Logistic regression analysis to estimate contaminant sources in water distribution systems

Li Liu, A. Sankarasubramanian and S. Ranji Ranjithan

ABSTRACT

Accidental or intentional contamination in a water distribution system (WDS) has recently attracted attention due to the potential hazard to public health and the complexity of the contaminant characteristics. The accurate and rapid characterization of contaminant sources is necessary to successfully mitigate the threat in the event of contamination. The uncertainty surrounding the contaminants, sensor measurements and water consumption underscores the importance of a probabilistic description of possible contaminant sources. This paper proposes a rapid estimation methodology based on logistic regression (LR) analysis to estimate the likelihood of any given node as a potential source of contamination. Not only does this algorithm yield location-specific probability information, but it can also serve as a prescreening step for simulation–optimization methods by reducing the decision space and thus alleviating the computational burden. The applications of this approach to two example water networks show that it can efficiently rule out numerous nodes that do not yield contaminant concentrations to match the observations. This elimination process narrows down the search space of the potential contamination locations. The results also indicate that the proposed method efficiently yields a good estimation even when some noise is incorporated into the measurements and demand values at the consumption nodes.

Key words | contaminant source, logistic regression analysis, probabilistic characterization, water distribution systems

Li Liu (corresponding author)
School of Civil Engineering,
Hefei University of Technology,
230009 Hefei,
Anhui,
China
E-mail: *lliuncsu@gmail.com*

A. Sankarasubramanian
S. Ranji Ranjithan
Department of Civil,
Construction, and Environmental Engineering,
North Carolina State University,
Raleigh,
NC 27695,
USA

INTRODUCTION

The vulnerability of drinking water due to contamination within a WDS has received much attention in recent years. Contamination, either accidental or intentional, is a major issue associated with the security of drinking water quality in the system. To discover contaminants, a WDS must have a set of sensors installed that can detect a contamination event. However, the installation and operational costs limit the large-scale use of monitoring sensors in a WDS. Many researchers have focused on where to site sensors within a network for best detection. During the Water Distribution Systems Analysis (WDSA) conference (2006) in Cincinnati, there was a special theme entitled "Battle of the water sensor network" with the aim of objectively comparing the performance of contributed sensor network designs (Ostfeld *et al.* 2008). While real-time measurements are collected from the monitoring stations at the selected locations, the observed data must be processed real-time or near-real-time to rapidly identify the pollutant source. Solutions to this problem are needed to generate an effective threat management strategy that can mitigate the threat by taking appropriate actions, such as warning the impacted residents to take action against being affected by the contamination, isolating the malicious contaminant sources and flushing out the contaminant.

doi: 10.2166/hydro.2010.106

Contaminant source characterization is complicated not only by the limited observational data, but also by the arbitrary nature of the contaminants that potentially can be injected from any point accessible to the public and with varying levels of strength. Based on sensor observations, this characterization problem can be categorized as an inverse problem. The complexity caused by real inverse problems, coupled with limited available data, typically yields ill-posed solutions, including solution *non-existence*, *non-uniqueness* and *instability*. *Non-existence* refers to no solution, given the available observations. *Non-uniqueness*, caused by insufficient data, refers to different solutions that are identified to give similar explanations to the observations. *Instability* refers to inverse solutions that are sensitive to small perturbations in the observations. Thus, in the context of a WDS contamination event, the dynamic nature and uncertainties of the system and the need for rapid characterization contribute to the complexities inherent to the contaminants and their sources.

Previous efforts have concentrated on characterizing the contaminant by constructing it as an optimization problem (e.g. van Bloemen Waanders *et al.* 2003; Laird *et al.* 2005, 2006; Guan *et al.* 2006; Liu *et al.* 2006; Preis & Ostfeld 2007, 2008). These optimization approaches include direct methods and simulation–optimization approaches. In van Bloemen Waanders *et al.* (2003), a standard successive quadratic programming tool was applied to solve a small-scale problem. Laird *et al.* (2005) suggested an origin tracking algorithm to estimate time-dependent contaminant injections for every network node based on a nonlinear programming framework. Laird *et al.* (2006), built on the results of Laird *et al.* (2005), to resolve the non-uniqueness difficulty by including a mixed-integer quadratic program. Because of the discreteness, nonlinearity and nonconvexity, as well as the limiting assumptions of existing optimization formulations, indirect methods have recently attracted increasing attention. Taking advantage of a simulation–optimization approach, wherein the water distribution system simulation model EPANET was used as a simulator, Guan *et al.* (2006) demonstrated its applicability to nonlinear contaminant sources and release-history identification by incorporating the reduced gradient method. Another simulation–optimization approach, proposed by Liu *et al.* (2006), used a multiple population-based evolutionary algorithm to search for a set of contaminant source characteristics that may result in similar sensor observations. Preis & Ostfeld (2007, 2008) described a straightforward approach for contaminant source identification by coupling EPANET with a genetic algorithm. Nevertheless, computational efficiency remains of great concern because such methods often require numerous time-consuming simulation runs to evaluate potential solutions. It is especially difficult to obtain a good solution within a reasonable amount of computational time in a large network, even using parallel or distributed computing implementations. Computational requirements may be reduced by using a prescreening technique that eliminates infeasible solutions to reduce *a priori* the decision space in which the procedure must search. One such prescreening method is the back-tracking algorithm reported by De Sanctis *et al.* (2006), which is able to identify all possible locations and times that explain contamination incidents detected by water quality sensors. Another approach, proposed by Di Cristo & Leopardi (2008), makes use of the pollution matrix concept to determine a group of candidate nodes that could explain discrete solute concentration measurements. The focus of the study presented in this paper is to complement the available search methods by developing and testing a procedure for prescreening the network to assign a relative probability of each node being a candidate potential source. A statistical model is proposed to estimate the likelihood that a given node is the contaminant source. The estimated probability values are then used to rank or group the sources that present an overall explanation for water quality observations under various uncertain circumstances.

