Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Identification of sources of pollution and contamination in water distribution networks based on pattern recognition*

Tao TAO†[1], Ying-jun LU[1], Xiang FU[2], Kun-lun XIN[1]

([1]*College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China*)

([2]*State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China*)

†E-mail: taotao@tongji.edu.cn

Received Oct. 24, 2011; Revision accepted Mar. 27, 2012; Crosschecked May 29, 2012

Abstract: An intrusion of contaminants into the water distribution network (WDN) can occur through storage tanks (via animals, dust-carrying bacteria, and infiltration) and pipes. A sensor network could yield useful observations that help identify the location of the source, the strength, the time of occurrence, and the duration of contamination. This paper proposes a methodology for identifying the contamination sources in a water distribution system, which identifies the key characteristics of contamination, such as location, starting time, and injection rates at different time intervals. Based on simplified hypotheses and associated with a high computational efficiency, the methodology is designed to be a simple and easy-to-use tool for water companies to ensure rapid identification of the contamination sources, The proposed methodology identifies the characteristics of pollution sources by matching the dynamic patterns of the simulated and measured concentrations. The application of this methodology to a literature network and a real WDN are illustrated with the aid of an example. The results showed that if contaminants are transported from the sources to the sensors at intervals, then this method can identify the most possible ones from candidate pollution sources. However, if the contamination data is minimal, a greater number of redundant contamination source nodes will be present. Consequently, more data from different sensors obtained through network monitoring are required to effectively use this method for locating multi-sources of contamination in the WDN.

Key words: Contamination, Identification, Water distribution network (WDN)
doi:10.1631/jzus.A1100286 **Document code:** A **CLC number:** TU991.33

1 Introduction

The safety of drinking water is a high priority of water-packaging industry owners and customers alike. The origin of source water as well as treatment practices has a great impact on the quality of the supplied water. Although most quality tests are carried out on the water sample when it leaves the treatment plant, its quality may deteriorate/contaminate substantially through several mechanisms and pathways during its transport to the consumers. An intrusion of contaminants into the water distribution network (WDN) can occur through storage tanks (via animals, dust-carrying bacteria, and infiltration) and pipes. Intrusion through water mains may occur during or after maintenance and repair events, through broken or corroded (pinholes or cracks) pipes and joints/gaskets, and cross-connections (Kirmeyer *et al*., 2001). Therefore, prevention of accidental and intentional contamination of WDNs is becoming an increasingly critical issue. Contaminant intrusion depends on three elements—a pathway, a driving force, and a source of contamination (Lindley, 2001). Detection of contamination in the WDN with the aid of a sensor

* Project supported by the National Natural Science Foundation of China (No. 50908165), and the Fundamental Research Funds for the Central Universities (No. 0400219207), China.

network could help identify and manage such contamination-related threats (Liu *et al.*, 2011). Although many efforts have been made to develop guidelines to determine the ideal location and maintenance of monitoring stations, the identification of pollutant source characteristics is still facing many challenges. Information about the pollutant source is needed to effectively manage contamination-related threats and to propose a control strategy.

There are several studies which consider the development of some procedures to identify the characteristics of contamination source by using the information obtained from sensor networks (Laird *et al*., 2005; Guan *et al*., 2006; Preis and Ostfeld, 2006; 2007; Di Cristo and Leopardi, 2008; Deng *et al.*, 2011; Liu and Ranjithan, 2011). Overall, all these studies attempt to identify a single solution using a fixed set of observations rather than a dynamic data patterns, by minimizing the error between the predicted concentrations and the actual observations at the sensor nodes in the network. Many studies have established methods for solving such a problem in a dynamic environment, although they still attempted to identify a single source of contamination (Bar-Shalom and Fortmann, 1988; Kurien, 1990; Blackman and Popoli, 1999; Vermaak *et al*., 2005; Tricarico *et al*., 2007).

This paper proposes a methodology to identify the multi-sources of contamination in a WDN, which involves the identification of key characteristics of contamination, such as contamination location, initial contamination time, and contaminant injection rates. The methodology is designed to be a simple and easy-to-use tool for water utilities to ensure the rapid identification of sources of contamination, and is associated with a high computational efficiency. The methodology uses the demand coverage concept performed using a pathway analysis of the network introduced by Lee and Deininger (1992). Lee and Deininger (1992) used the measured solute-concentration data and the pollution-matrix concept, and selected a group of candidate nodes as possible source nodes of pollution. The methodology proposed herein identifies the characteristics of the pollution source through a comparison of the dynamic patterns of both the simulated and measured concentrations among all the candidate nodes; further, the applications of this methodology to a literature network and a real WDN are illustrated with the aid of an example.

2 Methodology

The proposed methodology considers the case where a kind of conservative pollutant enters the WDN through intrusion points, and the pollutant input concentration is constant during the release time. Continuous measurements from the initial time of contamination, obtained from an online water-quality monitoring system, are needed to identify the solute propagation sequence and phase. This methodology assumes complete mixing at each node, and no pollutant dispersion.

The key characteristics of contamination are contamination location, initial contamination time, and contaminant injection rates which have a linear relationship with each other under the assumptions mentioned above. The methodology is designed to identify accidental contamination by solving the linear equation. First, candidate nodes are selected using a pathway analysis in which the measured concentration data in the network is used to compute the pollution matrix $\boldsymbol{P}$ (Kessler *et al*., 1998; Di Cristo and Leopardi, 2008). The methodology used to identify the most probable pollution source from among the proposed candidate nodes is described as follows.

For any candidate node injecting contaminant at a constant concentration, the continuous phase of the sequences obtained from the simulated data can be written as a matrix with dimension equal to the total simulated step. For a single intrusion point, all measured contaminant concentration values will be 0 prior to the initial time t of contamination; thus, the matrix of the contaminant concentration can be obtained:

$$\boldsymbol{\alpha}_t^i=[0\ \cdots\ a_{tt}^i\ \cdots a_{tm}^i]^{\mathrm{T}},\ t=1,...,m,i=1,...,n, \quad (1)$$

where $\boldsymbol{\alpha}_t^i$ represents the monitoring vector while the sensor detects the contaminants in step t with contaminants injected at node i; $\alpha_{tt}{}^i$ is the concen-

tration values of the contaminant at time step t; i is the index of node injected with contaminant; t is the time step of contaminant injection; n is the number of polluted nodes; and m is the total number of time steps.

