Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2019.24204

168 (2019) 123–135
November

Collaborative based pollution sources identification algorithm in water supply sensor networks

Jinyu Gong[a], Xuesong Yan[a,b], Chengyu Hu[a,*], Qinghua Wu[c]

[a]*School of Computer Science, China University of Geosciences, Wuhan 430074, China, emails: huchengyu@cug.edu.cn/ 2245226564@qq.com (C. Hu)*
[b]*State Key Lab of Digital Manufacturing Equipment & Technology, Huazhong University of Science & Technology, Wuhan 430074, China*
[c]*Faculty of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China*

Received 3 September 2018; Accepted 1 February 2019

ABSTRACT

Arranged water quality sensors at key nodes or water sources in water supply network to achieve real-time monitoring can prevent water pollution incident. However, when pollution incident occurs, to use the information collected by the water quality sensors to accurately locate and predict the location of pollutants (injection time, injection duration and injection quality) is a challenging problem. In this paper, the simulation optimization method which is the currently popular pollution source identification method is conducted an in-depth analysis and found that identification of pollution source in water supply network is an expensive optimization problem. So, the pollution source identification problem is converted into an expensive optimization problem in this paper. According to the characteristics of the problem, a collaborative based expensive optimization algorithm is proposed. According to the characteristics of water supply network, this algorithm proposes a more suitable strategy for each different variable to guide the search direction of the algorithm. The algorithm uses Gaussian surrogate model as much as possible to ensure the identification accuracy. Finally, through simulation experiment, the validity, efficiency and stability of the proposed algorithm is verified.

Keywords: Pollution sources identification; Simulation optimization; Gaussian surrogate model; Collaborative based algorithm

1. Introduction

In recent years, sudden and unexpected water pollution incidents often occur in China. Some drinking water incidents and malicious attack on water supply network have caused great economic loss and adverse social impact to our country [1,2]. To prevent water pollution incidents from causing major disasters and losses, urban water supply network needs to be equipped with real-time monitoring system for drinking water safety. In this system, the real-time monitoring can be achieved by arranging water quality sensors at key nodes or water sources. However, when pollution incident occurs, it is a challenge to use the information collected by the water quality sensor to accurately locate the pollution source and predict the location of pollutants, injection time, injection duration and injection quality.

In recent years, many scholars are trying to use the simulation–optimization model to transform pollution sources identification problems into optimization problems, and

[*] Corresponding author.

then use the evolutionary computation method to get the optimal solution. A group researcher proposed a simulation–optimization method to solve the problems of nonlinear pollution sources identification. By continuously reading the sensor data, the pollution sources are predicted and corrected [3–5]. Finally, pollution sources and pollutant release history are identified and proposed an adaptive dynamic optimization procedure based on evolutionary algorithms to search for the pollution sources properties (start time, position, release history) and slowly converge to obtain the only optimal solution by adding new available sensors. In the hybrid optimization model proposed by Ayvaz [1], binary genetic algorithm (GA) and generalized reduced-order gradient method are used to locate the information of pollution sources in underground water network. Sudhakaran et al. [6] considered the uncertainty of water demand of users, introduced different models to simulate water demand of users, and used GA to solve pollution sources identification problems. Deng et al. [7] proposed a Map Reduce based parallel niche GA for contaminant sources identification in water distribution network, the niche GA as an optimizer and EPANET as a simulator. Feng et al. [8] proposed a cultural algorithm for this problem. They also convert the pollution sources identification problem into a multimodal optimization problem and proposed a niching genetic algorithm to solve it.

In the simulation–optimization method, optimization algorithm is used as an optimizer. In the optimization algorithm, each individual needs to use EPANET as a simulator to simulate the pollution incident, so as to calculate the fitness value. Taking BWSN2 as an example (the network contains 12,527 nodes, 2 reservoirs, 2 ponds and 20 sensors) to simulate a pollution sources event, it takes about 3 s to calculate the fitness value. When using a GA (population size of 100, running 100 generations), it took 329 min which are nearly 5.5 h. In the current research, the time required for small pipe networks such as Rossman 2000 (92 nodes) is small, and the EPANET simulation time is small, which is not expensive under some set parameters [9–15]. However, in a real environment, the number of urban pipe network nodes is usually huge, so there is bound to be more time consumed. Therefore, identification of pollution sources problems in town water supply network is an expensive optimization problem; so in the optimization process, EPANET simulator consumes a large amount of time cost. In order to minimize the harm of pollutants to public health, when a certain amount of water quality information is obtained, the pollution sources needs to be located as fast as possible.

In many practical engineering optimization problems, the objective function cannot be expressed clearly, and the optimization model is also complicated. Time-cost simulation software is required for simulation and evaluation. Each calculation takes time and costs, hence are expensive optimization problems. In 1998, some researcher used Gaussian random model in the branch-and-bound algorithm to provide the expected target values for non-sampling points [16–20]. The validity of the random model was analyzed and proved to be an efficient global optimization algorithm. In 2002, a group research introduced random models to evolutionary algorithms and established global random models for global prediction. In 2004, others researcher established a local model using a random model to perform local prediction in evolutionary algorithms. In 2007, others researcher introduced a random model into the evolutionary algorithm to both establish a global model and establish a local model to accelerate the evolutionary efficiency.

Literature [10,33] are research papers on model-based single-objective evolutionary algorithms. Literature [26] which is a research paper on model-based expensive multi-objective evolutionary algorithms. Literature [17,21,36,51], on model-based multi-objective evolutionary algorithms. Literature [42] on application of intelligent calculation methods in expensive optimization problems. Keane [21] and Li et al. [22] embedded a meta-modeling mechanism into the global search algorithm to achieve a balance between the prediction model and the global search algorithm. In 2014, with combined the Gaussian prediction model and the optimization algorithm to solve the high-dimensional global optimization problem [23]. In 2015, a group researcher embedded selection evaluation strategies into support vector machine prediction models to classify and predict constrained optimization problems [4]. In 2017, others researchers solved high-dimensional expensive optimization problems by using cooperative swarm optimization [41].

An ordinary optimization algorithm requires a lot of iterative calculations and tens of thousands of evaluation times in order to obtain better results. In solving expensive optimization problems by objective functions, the general optimization algorithm with many times of iterations in the optimization process will lead to the use of a large number of expensive simulation models, seriously affecting the performance and efficiency of the algorithm [23–26]. The key to solving expensive optimization problems is to reduce the use of expensive simulation models as much as possible without affecting the accuracy of the algorithm. Therefore, an appropriate surrogate model is introduced in the expensive optimization algorithm to replace the expensive evaluation function for calculation. There are two main difficulties of the expensive optimization algorithm, one is how to set up a suitable surrogate model according to the sample points. The other is how to balance the use between the surrogate model and the expensive evaluation function so that the algorithm can search the optimal solution both quickly and accurately.

Pollution sources identification problem is a specific practical application with its own characteristics.

However, most evolutionary algorithms used in simulation optimization are highly random heuristic algorithms, which are often not effective in solving such problems with many characteristics. Specific problems require specific analysis, and the search algorithm needs to be guided by the characteristics. On the other hand, in the expensive optimization algorithms, the accuracy of the surrogate model is often related to the sample selection [27–29]. In the complex problems, the smaller the range of the sample, the higher the prediction accuracy of the model. For these considerations, this paper uses collaborative algorithm, adopting different strategies for different populations, achieving effective guidance for algorithm search. Moreover, only one variable will be changed for each population, and the other variables will not be changed. This produces a small change range of the solution, and the prediction accuracy of the expensive model is improved. Collaboration exists widely in nature and social systems, such as parasitism, competition and predation of species in ecosystems, and competition and cooperation among groups in life working together to further the overall evolution.

In the field of intelligent computing, co-evolution refers to the evolutionary technology in which multiple objects conduct collaborative search through certain mechanisms and strategies. The study of cooperative co-evolutionary approach began in the 1990 from the ranking network of Hills. Potter and De Jong [37] studied the mechanism of populations' co-evolution selection among multiple interacting sub-populations. Subbu et al. [40] proposed a distributed cooperative co-evolutionary approach model and applied it to optimization problems. Gu et al. [11] proposed a bi-population based competitive co-evolutionary quantum GA, and gave three competitive strategies to dynamically adjust the population size. Wang et al. [43] proposed a bi-population based EDA algorithm. Sun et al. [41] proposed a co-evolution framework called mutable interactive learning that puts all variables independently in separate groups, iteratively discovers their relationships, merges the groups and groups the variables reasonably.

