S, Li C (2010) A clustering particle swarm optimizer for locating and tracking multiple optima in dynamic environments. IEEE Ttrans Evolut Comput 14(6):959–974. https://doi.org/10.1109/TEVC.2010.2046667

Yerlikaya-Zkurt F, Taylan P (2020) New computational methods for classification problems in the existence of outliers based on conic quadratic optimization. Commun Stat 49(3/4):753–770. https://doi.org/10.1080/03610918.2019.1661477

Zechman EM, Ranjithan SR (2009) Evolutionary Computation-Based Methods for Characterizing Contaminant Sources in a Water Distribution System. J Water Resour Plan Manag 135(5):334–343. https://doi.org/10.1061/(ASCE)0733-9496(2009)135:5(334)