Neural Computing and Applications (2021) 33:209–222
https://doi.org/10.1007/s00521-020-05000-8

ORIGINAL ARTICLE

Pollution source intelligent location algorithm in water quality sensor networks

Xuesong Yan[1,2] · Jingyu Gong[1] · Qinghua Wu[3]

Received: 7 February 2020 / Accepted: 2 May 2020 / Published online: 15 May 2020

Abstract
Water is the source of human life and water pollution is becoming more and more serious with the development of cities. The supervision and treatment of water resources have become a big problem of urban development. Water quality monitoring is not timely, flood warning is not timely is directly related to the livelihood of the people. And the development of smart water utilities can solve problems timely and accurately. By placing water quality sensors in the urban water supply network, real-time monitoring of water quality can be performed to prevent incidents of drinking water pollution. After an incident of drinking water pollution occurs, reverse locating the pollution source through the information detected by the water quality sensors represents a challenging problem because in the actual water supply network, the direction and speed of the water flow will change with the water demand of the residents, thus leading to uncertainty in this problem. In conventional studies of pollution source location problems, it is often assumed that the water demand is fixed. However, due to the variability of the water demand of residents, this problem is actually a dynamic change problem and thus can be considered as a dynamic optimization problem. In this study, a Poisson distribution model was used to simulate the change of water demand among urban residents. On this basis, we proposed an improved genetic algorithm to solve the pollution source location problem and implemented two different water supply networks to perform the simulation experiments, which could accurately locate the pollution sources. The simulation results were compared with the standard genetic algorithm to verify the accuracy and robustness of the proposed algorithm.

Keywords Pollution source location · Water quality monitoring · Sensor networks · Poisson distribution · Simulation optimization · Genetic algorithm

1 Introduction

Smart city is inseparable from the support of smart water utilities. The core issue of smart water utilities is the big data processing, which determines the degree of smart. Smart water utilities by counting instrument, the wireless network, water quality on-line monitoring equipment such as hydraulic pressure gauge real-time running status of urban water supply and drainage system, perception and adopt the way of visualization of organic integration of water management and water supply and drainage facilities, form the "Internet of things" urban water affair, and massive amounts of water can be timely analysis and processing of information, and make corresponding processing results auxiliary decision-making recommendations, in a more elaborate and dynamic way of the water management system of the whole process of production, management and service, so as to achieve the status of "smart".

Urban water supply networks are large and open and thus vulnerable to destruction caused by accidents or deliberate pollution incidents. In order to preventing significant disasters and losses caused by drinking water pollution incidents from impacting society and residents, it is necessary to set up the sensors at the water sources or the

✉ Qinghua Wu
wuqinghua@sina.com

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

[2] Hubei Key Laboratory of Intelligent Geo-Information Processing, Wuhan 430074, China

[3] Faculty of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China

key nodes in the water supply networks for real-time monitoring [1–4]. With the collected information by the water quality monitoring sensors, the pollution source can be located and the precise position of the pollution source as well as the injection time can be determined; subsequently, the water supply network can be partially blocked, thus reducing the economic losses and social impacts of the pollution. Therefore, it is of great practical significance to carry out research on pollution source location problems. Locating the pollution source is the technical premise of real-time monitoring and early warning for drinking water safety. Based on the detected information from water quality sensors, the possible location, time, current state and diffusion trend of the pollutant can be obtained according to the water quality monitoring and information feedback, thus allowing for the pollution to be treated in a timely manner and reducing the associated harm.

The research of pollution source location problem primarily used particle trace-back method in early stage, and apportioned each node's state according to the information from the monitoring sensors in the former time. Shang et al. presented particle trace-back in accordance with the time sequence. The in reverse time and the pollutants as a particle from the source node which was detected to be polluted, and then reverse traced the pollution source location [5]. Laird et al. used the source position trace-back algorithm to identify the multiple pollution sources [6]. De Sanctis et al. used particle trace-back algorithm, which real-time identified possible pollution sources by comparing the hydraulic and water quality consistency, this algorithm has achieved good results for the pollution sources identification problem [7]. Costa et al. proposed the methodology, in his methodology the information detected by successive positive readings of sensors and the experimental results are not so good [8].

Using the machine learning algorithms to sort the network nodes with pollution possibilities, thus the nodes that are most susceptible by pollution incidents is given, which is a contaminant source identification method used by many scholars. Huang et al. developed a data mining method and use this method in conjunction with a maximum likelihood provides the means to identify the location and time of an intrusion event. This algorithm can find the pollution sources location soon, but cannot find the multi-sources [9]. Perelman et al. used Bayesian networks (BNs) statistics to estimate the possibility of the pollution injection and its propagation in the water supply networks. The authors developed the clustering method and applied it to formulate a simplified expression of the water supply networks with nodal connectivity properties [10]. Wang et al. investigate the support vector regression (SVR) in order to speed the likelihood evaluation of the Markov Chain Monte Carlo (MCMC) methods, because the MCMC for Bayesian analyses allow for the characterization of the uncertainty in the pollution event profile, but the MCMC implementation is most computationally expensive [11]. Wang et al. present a dynamic framework combined with Monte Carlo simulation model and the Bayesian approach, which is to couple with an application of a groundwater modelling scheme in pollution source identification of groundwater [12]. Machine learning as a predictive method based on probability.

In recent years, locating pollution sources in a water supply network based on simulation–optimization methods has become a hot research spot. Simulation–optimization methods mainly find the pollution node with the smallest error by comparing information from the monitoring point (the node where the water quality sensor is set) with the concentration information of various pollution events at the monitoring point to determine the source of the pollution (including the attributes of injection position, time, etc.). The problem of pollution source location is based on the assumption that the pollutant is injected into any node of the water supply networks, and it is constructed through simulating the cumulative concentration of the monitoring points. Supposing a forward simulation model, the known pollution source information can be used to simulate the cumulative pollution concentration of the monitoring points. The pollution source location problem is a reverse problem of the above process, and finding the source of pollution with the minimal difference value between the simulated cumulative concentration and the actual detected cumulative concentration at the monitoring point is an optimization problem. Simulation–optimization methods are problem solving approaches that use an optimization algorithm driven simulator (i.e. EPANET). Optimization algorithms have strong local and global optimisation ability and good convergence, the unique advantages and mechanisms of optimisation algorithms have attracted the attention of scholars and have been successfully applied in many fields [13–37].