Tenne and Goh [42] proposed an automatic decomposition strategy with differential grouping, revealing the underlying interaction structure and forming subcomponents of decision variables to keep their interdependencies to a minimum, which is a good solution to the decomposing of variables of co-evolutionary algorithm being applied to large-scale global optimization. Wang et al. [43] proposed a particle swarm optimization with information sharing mechanism, the competitive and cooperative particle swarm optimization algorithm solves the problem of premature convergence in global optimization problems to a certain extent. Deng et al. [7] proposed a multi-swarm self-adaptive cooperative optimization algorithm for genetic and ant colony. It introduced chaos optimization algorithm, multi-population cooperative strategy and adaptive control parameters into GA and ACO algorithms, enhancing the performance in solving complex optimization problems. Peng et al. [34] introduced niching-based multi-modal optimization in the standard co-evolutionary framework [44–46]. By providing more information partners to subcomponents for information compensation, a simple and efficient clustering method is added to prevent combinatorial explosion and effectively reduce information loss during co-evolution and prevent sub-optimization problems.

The pollution sources identification problem is converted into an expensive optimization problem in this paper. First, the Gaussian random process is used to establish the surrogate model. Due to the characteristics of Gaussian random process and water supply network, this paper establishes a sub-model for each node of the pipe network and verifies its validity [47]. Considering the problem of getting trapped in local optimum of large pipe network and the range of samples predicted by the surrogate model, this paper proposes solving pollution sources identification problems based on a collaborative expensive optimization algorithm, using different strategies to guide the population or variable for population search. Finally, the validity and efficiency of the proposed algorithm are verified.

2. Material and methods

2.1. Pollution sources identification model

To ensure the safety of drinking water, a variety of water quality sensors are installed at the important nodes of the town water supply network for real-time monitoring of water quality information. Once pollution incidents occur, timely warning and appropriate treatment measures can be made. In order to get accurate pollution sources information, many researchers adopt the simulation optimization method to effectively locate the sources. After the pollution occurs, the water quality sensor detects the occurrence of the pollution and records a series of information such as the concentration of the pollutants [48]. Upon getting the information, a series of pollution incident will quickly be produced through the EPANET simulator. At the same time, the sensor of the simulation software also records the corresponding data. By comparing the information recorded by the real sensor and by the simulation software, the most likely pollution incident is selected, that is, the located pollution sources information.

Simulation–optimization method is adopted in this paper, in which the pollution sources identification problem is converted into an expensive optimization problem for solving, and then evolutionary computation is used for optimized solution [49]. From the optimization point of view, when the minimum variance of the cumulative concentration and the actual cumulative detected concentration of the pollution incident at the sensor is 0 or less than a threshold value *e*, it is considered that the node injected by the pollution incident is the actual pollution sources. The optimization problem can be expressed as Eq. (1).

$$\begin{aligned} &\underset{\{M,n,t_I\}}{\text{Minimize}} \quad f = \sum_{j=1}^{N_s}\sum_{t=1}^{T_s}\left(c_j(t) - c_j^*(t)\right)^2 \\ &\text{S.T.} \quad M = \{m_1, m_2, \cdots, m_k\}; m_i \geq 0 \\ &\qquad\;\; n \in \{1, N\} \\ &\qquad\;\; t_I \leq T_s \end{aligned} \tag{1}$$

where N is the total number of nodes in the network, N_s is the number of sensors, T_s is the simulation period, M is the pollutant injection vector, n is the network node number injected by pollution sources, t_I is the initial time of pollutant injection, $c_j(t)$ is the pollutant concentration of sensor j at time t, which is a function of (M,n,t_I), and $c_j^*(t)$ represents the actual pollutant concentration measured by sensor j at time t. The goal of optimization is to find (M,n,t_I) to minimize the variance.

The most important information in pollution sources identification problems is location and time. In other words, if a pollution incident occurs in the real environment, the most important thing is to obtain the location and time of the injection so as to facilitate timely investigation and handling by relevant personnel. In the actual monitoring process, there will be errors in the quality information obtained by the sensor, the superposition of the final fitness value square error will be large [50]. So the quality information does not necessarily need to be accurate, the most important thing is to get the exact location and time. In this paper, when evaluating the stability of the algorithm, whether the optimal solution can obtain the precise position and time are taken as the criteria for evaluation.

2.2. Collaborative based pollution sources identification algorithm

2.2.1. Collaborative approach

Collaborative algorithm adopts a "divide and rule" idea, which decomposes the complex problem into several sub-problems, and then solves each sub-problem separately, exchanges between the populations and cooperates to optimize the collaborative algorithm. Distributed evolutionary algorithms in collaborative algorithm search for different sub-populations separately and share information through population migration. Different populations can use different strategies and algorithms to achieve co-evolution [51,52]. This paper combines the characteristics. While decomposing the problem, the characteristics of the specific problem of pollution sources are combined, different strategies are adopted to search for the variable space with different meanings. Under normal circumstances, decomposing the problem is a difficult but important issue. This paper is to solve pollution sources problem of four variables, it is clear that each variable is given the meaning, so each variable is divided into a sub-population. As mentioned in the above that the two variables, start time and duration, will jointly determine a continuous time series, they cannot be split. The specific framework is shown in the following figure:

As shown in Fig. 1, the populations are divided into three sub-populations after initialization—location, time and mass populations. They have the same decision variables with different search space. In other words, each population searches for the corresponding decision space (underlined), while the other decision space (not underlined) does not change. After searching reaches a certain number of iterations, the groups exchange and share information.

2.2.2. Improved strategy of pollution sources identification problems

Pollution sources identification problem is an engineering problem, not a purely theoretical problem. It has the characteristics of itself. In a GA based on hybrid coding, the problem may not be prominent if there are fewer network nodes. When there are more nodes, it is easy to fall into the local optimum, and the performance of the algorithm becomes unstable. To solve these problems, this paper proposes different improvement strategies to guide the individual search according to the characteristics of pollution sources problems.

2.2.3. Improved proximity search strategy

In many pollution sources identification problems, the time and duration of pollutant injection are usually coded separately to form two independent variables, corresponding to the positions of the genes. For example: (3, 3) represents the start time and duration, when the location and the quality

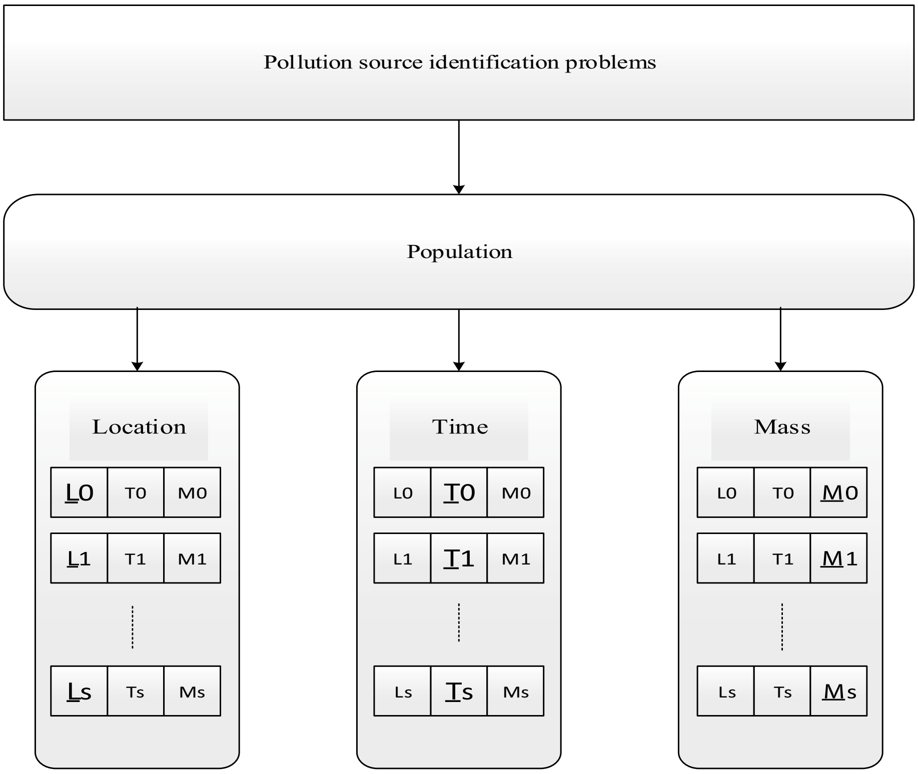


Fig. 1. Framework of co-evolutionary population division.

curve are also determined, an individual is identified, a pollution sources event can be simulated by substituting these parameters into the EPANET. In these treatments, the start time and the duration are separated independently, without noticing that the two actually determine a continuous time series, and the quality corresponds to the time series one by one. With this knowledge, we can see that (2, 4) represents four consecutive time series of 2, 3, 4, 5, (3, 3) represents 3, 4, 5, (5, 4) represents 5, 6, 7, 8. In fact, (2, 4) and (3, 3) are closer than (2, 4) and (5, 4), although the latter group has been consistent in the variable of duration. The more coincident the time series, the smaller the difference between the concentration difference detected by the sensor and the simulative sensor.