While the knowledge of an existing water network and its sensor placements allows simulations of various hypothetical contamination events, the relationship between contaminant source characteristics and their resulting sensor observations may be pre-established through the simulation of a large set of potential contamination events. The prescreening procedure presented here is built upon a large number of contamination simulations that are then processed to develop a probabilistic depiction of contaminant sources as a function of concentration observations at the sensors. This approach is expected to reduce online computational time and statistically characterize contaminant sources based on the currently available concentration data. The use of the developed method is demonstrated for two WDS networks.

PROBLEM DESCRIPTION

Numerous possible injection scenarios, unknown water consumption at demand nodes, and errors inherent to measurements and models contribute collectively to the high degree of uncertainty in identifying the source during a WDS contamination event. Because of these uncertainties, it is essential to provide a statistical characterization of the possible contaminant sources. Although a contaminant source is typically characterized by its location and corresponding mass loading history, just knowing the location helps isolate quickly the area in the network where the real source may reside. This study concentrates on estimating the probability of each node being a candidate source location based on the sensor observation data obtained from the first detection to the current time. As the contamination event is a dynamic process in which the set of observations changes, the estimation of the likelihood that any given node is the contaminant source is updated according to the varying number of sensor observations.

LOGISTIC REGRESSION ANALYSIS FOR THE RAPID DETERMINATION OF A CONTAMINANT SOURCE

Logistic regression (LR) analysis

An LR model (LRM) (Hosmer & Lemeshow 1989) can be used to estimate the probability of the presence of an event, given information about predictors that can potentially influence the outcome. As a class of generalized linear models, LRMs are distinguished from ordinary linear regression models by the range of their predicted values, the assumption of the variance of the predicted response and the distribution of the prediction errors. The general LRM formulation is

$$\log\left(\frac{p}{1-p}\right) = b_0 + bX \tag{1}$$

where p represents the probability of a response of 1 (i.e. the presence of an event); $\{b_0, b\}$ are the regression coefficients and X is a vector of the k explanatory variables. In the above formulation (Equation (1)), the term $\log(p/(1-p))$ is called a logit function, which is used to transform the predicted value between 0 and 1 to a response ranging from $-\infty$ to $+\infty$. This mathematical formulation assumes that a linear relationship exists between the logit function and the predictors.

LRMs have been used successfully in the field of water resources as predictive models to obtain categorical forecasts or estimates. The strength of an LRM lies in its ability to directly provide a categorical forecast (i.e. the probability of occurrence of a particular event) with low computational costs. The implementation of LRMs is simple and flexible in comparison to some other predictive methods. Lu *et al.* (2006) investigated the use of an LRM in the relationship between the presence of dehalococcoides DNA in groundwater from monitoring wells and the values of selected biogeochemical parameters. Also, Regonda *et al.* (2006) obtained categorical probabilistic forecasts from an LRM using a large-scale climate predictor to estimate the probability of the leading mode of a basin stream flow above a given threshold.

LRM construction

A linear LRM-based approach is employed to model the likelihood that any given node is the contaminant injection location, and is driven by the sensor measurements. The appropriate inclusion of the predictors is a major challenge, particularly in the LRM construction. With respect to model stability, the criterion of predictor selection can minimize the number of predictors, whereas incorporating more predictors into the model aids in an overall understanding of the problem. Unfortunately, a large number of predictors may result in an over-fitting of the model. Because a contaminant may be introduced arbitrarily into a network, the randomness of the contaminants and the resulting water quality data also pose challenges to the LRM construction. Given these considerations, to predict the likelihood that any given node is the source at time t, an LRM is constructed using the observations at the current time as predictors. This model construction approach yields one LRM for each node at each measurement time step. Thus, the total number of LRMs for the whole network is the number of potential source nodes multiplied by the number of time steps for observation.

The following mathematical formulation (Equation (2)) is defined to determine, at time t after the contamination is first detected at one or more sensors, the probability that node i is a contaminant source location based on the observation at

N sensors at current time t:

$$\pi_{it} = \log\left(\frac{p(A_i|C_1(t),\cdots,C_N(t))}{1-p(A_i|C_1(t),\cdots,C_N(t))}\right) = b_0(i,t) + b_1(i,t)C_1(t) + \ldots + b_j(i,t)C_j(t) + \ldots b_N(i,t)C_N(t) \quad (2)$$

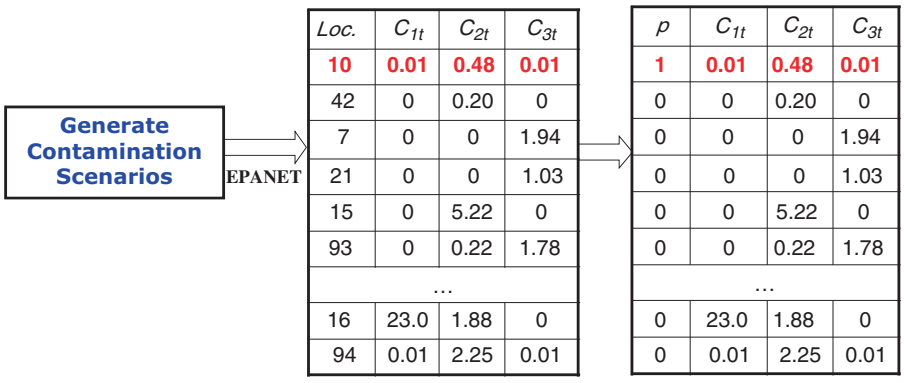
where $p(A_i|C_1(t),\cdots,C_N(t))$ denotes the likelihood of the contaminant introduced at node i given the observations at time t; A_i represents the contaminant entering through node i; ($C_1(t)$, …, $C_N(t)$) are the sensor observations at time t; and ($b_0(i,t)$, …, $b_N(i,t)$) are regression coefficients for node i at time t obtained by the maximum likelihood procedure. A detailed description of the maximum likelihood estimation can be found in Hosmer & Lemeshow (1989). In this paper, the maximum likelihood estimation method implemented within MATLAB is used to estimate the LRM coefficients. From this formulation, the probability that node i is the source location can be calculated from the observed concentration at time t as

$$p(A_i|C_1(t),\cdots,C_N(t)) = \frac{\exp(\pi_{it})}{1+\exp(\pi_{it})} \quad (3)$$

Ideally, it is expected that the LRM can identify the true source node with the greatest probability value compared to other nodes in the network. Several factors potentially impact the accuracy of the probability estimates, including the precision of the measurements, hydraulic variability, the degree of non-uniqueness (as multiple locations could potentially yield similar observations at the sensors) and assumptions of linearity in the regression function form that may be resolved by dividing one LRM into several to fit the observation data at different levels. Nevertheless, the estimated likelihood values are expected to be favorable in creating an effective control strategy in the event of contamination. Additionally, this analysis can serve as a prescreening step for some other methods, such as heuristic searches, to discover the optimal mass loading profiles at potential nodes.