In 2005, by matching monitoring point information with pollution events in a random pollution matrix, the location of the pollution, the injection speed, etc. were backtracked [38]. Guan proposed a simulation–optimization method for the nonlinear problem of pollution source location. By continuously reading the sensors data to optimize the predictions and correct the pollution source, the pollution source and pollutant release history could be finally identified [39]. Preis and Ostfeld present a simple, straightforward genetic algorithm (GA) method to enhance the security of water supply networks for pollution source identification and the limitations of the related previous works coupled with the proposed method [40]. Preis and Ostfeld also developed the pollution source detection model and proposed an improved GA method for pollution

source detection. The proposed algorithm's performance is demonstrated using two application examples, demonstrating the tradeoff between sensor types [41]. Zechman et al. proposed an evolutionary strategy-based method to find the most suitable pollution source using global heuristic search algorithm based on monitoring point information [42]. Vankayala et al. think the water demand is uncertainty in the water distribution system, so the pollution source identification problem maybe uncertainty too. The authors using GA as the optimizer and EPANET tool as the simulator to solve the pollution source location problem. In this method, by minimizing the difference between the simulation and observation concentrations at the sensor nodes to find the pollution source location and concentration [43]. In 2010, a simulation–optimization method was also applied to track pollution sources. Sodium hypochlorite instead of a pollution source was used in that study for the simulation, and then different input parameters were compared [44]. Drake and Zechman think the pollution source maybe multiple in water distribution system. The author proposed Niched Co-Evolution Strategies (NCES) to prevent the error location of a pollution source and improve response strategy [45]. Liu et al. used a self-adaptive dynamic optimization algorithm to location the pollution source with searching for the pollution source features (start time, position, and release history) and by adding new sensors continuously to get the only optimal solution through slow convergence [46]. Hu et al. proposed a mapreduce-based parallel niche GA for the pollution source location problem, they used the niche GA as the optimizer and EPANET software as the simulation tool [47]. Yan et al. have used intelligent optimization algorithms to study the pollution source location problem and achieved a series of results [48–54].

Currently, most related research assumes that the urban residents' water demand is a known constant input, whereas in reality, due to dynamic changes of the water demand, the model has dynamic variability at the input. To increase the accuracy of the positioning model, the change in water demand needs to be defined as an unknown input during the modelling process. This study used the Poisson distribution model to simulate the dynamic change of urban residents' water demand, which is used as the input for the problem model. This study also proposed an improved GA as the optimization algorithm for the pollution source location problem under variety water demand. Finally, the accuracy and robustness of the propose algorithm is verified via simulation experiments.

2 Pollution source location problem model

Pollution of the water supply network associated with the injection of pollutants is a complicated and random process involving the type of pollutants, the injection point, the injection amount, the injection time, the injection duration, etc. In the process of establishing actual modelling, it is generally assumed that the pollutant is a conserved substance that does not react with other substances in the pipe network and only gradually dilutes with the water flow. Moreover, the pollutant running speed is considered consistent with the average flow velocity of the pipe segment. Pollutants are injected into the pipe network only through the nodes, and the probability of injection for all nodes is either equal or set according to the population covered by the nodes. The injection from the pipe segment to the pipe network is not considered, and the injection of important facilities, such as water sources and water towers, is not considered. The sensors can monitor the concentration of any pollutant in real time. When the concentration of a pollutant exceeds the set threshold, then this pollution event can be detected at the monitoring point. The injection of the pollutant may occur at any node and any time, and intermittent injections are not considered for the injection point, injection amount, injection time, and injection duration of the pollutant. The research object of this paper is the case of single node pollution injection and the simulator is EPANET, EPANET simulators can simulate the characteristics of water supply network and the simulation process of pollutant intrusion, and can track the chemical concentration, pipeline flow and node pressure in the entire water supply network [55].

From the perspective of optimization, the pollution source location problem is to attain the minimum difference of the cumulative concentration by simulating and the actual cumulative concentration of the pollution event at the monitoring sensor by detecting. When the variance is 0 or less than a given threshold ε, the pollution event is considered to be an actually occurring pollutant injection event. Through the analysis of the pollution source location problem, we found that the pollution source location problem is uncertain. The uncertainty of the water supply network pollution source location system includes three aspects: (1) observation and sensor measurement errors due to sampling deviation and detection error; (2) model errors due to simplification and incorrect assumptions in the water quality model; and (3) the uncertain water demand due to the unknown real-time water demand fluctuations, which is mainly caused by inherent changes in water consumption levels and the water demand of the consumers at the node. Due to the above uncertainties, the problem of locating the pollution source in a water supply network cannot be

solved using a deterministic method (i.e. assuming the system operates under fully known conditions). This study did not consider observation and model errors. The uncertainty in the water demand for residents is because the actual water consumption of residents is uncontrollable. The data of when and how much water is needed cannot be fixed. When the pipe network is polluted, the uncertain water use of the residents will result in different water flow paths, thus leading to changes in the flowing path and concentration of the pollutants and causing uncertain characteristics in the location pollution source problem.

In this study, the urban residents' water demand data are uncertain, and the pollution injection start time and duration are assumed that the known, that is, the pollution source location problem can be simplify to locating the pollution source position and the cumulative concentration of the injected pollutant. The pollution source location problem model is shown in Eq. (1).

$$\begin{aligned} &\text{Find}\{L, C_0\} \\ &\text{minmize } F = \max_{k=1,\ldots,N_s} \left\{ \sum_{t=1}^{T} \left(C_k^{\text{obs}}(t) - C_k^* L, C_0, t \right)^2 \right\} \end{aligned} \tag{1}$$

where F refers to the prediction error, L refers to the node of location point of the pollution source, C_0 refers to the concentration of the pollutant in the known pollution injection time, t refers to the current time step, T refers to the total number of sampling time steps, k refers to the position of a certain sensor, N_S refers to the total number of sensors, C_k^{obs} refers to the true cumulative concentration at the sensor position k, and $C_k^*(L, C_0)$ refers to the cumulative simulated concentration observed at the sensor of position k. Here, the dynamic change in water demand is the cumulative simulated concentration C_k^* applied to the calculation rather than the true cumulative concentration C_k^{obs}, and the threshold is an empirical value.

3 Resident water demand model based on the Poisson distribution

3.1 Resident water demand model

Residents' water demand has a certain regularity. Vankayala et al. in Ref. [43] mentioned that to obtain a time-varying, randomized water demand model for residents, Buchberger and Wells statistically analysed the 30-day water demand data of 21 households in Milford Township in Ohio (USA) in 1996 to determine the regularity of water demand based on the statistics of the observed data. In this study, the water demand data of Milford Township was used for the experimental testing. The hourly water weighted factor can be obtained from the daily total water demand data. The hourly total water demand data can be obtained using the weighted factor and the daily basic water demand data in a given pipe network. The total water demand per hour is then divided into nodes according to the weighted factor of each node. Using the water demand data of Milford Township from Ref. [43], we can get the mean hourly water demand with MATLAB 2014a as shown in Fig. 1. Although the water demand of residents is dynamic, it follows a certain rule. We can give the instantaneous residents' water demand of any node in a water supply networks using the Poisson distribution model, so the Poisson distribution model is used to simulate the residents' water demand rule in this study.

3.2 Poisson distribution model

The Poisson distribution model [56–58] is a discrete probability distribution commonly found in statistics and probability. The Poisson distribution model can give the instantaneous residents' water demand of any water supply networks node. According to the water consumption of residents of Milford over a month, the instantaneous time and spatial variability of the water flow were quantified. In this study, we used Poisson rectangular pulse process to characterize the duration, frequency and intensity of residents' water demand with a single household. These characteristics can be modelled as a non-homogeneous Poisson rectangular pulse process (PRP). Assuming the frequency of water demand follows the Poisson distribution, i.e. an arrival process with a rate parameter that varies with time. When used in a single household, the water use frequency is approximated as a rectangular pulse with random time and random intensity as shown in Fig. 2. According to the Poisson distribution model, multiple pulses cannot occur at the same starting time. Due to the limited duration of each pulse, two or more pulses in different start times maybe overlap in a limited time. When this happens, the total amount of water used by multiple residents in their homes is the sum of the individual pulse intensities.