When the start time and the duration fall into the local optimum, it is difficult for ordinary crossover operators to jump out. For example, (2, 4) is the start time and duration of a real pollution sources event, but it is very difficult to achieve (2, 4) when it is in (3, 3) local. Since falling into the local optimum solution usually means that most individuals have this value, crossover has been very difficult at this time, and the probability of mutation operator from (3, 3) to (2, 4) is very small. In this paper, a proximity search mechanism that can jump out of the local optimum while maintaining individual diversity is proposed. The start time and the duration are taken as a whole. In each iteration, they collectively determine a time series to search for the time series with the highest degree of coincidence. This paper imitates the principle of Harley coding, the searched time series and itself is only a period of time difference (Fig. 2).

In the above figure, it is assumed that the time interval is a total of 12 h from 0 to 11, and gray area represents the injection of pollution sources in this period of time. The second line shows the time series corresponding to (2, 4). It may be in different directions in proximity search, but only differ by four-time series of a time period from itself, that are (3, 3), (1, 5), (2, 3), (2, 5) represented by 3, 4, 5, 6, respectively. It randomly chooses one of the searches and forms a corresponding new start time and duration.

2.2.4. Improved mutation strategy

In experiments, we found that whether using surrogate model or not, upon the algorithm converges to the exact position, and upon the start time, duration and fitness value reach a small value, it is difficult to continue convergence, and the quality curve cannot be precise. After testing a variety of evolutionary calculations (at fixed position, start time and duration), we found that the fitness value of the GA drops very fast in the initial stage, but becomes slow or even remains the same in the following generations, which cannot reach a small value. The PSO happens to be the opposite, the early stage changes slowly, but can reach a very small value. In the collaborative algorithm, since the exchange of information is to produce new individuals through the combination of gene fragments, it is a good property to decrease the fitness value quickly, but it is also necessary to achieve a small value. So, keeping the crossover algorithm of GA, improving the mutation strategy, in combination with the PSO idea, we can achieve the best individual variation. By adding an adaptive factor, we can ensure the diversity of individuals. During the operation of the algorithm, the adaptive factor will gradually become smaller as the algorithm converges, so that the range of variation can be adaptively adjusted.

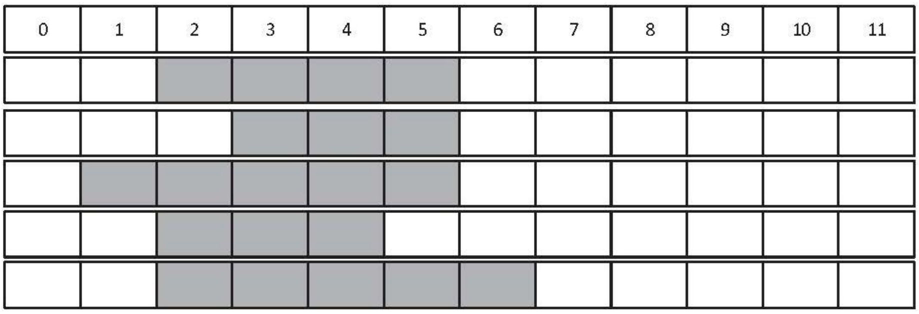
$$\text{indi}\left(i_j\right) = \frac{\left(\text{indi}\left(i_j\right) + \text{Bestindi}\left(i_j\right)\right)}{2 \pm \text{rand} \times \text{Maxd}} \tag{2}$$

Here 'indi' represents an individual, i represents the number of the mutation that needs to be mutated, that is, the dimension, 'Bestindi' represents the best individual in the 'Pm' population, 'rand' represents a random number between (0,1), and 'Maxd' represents the number of maximum spacing of all individuals in the jth dimension.

2.2.5. Improved perturbation method

The search space for some problems is very rugged and unsmooth in some parts, making it easy to fall into the local optimum in solving problems. However, under normal circumstances, the global optimal solution is also near the local optimal solution. This is because of the mutation points in the search space, making it difficult to get the global optimal solution. The perturbation method slightly modifies the value of variables. This method can increase individual diversity and can effectively jump out of the local optimum in solving the problems with a very rough search space.

In pollution sources identification problems, especially in large pipe networks, it is easy to locate a good area, but it is also easy to fall into the local optimum. After analyzing the entire network topology, it is found that its location is not continuous, that means the topology of other nodes



0	1	2	3	4	5	6	7	8	9	10	11
		▨	▨	▨	▨						
			▨	▨	▨						
	▨	▨	▨	▨	▨						
		▨	▨	▨							
		▨	▨	▨	▨	▨					

Fig. 2. Proximity search.

around one node is similar, but their location labels are not continuous (Fig. 3). In this way, when a good solution is found, iteration gradually converges toward this point, but when the location label gradually converges and maps to the topological structure of the pipe network, it searches toward other areas. So that good areas around good individuals do not conduct very efficient searches, resulting in easy falling into the local optimum, maintaining local optimum for many generations without change.

In order to solve this problem, perturbation strategy is added in the search of pollution sources position. In general, when an individual appears the same, a small perturbation such as ceil (Normrnd (0, 1)) is added to a certain one-dimensional variable. Normrnd (0, 1) means standard normal distribution, ceil means rounding. In this paper, we find that in the elite strategy, in case of the same number of individuals, the same individuals are perturbed. The perturbation, combined with the characteristics of pollution sources problem, searches the nodes with similar topological structure in the position. The same elite individuals are replaced when there are better individuals (evaluation of evaluation function), and individuals who are different from the same elite individuals in the population (fitness values) are selected directly if there are no better individuals. The proposed strategy can further maintain the diversity of the population and largely solve the problem of falling into the local optimum caused by the discontinuity of the continuous position labels of the topological structure in the pipe network.

Predictive simulation has the most direct impact on the evaluation of individuals, so an appropriate prediction model is the key to solving expensive optimization problems. The Gaussian random process [8] model is a method of establishing surrogate models. Gaussian random process simulation has fewer parameters and is easily solved by maximum likelihood and optimization algorithms. Gaussian random process is taken as a surrogate model because of its three characteristics. (1) Gaussian random process has strong ability of overcoming over-fitting; (2) the parameters of Gaussian random process model are limited and adjustable; (3) Gaussian random process model can add samples to update model in real time after modeling, which is more beneficial to improve the model accuracy.