When injecting the contaminant at the candidate node i at different starting times with a constant concentration, the influence matrix $\boldsymbol{A}_i$ ($m\times m$) is written as

$$\begin{aligned}\boldsymbol{A}_i &= \begin{bmatrix} a_{11}^i & & & & \\ \vdots & \ddots & & & \\ a_{t1}^i & & a_{tt}^i & & \\ \vdots & & & \ddots & \\ a_{m1}^i & \cdots & a_{mt}^i & \cdots & a_{mm}^i \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1^i & \cdots & \alpha_t^i & \cdots & \alpha_m^i \end{bmatrix}, \quad i=1,\ldots,n,\end{aligned} \tag{2}$$

where $\boldsymbol{A}_i$ represents the $m\times m$ coefficient matrix for injections at concentration c_i at the candidate node i for different initial injection time.

The sensor data can be written as a vector:

$$\boldsymbol{\beta}_1 = \begin{bmatrix} 0 & \ldots & b_t & \ldots & b_m \end{bmatrix}^{\mathrm{T}}, \tag{3}$$

where the subscript 1 represents the number of initial sources of contamination, and b_t is the recorded contaminant concentration at time step t. Note that the entries of $\boldsymbol{\beta}_1$ are 0 prior to the initial contaminant concentration occurring at time step t.

For contaminants that do not react with each other, the monitoring data $\boldsymbol{\beta}_1$ and the influence matrix $\boldsymbol{A}_i$ should satisfy Eq. (4):

$$\sum_{i=1}^{N} k_i \boldsymbol{A}_i \boldsymbol{x}_i = \boldsymbol{\beta}_1, \tag{4}$$

where $\boldsymbol{x}_i$ is a vector indicating the pollution source, N represents the total number of candidate nodes, and k_i is the linear coefficient that depends on the real contaminant concentration as well as the influence matrix. For a dosage c_i and a real contaminant concentration c_i', we have $k_i=c_i'/c_i$.

The location of the source of contamination, the input concentration, and the initial time of contamination can be obtained by solving linear equations. The coefficient matrix of these linear equations, $\boldsymbol{A}_i$, can be obtained through simulation using EPANET for different nodes, steps, and concentrations (or by using the same concentration at different candidate nodes to simplify the simulation). The vector $\boldsymbol{\beta}_1$ can be obtained from sensor data, and $\boldsymbol{x}_i$ and k_i are unknown values.

To solve $\boldsymbol{x}_i$ and k_i, the relationship between the modeled concentrations and the sensor data can be given as

$$\sum_{i=1}^{v} k_i' \boldsymbol{A}_i' \boldsymbol{x}_i' = \sum_{i=1}^{v} k_i' \begin{bmatrix} 0 \\ \vdots \\ l_{t+i-1}^i \\ \vdots \\ l_m^i \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ b_t \\ \vdots \\ b_m \end{bmatrix} = \boldsymbol{\beta}_1, \tag{5}$$

where v can be determined by the number of non-zero entries in the $\boldsymbol{x}$ vector, $v\geqslant v_{\mathrm{ture}}$ (v_{ture} represents the total number of contaminated nodes); $\boldsymbol{A}_i'$ represents the $m\times m$ coefficient matrix of the real pollution source, where $\boldsymbol{A}_i'\in\{\boldsymbol{A}_j|j=1, \ldots, n\}\cup\{\boldsymbol{A}|\boldsymbol{A}=0\}$, and l_u^i represents the influence of node i on the uth component with a standard injection rate ($u=1, 2, \ldots, t+i-1, \ldots, m$).

Thus, Eq. (5) can be solved as follows:

$$\begin{aligned}\sum_{i=s+1}^{v} k_i' \begin{bmatrix} 0 \\ \vdots \\ l_{t+i-1}^i \\ \vdots \\ l_m^i \end{bmatrix} &= \boldsymbol{\beta}_1 - \sum_{i=1}^{s} k_i' \boldsymbol{\eta}_i' = \boldsymbol{\beta}_1 - \sum_{i=1}^{s} k_i' \begin{bmatrix} 0 \\ \vdots \\ l_{t+i-1}^i \\ \vdots \\ l_m^i \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ \vdots \\ b_t - k_1' l_t^1 \\ \vdots \\ b_{t+s} - \sum_{i=1}^{s} k_i' l_{t+s}^i \\ \vdots \\ b_m - \sum_{i=1}^{s} k_i' l_m^i \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ b_{t+s} - \sum_{i=1}^{s} k_i' l_{t+s}^i \\ \vdots \\ b_m - \sum_{i=1}^{s} k_i' l_m^i \end{bmatrix} = \boldsymbol{\beta}_{1+s}.\end{aligned} \tag{6}$$

There is at most one element equal to 1 in the vector $\boldsymbol{x}_i$, and $\boldsymbol{A}_i'\neq 0$, therefore $\boldsymbol{A}_i'$ satisfies

$$\boldsymbol{A}_i' \in \{\boldsymbol{A}_j \mid \boldsymbol{A}_j[0\cdots 1^w\cdots 0]^{\mathrm{T}} = [0\cdots a_{t+i-1}\cdots a_m]^{\mathrm{T}}, \; j=1,...,m\}, \tag{7}$$

where w represents the location of the element equal to 1 in the vector, and $a_u>0$, $u=t+i-1, t+i, \ldots, m$.

Using Eq. (6), the coefficient k_i is solved explicitly as follows:

$$\begin{cases} k_1' = b_t / l_t^1, \\ k_i' = \left(b_{t+i-1} - \sum_{j=1}^{i-1} k_j' l_{t+i-1}^j \right) \Big/ l_{t+i-1}^i, \; i \geq 2. \end{cases} \tag{8}$$

The algorithm used to obtain the solution of the model is described as follows:

1. Set i=1 as the initial index of the pollution source node.

2. Set j=1 as the initial index of the candidate nodes.

3. Starting j, let the matrix $\boldsymbol{A}_i'$ be described as shown in Eq. (7), set the corresponding w in $\boldsymbol{x}_i'$ equal to 1, and calculate $\boldsymbol{\eta}_i'$, k_i', $\boldsymbol{\beta}_{i+1}$ using Eqs. (6)–(8).

4. Consider whether $\boldsymbol{\beta}_{i+1}$ has any negative or zero components. If the answer is "Yes", then it means that the nodes selected in Step 2 are not the real pollution source node.