Data generation

To develop the LRMs as described above, first a large set of contamination scenarios is generated to represent the sensor observations at various intervals in response to possible contamination events. Each contamination scenario includes at least one non-zero sensor observation. Contaminants vary according to the injection location, starting time, duration and mass injection rates. The injection location could be any of the network nodes, and the starting time and duration, as well as the mass injection rates, are randomly selected from a uniform distribution, bound by the specified values. Accordingly, a large set of sensor measurements is produced using EPANET simulations for the randomly generated events; these measurements are then used as inputs for developing the LRM. During the training of the LRM for each node, the probability value (i.e. the output of the LRM) is assigned a value of 1 (or 0) if the contamination occurs at this location. Figure 1 shows an example of training data generation of a given node (e.g. node 10) at time t, starting with creating a large set of contamination scenarios. For each scenario, the sensor measurements at time t (three sensors are assumed in this example) can be obtained by running the EPANET



Loc.	C_{1t}	C_{2t}	C_{3t}
10	0.01	0.48	0.01
42	0	0.20	0
7	0	0	1.94
21	0	0	1.03
15	0	5.22	0
93	0	0.22	1.78
…			
16	23.0	1.88	0
94	0.01	2.25	0.01

p	C_{1t}	C_{2t}	C_{3t}
1	0.01	0.48	0.01
0	0	0.20	0
0	0	0	1.94
0	0	0	1.03
0	0	5.22	0
0	0	0.22	1.78
…			
0	23.0	1.88	0
0	0.01	2.25	0.01

Figure 1 | Illustration of training data generation for node 10 at time t.

simulation model and the simulated sensor concentrations, including at least one non-zero value, are recorded. Then, the source location is converted to a probability value (0 or 1). Although multiple nodes could have been the source due to the limited observations in a contamination event, only the correct source node (e.g. node 10) receives the probability of 1 and the other nodes get the probability of 0. Thus, the training data for a given node incorporates a set of probability values of being the true source and sensor measurements. To save computation time on the EPANET simulation runs, the same set of contamination scenarios is used for creating the LRMs for all the nodes in a network.

Performance evaluation

To assess the performance of an LRM, a validation dataset is generated as well. Using the data generation approach described for LRM construction, a different set of injection scenarios is created for validation purposes. The following three performance evaluation criteria are used: (1) the frequency with which the true source location obtains a non-zero probability that it is a candidate source location; (2) a cumulative distribution function of the number of candidate nodes among a large set of scenarios and (3) the frequency with which the true source location is identified as the most likely source of contamination based on the LRM predictions.

APPLICATIONS AND RESULTS

In this section, two WDS networks with different levels of complexity are used to test and demonstrate the predictive potential of the LRM. In the small network, a large number of hypothetical contamination events are examined to assess the identification ability of the LRM. This investigation is furthered by varying the source parameter range, training dataset size, and the number and quality of the measurements as well as by giving consideration to water consumption uncertainty. In the second larger network application, the development of the LRMs, considering their similarities between two consecutive time steps, is assessed.

The hydraulic and water quality simulations are executed by running EPANET during the generation of the dataset. The hydraulics remains at a steady state during hourly simulations and has a periodicity of 24 h. A conservative contaminant is assumed to be injected at a single location where the hydraulic conditions are known. Although the varying parameters that are used to create numerous scenarios serve as the characteristics of the contaminant sources in this study, the suggested approach can be extended further to incorporate system uncertainties when building the LRMs.

Small example network

The first illustrative example uses a small network, which is one of the problem scenarios available as a tutorial within EPANET (Rossman, 2000). This network consists of 97 nodes, 2 sources, 3 tanks and 117 pipes. The configuration of the network is depicted in Figure 2 and further details can be found in the EPANET user's manual. The contaminant transport is simulated in 10-min intervals and the concentration values at the sensors are observed at 10-min increments.

To demonstrate the algorithm's performance, a set of LRMs that corresponds to each node at each time interval is built upon the generated training datasets. Table 1 lists the parameters and their values that are used for the simulations of the hypothetical events. Here, LRMs that span 12 h, which correspond to 72 time steps (each represents a 10-min interval), are chosen for the investigation. The computational time for

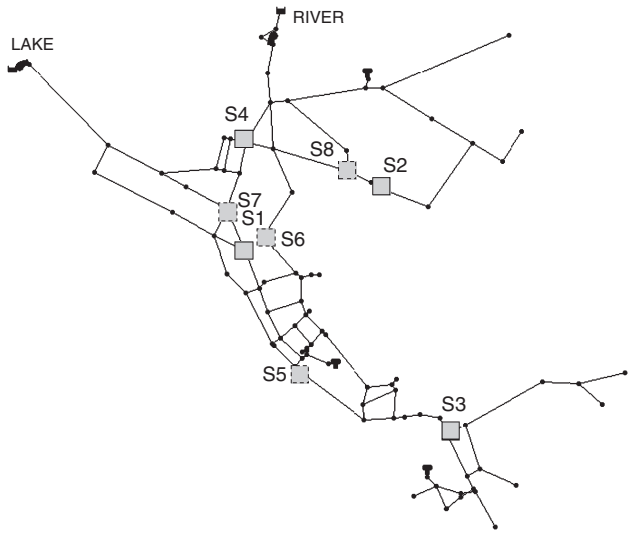


Figure 2 | Water distribution network schematic (small network example). Squares designate sensor locations. For Scenario 1, the sensor network is composed of S1, S2, S3 and S4, Scenario 2 incorporates S1 and S3, only sensor S3 is incorporated in Scenario 3 and Scenario 4 includes S5, S6, S7 and S8.