When a server is busy, the water use demand is assumed to be a rectangular pulse of random duration and random intensity. Assume that Y_1 is the water usage intensity of the Resident 1 in the water server, regardless of the type or amount of the water used. When the server is busy, Y_1 is the mean and difference of an independent, positive continuous and identically distributed random variable; and when the server is idle, the intensity of water use is zero. In any instant use of water, one or more servers can be very busy. To consider these possibilities, let Y_1 representing the sum of the water pulse intensities at the same time for Resident 1 as shown in Eq. (2).

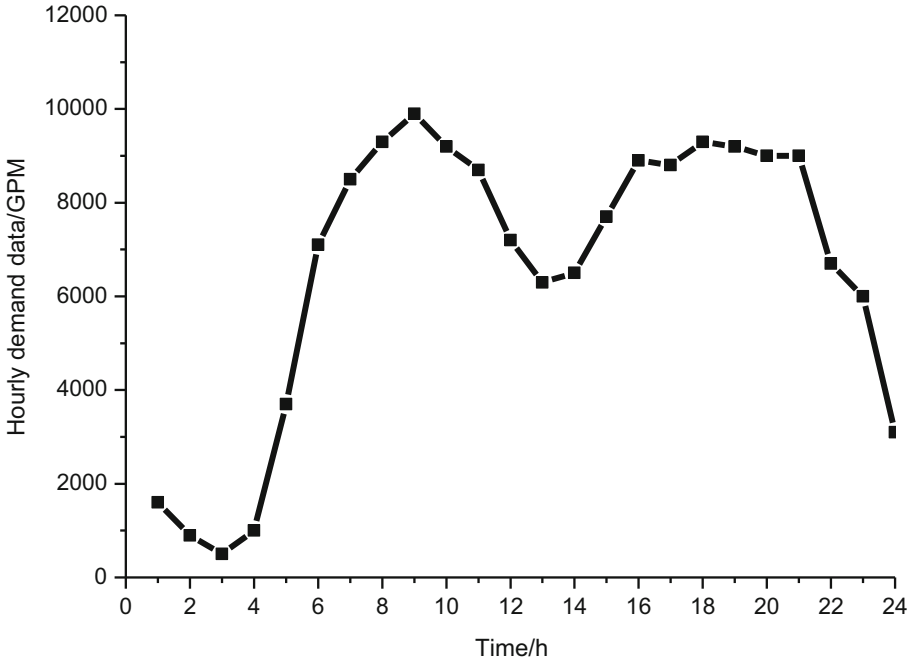

Fig. 1 Mean hourly water demand of Milford Township

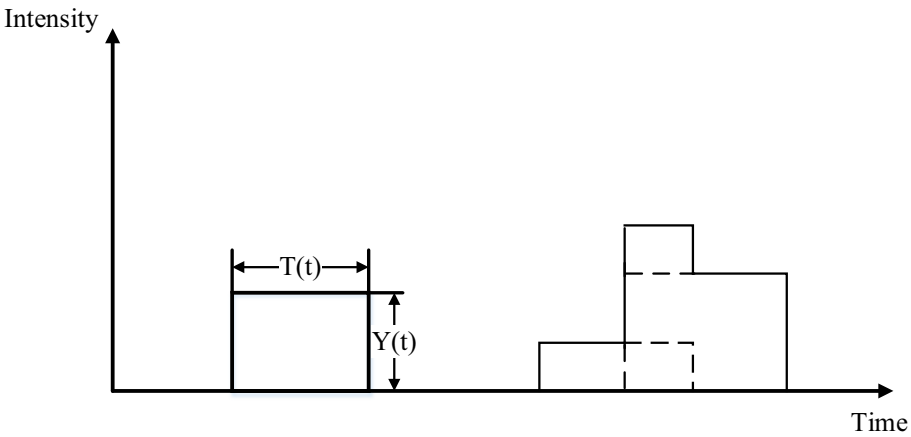

Fig. 2 Rectangular pulse of water demand

$$Y_1(k) = \sum_{i=0}^{k} (Y_1)_i; \quad k = 0, 1, 2, \ldots \tag{2}$$

where $(Y_1)_0 = Y_1(0) = 0$ because when all the water servers are idle, there is no water flow. Since k is a given fixed sum and the water use intensity is assumed to be independent of each other, the mean of $Y_1(k)$ is $k\alpha_1$ and the variance is $k\beta_1^2$. The cumulative distribution function is shown in Eq. (3).

$$G_1(q, k) = P[Y_1(k) \leq q] \tag{3}$$

Equation (3) represents the multiple convolution of the cumulative distribution function for each intensity. When all water servers are idle, no water is used, or $G_1(q, 0) = 1, q \geq 0$; whereas if at least one water server is busy, the water usage must exceed zero, or $G_1(0, k) = 0, k \geq 1$. Where $\lim_{t->\infty} P_1(k, t) = P_1(k) = P(K_1 = k)$ is recorded as the equilibrium probability, which is obtained based on the Erlang loss Eq. (4):

$$P_1(k) = P_1(0)\left(\frac{\lambda_1}{\mu_1}\right)^k \frac{1}{k!}; \quad k = 0, 1, 2, \ldots, m \tag{4}$$

Because the probability is constant $\sum_k P_1(k) = 1, k = 0, 1, \ldots, m$, the probability that Resident 1 has no busy server can be calculated by Eq. (5).

$$P_1(0) = \frac{1}{\sum_{k=0}^{m} \left(\frac{\lambda_1}{\mu_1}\right)^k \frac{1}{k!}} = \exp\left(-\frac{\lambda_1}{\mu_1}\right) \tag{5}$$

The dimensionless term $\frac{\lambda_1}{\mu_1}$ is often used as a factor of utilization rate in queuing theory and gives a measure of a server's average use, which is represented as $\rho_1 = \lambda_1/\mu_1$. Equation (5) is substituted into Eq. (4) to obtain Eq. (6).

$$P_1(k) = \frac{\rho_1^k e^{P_1}}{k!}; \quad k = 0, 1, 2, \ldots \tag{6}$$

Assume that Q_1 is the flow of water through the reservoir in the amount of water used by Resident 1. Because the flow is an aggregate of K_1 busy server responses, Q_1 can be described via Eq. (7).

$$Q_1 = \sum_{i=0}^{K_1} (Y_1)_i; K_1 \geq 0 \tag{7}$$

In Eq. (6), K_1 is fixed and the busy servers' number K_1 here is random. Because K_1 obeys the Poisson distribution and Y_1 is an independent random variable with identically distributed, Q_1 becomes a complex Poisson distribution process.

The Q_1 at time n is presented as $(Q_1|K_1 = k) = Y_1(k)$, and then the number of busy servers used by Resident 1 is adjusted to $E(Q_1^n) = \sum_{k=0}^{\infty} E[Y_1^n(k)] \cdot P$.