In this paper, we build expensive optimization model by using the modeling method of a paper on expensive optimization to be prepared. The sample used in establishing a Gaussian prediction model in a thesis is obtained before, and the sample has not changed during the entire algorithm running. As for the locations in the pollution sources identification problem, some of those continuous in the topology are not continuous on the location label, which leads to the unobvious position correlation. When it comes to finding the correlation matrix, the relevance of the position is not obvious or there is no rule. Therefore, when selecting the most recent points of the current sample and the previous sample as the training set of the model, the properties or accuracy are not greatly improved. The variables in time and mass remain unchanged, the relevance at this time is very high, if the data in the program is retained, the most recent points of the current data and the original sample data are taken as the training points to greatly improve the accuracy of the model. The specific approach is shown in Fig. 4:

The DB in the figure above represents the previous sample, and DBt represents the temporary data sample, which is used to save the data evaluated by the real evaluation function during the operation of the algorithm. According

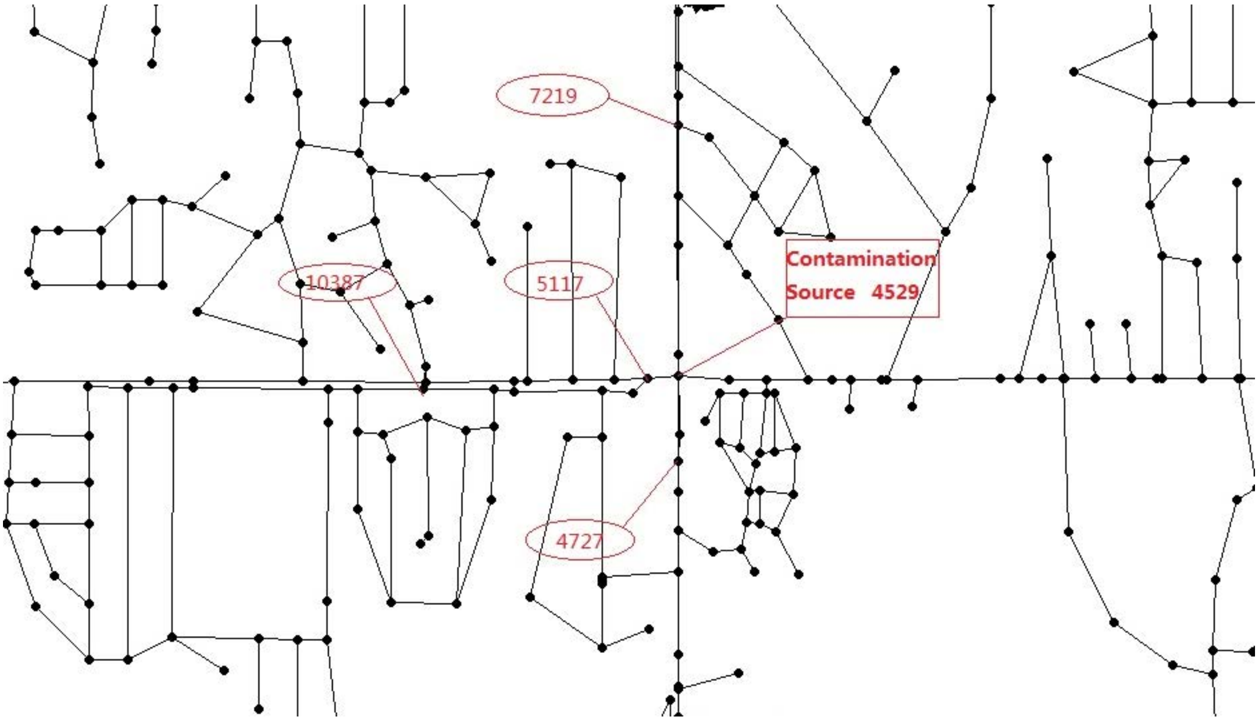


Fig. 3. A partial enlarged view of the pipe network BWSN2 [32] *Solution Model of Expensive Optimization Algorithm.*

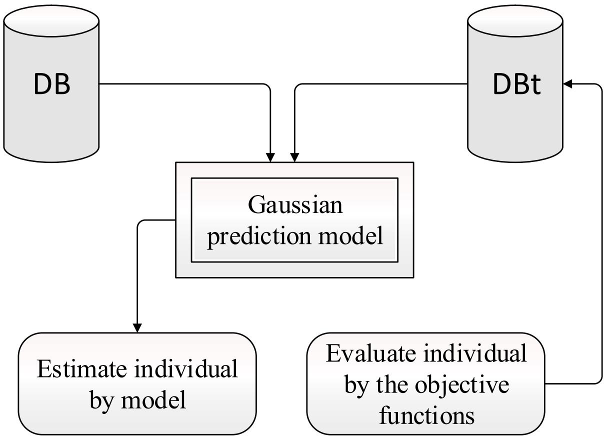

DB
DBt
Gaussian prediction model
Estimate individual by model
Evaluate individual by the objective functions

Fig. 4. Individual evaluation process.

to (in a thesis) the condition of using Gaussian prediction model, the unqualified ones are assessed using real evaluation function and the data are saved to DBt, the qualified ones are established with a Gaussian prediction model using the appropriate (nearest) data selected from the DB and DBt, and this model is used to evaluate the individual.

When using simulation–optimization model to solve pollution sources identification problems, EPANET is used as a simulator and optimization algorithm as an optimizer. Unlike the general simulation–optimization models, the EPANET simulator or Gaussian process surrogate model is used to calculate individual fitness values. The introduction of Gaussian process surrogate model in the optimization algorithm can reduce the use of EPANET simulator and improve the efficiency of the algorithm. The solution framework based on the expensive optimization algorithm is shown in Fig. 5.

In the expensive optimization algorithm, Gaussian random process modeling is adopted, collaborative algorithm is used as an optimization algorithm. Each individual in the population represents a pollution incident. The EPANET simulator can be used to simulate a pollution incident and output the actual pollutant concentration information of the network node. By comparing with the information actually detected by the sensor, the individual fitness value can be calculated, and also the fitness value of the individual can be predicted by the Gaussian process surrogate model. By reasonably balancing the use of EPANET with the Gaussian process surrogate model, the EPANET simulator can be used as little as possible while guaranteeing positioning accuracy, reducing the time cost of the algorithm. Therefore, the algorithm has two main problems. One is how to build an appropriate Gaussian process surrogate model, and the other is how to balance the use of the Gaussian process surrogate model with the EPANET simulator.

2.2.6. Based on collaborative expensive optimization algorithm

Based on the classic solving process of expensive optimization, Gaussian process surrogate model is introduced in the process of optimization to reduce the number of real evaluation functions. In the process of optimization, the strategy is constantly adjusted to balance the use of the Gaussian prediction model with the EPANET simulator so that the algorithm reduces the use of the EPANET simulator while achieving the required accuracy. According to the solution framework of pollution sources identification based on expensive optimization algorithm shown in Fig. 1,

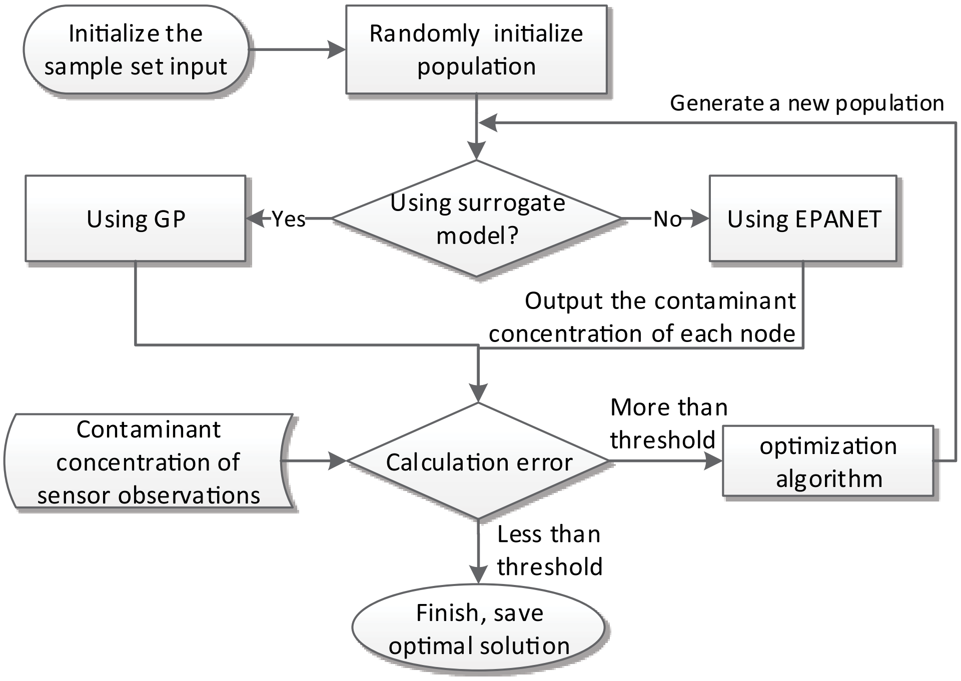

Initialize the sample set input
Randomly initialize population
Generate a new population
Using GP
Yes
Using surrogate model?
No
Using EPANET
Output the contaminant concentration of each node
Contaminant concentration of sensor observations
Calculation error
More than threshold
optimization algorithm
Less than threshold
Finish, save optimal solution

Fig. 5. Solution framework diagram based on expensive optimization algorithm.

this paper proposes a collaborative expensive optimization algorithm, the flow chart is shown in Fig. 6.