5. Calculate f using Eq. (9). The purpose of calculating f is to determine the congruence between the candidate nodes and the real source nodes of contamination, in addition to estimating the condition for terminating the computational process. The role of f is consistent with the fitness function proposed by Di Cristo and Leopardi (2008). Theoretically, if the matrix $\boldsymbol{\beta}_{i+1}$ for the candidate nodes correctly identifies all source nodes, then the result of Eq. (9) will be zero. However, to account for computational rounding error and measurement error, we require only that f be less than a limiting value ε to claim correct source node identification. We have

$$f = |\boldsymbol{\beta}_{i+1}|^2 = \sum_{k=i}^{m-t} \left(b_{t+k} - \sum_{j=1}^{i} k_j' l_{t+i}^j \right)^2 \leq \varepsilon, \; i \geq 1. \tag{9}$$

Errors are mainly caused by the measurement data b_{t+i-1} received from the sensor station; therefore, a simple sum method to calculate the limiting value is proposed based on the error Δb_{t+k}, that is,

$$\varepsilon = \left| 2 \times \sum_{k=i}^{m-t} \left(b_{t+k} - \sum_{j=1}^{i} k_j' l_{t+i}^j \right) \times \Delta b_{t+k} \right|.$$

When $|\boldsymbol{\beta}_{i+1}|^2$ does not satisfy Eq. (9), it means that there are other pollution sources which need to be identified. In this case we let $i=i+1$. If the next $i<v$ (v can be calculated by Eq. (10)), then let $k_i'=0$, $\boldsymbol{A}_i'=0$, and $\boldsymbol{x}_i'=0$. Otherwise, go back to Step 2. The purpose is to account for the case where the initial contamination times corresponding to different sources of contamination are varying in the first sensor.

$$v = -d + m - t + 2, \quad d = 1, ..., m - t + 1, \tag{10}$$

where d represents the number of non-zero positive components of the current $\boldsymbol{\beta}_{i+1}$, and v represents the index of the next possible contaminated node for which the coefficient k_i is determined using Eq. (8).

When $|\boldsymbol{\beta}_{i+1}|^2$ satisfies Eq. (9), it means that one of the possible pollution sources has been located. Using Eq. (11), we can identify the location of the one source of contamination, the input concentration, and the initial time of contamination. Then, we let $j=j+1$ and go back to Step 3 to determine if there are any other source nodes which also satisfy Eq. (9).

A flow chart describing the process used to identify the contaminated nodes in a WDN is presented in Fig. 1. At this point, a series of k_i', $\boldsymbol{A}_i'$, $\boldsymbol{x}_i'$ has been calculated. While $k_i'\neq 0$, $\boldsymbol{A}_i'\neq 0$, and $\boldsymbol{x}_i'\neq 0$, the variables k_i', $\boldsymbol{A}_i'$, and $\boldsymbol{x}_i'$ represent only one of the possible pollution sources. The contamination characteristics of this node, including the location of the source, the initial time of contamination, and the injected concentration, can be calculated using Eq. (11).

We have

$$\begin{cases} \mathrm{PN}_i = \sigma_{\boldsymbol{A}_i'=\boldsymbol{A}_j}(j), \\ \mathrm{PT}_i = T_0 + (w(\boldsymbol{x}_i) - 1)\Delta t, \\ \mathrm{PC}_i = k_i' c_i = k_i' c_0, \end{cases}$$

$$i=1,...,v,\quad j\in\{1,...,n\}, \tag{11}$$

where PN_i represents the index of pollution source node i; $\sigma_{A_i'=A_j}(j)$ represents the index of the node from the candidate node j which satisfies $\boldsymbol{A}_i'=\boldsymbol{A}_j$; PT_i represents the injection time of the pollutant from source node i; T_0 represents the starting time of the simulation period; $w(\boldsymbol{x}_i)$ represents the position of the element equal to 1 in $\boldsymbol{x}_i$; Δt represents the distance between time steps; PC_i represents the injected concentration at node i; and c_i represents the concentration for building the contamination matrix. To simplify the calculation, the constant c_0 is used to build every contamination matrix.

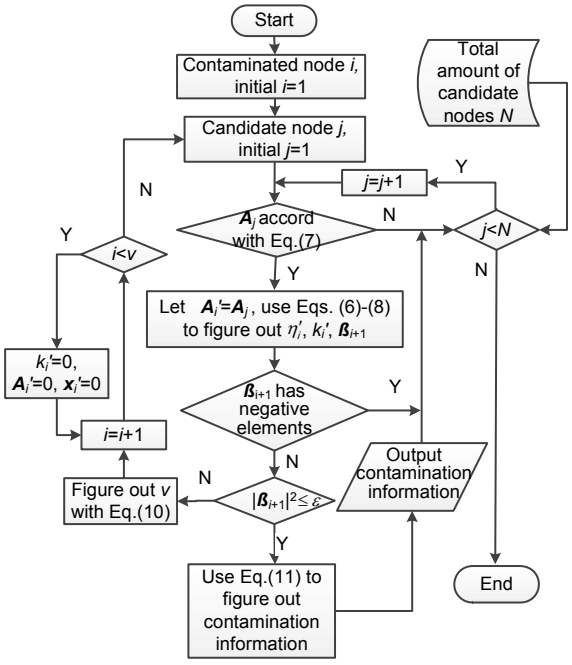


Fig. 1 Flow chart of identifying contaminated nodes

3 Case study

The methodology described two cases to highlight its effectiveness.

3.1 Case 1

The first case considers synthetic data from the literature network of Anytown USA. The Anytown network scheme is shown in Fig. 2. The network data as well as the 24 h demand pattern assigned to all nodes are reported in (Kessler *et al*., 1998). The monitoring stations are selected using the Maximum Coverage Criterion proposed by Lee and Deininger (1992). According to the results of Kessler *et al*. (1998), the best locations for the three sensors on the Anytown network are Nodes 90, 140, and 160.

The same contaminant injection was performed for comparing results with those presented by Di Cristo and Leopardi (2008). That is, contaminant was injected at Node 70, realized using the mass point booster option set at a steady input rate of 25 g/min, and the injected time is 0.00. The sensor data obtained is reported in Table 1. The influence matrix of Sensor 90 is shown in Table 2, which was built as the contaminant was injected at Node 70 with a steady input rate of 100 g/min.