Table 1 | Contaminant source parameters and ranges for generating training dataset

Source parameter	Small network	Micropolis network
Location	Any node (1–97)	Any node (1–1577)
Starting time	Within simulation 24 h	Within simulation 48 h
Duration (h)	0–24	0–24
Mass injection rate (g/min)	0–400	0–400

training these LRMs was approximately 15 min on a 2.20 GHz Core™ 2 Duo machine.

Effect of monitoring sensors

In this subsection, the predictive capability of the LRMs is examined when the varying number of observations changes. In addition to determining whether the true source is recovered as a potential solution, the relative rank of the true injection location compared to that of other nodes is evaluated, and the number of potential solutions is determined. Four scenarios employed here incorporate four sensors, two sensors and a single sensor, respectively. The locations of the selected sensors are shown in Figure 2. Scenario 1 is composed of sensors S1, S2, S3 and S4, which are used in subsequent analyses, Scenario 2 incorporates S1 and S3, Scenario 3 includes S3 only, while Scenario 4 includes the observations obtained from sensors S5, S6, S7 and S8. The same set of contamination scenarios is used to enable a meaningful comparison. Each scenario contains at least one non-zero concentration data point and the total number equals 1000 at each time interval.

The generated results of the four scenarios indicate that the established LRMs are capable of recovering the true source node as a candidate solution, with an estimated non-zero probability. For each scenario, all the candidate source nodes are ranked according to the calculated probabilities in descending order, whereby the node with the largest value is ranked first. The number of times that true source is ranked highest can be used as a measure of performance of the LRMs. Figure 3 (left) shows the variation in the frequency with which the true source was ranked top to the lowest rank. Overall, a high frequency corresponds to a top rank for the true source node. Indeed, the frequency trend greatly depends on the amplitudes of the variations of the parameter values when building the LRM. From Figure 3 (left), it can be seen that the correct node is ranked first 17% of the time and the correct node is ranked 40th only 0.1% of the time in Scenario 1. Compared to Scenarios 2 and 3, an increase in the number of measurements improves the rank of the true source nodes. Further, the estimation uncertainty is measured as the number of potential solutions. The cumulative distribution function (CDF) of the number of potential solutions is shown in Figure 3 (right). Due to the additional measurements in Scenario 1 in comparison with Scenarios 2 and 3, the probability of identifying a smaller set of candidate solutions increases (nearly 40% of the time that 20 possible solutions are identified in Scenario 1), indicating that a large number of observations aid in reducing the uncertainty in identifying candidate solutions. It is also noted that, although Scenarios 1 and 4 incorporate the same number of sensors, the sensor network in Scenario 4 yields better performance since it achieves a higher rank of the true source node and identifies a smaller set of candidate solutions more frequently as a result of different sensor locations. This indicates that the LRMs could help better locate sensors in the network; thus the performance of LRMs can be improved accordingly.

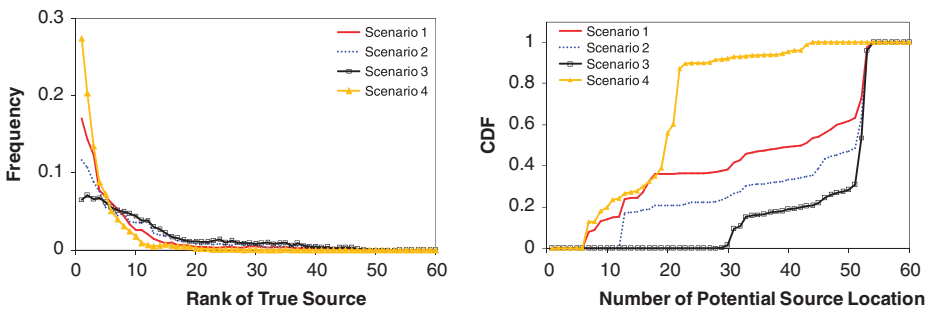


Figure 3 | Comparison of performance of LRMs for four scenarios: (left) rank of true source location vs. frequency; (right) CDF for number of potential source locations.

Table 2 | Allowable source parameter ranges for generating training dataset

Source parameter	Set A	Set B	Set C
Location	Any node (1–97)	Any node (1–97)	Any node (1–97)
Starting time	Within simulation 24 h	Within simulation 24 h	Within simulation 24 h
Duration (h)	0–12	0–24	0–24
Mass injection rate (g/min)	0–200	0–400	0–600

Effect of training data generation

In this subsection, first the influence of contaminant source parameter ranges in the model-building process was examined. Three different training data vary by the contaminant source parameter ranges for the small network example, listed in Table 2. To make the comparison meaningful, the same set of contamination scenarios was used, which contains 1000 samples bounded by the set *C* and are different from the data for the model training. Figure 4 presents the frequency versus the rank of the true source. The parameter ranges greatly influence the predictive power of the LRMs, as shown by Figure 4. The LRMs developed on set *C* exhibit a better performance, reflecting that the more restrictive data does not capture the full range of the contaminant source and reduces the applicability of the LRMs for predictive purposes, although the LRMs built on more restrictive data may perform better for a particular contamination event. Nevertheless, the LRMs developed on different allowable ranges can always identify the true source node as a potential solution. The second investigation was carried out to study the effect of the sample size on the LRMs' performance. Various training datasets are generated, differing in the number of training samples, ranging from 500 to 20 000. Again, the established LRMs were tested on the same set of hypothetical contamination scenarios for different cases. The experimental results with respect to the frequency of the rank of the true source are shown in Figure 5. Increasing the dataset size is beneficial in improving the rank of the true source node. However, this advantage tends to decline as the size increases. Also, we note that a minimum of 5000 data points is sufficient for the LRMs to predict the true source node as a potential solution, since the established LRMs with no less than 5000 training samples are always capable of identifying the true source node as a candidate solution with an estimated non-zero probability.