According to the above strategy using surrogate model, the detailed steps of the algorithm are as follows:

Step 1: Initialize the population. In order to minimize the use of EPANET, population P is initialized directly from the sample set.

Step 2: The population P is evenly divided into three sub-populations, namely Pl, Pt, Pm.

Step 3: Improve roulette options. In order to make the algorithm fall into the local optimum, this paper uses an improved roulette selection method.

Step 4: Use the following strategies for the three sub-populations.

Step 4.1: For population Pl, the crossover operator adopts two-point crossover, and the mutation operator adopts single-point mutation and Gaussian mutation.

Step 4.2: For population Pt, the crossover operator adopts two-point crossover followed by the neighbor search strategy presented above.

Step 4.3: For population Pm, the crossover operator uses real-numbers crossover, and the mutation adopts the improved mutation strategy presented above.

Step 5:

Step 5.1: The individual fitness values μ (predicted values) and forecast error σ of the crossover or mutation-producing individuals are predicted using a Gaussian process surrogate model. If the trigger coefficient $3\sigma/\mu < 0.2$ (0.2 is obtained by experimental analysis), predicted values are used directly as the fitness values of new individuals, otherwise enter Step 5.2.

Step 5.2: Generate a probability P_* randomly. If $P_* < P$, $P = t/x$, use EPANET to calculate the fitness value, where t is the number of iterations and x is the radix, otherwise the individual fitness value is calculated using the Gaussian process surrogate model.

Step 6: After each iteration, the population is sorted by fitness values. As for the first N individuals of which the fitness values are calculated using the Gaussian process surrogate model, the real model EPANET is used for correcting the fitness values.

Step 7: Use elitist strategy, retention strategy and elitist strategy for the three sub-populations, respectively. If the elite individuals in population Pl are the same, then the above-mentioned perturbation strategy is used, while in Pm, the disturbance strategy is used directly for an elite individual.

Step 8: After an integer multiple of S, the three sub-populations combine the genes of the optimal individuals to generate new individuals and uniformly add them to the three sub-populations (directly replace the corresponding worst individuals) Otherwise skip to Step 1.

Step 9: Determine whether the stop condition is satisfied. If the condition is satisfied, the algorithm ends; otherwise, skip to Step 2.

3. Results

3.1. Experimental simulation and analysis

3.1.1. Water supply network parameters and algorithm parameters setting

To show the necessity of the surrogate model, a large-scale pipe network BWSN2 was used in this paper. As shown in Fig. 7, the pipeline network contains 12,527 nodes, 2 reservoirs and 2 ponds. In the water network, 20 sensors are arranged {7,626, 8,912, 5,363, 6,632, 6,725, 4,889, 10,861, 2,372, 8,820, 3,070, 6,840, 11,550, 3,430, 7,959, 6,744, 9,488, 11,330, 7,211, 6,006, 5,890}. The total simulation time of the pipe network was 48 h, the simulation hydraulic time step took 1 h, the water quality time step took 5 min, and the real pollution scene was the continuous injection of pollutants from node 4,529 for 4 h which started 2 h after the simulation. The relevant parameters of the expensive optimization algorithm based on the Gaussian process surrogate model are shown in Table 1.

Experimental environment: the processor of the simulation machine is Intel Core i5–6500 @ 3.20 GHZ, RAM is 8.0 GB, OS (Operating System) is Windows 7 Professional 64-bit. The experiments in this paper are algorithmic performance analysis, and then comparison of algorithms using the surrogate model and not using the surrogate model. The effectiveness and efficiency of the expensive optimization algorithm based on the Gaussian process surrogate model are verified by analyzing the number of EPANET evaluations and the time cost of the algorithm.

3.1.2. Algorithm performance analysis

The expensive optimization algorithm uses a surrogate model with little computation or time consuming to replace the computationally intensive real objective function. In this paper, Gaussian random process is used to obtain the approximation of the fitness value of the EPANET simulator to reduce the time cost. In order to balance the accuracy of pollution sources identification while reducing the time, this paper follows the usage strategy of Gaussian process surrogate model and the dynamic selected sample of experimental operating data to update model. In this section, through the simulation experiment, the comparison between the established Gaussian process surrogate model and EPANET is carried out first, and then the validity of the collaborative expensive optimization algorithm used in this paper is verified through many experiments.

In the process of algorithm convergence, extensive use of Gaussian process surrogate model can greatly reduce the cost of time. Fig. 8 shows a comparison of the time required to predict individual fitness values using a single EPANET simulator and using a one-time Gaussian process surrogate mode. The one-time Gaussian process surrogate model reduces the time by approximately 20 times as compared with the EPANET simulator. It can be proved that the extensive use of Gaussian process surrogate model can greatly reduce the time cost of the algorithm.

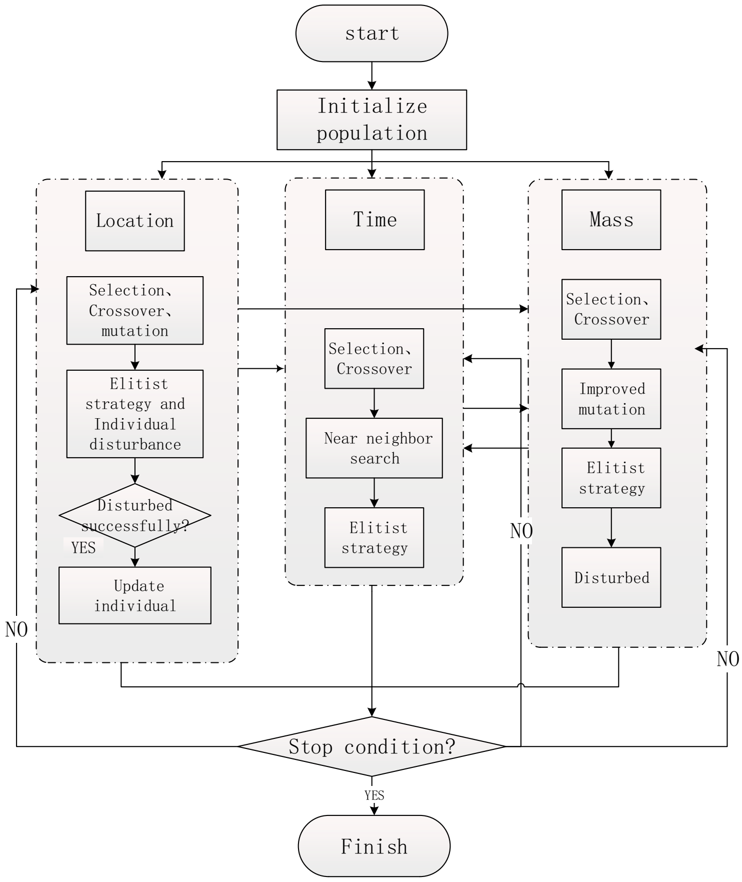


Fig. 6. Collaborative expensive optimization algorithm.