The proposed methodology requires data from only a single sensor to identify the source. The computed value of $\boldsymbol{\beta}_2$ and the fitness functions based on Sensors 90 and 160 are shown in Table 3. It can be concluded using Eq. (11) with data from only Sensor 90 that the pollution source was located at Node 70, the injection time was 0:00−1:00, and the input rate was 25.03 g/min.

The final results obtained from data from three sensors are shown in Table 4. In this case, it can be concluded that the source is located at Node 70,

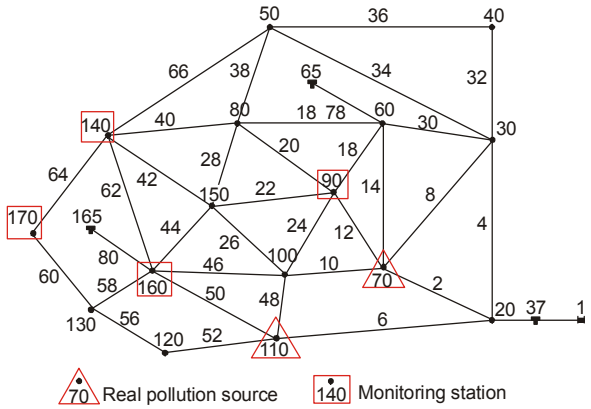


Fig. 2 Real pollution sources in Anytown network scheme and monitoring stations (Di Cristo and Leopardi, 2008)

the injection time is 0:00–1:00, and the average input rate is 24.81 g/min ((25.03+24.39+25.00)/3).

Table 1 Sensor data in Anytown (Di Cristo and Leopardi, 2008)

Sensor station	Concentration from sensor station (mg/L)							
	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00
90	0	0.88	1.77	2.24	2.34	2.89	3	2.25
140	0	0	0	0	0	0	0.13	0.23
160	0	0	0	0.51	0.51	0.51	0.51	0

Table 2 Elements of Influence Matrix of Node 90 (injecting from Node 70 at different time)

Time	Concentration (mg/L)					
	7:00–8:00 …		0:00–1:00 …		6:00–7:00	
8:00 (pre.*)	4.278					Row 1
…	…	…				…
1:00	8.157	⋯	3.515			Row 18
…	…	…	…	…		…
7:00	8.228	⋯	9.012	⋯	4.277	Row 24
	Column 1 ⋯		Column 18 ⋯		Column 24	

* pre. means the previous day

Table 3 Values of β_2 and fitness functions (Case 1)

Sensor station	Node	β_2							k	x'	f	ε	$f<\varepsilon$	$\beta_2<0$
		1:00	2:00	3:00	4:00	5:00	6:00	7:00						
90	1	0.000	0.888	0.828	0.773	1.013	0.915	0.223	0.3420	$(0\ldots1^{17}\ldots0)$	3.986	4.641×10^{-1}	False	False
	20	0.000	0.764	0.756	0.773	1.013	0.897	0.223	0.3420	$(0\ldots1^{17}\ldots0)$	3.634	4.427×10^{-1}	False	False
	30	0.000	0.890	1.360	1.209	1.759	1.869	1.091	0.4256	$(0\ldots1^{16}\ldots0)$	1.188×10	8.176×10^{-1}	False	False
	60	0.000	0.639	1.109	1.209	1.759	1.680	1.091	0.1373	$(0\ldots1^{17}\ldots0)$	1.020×10	7.485×10^{-1}	False	False
	70	0.000	−0.004	−0.003	0.001	−0.008	−0.008	−0.006	0.2503	$(0\ldots1^{18}\ldots0)$	1.926×10^{-4}	2.832×10^{-3}	True	False
	90	0.000	0.890	1.360	1.313	1.863	1.973	1.619	0.0253	$(0\ldots1^{18}\ldots0)$	1.435×10	9.019×10^{-1}	False	False
	110	0.000	0.886	1.360	1.313	1.863	1.973	1.619	1.9028	$(0\ldots1^{14}\ldots0)$	1.435×10	9.015×10^{-1}	False	False
	Others	–	–	–	–	–	–	–	–	–	–	–	–	–
160	1	0.000	0.000	0.000	−0.004	−0.158	−0.276	−0.232	0.1657	$(0\ldots1^{18}\ldots0)$	1.549×10^{-1}	6.701×10^{-2}	False	True
	20	0.000	0.000	0.000	−0.054	−0.158	−0.276	−0.232	0.1657	$(0\ldots1^{18}\ldots0)$	1.579×10^{-1}	7.202×10^{-2}	False	True
	40	0.000	0.000	0.000	−39.305	−113.815	−113.064	0.000	81.1231	$(0\ldots1^{13}\ldots0)$	2.728×10^{4}	2.662×10	False	True
	50	0.000	0.000	0.000	−0.853	−0.835	−1.059	0.000	0.5118	$(0\ldots1^{14}\ldots0)$	2.547	2.747×10^{-1}	False	True
	70	0.000	0.000	1.743×10^{-4}	1.743×10^{-4}	1.743×10^{-4}	1.743×10^{-4}	0.000	0.2500	$(0\ldots1^{18}\ldots0)$	1.215×10^{-7}	6.970×10^{-5}	True	False
	80	0.000	0.000	0.000	0.000	0.000	−0.276	0.000	0.2517	$(0\ldots1^{15}\ldots0)$	7.628×10^{-2}	2.762×10^{-2}	False	True
	90	0.000	0.000	0.000	−0.283	−0.318	−0.318	0.000	1.0540	$(0\ldots1^{13}\ldots0)$	2.826×10^{-1}	9.194×10^{-2}	False	True
	100	0.000	0.000	0.000	−0.137	−0.137	−0.245	0.000	0.0997	$(0\ldots1^{19}\ldots0)$	9.756×10^{-2}	5.190×10^{-2}	False	True
	110	0.000	0.000	0.000	0.000	−0.085	−0.119	−0.232	0.0390	$(0\ldots1^{19}\ldots0)$	7.494×10^{-2}	4.352×10^{-2}	False	True
	150	0.000	0.000	0.000	−0.004	−0.158	−0.276	−0.232	0.1657	$(0\ldots1^{18}\ldots0)$	1.549×10^{-1}	6.701×10^{-2}	False	True
	160	0.000	0.000	0.000	−0.085	−0.262	−0.291	−0.212	0.0133	$(0\ldots1^{20}\ldots0)$	2.058×10^{-1}	8.506×10^{-2}	False	True
	Others	–	–	–	–	–	–	–	–	–	–	–	–	–