Let $n = 1$, then the average value of Q_1 is obtained by combining Eq. (7) as shown in Eq. (8).

$$E(Q_1) = \sum_{k=0}^{\infty} (k\alpha_1) \cdot \frac{\rho_1^k e^{P_1}}{k!} = \rho_1 \alpha_1 \tag{8}$$

The variance of Q_1 can be obtained using the mean value of Q_1 as shown in Eq. (9).

$$\mathrm{var}(Q_1) = E(Q_1^2) - E^2(Q_1) = \rho_1(\alpha_1^2 + \beta_1^2) \tag{9}$$

Combined with the mean and variance of the Poisson process [Eqs. (8) and (9)], the model for a single household water demand can be obtained, and it can be used to obtain the data of a single-family residential water use at any time point.

For multiple household water demand, the water demand of a single household,Q_j, is expanded to represent the water demand of the resident j. The flow of n residents through reservoir n in K_n^* busy servers is: $Q_n^* = \sum_{i=0}^{n} Q_j = \sum_{i=0}^{K_n^*} (Y_N^*)_i$; $K_n^* \geq 0$. The mean and variance are presented in Eqs. (10) and (11), respectively.

$$E(Q_n^*) = E\left[\sum_{i=1}^{K_n} (Y_N^*)_i\right] = E(K_n^*)E(Y_N^*) = \sum_{j=1}^{n} \rho_j \alpha_j \tag{10}$$

$$\begin{aligned} \mathrm{var}(Q_n^*) &= \mathrm{var}\left[\sum_{i=1}^{K_n} (Y_N^*)_i\right] = E(K_n^*)E(Y_N^*)^2 \\ &= \sum_{j=1}^{n} \rho_j(\alpha_j^2 + \beta_j^2) \end{aligned} \tag{11}$$

According to the mean and variance of the Poisson process [Eqs. (10) and (11)], the water demand model for multiple households can be obtained, and it can be used to obtain the data of a multi-family residential water use at any time point.

Using the Poisson distribution model, the dynamically changing water demand in the experimental water supply networks can be obtained. Since the Poisson process is discrete, the obtained data represent the water demand of the residents at each time point, and the demand data curve can be obtained by connecting the water demand data of each moment as shown in Fig. 3, the grey line represents the water demand generated by the Poisson distribution model, and the black line represents the water demand shown in Sect. 3.1.

4 Pollution source location method based on the improved genetic algorithm

When using the simulation–optimization model for the pollution source location problem, the state of the hydraulic water quality is outputted positively through the EPANET simulator, and the results are compared with the sensor's information actually detected by the monitoring sensor. If the difference value is less than the given threshold, then the optimization algorithm's optimal solution is the pollution source position; otherwise, the iteration of the optimization algorithm continues to perform until the stop condition is met, and then the algorithm ends. The overall framework of the simulation–optimization is shown in Fig. 4.

As shown in Fig. 4, the optimization algorithm acts as an optimizer to generate the pollution event and the simulation software EPANET uses to simulate the pollution event and output the actual pollutant concentrations at each node. In this study, an improved genetic algorithm is used as the optimization algorithm. In the population, each individual represents a pollution event. The EPANET simulator can simulate the pollution event and output the actual pollutant concentration information of the nodes in the water supply networks. Via comparisons with the real information detected by the monitoring sensor, the individual fitness can be calculated. A smaller fitness value corresponds to a greater likelihood that the pollution event is a real source of pollution.

4.1 Problem code

To solve the above framework model, this paper proposed a genetic algorithm based on hybrid coding. The genetic algorithm is based on Darwin's theory of biological evolution. By simulating the evolutionary process of living creatures in nature, the solution space is filtered and searched by the rule of "survival of the fittest", and the optimal solution is then obtained. In nature, species wither greater adaptability present more appropriate adaptations to the environment and a greater likelihood of survival. The

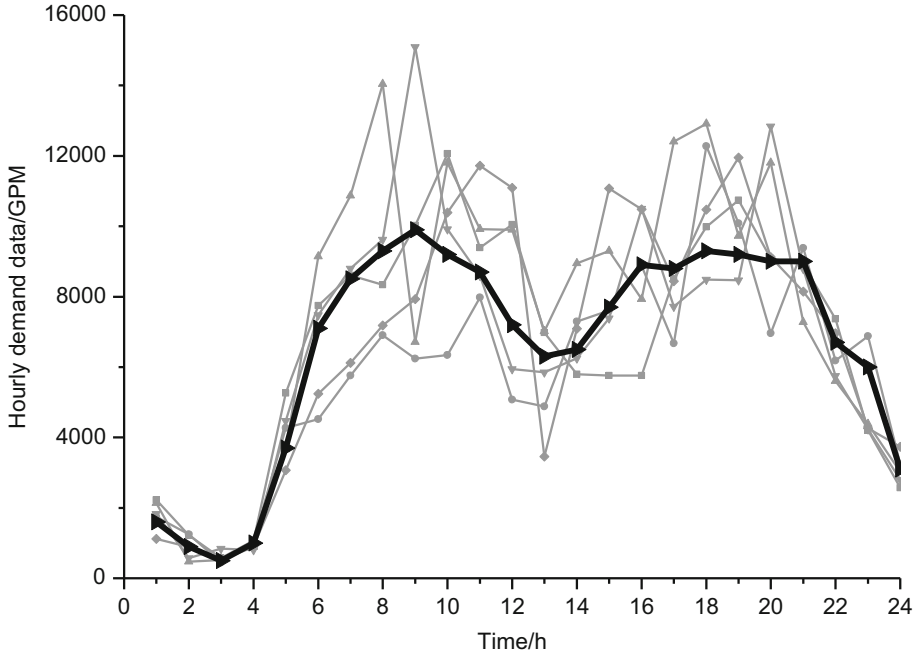

Fig. 3 Water demand generated by the Poisson model

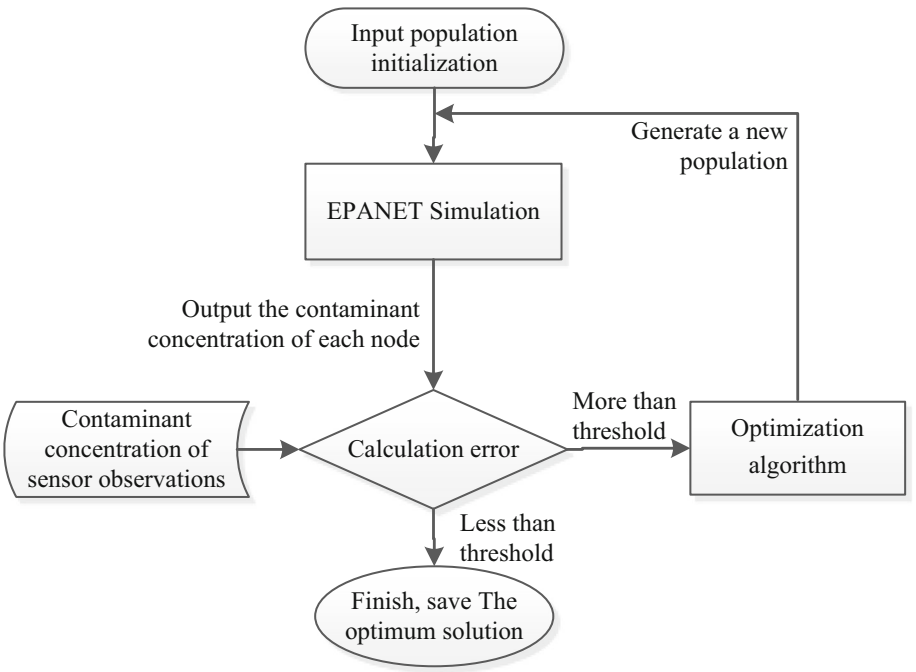

Fig. 4 Framework of simulation–optimization method

variables of the pollution source location problem include the pollution source location, start time, duration, and pollutant injection vector. Based on the properties of the variable, this study uses the combination of integer coding and real coding to encode the pollution source location problem, and the genetic operation of the algorithm was modified according to the corresponding coding mode, which accelerated the speed at which the optimal solution was obtained and improved the convergence speed of the algorithm.