Impact of measurement errors and demand uncertainties

Poor performance may occur as a result of the errors that are related to the information, such as imperfect measurements or uncertain amounts of water consumption, that affects the simulation of the contamination event. To understand the effects of these uncertainties on LRM solutions, a normally

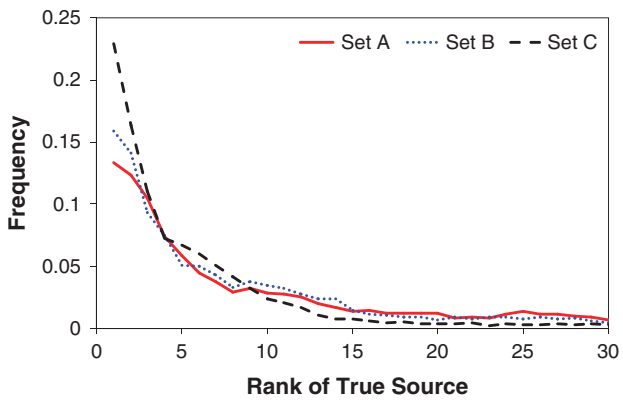


Figure 4 | Comparison of the LRM performance with different settings of parameter ranges.

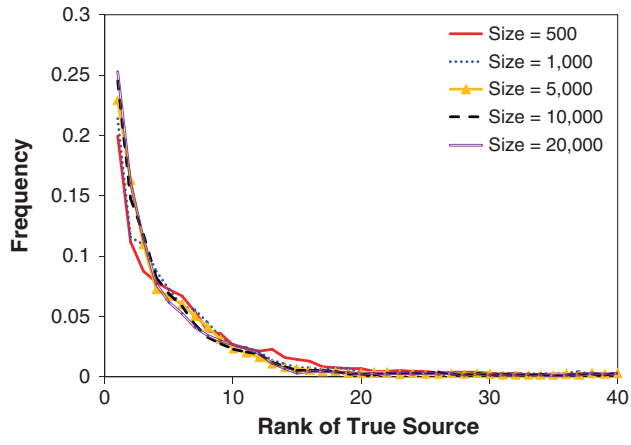


Figure 5 | Comparison of the LRM performance with different sizes of training datasets.

distributed white noise was added to each factor. The mathematical formulation for modeling these factors with perturbation is expressed as

$$y_{it}^{err} = y_{it} + \alpha^* y_{it}^* N(0,1) \quad (4)$$

where y_{it}^{err} denotes the perturbed measurement of sensor i or the demand multiplier of node i at time t; y_{it} denotes the true measurement of sensor i or the demand multiplier of node i at time t; and α represents the error level added to the perturbed factor.

The results of incorporating different levels of either measurement errors or demand uncertainties are summarized in Table 3. Making use of the same set of hypothetical events as demonstrated above, the performance is evaluated in terms of the frequency that the LRM predicts the true source as a candidate solution. As shown in Table 3, the 50% uncertainty level in the measurements causes a small number (1%) of cases in which the true source is not identified as a candidate solution. This result implies that the performance is highly independent of measurement errors (as modeled using Equation (4)) in the observations. An explanation for this behavior is that a conservative contaminant, which is what was assumed in these analyses, leads to a linear relationship between the sensor observations and the contaminant loading. Therefore, the estimation of the contaminant source location depends on the presence/absence status of contamination rather than the magnitude of observations at the sensors. The same level of uncertainty associated with water consumption, however, yields poor performance. In reality, if the demands are highly uncertain, changes in the flow direction in the network may occur, thereby biasing the prediction of the LRMs developed under normal conditions.

Table 3 | Summary of results under various uncertain conditions

Scenario	Frequency of true source predicted to have a non-zero probability
Ideal condition	100%
Measurement error ($\alpha = 10\%$)	100%
Measurement error ($\alpha = 50\%$)	99.9%
Uncertain demand ($\alpha = 10\%$)	99.7%
Uncertain demand ($\alpha = 50\%$)	97.2%

Binary sensor condition

In addition to the above analysis, which assumes the use of chemical-specific probes in the sensor network, the LRM performance is further investigated if the observations are reported as a binary signal (detection or no detection). A binary sensor signal filters the original chemical signal, resulting in potential reduction in information quality. In reality, the deployment of binary sensors allows water utility operators to access merely the status of the contamination, which may be specified by the level of water quality indicators (e.g. pH, chlorine, conductivity). For simplicity, a concentration level is set as the detection threshold. The observation data are converted to 1 (presence of contamination) if the reading exceeds this threshold and 0 (absence) otherwise. Thus, a set of LRMs is built using these binary observations as inputs to the LRMs. Although the same set of validation data is used as demonstrated above, the set must be converted to 0/1 according to the given threshold. The performance results for 0.01, 0.1 and 1.0 mg/L detection thresholds are compared in Figure 6.

Figure 6 (left) shows that the frequency with which the true source is ranked high decreases as the threshold value increases. As more data filtering takes place with increasing threshold value, the reduced data quality leads to relatively poor performance. The cumulative probability of the number of potential solutions is shown in Figure 6 (right), which suggests that a larger number of unlikely nodes are eliminated when the detection limit is lower. The frequency of the LRMs that recognize true locations as candidate solutions exceeds 99.9% among all scenarios, even with a detection limit of 1.0 mg/L. These results indicate the effectiveness of the LRMs in ruling out unnecessary nodes as candidate source locations with extremely coarse data.

Real-time updates of probabilities

During a contamination event, sensor monitoring data are collected dynamically as time progresses. While the LRMs offer the capability to predict the probability given the observations at the current time, the time series of the observed data, from the first detection to the current time, can be used collectively to recover the source of contamination. A joint probability that a node is not the source can be specified as a

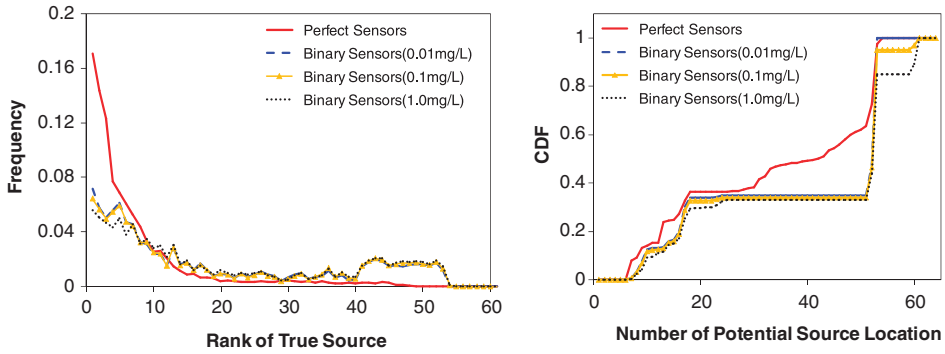


Figure 6 | Comparison of results between perfect and binary sensor conditions: (left) rank of true source location; (right) CDF of number of potential source locations.