Note: $(0\ \ldots\ 1^{18}\ \ldots\ 0)$ means the 18th element of solution vector is 1, the rest is 0

Table 4 Simulation results of locations of contamination source (Case 1)

Sensor station	Pollution source node		Injecting rate		Injecting time		f	ε
	A_1'	Node	k_1'	Rate (g/min)	x_1'	Time		
90	A_{70}	70	0.2503	25.03	$(0\ldots1^{18}\ldots0)$	0:00–1:00	1.93×10^{-4}	2.83×10^{-3}
140	A_{70}	70	0.2439	24.39	$(0\ldots1^{18}\ldots0)$	0:00–1:00	4.06×10^{-5}	6.37×10^{-4}
	A_{60}	6	0.0454	4.54	$(0\ldots1^{19}\ldots0)$	1:00–2:00	6.04×10^{-3}	7.77×10^{-3}
160	A_{70}	70	0.2500	25.00	$(0\ldots1^{18}\ldots0)$	0:00–1:00	1.22×10^{-7}	6.97×10^{-5}

In cases with multiple pollution sources, the method proposed by Di Cristo and Leopardi (2010) is used to simplify the problem and reduce the number of potential source locations. This method is also applicable in the case where we need to determine the locations of multiple sources of contamination, and details are shown in Case 2.

3.2 Case 2

In Case 2, the methodology is applied to a real-life WDN. The real-life WDN is composed of three water sources and provides more than 4×10^5 m^3 of water per day. A WDN model for this real-life network has been developed, which has 77 nodes and 108 pipelines with diameters larger than 500 mm. The pipe roughness values range from 100 to 120 according to the pipe material and service life. With the Maximum Coverage Criterion proposed by Lee and Deininger (1992), six water-quality monitors were selected (Wang, 2010), as depicted in Fig. 3.

The synthetic concentration data are generated by simulating hydraulics and solute transport, with the contaminants injected at different nodes using the mass point booster option set at different input rates. To generate the simulated concentrations, and in the application of our methodology, EPANET is run with a 15-min hydraulic time step and a 5-min water quality time step. According to method used to construct our influence matrix, a time series with 96 intervals (Δt=15 min) was created by counting back from the current moment (12:00) to the previous day (12:00).

Several scenarios have been considered to investigate the performance and applicability of the proposed method. In all scenarios, initial time of contamination in the monitoring point 17 is in the range of 10:45 to 12:00 (Table 5). The set of "candidate" nodes is selected by considering the measured concentration data collected in the network and computing the pollution matrix $\boldsymbol{P}$ (Kessler *et al.*, 1998); thus, the candidate set is

$$CN\in\{4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 83\ \ 17\}.$$

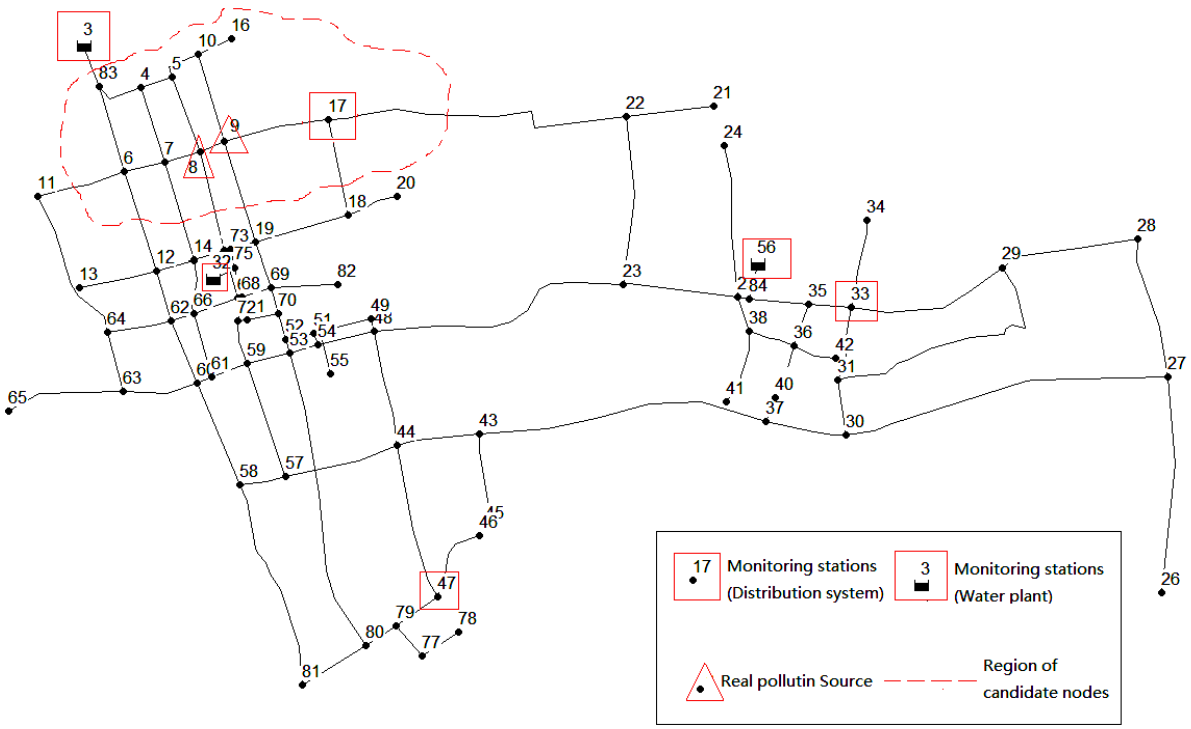


Fig. 3 Case study area (Wang, 2010)

Some crucial features of these scenarios are summarized:

Scenario 1: Contaminant was injected at Nodes 8 and 9 at 9:30, and Sensor 17 detected the contamination in the time interval 10:45 to 12:00, as shown in Table 5.

Scenario 2: The contaminant was injected at Nodes 7, 8, and 9 at 9:30, and the concentrations injected were 30, 50, and 90 g/min, respectively. Sensor 17 detected the contamination in the time interval 10:45 to 12:00.

Scenario 3: The contaminant was injected at Node 8 at 10:15, and Sensor 17 detected only two contaminant data in the time interval 11: 30 to 12:00, as shown in Table 5.