In the population, each individual represents a pollution event, and an individual contains four variables: {node position, pollution injection time, duration, pollutant injection mass}, where the first three variables are integer variables using integer coding and the fourth variable is a real number vector encoded in real numbers. For example, if the pollution source node is 73, the pollutant injection

time is 2:00, the pollutant duration is 4 h, and the pollution source injection mass is (200, 216.5, 310.9, 300), then the individual code corresponding to this pollutant injection event is {00073024} and {200, 216.5, 310.9, 300}.

4.2 Genetic operation

The problem code in this paper combines the integer and real number coding. It is also a combination of two methods in the corresponding cross operation and mutation operation. In the cross operation, the integer coding adopts double point cross transformation and the real number coding adopts real number recombination; in the mutation operation, the integer coding adopts a single point mutation and Gaussian mutation.

Cross operation Assume that in individual 1, the pollution source position is 125, the starting injection time is 10:00, the injection duration is 3 h, the injection concentration changes once every hour, and the injection mass vector is (215.1, 345.8, 457.8); and assume that in individual 2, the pollution source position is 96, the starting injection time is 3:00, the injection duration is 4 h, the injection concentration changes once every hour, and the injection mass vector is (300.1, 123.4, 39.1, 356.8). According to the coding method proposed in this paper, the encoded individuals are as follows:

Individual 1: {00125103} {215.1, 345.8, 457.8}
Individual 2: {00096034} {300.1, 123.4, 39.1, 356.8}

As shown above, the former integer is encoded as an array of integers with a fixed length of 8 bits using two-point cross transformation, that is, randomly selecting two positions for cross transformation. Suppose that two numbers 3 and 6 are randomly selected, meaning that the genes at the position 3 to position 6 genes of Individual 1 are exchanged with those of Individual 2, thus obtaining two new individuals after the cross transformation:

Individual 3: {00096003} {215.1, 345.8, 457.8};
Individual 4: {00125134} {300.1, 123.4, 39.1, 356.8}.

The latter real number is encoded as an array of real number with variable lengths, which varies according to the length of time. The algorithm for real reorganization is as follows:

Sub-individual 1 = $a \times$ parent individual 1 + $(1 - a) \times$ parent individual 2;
Sub-individual 2 = $(1 - a) \times$ parent individual 1 + $a \times$ parent individual 2;

Where a is a random scale factor. After the real number recombination of the real number coding of the above Individual 3 and Individual 4, the crossing sub-individual can be obtained (assuming a is randomly selected as 0.7):

Sub-individual 1: {00096003} {240.6, 279.08, 332.19}
Sub-individual 2: {00125134} {274.6, 190.12, 164.71, 249.76}

Since the real part of Individual 3 has only three digits and that of Individual 4 has four digits, the corresponding real number is 0 when the fourth bit of Individual 4 is calculated.

Mutation operation Similar to the cross operation, the mutation operation is also performed separately in two parts. The integer part is a single point mutation, and the real part is a Gaussian mutation. Assuming that the Sub-individual 1 after the above cross operation is retained for the mutation operation, the former part is first subjected to a single point mutation, and the eighth bit is randomly selected for mutation, resulting in {00096005} after the mutation.

The latter part is a Gaussian mutation in which a random number obeying Gaussian distribution is generated to replace the mass in the current mass vector. The mathematical expectation of the random number generated by the algorithm should be the mass value of the current mutation. In this study, six random numbers obeying $U(0,1)$ are generated by simulation, and their mathematical expectation is considered an approximation of the random numbers in Gaussian distribution. Assuming that the approximate Gaussian distribution random number is 0.85, the upper bound of the injection mass is 500 and the lower bound is 10; therefore, the mutation is as follows:

Real number after mutation = $[(0.85 \times (500 - 10) + 10) +$ Sub-individual 1$]/2$

The individual after the mutation is {00096005}{333.55, 352.79, 379.34, 213.25, 39.8}. Since the real part of the original individual has only 3 bits and the 8th bit of the integer part is 5 after the mutation, random initialization is performed for the last two bits.

4.3 Procedures of the algorithm

In this study, we using the simulation–optimization scheme for the pollution source location problem and proposed an improved GA. The improved GA is used as the optimizer, and simulation software EPANET is used as the simulator to simulate the pollution events. The detail procedures of the improved genetic algorithm are as follows:

Step 1 Coding Assuming that the pollution position is 121, the injection starting time is 2, the duration is 2 h, the time step is 30 min, and the pollutant injection quality is 300.5, 180.7, 240.9, 300.0. The coding is two arrays: array 1 {00121, 300, 02, 2} and array 2 {300.5, 180.7, 240.9, 300.0};