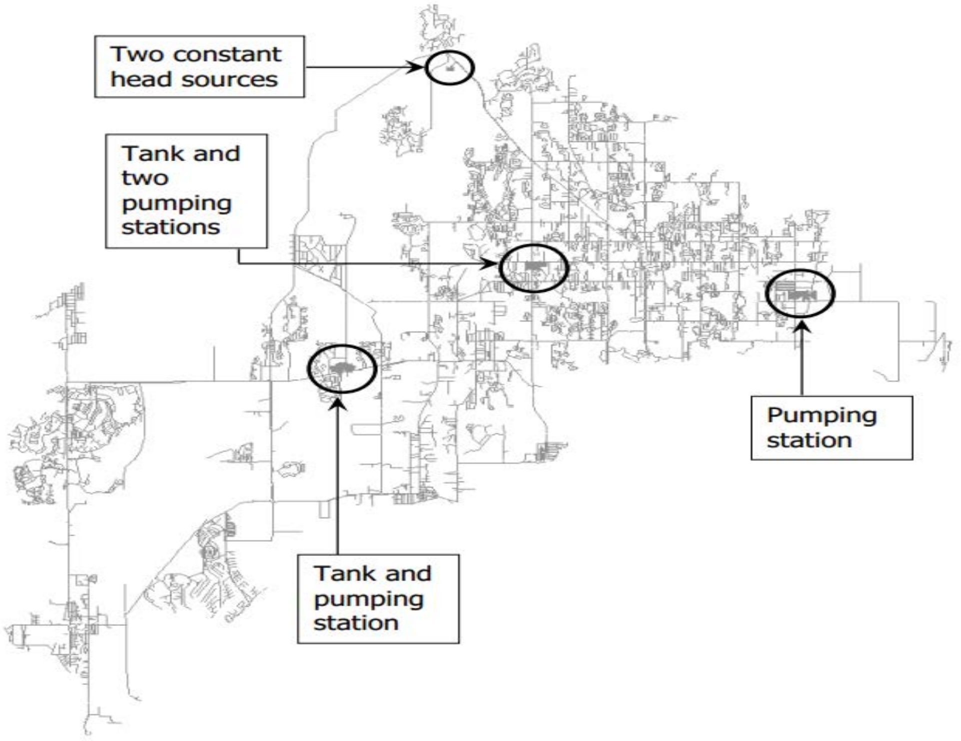


Fig. 7. BWSN2.

Table 1
Algorithm parameters setting

Parameter	Description	Value
POP_SIZE	Population size	100
NUM_ITRE	Number of iterations	100
Pc	Crossover probability	95%
Pm	Mutation probability	70%
M	Individual selected by the elite strategy	5
n	Improved roulette parameter	6
N	Number of updated individuals	10
S	Number of iterations of information exchange	5

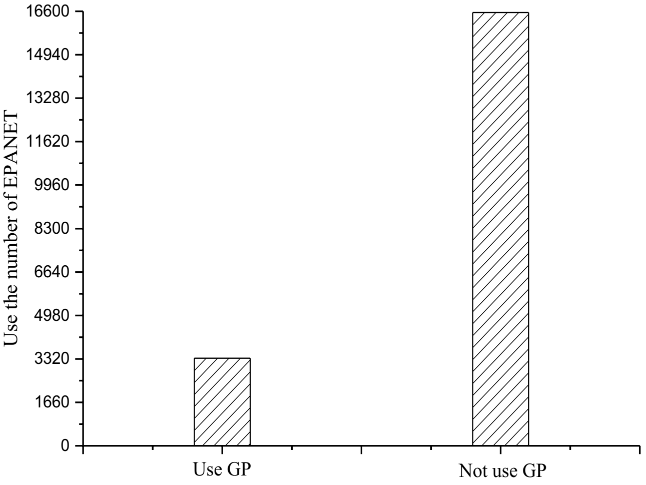


Fig. 8. Time spent in a single evaluation.

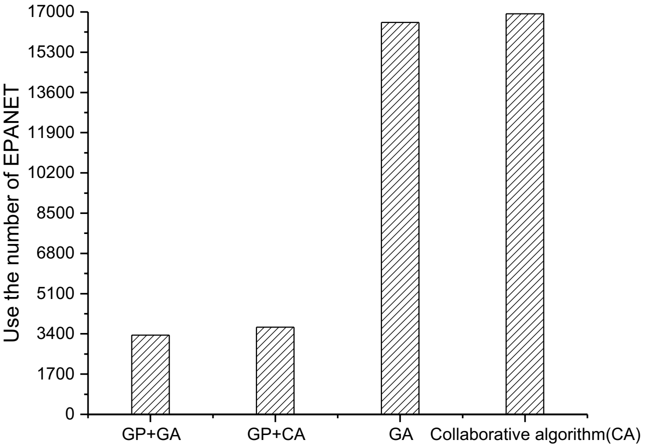


Fig. 9. EPANET using the Gaussian model and not using the Gaussian model under the same result.

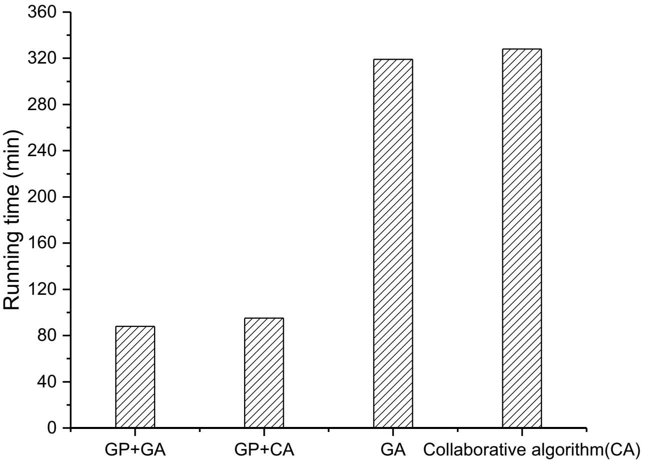


Fig. 10. Time consumed using the Gaussian process surrogate model and not using the Gaussian process surrogate model under the same result.

After verifying that the Gaussian process, surrogate model can reduce the time by theory, we adopted the appropriate strategy and used surrogate model reasonably. In this paper, we used ordinary GA and the algorithm without surrogate model to compare the time consumed in using EPANET and the running time of the algorithm under the same pollution incident. Then compare the number of EPANET uses and the time spent on the program after adding the surrogate model. After 20 iterations, the number of EPANET uses and the time spent on using the Gaussian process surrogate model and not using the Gaussian process surrogate model are shown in Figs. 9 and 10. As can be seen from the figure, after using the surrogate model reasonably, both algorithms can significantly reduce the number of EPANET uses and the time spent on the algorithm.

However, it is clear from the experimental results that the time between the two algorithms and the number of EPANET uses are not significantly different, and even higher in this paper. The main reason for this problem is that in the collaborative algorithm, after a certain number of iterations, the population needs the exchange of information, so there will be a certain EPANET calling, so the time and the use of EPANET are higher. In case when the pollution sources can be identified, EPANET use in Gaussian process surrogate model greatly reduces by nearly 2/3 times, so that the algorithm consumes less time, greatly improving the efficiency of the algorithm. It further proves the efficiency of GA added with a surrogate model and the proposed algorithm.

In the above experiment, we only analyzed the difference of time and EPANET uses, but the goal of expensive optimization is to reduce the time consumption as much as possible while getting accurate solutions. Combined with the pollution sources identification problems, whether the algorithm can find the exact information of pollution sources reflects the stability of the algorithm. Evolution algorithms are not widely applied in industrial engineering mainly due to the poor stability, so the stability analysis is very important. Previous work has found that the general algorithms such as ordinary GA is more stable in small pipe networks, but due to the large solution space in large pipe networks, the algorithm is very unstable, easily falling into the local optimum.

This paper has done some work to avoid falling into the local optimum in the actual pollution sources identification problems, so as to verify its effectiveness. By comparing the complete algorithm of this paper, the ordinary GA with an expensive model, the ordinary GA and the algorithm of this paper without an expensive model, the results were obtained after 20 iterations. As shown in Table 2, as mentioned above, either the proposed algorithm or the GA with an expensive model can both effectively reduce the use of EPANET and reduce the time after adding an expensive model.

However, as shown in Table 3, after repeated experiments, the proposed algorithm is obviously superior to the ordinary GA with Gaussian process surrogate model, both in the identification of the pollution sources and the start time and the duration. Combined with Table 2, compared with the algorithm without an expensive model, the proposed algorithm greatly reduces the calculation time. Comparing the GA with a Gaussian process surrogate model, although the average time increased by 6 min, the stability of the algorithm has been greatly improved. In pollution sources identification problems, the location and time are the most important. In a real environment, in the event of a pollution incident, an accurate location and time can help to effectively deal with and control the pollution. Therefore, the stability of the algorithm is very important. Although the poor stability of the algorithm may sometimes lead to a feasible solution with low fitness value, it is not advisable given its high uncertainty.