Scenario 4: The contaminant was injected at Nodes 5, 7, and 10 at 9:15; Sensor 17 detected only three contaminant data in the time interval 11:15 to 12:00. There was no initial contamination time interval when the monitoring station first detected the pollutants from Nodes 5 and 10. Nine influence matrixes (96×96) of the candidate nodes were created using EPANET for different contamination time.

In Scenario 1, our approach identified six nonzero elements in vector $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_1$=[0… 1.65 2.74 3.45 3.52 3.48 3.52]$^{\mathrm{T}}$, and nine matrixes satisfied the constraints of Eq. (8). For the first source node of contamination, the time when the pollutants were detected was 10:45. The results of the first source node of contamination are shown in Table 6.

Nodes 4, 5, 6, 8, 10, and 83 are excluded as the first-contaminated nodes. Meanwhile, the values of the remaining nodes do not satisfy $f<\varepsilon$, which indicates that there are multiple sources of contamination. The possible values of ($\boldsymbol{A}_1'$ k_1' $\boldsymbol{x}_1'$) in Scenario 1 are shown in Table 7.

All possible values of ($\boldsymbol{A}_i'$ k_i' $\boldsymbol{x}_i'$) can be obtained from Table 7. A phylogenetic tree that describes the source of contamination is used to describe the evolution process of the possible contaminated nodes, as shown in Fig. 4. When all $\boldsymbol{\beta}_{i+1}<0$ ($\forall\boldsymbol{\beta}_{i+1}<0$), a branch stops growing, which means that there is no contaminated node located in this branch.

In all branches, only the growth from A_9 to A_8 stops because $f\leq\varepsilon$, which means that the contamination must be caused by Nodes 9 and 8. With the values of ($\boldsymbol{A}_i'$ k_i' $\boldsymbol{x}_i'$) calculated in each step, the source of contamination can be located using Eq. (11), and the result is shown in Table 8 (p.568).

In Scenario 2, the applicability of this method is shown, assuming that there are several time intervals when the sensor station detects the contaminants from different pollution sources. There are three time intervals (45 min) when Sensor 17 detects the contaminants from Nodes 7 and 8. The result shows that this method is also applicable to the multiple detection intervals of contamination by introducing ($\boldsymbol{A}_3'$ k_3' $\boldsymbol{x}_3'$) and ($\boldsymbol{A}_4'$ k_4' $\boldsymbol{x}_4'$) for auxiliary operations.

Theoretically, this method can be used to efficiently identify the multi-source contamination when there are data from more than two sensor stations. When the information obtained from monitoring sensors is not sufficient, the computational process

Table 5 Real pollution in water distribution system

Scenario number	Pollution source	Input rate (g/min)	Injecting time	Concentration of Sensor station 17 at different time (mg/L)						
				10:30	10:45	11:00	11:15	11:30	11:45	12:00
1	9	100	9:30	–	1.65	2.74	3.45	3.52	3.48	3.52
	8	50	9:30							
2	9	30	9:30	–	0.49	1.06	2.21	2.28	3.21	3.64
	8	90	9:30							
	7	150	9:30							
3	8	30	10:15	–	–	–	–	–	0.08	0.49
4	10	150	9:15	–	–	–	–	0.65	0.87	0.93
	7	100	9:15							
	5	50	9:15							

Table 6 Results of the first contamination source node (Scenario 1)

Candidate node	Possible column*	η_1' (column values of the matrix)								k	β_2**							f	ε	$f<\varepsilon$
4	80	0.00	…	0.04	0.63	0.79	0.75	1.13	1.21	41.25	0.00	…	0.00	−23.25	−29.14	−27.42	−43.13	−46.39	6.15×10^3	16.9	N
5	81	0.00	…	0.18	0.75	0.71	1.11	1.62	1.61	9.167	0.00	…	0.00	−4.14	−3.06	−6.66	−11.37	−11.24	3.3×10^2	3.65	N
6	74	0.00	…	0.01	0.07	0.07	0.07	0.07	0.07	165.0	0.00	…	0.00	−8.81	−8.10	−8.03	−8.07	−8.03	3.37×10^2	4.10	N
7	83	0.00	…	0.78	0.90	0.91	0.88	0.87	0.91	2.115	0.00	…	0.00	0.84	1.53	1.66	1.64	1.60	1.11×10	0.73	N
8	86	0.00	…	0.62	1.54	1.57	1.64	1.69	1.63	2.661	0.00	…	0.00	−1.36	−0.73	−0.84	−1.02	−0.82	4.80	0.48	N
9	87	0.00	…	1.65	2.58	2.66	2.70	2.64	2.70	1.000	0.00	…	0.00	0.16	0.79	0.82	0.84	0.82	2.70	0.34	N
10	81	0.00	…	0.83	1.96	1.87	1.91	2.02	2.01	1.988	0.00	…	0.00	−1.16	−0.27	−0.28	−0.54	−0.48	2.01	0.27	N
17	91	0.00	…	3.96	4.18	4.09	3.89	3.93	4.15	0.417	0.00	…	0.00	1.00	1.75	1.90	1.84	1.79	1.43×10	0.83	N
83	78	0.00	…	0.01	0.22	0.33	0.33	0.45	0.49	165.0	0.00	…	0.00	−33.56	−51.00	−50.93	−70.77	−77.33	1.73×10^4	28.4	N

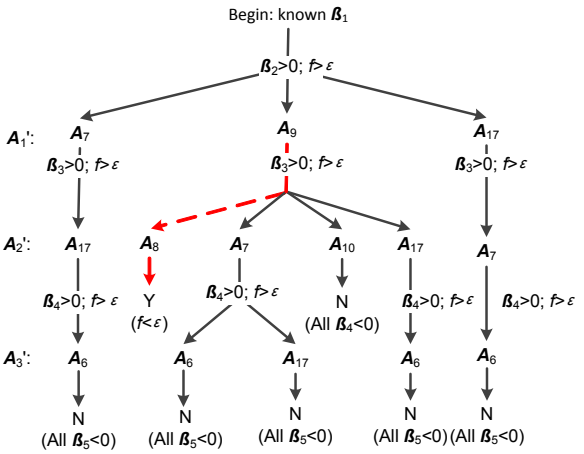
* Column in matrix which has the same positive elements with β_1; ** β_1=[0.00...0.00 1.65 2.74 3.45 3.52 3.48 3.52], $\beta_2=\beta_1-k\eta_1'$ (more details refer to Eq.(6)).