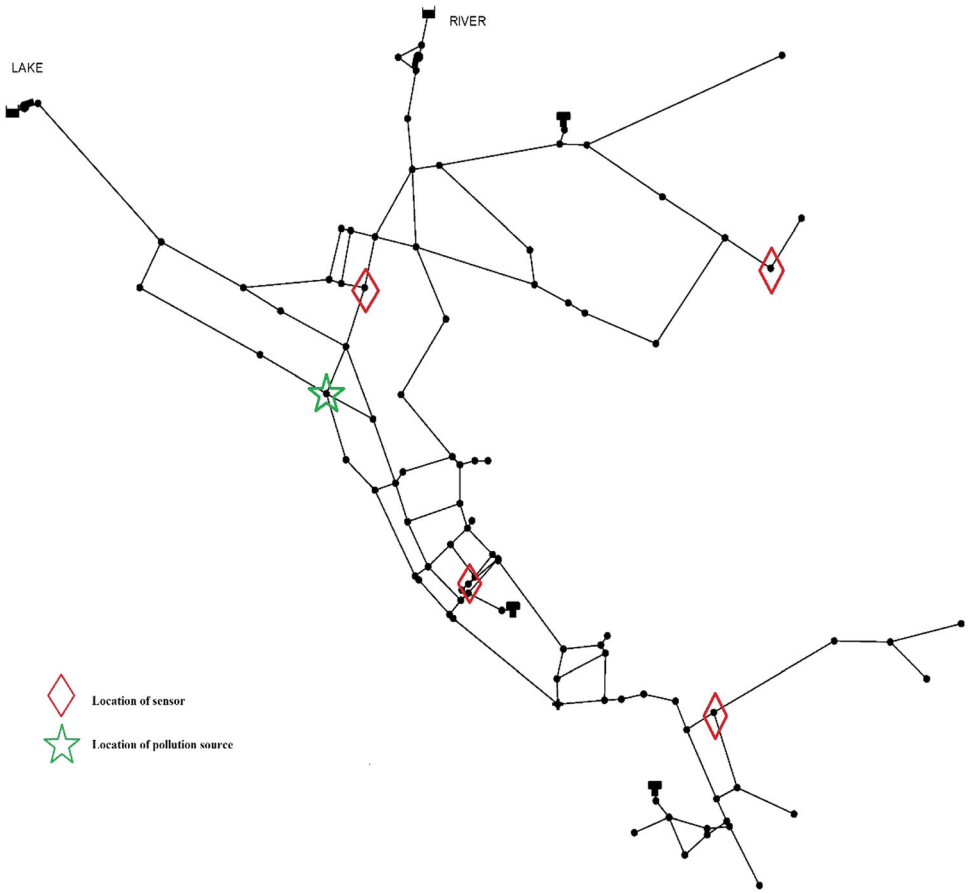

Fig. 5 Position of the sensors and pollution source location in test networks 1

Step 2 Initialization The population size is *N*, and each individual is initialized with array 1 and then initialized with array 2 according to the duration, where the length of array 2 = the duration × (1 h/time step);
Step 3 Selection operation N individuals are selected by roulette;
Step 4 Cross operation As mentioned above, a double-point cross operation is applied for array 1, and real number reorganization is applied for array 2;
Step 5 Mutation operation As mentioned above, a single point mutation is applied for array 1, and Gaussian variation is applied for array 2;
Step 6 Judgement Whether the stop condition is met is determined. If not, go to Step3; otherwise, the program ends.

5 Experimental simulation and analysis

5.1 Parameter setting

In this study, we use two test water supply networks for the algorithm comparison experiment and verification of the algorithm's performance. The dataset of these two water supply networks are from the website http://emps.exeter.ac.uk/engineering/research/cws/downloads/benchmarks/expansion/. One network is test network 1, which contains 92 nodes, 3 reservoirs, 2 pools, 2 pumps, and 4 sensors {18, 29, 47, 67} as shown in Fig. 5; and the other is test network 2, which contains 430 nodes, 4 reservoirs, 3 pools, 11 pumps, and 10 sensors {6, 22, 30, 34, 40, 42, 43, 76, 80, 87} as shown in Fig. 6. The threshold of the concentration detected by the sensor is assumed to be 0.0001 mg/l. All the nodes in the network are assumed to provide water only for the residential water supply area. Simulation software is EPANET 2.0, and the specific parameters of the two networks are shown in Table 1. The parameters for the genetic algorithm used in the experiment are shown in Table 2.

In this study, assuming that the actual pollution has a single source, two sets of experiments are mainly performed in this study. In order to verify the accuracy of the proposed algorithm, we design the experiment 1 on the benchmark water supply networks. In experiment 1, the proposed algorithm is compared with the algorithm in the study by Praveen [37]. Experiment 2 is to verify the proposed algorithm's robustness via comparisons among experiments on a large size of water supply network.

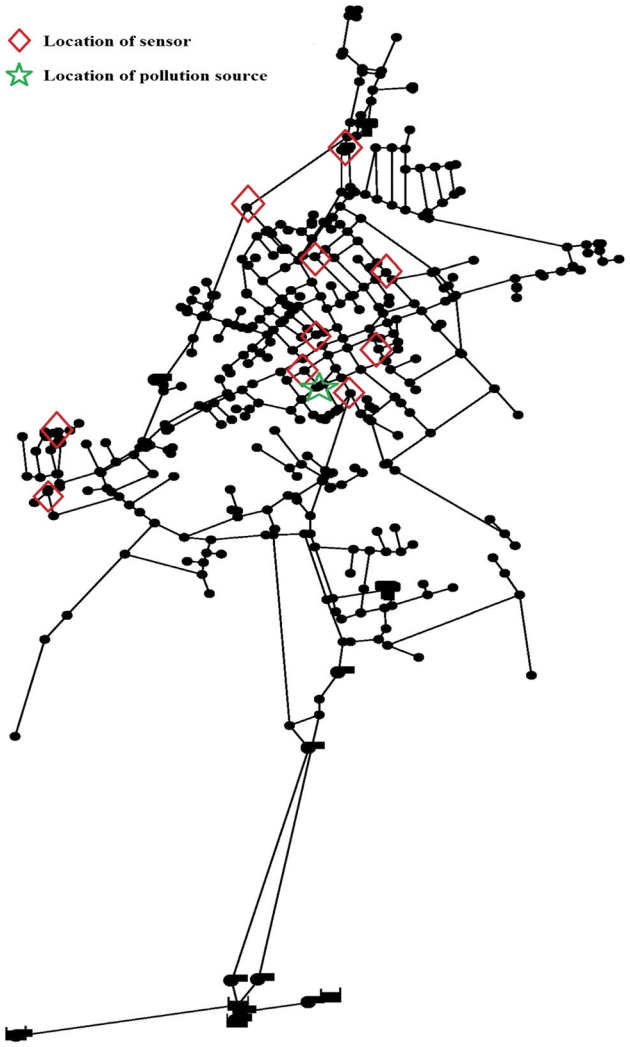

Fig. 6 Position of the sensors and pollution source location in test networks 2

5.2 Verification of the algorithm accuracy

The goal of the pollution source location algorithm is to accurate locate of the pollution source and get the concentration of the injected pollution. In test network 1, four sensors are set up, and the Poisson model as water demand model was used to test the standard GA and the improved GA to verify the accuracy of the improved algorithm. The two sets of experimental scenarios are as follows:

- *Scenario 1* Poisson model and standard GA
- *Scenario 2* Poisson model and improved GA

It should be noted that the concentration of the pollution source in Scenario 1 is maximized. For example, if the injection concentration is 2674.21, then it is set to 3000 when the experiment is performed. However, this assumption is not applied to Scenario 2.

For each scenario, the Poisson model was used to generate a random water demand of 800 groups to conduct experiments on network 1 using the corresponding optimization method. In this study, using the hit probability to describe the algorithm's accuracy, that is, the amount of water demand in a real pollution scenario obtained by optimizing and simulating the water demand for several times divided by the total number of residents' water demand. The calculation formula of the hit probability is as Eq. (12).

$$P = \frac{N_r}{N_s} \tag{12}$$

In Eq. (12), P indicates the hit probability, N_r indicates the number of residents' water demand to find a true pollution source, and N_s indicates the total number of residents' water demand in the simulation.