The proposed algorithm has a very small fitness value when getting the optimal solution but may be larger than that of other optimization problems. This is mainly due to the superposition of errors in the objective function by using simulation and optimization to solve the pollution sources identification problems, aiming for a more vivid representation of the pollution sources concentration accuracy solved by the proposed algorithm. By comparing the information obtained by the sensor of which the optimal solution is taken as the pollution sources and by the real sensor, the pollution

Table 2
Experimental results comparison

	Time costs	Number of EPANET calls	Optimal solution fitness
Collaborative algorithm (CA) + GP	94 min	3,681 times	9.22
GA + GAM	88 min	3,550 times	17.81
CA	3,325 min	16,716 times	5.1
GA	319 min	16,555 times	7.8

Table 3
Experimental results

Trials	CA + GP			GA + GP		
	Location + 222222222	Start Time	Duration	Location	Start Time	Duration
1	4,529	2	4	4,529	3	3
2	4,529	2	4	4,531	2	4
3	4,529	2	4	4,531	3	3
4	4,587	1	5	10,386	3	5
5	4,529	2	4	4,531	2	5
6	4,529	2	4	4,529	2	4
7	4,529	2	4	4,529	2	4
8	4,529	2	4	4,739	1	5
9	5,116	4	4	4,529	2	4
10	4,529	2	4	4,529	2	4
11	4,529	2	4	4,529	2	4
12	4,529	2	4	4,531	0	5
13	4,531	2	4	4,529	1	4
14	4,529	2	4	4,587	3	2
15	4,529	2	4	4,529	2	4
16	4,529	2	4	4,529	1	5
17	4,529	2	4	4,529	1	4
18	4,529	2	4	4,529	2	4
19	4,529	2	4	8,139	1	4
20	4,529	2	4	4,529	1	5

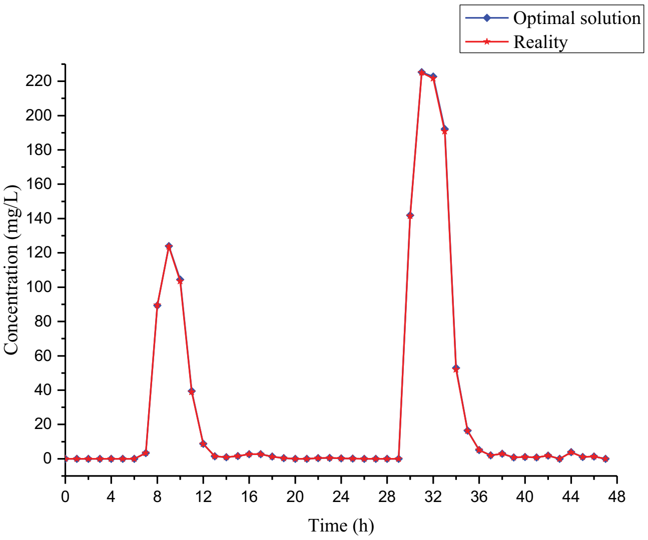


Fig. 11. Concentration information detected by the sensor 7626.

sources information obtained from the proposed algorithm is further verified. As shown in Fig. 11, which is a concentration curve of one of the sensors 7,626 under the EPANET simulation, it is seen that the curve of the concentration information of pollutant detected by the sensor 7,626 corresponding to the optimum solution is consistent. It shows that the proposed algorithm can effectively solve the feasible solutions and basically match with the actual pollution sources information.

4. Conclusions

Pollution sources identification problem is an interdisciplinary problem in the field of environmental science and computational science. In this paper, pollution sources identification problems are transformed into function optimization problems by using simulation–optimization model. According to the expensiveness of pollution sources identification problems, these problems are transformed into expensive optimization problems, which can easily fall into the local optimum when the numbers are not actually numbered in spatially adjacent positions and are not processed in time. To solve this problem, based on the collaborative expensive optimization algorithm, according to the characteristics of water supply networks and different variables, this paper proposes a strategy that is more suitable for variables for each variable so as to guide the search direction of the algorithm. The algorithm uses as many Gaussian process surrogate models as possible while ensuring the positioning accuracy. Finally, the simulation experiment is conducted to verify the effectiveness, efficiency and stability of the proposed algorithm.

In the study of pollution sources identification problems, when the urban pipeline network nodes exceed 1,000 and the water requirement of users' changes in real time, the problem can be abstracted as an optimization problem of dynamic, expensive and multimodal functions. Therefore, it is necessary to further propose a solution to the dynamic multimodal expensive optimization problems, which is also the follow-up research work of this paper. Considering the importance of location for pollution sources identification problems, changing the method of evaluation instead of just by the sum of squares of errors requires further consideration in the future.

Acknowledgments

This paper is supported by Natural Science Foundation of China. (No. 61673354), the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (CUGGC03), Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing and the State Key Lab of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology (DMETKF2018020).

References

[1] M.T. Ayvaz, A hybrid simulation–optimization approach for solving the areal groundwater pollution source identification problems, J. Hydrol., 538 (2016) 161–176.

[2] J. Gao, L.-L. Yue, X. Jiang, L. Ni, M.F. Saleem, Y. Zhou, K. Li, J. Xiao, Phylogeographic patterns of *Microtus fortis* (Arvicolinae: Rodentia) in China based on mitochondrial DNA sequences, Pak. J. Zool., 49 (2017) 1185–1195.

[3] A.J. Ahamed, K. Loganathan, S. Ananthakrishnan, J.K.C. Ahmed, M.A. Ashraf, Evaluation of graphical and multivariate statistical methods for classification and evaluation of groundwater, Appl. Ecol. Environ. Res., 15 (2017) 105–116.

[4] K.S. Bhattacharjee, T. Ray, An Evolutionary Algorithm with Classifier Guided Constraint Evaluation Strategy for Computationally Expensive Optimization Problems, Australasian Joint Conference on Artificial Intelligence, Springer, Cham, 2015, pp. 49–62.

[5] W.X. Chen, T. Weise, Z. Yang, K. Tang, Large-scale Global Optimization Using Cooperative Coevolution with Variable Interaction Learning, International Conference on Parallel Problem Solving from Nature, Springer, Berlin, Heidelberg, 2010, pp. 300–309.

[6] M. Sudhakaran, D. Ramamoorthy, V. Savitha, S. Balamurugan, Assessment of trace elements and its influence on physicochemical and biological properties in coastal agroecosystem soil, Puducherry region, Geol. Ecol. Landscapes, 2 (2018) 169–176.

[7] W. Deng, H. Zhao, L. Zou, G. Li, X. Yang, D. Wu, A novel collaborative optimization algorithm in solving complex optimization problems, Soft Comput., 21 (2017) 4387–4398.

[8] Z. Feng, Q. Zhang, Q. Zhang, Q. Tang, T. Yang, Y. Ma, A multiobjective optimization based framework to balance the global exploration and local exploitation in expensive optimization, J. Global Optim., 61 (2015) 677–694.

[9] F. Qiao, Research on design principles of visual identity in campus environment, Sci. Heritage J., 2 (2018) 1–3.

[10] J.E. Fieldsend, R.M. Everson, On the Efficient Use of Uncertainty When Performing Expensive ROC Optimization, Evolutionary Computation, CEC 2008 (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 3984–3991.

[11] J. Gu, M. Gu, C. Cao, X. Gu, A novel competitive co-evolutionary quantum GA for random job shop scheduling problem, Comp. Oper. Res., 37 (2010) 927–937.

[12] N.S. Ramli, N.H.M. Zin, Alpha-amylase inhibitory activity of inhibitor proteins in different types of commercial rice, Sci. Heritage J., 2 (2018) 27–29.

[13] J. Guan, M.M. Aral, M.L. Maslia, W.M. Grayman, Identification of contaminant sources in water distribution systems using simulation-optimization method: case study, J. Water Resour. Plann. Manage., 132 (2006) 252–262.

[14] W.D. Hillis, Co-evolving parasites improve simulated evolution as an optimization procedure, Physica D, 42 (1990) 228–234.

[15] C. Hu, J. Zhao, X. Yan, A Map, Reduce based Parallel Niche GA for contaminant source identification in water distribution network, Ad Hoc Networks, 35 (2015) 116–126.

[16] T.D.T. Oyedotun, L. Johnson-Bhola, Beach litter and grading of the coastal landscape for tourism development in sections of Guyana's coast, J. CleanWAS, 3 (2019) 1–9.