After the first-round calculation, the first contaminated source node is probably among the following nodes:

1. Node 7, injecting time: 8:30, injecting rate: 211.5 g/min, influence value to monitoring point (0.00,…,1.65,1.60,1.92,1.86,1.84,1.92), the next contaminated source should be located with these conditions.

2. Node 9, injecting time: 9:30, injecting rate: 100 g/min, influence value to monitoring point (0.00,…,1.65,2.58,2.66,2.70,2.64,2.70), the next contaminated source should be located with these conditions.

3. Node 17, injecting time: 10:30, injecting rate: 41.7 g/min, influence value to monitoring point (0.00,…,1.65,1.74,1.70,1.62,1.64,1.73), the next contaminated source should be located with these conditions



Note:

1. N represents that deduction stops since all β_{i+1} contains negative element (all β_{i+1}<0), and nodes of this branch are invalid contamination combination.

2. Y represents that deduction stops since $f<\varepsilon$, nodes of this branch are valid contamination combination

Fig. 4 Phylogenetic tree of pollution sources

cannot be terminated, because $\beta_{i+1}>0$ during the entire process, which might result in some redundant contamination combinations other than the real contaminated node. To reduce redundancy, we need to assume that for each source node, at least two

Table 7 Possible values of (A_1' k_1' x_1') (Scenario 1)

Node	A_1'	k_1'	x_1'
7	A_7	2.115	$(0\ \ldots\ 1^{83}\ \ldots\ 0)$
9	A_9	1.000	$(0\ \ldots\ 1^{87}\ \ldots\ 0)$
17	A_{17}	0.417	$(0\ \ldots\ 1^{91}\ \ldots\ 0)$

Note: $(0\ \ldots\ 1^{83}\ \ldots 0)$ means the 83rd element of solution vector is 1, the rest is 0

contamination data are detected by the sensor station. Under this assumption, the results of Scenario 3 (a single source of contamination) prove that less redundant sources of contamination will be identified based on two or more sensor contamination data.

In Scenario 4 it analyzes the situation where no interval exists when the sensor stations detect the contaminants. The sensor station obtains the data from Node 7 at 11:45, but at 12:00 the data are from Nodes 7, 5 and 10. This means that there is no interval in which contaminants are transported from Nodes 10 and 5 to Node 7. The results shows that not all the pollution sources are identified according to Eq. (6) and the real pollution source cannot be located.

Table 8 Results of locations of pollution sources in water distribution system

Scenario	Pollution source		Injecting rate		Injecting time		ε	f
	Value of matrix	Node	Coefficient	Rate (g/min)	$\boldsymbol{x}_i'$	Time		
1	$\boldsymbol{A}_1'=A_9$	9	$k_1'=1.000$	100	$\boldsymbol{x}_1'=(0\ldots1^{87}\ldots0)$	9:30–9:45	0.343	2.7
	$\boldsymbol{A}_2'=A_8$	8	$k_2'=0.500$	50	$\boldsymbol{x}_2'=(0\ldots1^{87}\ldots0)$	9:30–9:45	3.00×10^{-3}	2.00×10^{-4}
2	$\boldsymbol{A}_1'=A_9$	9	$k_1'=0.297$	29.7	$\boldsymbol{x}_1'=(0\ldots1^{87}\ldots0)$	9:30–9:45	8.46×10^{-1}	1.82×10
	$\boldsymbol{A}_2'=A_8$	8	$k_2'=0.906$	90.6	$\boldsymbol{x}_2'=(0\ldots1^{87}\ldots0)$	9:30–9:45	2.25×10^{-1}	2.66
	$\boldsymbol{A}_3'=0$	–	$k_3'=0.000$	–	$\boldsymbol{x}_3'=(0\ldots0\ldots0)$	–	–	–
	$\boldsymbol{A}_4'=0$	–	$k_4'=0.000$	–	$\boldsymbol{x}_4'=(0\ldots0\ldots0)$	–	–	–
	$\boldsymbol{A}_5'=A_7$	7	$k_5'=1.500$	150	$\boldsymbol{x}_5'=(0\ldots1^{87}\ldots0)$	9:30–9:45	1.00×10^{-2}	2.50×10^{-5}
3	$\boldsymbol{A}_1'=A_8$	8	$k_1'=0.308$	30.8	$\boldsymbol{x}_2'=(0\ldots1^{90}\ldots0)$	10:15–10:30	1.15×10^{-3}	1.30×10^{-4}
4	$\boldsymbol{A}_1'=A_7$	7	$k_1'=1.000$	100	$\boldsymbol{x}_1'=(0\ldots1^{86}\ldots0)$	9:15–9:30	2.00×10^{-3}	4.00×10^{-4}
	$\boldsymbol{A}_1'=A_{17}$	17	$k_1'=0.167$	16.7	$\boldsymbol{x}_1'=(0\ldots1^{94}\ldots0)$	11:15–11:30	4.50×10^{-2}	1.01×10^{-1}
	$\boldsymbol{A}_2'=A_6$	6	$k_2'=3.047$	304.7	$\boldsymbol{x}_2'=(0\ldots1^{77}\ldots0)$	7:00–7:15	2.99×10^{-3}	7.16×10^{-4}

Uncertainty of input data such as measurement error and node demand may be of great value for practical applications. As well as the input data, the presented methodology requires water quality measurements and the knowledge of the flow pattern at each node. However, measured concentrations are affected by measurement errors and water demands. Due to the random nature of these uncertainties, only statistical properties of uncertainty can be indirectly estimated (Di Cristo and Leopardi, 2008).

The measurement error or demand uncertainty ultimately will be reflected in the measurement of the true value b_{t+k}, which means that the rounding error is a function of measurement error and demand uncertainty, that is $\Delta b_{t+k}=f(\Delta P, \Delta m)$. Due to this functional relationship, the proposed method not only considers the minimum of the fitness function, but allows error ε determined by Δb_{t+k} to ensure all possible pollution sources are located.

The relationship among the pollution sources number, the percent of pollution nodes to candidate nodes and Δb for Sensor station 90 in Case 1 is shown in Fig. 5. Based on these results, it can be concluded that the average contamination concentration is 2.2 mg/L at sensor station 90. When the average error increased from 0 mg/L (0% of the average concentration at Sensor station 90) to 1.1 mg/L (50% of the average concentration at Sensor station 90), the sources located by this method increased from 1 to 7. Therefore, this method requires that Δb be determined as accurately as possible based on different pipe network characteristics. For example, when there is a greater difference between the actual demand patterns and model demand patterns, or when there is low measuring accuracy, Δb should be large to ensure that the group of estimated pollution sources contains the real sources; otherwise, Δb should be chosen to be small to minimize the number of estimated pollution sources.