As shown in Fig. 7, Scenario 1 uses the Poisson model and the standard genetic algorithm to identify possible pollution source nodes, which are marked with the diamond and triangle in the figure. The triangle is the

Table 1 Parameters of the water networks

Parameter	Test network 1	Test network 2
Size of the network	97	430
Type of the pollution source	Single source	Single source
Water demand	Generated by the Poisson distribution model	
Hydraulic time step	1 h	1 h
Good water quality time	5 min	30 min
Total simulation time	24 h	48 h
Pollution injection node	15	374
Pollution injection time	–	0:00
Pollution injection duration	–	1 h
Number of sensors	12/12/4	4
Pollution injection mass	2674.21	300

Table 2 Parameter settings for the experimental comparisons of the algorithm

Parameter	GA	Improved GA
Size of the population c	80	80
Number of iterations m	40	40
Amount of water demand realized	800	800
Cross probability	0.8	0.8
Mutation probability	Gaussian mutation	Gaussian mutation

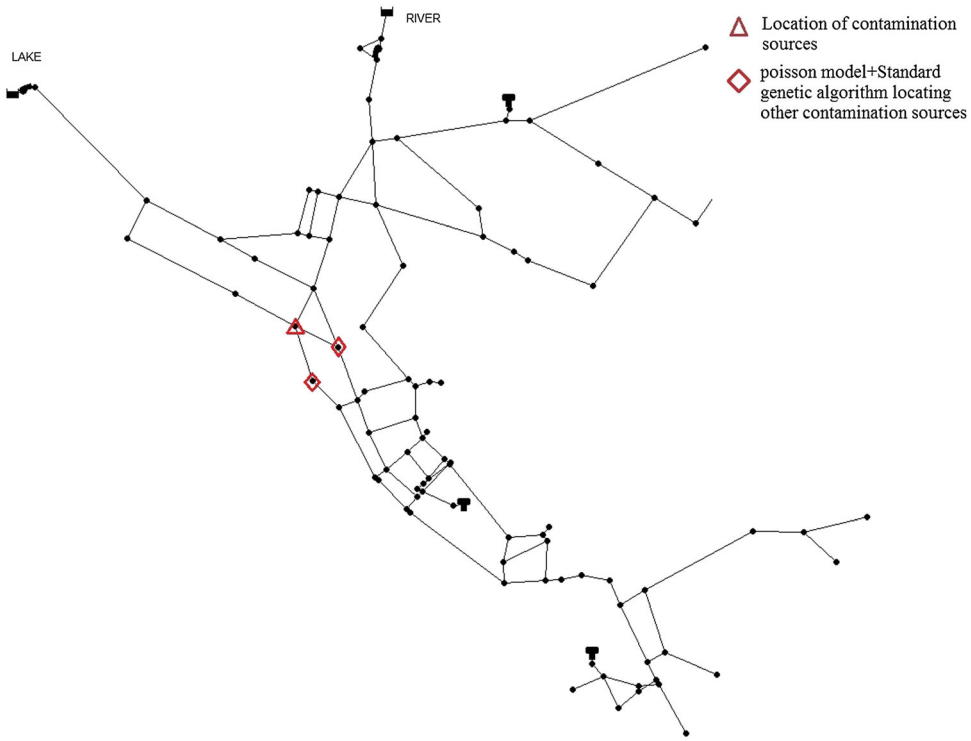

Fig. 7 Experiment 1 results

pollution source's real location, and the diamond is the interference of the pollution source node introduced by the dynamic variability of the algorithm. The hit probability for the pollution source location is 56%. The experimental results of Scenario 2 showed that the hit probability for the pollution source location is 100%; therefore, the obtained result only showed the node position marked with a triangle. By adding the improved strategy in the standard GA, the algorithm can effectively jump out of the local optimal and the water supply networks is relatively small and does not contain many nodes; therefore, the interference term can be eliminated, and the real pollution source can be found. The comparison of Scenario 1 with Scenario 2 shows that the proposed algorithm has a better hit probability.

5.3 Verification of the algorithm's robustness

Experiment 1 showed that the accuracy of the improved algorithm for test network 1 was higher between the two tested algorithms; thus, whether the algorithm is still valid for the large size of water supply networks is now confirmed. In Experiment 2, 10 sensors were set up in test networks 2 and two scenarios were implemented:

- *Scenario 3* Poisson model and standard GA
- *Scenario 4* Poisson model and improved GA

The experimental results are shown in Fig. 8 for Scenario 3 and Scenario 4 of test network 2. For test networks 2, the hit probabilities of the pollution source location for Scenario 3 and Scenario 4 are 30% and 61%, respectively. The triangle represents the location of the true pollution source, and the circle represents the locations found by the improved algorithm that may be pollution sources other than the true pollution source.

The comparison with the results of experiment 1 show that the standard GA is not suitable for locating the pollution source in a large size of water supply networks. However, for the improved algorithm, although it could not find the exact position of the pollution source with the hit

Fig. 8 Experiment 2 results

probability 100%, the probability of finding the true location of pollution source was twofold higher than that of the standard GA; therefore, the improved GA can solve the problem of pollution source location in a large size of water supply networks.

6 Conclusions

With the rapid development of economy, the quality of people's life has been continuously improved, and the ecological environment has been deteriorating at an unprecedented speed, which leads to the frequent occurrence of large-scale water pollution incidents in cities and towns by accident or on purpose. The open operation and easy intrusion of water supply networks have led to the frequent occurrence of sudden pollution incidents in a water supply networks. To reduce the significant economic losses and adverse social impacts caused by pollution of the water supply network, the pollution source must be located, and studies of this problem have an important practical significance. In this paper, an urban resident water supply network was used as the research object. EPANET 2.0 software was used as the simulation platform to study the pollution source location problem under variety water demand. According to the dynamic change of residents' water demand, the pollution source location problem was transformed into a dynamic optimization problem, and used the optimization algorithm to solve this problem. In order to solve the variety water demand of residents, the Poisson distribution model was used to simulate the variety water demand of residents and an improved genetic was proposed as the optimization algorithm. The accuracy of the proposed algorithm was verified by simulation experiments on benchmark water supply networks, and the robustness of the proposed algorithm was verified by two different size of water supply networks.

Acknowledgements This paper was supported by National Natural Science Foundation of China (61673354 and U1911205), the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (CUGGC03) and Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences (Wuhan) (KLIGIP-2018B13).

Compliance with ethical standards

Conflict of interest The authors declare there is no conflict of interest.

References

1. Najah A, El-Shafie A, Karim OA et al (2013) Application of artificial neural networks for water quality prediction. Neural Comput Appl 22:187–201
2. Hameed M, Sharqi SS, Yaseen ZM et al (2017) Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. Neural Comput Appl 28:893–905
3. Kayaalp F, Zengin A, Kara R et al (2017) Leakage detection and localization on water transportation pipelines: a multi-label classification approach. Neural Comput Appl 28:2905–2914
4. Mohammadrezapour O, Kisi O, Pourahmad F (2020) Fuzzy c-means and K-means clustering with genetic algorithm for identification of homogeneous regions of groundwater quality. Neural Comput Appl 32:3763–3775
5. Shang F, Uber JG, Polycarpou MM (2002) Particle backtracking algorithm for water distribution system analysis. J Environ Eng 128(5):441–450
6. 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
7. De Sanctis AE, Shang F, Uber JG (2009) Real-time identification of possible contamination sources using network backtracking methods. J Water Resour Plan Manag 136(4):444–453
8. 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