[17] S. Jeong, S. Obayashi, Efficient Global Optimization (EGO) for Multi-Objective Problem and Data Mining, IEEE Congress on Evolutionary Computation, Edinburgh UK, 2005, pp. 2138–2145.

[18] Y. Jin, M. Olhofer, B. Sendhoff, A framework for evolutionary optimization with approximate fitness functions, IEEE Trans. Evol. Comput., 6 (2002) 481–494.

[19] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, J. Global Optim., 13 (1998) 455–492.

[20] J. Ali, A.A.J. Mohamed, M.S.A. Kumar, B.A. John, Organophosphorus pesticides toxicity on brine shrimp, Artemia, J. Clean WAS, 2 (2018) 23–26.

[21] A.J. Keane, Statistical improvement criteria for use in multiobjective design optimization, AIAA J., 44 (2006) 879–891.

[22] Y. Li, Z.H. Zhan, S. Lin, J. Zhang, X. Luo, Competitive and cooperative particle swarm optimization with information sharing mechanism for global optimization problems, Inf. Sci., 293 (2015) 370–382.

[23] B. Liu, Q. Zhang, G.G.E. Gielen, A Gaussian process surrogate model assisted evolutionary algorithm for medium scale expensive optimization problems, IEEE Trans. Evol. Comput., 18 (2014) 180–192.

[24] O.T. Joseph, O.O. Adeoti, A.A. Olufemi, Study of the phytodiversity along Antorun Reservoir, near Ogbomoso, Nigeria, Environ. Ecosyst. Sci., 3 (2019) 1–12.

[25] L. Liu, S.R. Ranjithan, G. Mahinthakumar, Contamination source identification in water distribution systems using an adaptive dynamic optimization procedure, J. Water Resour. Plann. Manage., 137 (2010) 183–192.

[26] W. Liu, Q. Zhang, E. Tsang, B. Virginas, Fuzzy Clustering Based Gaussian Process Model for Large Training Set and its Application in Expensive Evolutionary Optimization, IEEE Congress on Evolutionary Computation, Cec. Trondheim, Norway, 2009, pp. 2411–2415.

[27] M. Wilson, M.A. Ashraf, Study of fate and transport of emergent contaminants at wastewater treatment plant, Environ. Contam. Rev., 1 (2018) 1–12.

[28] C. Luo, S.-L. Zhang, C. Wang, Z. Jiang, A metamodel-assisted evolutionary algorithm for expensive optimization, J. Comput. Appl. Math., 236 (2011) 759–764.

[29] M.J. OmaraShahestan, S. OmaraShastani, Evaluating environmental considerations with checklist and delphi methods, case study: Suran city, Iran, Environ. Ecosyst. Sci., 1 (2017) 1–4.

[30] M.N. Omidvar, X. Li, Y. Mei, X. Yao, Cooperative co-evolution with differential grouping for large scale optimization, IEEE Trans. Evol. Comput., 18 (2014) 378–393.

[31] G.G. Mahmood, H. Rashid, S. Anwar, A. Nasir, Evaluation of climate change impacts on rainfall patterns in Pothohar Region of Pakistan, Water Conserv. Manage., 3 (2019) 1–6.

[32] A. Ostfeld, J.G. Uber, E. Salomons, J.W. Berry, W.E. Hart, C.A. Phillips, J.P. Watson, G. Dorini, P. Jonkergouw, Z. Kapelan, F.D. Pierro, S.T. Khu, D. Savic, D. Eliades, M. Polycarpou, S.R. Ghimire, B.D. Barkdoll, R. Gueli, J.J. Huang, E.A. McBean, W. James, A. Krause, J. Leskovec, S. Isovitsch, J. Xu, C. Guestrin, J. VanBriesen, M. Small, P. Fischbeck, A. Preis, M. Propato, O. Piller, G.B. Trachtman, Z.Y. Wu, T. Walski, The battle of the water sensor networks (BWSN): a design challenge for engineers and algorithms, J. Water Resour. Plann. Manage., 134 (2008) 556–568.

[33] I. Paenke, J. Branke, Y. Jin, Efficient search for robust solutions by means of evolutionary algorithms and fitness approximation, IEEE Trans. Evol. Comput., 10 (2006) 405–420.

[34] X. Peng, Y. Wu, Large-scale cooperative co-evolution using niching-based multi-modal optimization and adaptive fast clustering, Swarm Evol. Comput., 35 (2017) 65–77.

[35] N.S. Zafisah, W.L. Ang, A.W. Mohammad, Cake filtration for suspended solids removal in digestate from anaerobic digested palm oil mill effluent (POME), Water Conserv. Manage., 2 (2018) 5–9.

[36] W. Ponweiser, T. Wagner, D. Biermann, Multiobjective Optimization on a Limited Budget of Evaluations Using Model-Assisted *S*-Metric Selection, International Conference on Parallel Problem Solving from Nature: PPSN X, Springer, Berlin, Heidelberg, 2008, pp. 784–794.

[37] M.A. Potter, K.A. De Jong, A Cooperative Co-evolutionary Approach to Function Optimization, International Conference on Parallel Problem Solving from Nature, Springer, Berlin, Heidelberg, 1994, pp. 249–257.

[38] R.G. Regis, C.A. Shoemaker, Local function approximation in evolutionary algorithms for the optimization of costly functions, IEEE Trans. Evol. Comput., 8 (2004) 490–505.

[39] H.K. Singh, A. Isaacs, T. Ray, A Hybrid Surrogate-based Algorithm (HSBA) to Solve Computationally Expensive Optimization Problems, Evolutionary Computation (CEC), 2014 IEEE Congress, 2014, pp. 1069–1075.

[40] R. Subbu, A.C. Sanderson, Modeling and convergence analysis of distributed co-evolutionary algorithms, IEEE Trans. Syst. Man Cybern. Part B Cybern., 34 (2004) 806–822.

[41] C. Sun, Y. Jin, R. Cheng, J. Ding, J. Zeng, Surrogate-assisted cooperative swarm optimization of high-dimensional expensive problems, IEEE Trans. Evol. Comput., 21 (2017) 644–660.

[42] Y. Tenne, C.K. Goh, Computational Intelligence in Expensive Optimization Problems, Springer, Berlin, Heidelberg, 2012.

[43] L. Wang, S. Wang, Y. Xu, G. Zhou, M. Liu, A bi-population based estimation of distribution algorithm for the flexible job-shop scheduling problem, Comput. Ind. Eng., 62 (2012) 917–926.

[44] Q.X. Wei, X.F. Liu, Q. Huang, The comparison of selection methods in different GAs, J. Commun. Comput., Chinese/English Version, 8 (2008) 61–65.

[45] X. Yan, J. Zhao, C. Hu, Q. Wu, Contaminant source identification in water distribution network based on hybrid encoding, J. Comput. Methods Sci. Eng., 16 (2016) 379–390.

[46] X. Yan, J. Sun, C. Hu, Research on contaminant sources identification of uncertainty water demand using GA, Cluster Comput., 20 (2017) 1007–1016.

[47] X. Yan, Z. Zhu, T. Li, Pollution source localization in an urban water supply network based on dynamic water demand, Environ. Sci. Pollut. Res., 26 (2019) 17901–17910.

[48] X. Yan, T. Li, C. Hu, Real-time localization of pollution source for urban water supply network in emergencies, Cluster Comput., (2018). https://doi.org/10.1007/s10586-018-1725-y.

[49] X. Yan, W. Gong, Q. Wu, Contaminant source identification of water distribution networks using cultural algorithm, Concurrency Comput. Pract. Experience, (2017). doi: 10.1002/cpe.4230.

[50] X. Yan, J. Zhao, C. Hu, D. Zeng, Multimodal optimization problem in contamination source determination of water supply networks, Swarm Evol. Comput., 47 (2019) 66–71.

[51] A. Zhou, Q. Zhang, Y. Jin, Approximating the set of Pareto-optimal solutions in both the decision and objective spaces by an estimation of distribution algorithm, IEEE Trans. Evol. Comput., 13 (2009) 1167–1189.

[52] Z. Zhou, Y.S. Ong, P.B. Nair, A.J. Keane, K.Y. Lum, Combining global and local surrogate models to accelerate evolutionary optimization, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., 37 (2007) 66–76.