product of the likelihood of the contaminant not being introduced at this node through a sequence of time intervals. Specifically, if a series of available observations collectively indicate that a node is not the source, it is concluded that the contaminant is not introduced at this location. Thus, the probability that a given node is a source can be updated in real-time as follows:

$$P(A_i|C_0,\cdots,C_t) = 1-(1-p(A_i|C_0))\\ (1-p(A_i|C_{t_0+1}))\cdots(1-p(A_i|C_t)) \quad (5)$$

where $P(A_i|C_0,\cdots,C_t)$ represents the updated probability of the contaminant injected at node i at time t given currently available observations $\{C_0,\cdots,C_t\}$; A_i represents that contamination occurs at node i; $p(A_i|C_t)$ denotes the predicted probability of node i as a source using the observation C_t from all the sensors in a network at current time t, which is estimated directly by LRMs; and t_0 refers to the first detection time.

A total of 1000 contamination events are considered to achieve statistical significance. The 95% confidence interval of updated probabilities and the rank of true nodes are used to indicate the level of robustness of the results, as listed in Table 4. As is the case with increased measurements, a longer observation period yields a higher likelihood of selecting the true source node as the most likely candidate contamination source. When measurements up to three hours are included, on average the true source node is predicted as a candidate source node with over 50% likelihood, with a small confidence interval. However, this occurrence does not mean that true source nodes must be increasingly dominant over other nodes with more measurements. As shown in Table 4, the rank of the injection node shows a slight increase with time. An explanation for this behavior is that more nodes become incorporated into the candidate set due to increasingly available measurements. This observation also indicates the complexity of the source identification, which results from the high levels of uncertainty associated with such a problem.

Micropolis example network

To evaluate a more general effectiveness of the LRMs, a relatively large Micropolis network is examined. In addition to studying the effects of the increased problem complexities on performance of the LRM approach, a strategy to reduce the number of LRMs is evaluated. The configuration of the water network is depicted in Figure 7, which is composed of 1574 junctions, 1415 pipes, 8 pumps, 2 reservoirs and 1 tank. This example was developed for the Micropolis virtual city with 5000 residents, further details of which can be found in Brumbelow *et al.* (2007). The locations of five sensors are randomly selected within the entire network (see Figure 7).

In a real network with a large number of nodes and long event simulation periods, the number of LRMs (one for each node at each time step) needed to be developed is high.

Table 4 | Statistical summary of the LRM results that correspond to the true source node

	Confidence interval (95%)	
Elapsed time (h)	Probability (%)	Rank
1	[33, 36]	[6.83, 7.92]
3	[56, 59]	[7.53, 8.69]
6	[70, 73]	[8.24, 9.47]
12	[77, 80]	[8.54, 9.79]

Although the LRMs are developed offline *a priori*, the computational demand for developing this large set of LRMs can become computationally intensive. Alternatively, one may be able to use fewer LRMs, as some similarity may exist between an LRM at a current time step, say t, and an LRM at $t-1$. The need to regenerate a new LRM at t could be determined depending on whether the model at the previous time fits the current observation data. If the fit with the LRM for the prior time step is poor, then a new LRM corresponding to the current time is created. This process is expected to reduce the computational costs during the model-building process. The fit of the LRMs is evaluated in terms of the following mathematical constraint:

$$\sum_{s=1}^{M} p(A^s_{true}|C^s_1(t), \cdots, C^t_N(t)) \geq M * p^{\min} \quad (6)$$

where M is the number of simulated contamination scenarios used for evaluation; $(C^s_1(t), \cdots, C^s_N(t))$are the sensor measurements at time t for the sth scenario; $p(A^s_{true}|C^s_1(t), \cdots, C^s_N(t))$ is the estimated likelihood of the contaminant introduced at the true source node given the sensor measurements at time t, which is calculated by the LRMs at previous time t-1; and p^{min} denotes the minimal probability value allowed for the true source node. If the training data corresponding to the observation time t meets the constraint, reuse the LRMs at time t-1; otherwise, create new LRMs for time t.

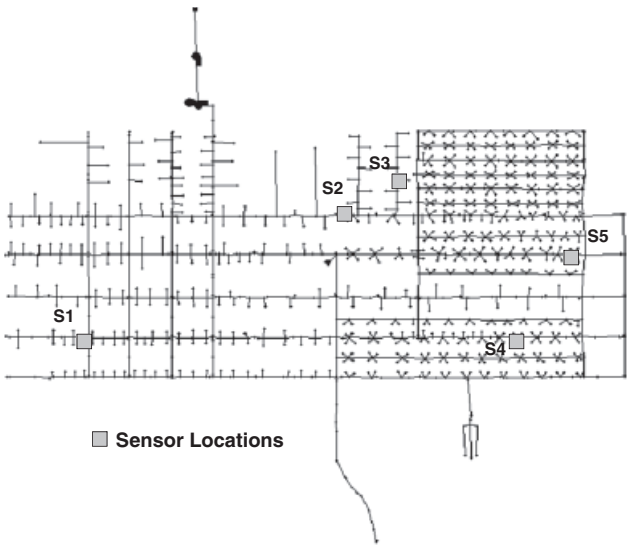


Figure 7 | Layout of Micropolis water distribution network. The sensor network is composed of S1, S2, S3, S4 and S5, denoted by squares.