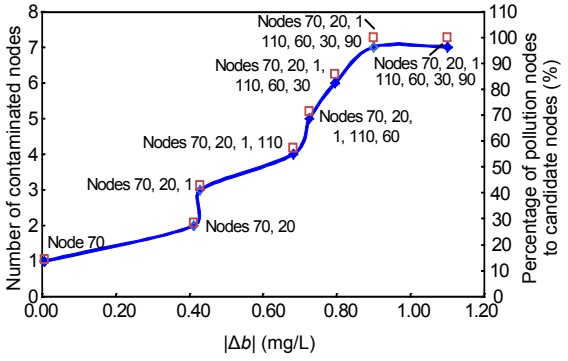


Fig. 5 The number of contaminated nodes on different Δb

4 Conclusions

In this paper, we proposed a simple method for the identification of the source of contamination in a

WDN. The location of the source of contamination, the input concentration, and the initial time of contamination can be obtained by solving linear equations. Both the theoretical analysis and the application of our approach to case studies indicate that this method can efficiently identify multiple sources of contamination on the condition that contaminant from each source of contamination is detected at a sensor for some time interval. Four scenarios are considered to determine the efficiency of this method. Results show that if the pollutants emitted from the sources of contamination are transported to the sensors at intervals, then this method can identify the most possible pollution sources among the candidate source nodes of pollution. However, in cases where there is limited contamination information and no time intervals, there will be a greater number of redundant source nodes of contamination. This approach works best when there is sufficient data from different sensors. In particular, we recommend at least two contamination concentrations for each node.

Both measurement error and demand error are taken into consideration in the case study. To reduce redundancy without sacrificing accuracy and efficiency of this approach to identify multiple contamination nodes, the goal of the next stage research is to find an efficient way to determine the limiting value.

References

Bar-Shalom, Y., Fortmann, T.E., 1988. Tracking and Data Association. Academic Press, San Diego.

Blackman, S.S., Popoli, R., 1999. Design and Analysis of Modern Tracking Systems. Artech House, Norwood, MA.

Deng, Y., Jiang, W., Sadiq, R., 2011. Modeling contaminant intrusion in water distribution networks: A new similarity-based DST method. *Expert Systems with Applications*, **38**(1):571-578. [doi:10.1016/j.eswa.2010.07.004]

Di Cristo, C., Leopardi, A., 2008. Pollution source identification of accidental contamination in water distribution networks. *Journal of Water Resources Planning and Management*, **134**(2):197-202. [doi:10.1061/(ASCE)0733-9496(2008)134:2(197)]

Di Cristo, C., Leopardi, A., 2010. Closure to "DI CRISTO C, LEOPARDI A. (2008). Pollution source identification of accidental contaminations in water distribution networks. *Journal of Water Resources Planning and Management*, 134(2)". *Journal of Water Resources Planning and Management*, **136**(5):292-293. [doi:10.1061/(ASCE)WR.1943-5452.0000048]

Guan, J., Aral, M.M., Maslia, M.L., Grayman, W.M., 2006. Identification of contaminant sources in water distribution systems using simulation—Optimization method: Case study. *Journal of Water Resources Planning and Management*, **132**(4):252-262. [doi:10.1061/(ASCE)0733-9496(2006)132:4(252)]

Kessler, A., Osfeld, A., Sinai, G., 1998. Detecting accidental contaminations in municipal water networks. *Journal of Water Resources Planning and Management*, **124**(4):192-198. [doi:10.1061/(ASCE)0733-9496(1998)124:4(192)]

Kirmeyer, G.J., Friedman, M., Martel, K., Howie, D., 2001. Pathogen Intrusion into Distribution System. Denver, CO: AwwaRF.

Kurien, T., 1990. Issues in the Design of Practical Multi-Target Tracking Algorithms. Multitarget-Multisensor Tracking: Advanced Applications, Bar-Shalom, Y. (Ed.), Artech House, Norwood, MA.

Laird, C.D., Biegler, L.T., van Bloemen Waanders, B.G., Bartlett, R.A., 2005. Contamination source determination for water networks. *Journal of Water Resources Planning and Management*, **131**(2):125-134. [doi:10.1061/(ASCE)0733-9496(2005)131:2(125)]

Lee, B.H., Deininger, R.A., 1992. Optimal location of monitoring stations in water distribution system. *Journal of Environmental Engineering*, **118**(1):4-16. [doi:10.1061/(ASCE)0733-9372(1992)118:1(4)]

Lindley, T.R., 2001. A Framework to Protect Water Distribution Systems against Potential Intrusions. MS Thesis, University of Cincinnati, Cincinnati.

Liu, L., Ranjithan, R.S., Mahinthakumar, G., 2011. Contamination source identification in water distribution systems using an adaptive dynamic optimization procedure. *Journal of Water Resources Planning and Management*, **137**(2):183-192. [doi:10.1061/(ASCE)WR.19435452.0000104]

Preis, A., Ostfeld, A., 2006. Contamination source identification in water systems: A hybrid model trees-linear programming scheme. *Journal of Water Resources Planning and Management*, **132**(4):263-273. [doi:10.1061/(ASCE)0733-9496(2006)132:4(263)]

Preis, A., Ostfeld, A., 2007. A contamination source identification model for water distribution system security. *Engineering Optimization*, **39**(8):941-951. [doi:10.1080/03052150701540670]

Tricarico, C., de Marinis, G., Gargano, R., Leopardi, A., 2007. Peak residential water demand. *Proceedings of the ICE-Water Management*, **160**(2):115-121. [doi:10.1680/wama.2007.160.2.115]

Vermaak, J., Godsill, S.J., Perez, P., 2005. Monte Carlo filtering for multi-target tracking and data association. *IEEE Transactions on Aerospace and Electronic Systems*, **41**(1):309-332. [doi:10.1109/TAES.2005.1413764]

Wang, F.X., 2010. The Software and Pre-Warning of Water Quality in Water Distribution Network. MS Thesis, College of Environmental Science and Engineering, Tongji University, Shanghai (in Chinese).