9. Huang JJ, McBean EA (2009) Data mining to identify contaminant event locations in water distribution systems. J Water Resour Plan Manag 135(6):466–474
10. Perelman L, Ostfeld A (2012) Bayesian networks for source intrusion detection. J Water Resour Plan Manag 139(4):426–432
11. Wang H, Harrison KW (2012) 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
12. Wang H, Jin X (2013) Characterization of groundwater contaminant source using Bayesian method. Stoch Environ Res Risk Assess 27(4):867–876
13. Guo Y-N, Pei Z, Cheng J, Wang C, Gong D (2018) Interval multi-objective quantum-inspired cultural algorithms. Neural Comput Appl 30(3):709–722
14. Yan X, Zhu Z, Hu C, Gong W, Wu Q (2019) Spark-based intelligent parameter inversion method for prestack seismic data. Neural Comput Appl 31(9):4577–4593
15. Gong W, Wang Y, Cai Z, Wang L (2018) Finding multiple roots of nonlinear equation systems via a repulsion-based adaptive differential evolution. IEEE Trans Syst Man Cybern Syst. https://doi.org/10.1109/TSMC.2018.2828018
16. Wu B, Qian C, Ni W, Fan S (2012) The improvement of glowworm swarm optimization for continuous optimization problems. Expert Syst Appl 39(7):6335–6342
17. Lu C, Gao L, Li X, Zheng J, Gong W (2018) A multi-objective approach to welding shop scheduling for makespan, noise pollution and energy consumption. J Clean Prod 196:773–787
18. Wu Q, Zhu Z, Yan X, Gong W (2019) An improved particle swarm optimization algorithm for AVO elastic parameter inversion problem. Concurr Comput Pract Exp 31(9):1–16
19. Yu P, Yan X (2020) Stock price prediction based on deep neural network. Neural Comput Appl 32(6):1609–1628
20. Gong W, Cai Z (2013) Parameter extraction of solar cell models using repaired adaptive differential evolution. Sol Energy 94:209–220
21. Wang F, Zhang H, Li Y, Zhao Y, Rao Q (2018) External archive matching strategy for MOEA/D. Soft Comput 22(23):7833–7846
22. Wu J, Zhu X, Zhang C, Yu PS (2014) Bag constrained structure pattern mining for multi-graph classification. IEEE Trans Knowl Data Eng 26(10):2382–2396
23. Wu G, Shen X, Li H, Chen H, Lin A, Suganthan PN (2018) Ensemble of differential evolution variants. Inf Sci 423:172–186
24. Wu J, Pan S, Zhu X, Zhang C, Wu X (2018) Multi-instance learning with discriminative bag mapping. IEEE Trans Knowl Data Eng 30(6):1065–1080
25. Wang R, Ishibuchi H, Zhou Z, Liao T, Zhang T (2018) Localized weighted sum method for many-objective optimization. IEEE Trans Evol Comput 22:3–18
26. Lu C, Gao L, Yi J (2018) Grey wolf optimizer with cellular topological structure. Expert Syst Appl 107:89–114
27. Yang P, Tang K, Yao X (2018) Turning high-dimensional optimization into computationally expensive optimization. IEEE Trans Evol Comput 22(1):143–156
28. Wang F, Zhang H, Li K, Lin Z, Yang J, Shen X-L (2018) A hybrid particle swarm optimization algorithm using adaptive learning strategy. Inf Sci 436–437:162–177
29. Guo Y-N, Yang H, Chen M, Cheng J, Gong D (2019) Ensemble prediction-based dynamic robust multi-objective optimization methods. Swarm Evol Comput 48:156–171
30. Yan X, Li P, Tang K, Gao L, Wang L (2020) Clonal selection based intelligent parameter inversion algorithm for prestack seismic data. Inf Sci 517:86–99
31. 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
32. Wu J, Pan S, Zhu X, Cai Z (2015) Boosting for multi-graph classification. IEEE Trans Cybern 45(3):430–443
33. Tang K, Yang P, Yao X (2016) Negatively correlated search. IEEE J Sel Areas Commun 34(3):1–9
34. Shi J, Lei Y, Wu J et al (2019) Uncertain active contour model based on rough and fuzzy sets for auroral oval segmentation. Inf Sci 492:72–103
35. Lei Y, Zhou Y, Shi J (2019) Overlapping communities detection of social network based on hybrid c-means clustering algorithm. Sustain Cities Soc. https://doi.org/10.1016/j.scs.2019.101436
36. 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 190:465–474
37. Wang F, Li Y, Zhang H, Hu T, Shen X-L (2019) An adaptive weight vector guided evolutionary algorithm for preference-based multi-objective optimization. Swarm Evol Comput 49:220–233
38. Ostfeld A, Salomons E (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
39. Guan J, Aral MM, Maslia ML, Grayman WM (2006) Identification of contaminant sources in water distribution systems using simulation–optimization method: case study. J Water Resour Plan Manag 132(4):252–262
40. Preis A, Ostfeld A (2007) A contamination source identification model for water distribution system security. Eng Optim 39(8):941–947
41. Preis A, Ostfeld A (2008) Genetic algorithm for contaminant source characterization using imperfect sensors. Civ Eng Environ Syst 25(1):29–39
42. 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
43. Vankayala P, Sankarasubramanian A, Ranjithan SR et al (2009) Contaminant source identification in water distribution networks under conditions of demand uncertainty. Environ Forensics 10(3):253–263
44. Lv M, Wang M, Liu J, Dong S (2010) Notice of retraction investigation on backward tracking of contamination sources in water supply systems-case study. Int Conf Environ Sci Inf Appl Technol 3:484–487
45. Drake K, Zechman E (2011) Using niched co-evolution strategies to address non-uniqueness in characterizing sources of contamination in a water distribution system. World Environ Water Resour Congr 2011:24–329
46. Liu L, Ranjithan SR, Mahinthakumar G (2010) Contamination source identification in water distribution systems using an adaptive dynamic optimization procedure. J Water Resour Plan Manag 137(2):183–192
47. Hu C, Zhao J, Yan X, Zeng D, Guo S (2015) A mapreduce based parallel niche genetic algorithm for contaminant source identification in water distribution network. Ad Hoc Netw 35(C):116–126
48. Yan X, Sun J, Hu C (2017) Research on contaminant sources identification of uncertainty water demand using genetic algorithm. Clust Comput 20(2):1007–1016
49. Yan X, Gong W, Wu Q (2017) Contaminant source identification of water distribution networks using cultural algorithm. Concurr Comput Pract Exp 29(24):1–11
50. Yan X, Yang K, Hu C (2018) Pollution source positioning in a water supply network based on expensive optimization. Desalin Water Treat 110:308–318
51. Yan X, Zhao J et al (2019) Multimodal optimization problem in contamination source determination of water supply networks. Swarm Evol Comput 47:66–71

52. Yan X, Zhu Z, Li T (2019) Pollution source localization in an urban water supply network based on dynamic water demand. Environ Sci Pollut Res 26(18):17901–17910
53. Gong Jinyu, Yan Xuesong, Chengyu Hu, Qinghua Wu (2019) Collaborative based pollution sources identification algorithm in water supply sensor networks. Desalin Water Treat 168:123–135
54. Yan X, Li T, Hu C (2019) Real-time localization of pollution source for urban water supply network in emergencies. Clust Comput 22:5941–5954
55. Rossman LA (2000) Epanet 2 users manual, vol 19(1). Laboratory Office of Research & Development United States Environmental Protection Agency, Cincinnati, pp 115–118
56. Haight FA (1967) Handbook of poisson distribution. Wiley, New York, pp 169–179
57. Consul PC, Jain GC (1973) A generalization of the Poisson distribution. Technometrics 15(4):791–799
58. Johnson NL, Kemp AW, Kotz S (2005) Poisson distribution. Univariate discrete distributions, 3rd edn. Wiley, New York, pp 156–207