The reuse of the LRMs is assessed by comparing their performance against that of an independent model generation. Again, the LRMs that span 12 h, thus representing 72 time intervals, are chosen for the investigation. The parameters and their values that are used for simulating hypothetical events are shown in Table 1. For evaluation purposes, a set of contamination scenarios is generated, which contains 5000 samples at each time interval and varies according to the injection location, starting time and duration as well as mass injection rates. The first interval is selected as the time immediately after the one-day simulation period, which allows enough time for contaminants to reach the downstream nodes. A comparison of frequency as a function of the true source ranking between the reuse strategy and the independent model generation strategy is presented in Figure 8. Although the reuse strategy is slightly worse than the independent model generation strategy in terms of the true source node ranking, both strategies capture the true source node as a candidate solution among all scenarios. The reuse method took only 5 h 20 min on a 2.20 GHz Core™ 2 Duo machine to train LRMs spanning 12 h, while the computation time for the independent LRM generation was approximately 86 h. Thus, the reuse strategy saves around 94% of the computational costs during the model-building process.

To examine the distribution of potential solutions obtained from the LRMs, one hypothetical contamination event is simulated. The contaminant, with a constant mass injection rate of 60 g/min, is introduced at the source node (labeled as IN 1646), shown in Figure 9. The detection occurred initially at 12:30 p.m. and lasted until 1:40 p.m. at sensor S5. The candidate source nodes determined by LRMs for the first observation are shown in Figure 9. The LRMs identified 167 solutions out of 409 nodes that could contribute to observations at the given sensor locations. These identified candidate locations have very similar likelihood values of being an injection location. As time goes on with more available information, some candidate nodes achieve a much higher probability value than others. Figure 10 shows the estimated probabilities of the network nodes being the source location up to 1:40 p.m. It is worth noting that the possible sources with a higher probability are relatively close to the true source. For the given sensor network, however, a large set of unknown nodes exists, because such nodes are

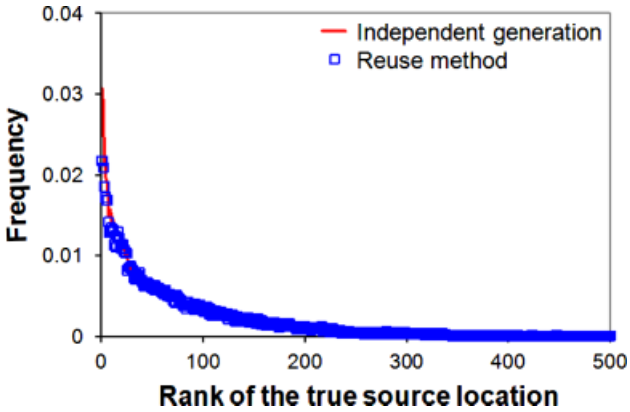


Figure 8 | Comparison of results between two model-building strategies (for the Micropolis network).

undetectable if the contaminant is injected at these nodes. It is also noted that the true source location is one of the candidate source nodes at every time step.

FINAL REMARKS

Identifying a contaminant source quickly is vital for creating an effective threat management strategy in the face of accidental or intentional contamination. A high level of uncertainty inherent to the contaminant and WDS complicates the characterization of the contaminant sources. The approach demonstrated here enables a probabilistic description of the

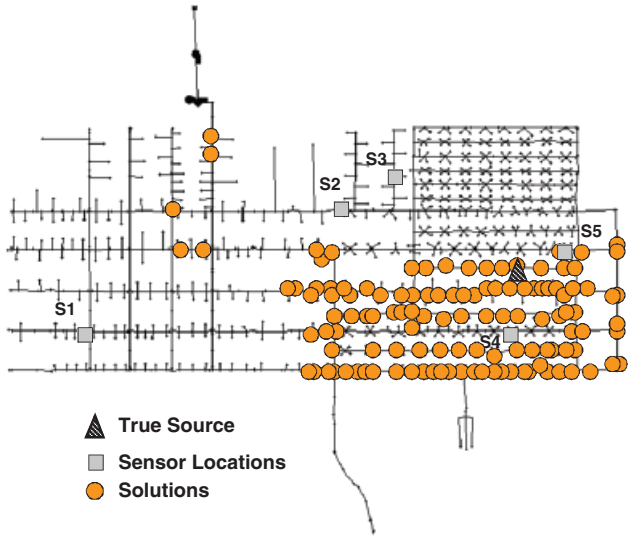


Figure 9 | Locations of candidate nodes at 12:30 p.m. (based on 10-min observations).

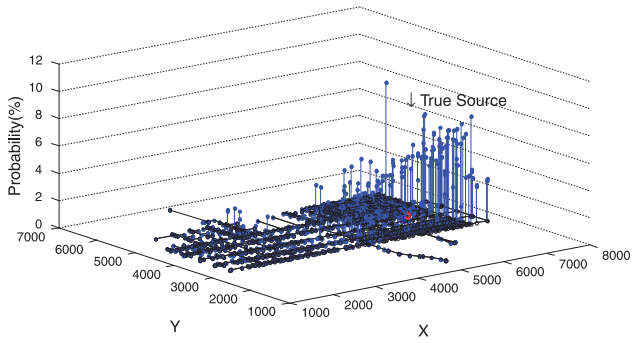


Figure 10 | The probability of the network nodes being the source location identified at 1:40 p.m. (based on 70-min observations).

source location that allows for considering various uncertainties associated with the contamination. Statistical models are built to describe the contaminant as a function of available measurements using a large number of hypothetical contamination simulations. Together, rapid prediction and simple implementation can be achieved through the use of LR analysis. In this work, the relationship between the likelihood that a given node is a source and the sensor observations is expressed by LRMs. Once established, using these LRMs can lead to fast estimation of candidate source nodes when contamination is detected.

The LRMs were applied to two WDS networks, and numerous contamination events were investigated. The smaller network application considers the effects of training data generation and the sensor measurements as well as the demand variations on source node identification as the LRMs are developed. A method to update the probabilities dynamically is proposed herein as well. The larger network application demonstrates the applicability of LRMs to a problem with more nodes and longer simulation periods. A procedure for reducing the number of necessary LRMs is developed and tested. This procedure checks for the similarities in the LRMs between two consecutive time steps and uses the LRM from the previous time step to avoid the creation of additional LRMs.

From the results and analysis described, this proposed LRM approach is able to determine candidate source locations, among which the true source node is included. This performance was consistently observed for numerous scenarios, including ones with coarse and noisy monitoring data. The results indicate that demand uncertainty has a larger impact than measurement errors due to the possibility of a

change in the flow direction in the WDS network. Additional measurements from more sensors and for longer observation times can improve the performance of such a problem with respect to the rank of the true source as well as the number of candidate solutions. The allowable source parameter range for generating training data has a great effect on the performance of the LRMs, while the model accuracy tends to increase at a decreasing rate as sample size increases. The reuse of LRMs can produce results that are comparable to those produced by the independent generation of LRMs, with a significant reduction in computational costs.

The proposed methodology showed an ability to estimate potential contaminant source locations, on the basis of the LR modeling using prior simulation results. The way in which the LRM is constructed can alleviate online computational burdens while the contamination event is occurring, although the process of model-building requires the time for model construction as well as a large number of offline simulation runs in which EPANET is used. A significant advantage of this approach is that the pre-established LRMs allow fast estimation of the contaminant source. For example, the calculation of probabilities by running the LRMs of the entire network nodes took only 0.1 s of CPU time on a 2.20 GHz Core™ 2 Duo machine for the small example network and 1 s for the Micropolis network. Therefore, the methodology described is applicable to a larger network than the examples presented here.

Although the proposed approach facilitates a probabilistic characterization of each node in a contamination event, the other characteristics associated with the contaminant (e.g. injection starting time, duration, mass flow rates) are underdetermined. Further work could consider the LRMs in combination with other methods, such as heuristic search approaches, to enhance the contaminant source characterization accuracy. For example, an LR analysis is performed prior to the heuristic search method. The location-specific probability information is then used to limit the potential source nodes, thus reducing the feasible solution space and yielding a fast convergence for the heuristic search. In addition, the selection pressure in the subsequent heuristic search may be assigned differently to diverse regions of the water distribution network based on the probability that any given source location is the true source. Moreover, future work is required to extend this approach to a more realistic condition, such as the likelihood of simultaneous multiple injection locations, unknown hydraulic conditions, false positives and false negatives from sensor readings. The location of monitoring sensors will unavoidably affect the contaminant source identification problem, which is arbitrarily selected in this study. It is valuable to investigate how LRMs can help better locate sensors in the network, so the performance of LRMs can be improved accordingly.

ACKNOWLEDGEMENTS

This work is supported by the Major Special Technological Program of Water Pollution Control and Management (Program No.2009ZX07106-001) and the National Science Foundation (NSF) under grant no. CMS-0540316 under the DDDAS program.

REFERENCES

Brumbelow, K., Torres, J., Guikema, S., Bristow, E. & Kanta, L. 2007 Virtual cities for water distribution and infrastructure system research. In: *Proc. World Environmental and Water Resources Congress, Tampa, Fl, 15–19 May*. ASCE, Reston, VA.

De Sanctis, A. E., Shang, F. & Uber, J. G. 2006 Determining possible contaminant sources through flow path analysis. In: *Proc. Water Distribution Systems Analysis Symposium, Cincinnatti, OH, 27–30 August*. ASCE, Reston, VA.

Di Cristo, C. & Leopardi, A. 2008 Pollution source identification of accidental contamination in water distribution networks. *J. Wat. Res. Plann. Mngmnt.* **134**(2), 197–202.

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. *J. Wat. Res. Plann. Mngmnt.* **132**(4), 252–262.

Hosmer, D. S. & Lemeshow, S. 1989 *Applied Logistic Regression*. Wiley, New York.

Laird, C. D., Biegler, L. T. & van Bloemen Waanders, B. G. 2006 Mixed-integer approach for obtaining unique solutions in source inversion of water networks. *J. Wat. Res. Plann. Mngmnt.* **132**(4), 242–251.

Laird, C. L., Biegler, L. T., van Bloemen Waanders, B. G. & Bartlett, R. A. 2005 Contamination source determination for water networks. *J. Wat. Res. Plann. Mngmnt.* **131**(2), 125–134.

Liu, L., Zechman, E. M., Brill, E. D., Mahinthakumar, G., Ranjithan, S. & Uber, J. G. 2006 Adaptive contamination source identification in water distribution systems using an evolutionary algorithm-based dynamic optimization procedure. In: *Proc. Water Distribution Systems Analysis Symposium, Cincinnatti, OH, 27–30 August*. ASCE, Reston, VA.

Lu, X., Wilson, J. T. & Kampbell, D. H. 2006 Relationship between geochemical parameters and the occurrence of dehalococcoides DNA in contaminated aquifers. *Wat. Res. Res.* **42**(16), 3131–3140.

Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., Phillips, C. A., Watson, J. -P., Dorini, G., Jonkergouw, P., Kapelan, Z., di Pierro, F., Khu1, S. -T., Savic, D., Eliades, D., Polycarpou, M., Ghimire, S. R., Barkdoll, B. D., Gueli, R., Huang, J. J., McBean, E. A., James, W., Krause, A., Leskovec, J., Isovitsch, S., Xu, J., Guestrin, C., VanBriesen, J., Small, M., Fischbeck, P., Preis, A., Propato, M., Piller, O., Trachtman, G. B., Wu, Z. Y. & Walski, T. 2008 The battle of the water sensor networks (BWSN): a design challenge for engineers and algorithms. *J. Wat. Res. Plann. Mngmnt.* **134**(6), 556–558.

Preis, A. & Ostfeld, A. 2007 A contamination source identification model for water distribution system security. *Engng. Optim.* **39**(8), 941–951.

Preis, A. & Ostfeld, A. 2008 Genetic algorithm for contaminant source characterization using imperfect sensors. *Civil Engng. Environ. Syst.* **25**(1), 29–39.

Regonda, S. K., Rajagopalan, B. & Clark, M. 2006 A new method to produce categorical streamflow forecasts. *Wat. Res. Res.* **42**(9), 9501–9506.

Rossman, L. A. 2000 *EPANET User's Manual*. Risk Reduction Engineering Laboratory, US Environmental Protection Agency, Cincinnati, OH.

van Bloemen Waanders, B. G., Bartlett, R. A., Bigler, L. T. & Laird, C. D. 2003 Nonlinear programming strategies for source detection of municipal water networks. In: *Proc. ASCE World Water and Environmental Congress, Philadelphia, PA, 23–26 June*. ASCE, Reston, VA.

First received 29 November 2009; accepted in revised form 22 February 2010. Available online 28 October 2010