Preprints of the 18th IFAC World Congress
Milano (Italy) August 28 - September 2, 2011

Fault Isolation and Impact Evaluation of Water Distribution Network Contamination ⋆

Demetrios G. Eliades and **Marios M. Polycarpou**

KIOS Research Center for Intelligent Systems and Networks,
Department of Electrical and Computer Engineering,
University of Cyprus, Nicosia CY-1678, Cyprus,
(e-mail: {eldemet, mpolycar}@ucy.ac.cy)

Abstract: The security of drinking water distribution operation is an important issue that has received increased interest within the last few years. The US Environmental Protection Agency (EPA) has issued guidelines for water utilities regarding which qualitative and quantitative metrics to monitor, as well as what response actions to take from the moment a contamination fault alarm has been triggered, until the fault has been accommodated and the system has returned to normal operation. "Expanded Sampling" is a type of response action, in which the water utilities examine the water quality at certain locations in the network after a contamination fault has been detected, to help evaluate the contamination impact and locate the source-area. In this framework, fixed quality sensors are used to detect the presence of any contaminant in the network, while and manual quality samplings are used for fault isolation and impact evaluation. In this work, a computational approach is proposed for choosing a sequence of nodes in the distribution network to perform expanded sampling, so that the water contamination impact is evaluated and the source-area is isolated, with as few manual quality samplings as possible and by minimizing the impact-risk. The proposed method is based on constructing a decision tree using multiple objectives. To illustrate the solution methodology, results are presented based on a water distribution system.

Keywords: Water Distribution Networks, Failure Isolation, Fault Location, Decision Trees.

1. INTRODUCTION

Water security is a challenging task and has become an important part of the drinking water distribution operation. To provide a framework for water security, the US Environmental Protection Agency (EPA) has published guidelines for water utilities describing a *consequence management plan*. The guidelines provide information on the qualitative and quantitative parameters that need to be monitored, as well as on the response actions to be taken from the moment a contamination fault alarm has been triggered until the fault has been accommodated and the system has returned to normal operation (U.S. Environmental Protection Agency, 2008a,b); these actions include: a) monitoring and surveillance of the system; b) event detection and determination if the contamination fault is "Possible" (i.e. if there are no strong indications of a false alarm); c) determination if the contamination fault is "Credible" (i.e. evaluating field results from the area where the contamination fault has occurred), d) determination if the contamination fault is "Confirmed" (i.e. evaluating laboratory results from multiple samples); e) implementing remediation and system recovery (U.S. Environmental Protection Agency, 2008a).

Part of the confirmation operation plan is to develop and implement "Expanded Sampling", i.e. to augment existing on-line sensor information with manual sampling at other parts of the distribution network, to determine the extend of the contamination (U.S. Environmental Protection Agency, 2008b). The EPA recommends the use of hydraulic models of the water distribution system to determine where to sample and to evaluate the spread of the contamination. The use of hydraulic models can reduce the time required for planning expanded sampling and assist in understanding the contaminant propagation path; in addition, operators can issue targeted water usage restriction notices as necessary (U.S. Environmental Protection Agency, 2008b).

Choosing where to perform expanded sampling can be a challenging task, due to the large-scale nature of the distribution network and the partially unknown hydraulic dynamics. In practice, water utilities may choose beforehand in an *ad hoc* manner certain locations in the network where expanded sampling should be conducted; sampling these locations however, may not provide adequate information regarding the contamination fault impact, or the location in the distribution network where the contamination originated. An operator may select additional

⋆ This work is supported by the Cyprus Research Promotion Foundations Framework Programme for Research, Technological Development and Innovation, co-funded by the Republic of Cyprus and the European Regional Development Fund.

nodes in the network to conduct manual sampling, based on higher-level reasoning. To reduce the bias due to human judgement, especially in large-scale networks, it is desirable to utilize a computational method for finding an optimal sequence of nodes where manual sampling should be conducted, taking into account the actual quality and hydraulic measurements as well as the topological model of the system.

The general problem of contamination source isolation has received some research attention in the last few years. A system modelling and control approach of water systems was first presented in (Brdys and Ulanicki, 1994). Analytical problem formulations for the contamination source isolation problem have been presented in an optimization framework (Uber, 2005; Laird et al., 2005, 2006; Guan et al., 2006); a graph-based method was presented in (Cristo and Leopardi, 2008) and a reverse-flow solution was presented in (De Sanctis et al., 2010). Computational and artificial intelligence techniques have been investigated, such as hybrid trees, genetic algorithms and other evolutionary algorithms (Preis and Ostfeld, 2007; Kumar et al., 2007; Liu, 2009; Zechman and Ranjithan, 2009; Vankayala et al., 2009), as well as some probabilistic methodologies (Ostfeld and Salomons, 2005; Huang and McBean, 2009).

In most of the previous work, fixed quality sensors are considered and based on their measurements, an optimization problem is solved to isolate the location where the contamination could have started. In the present work, we formulate and solve the problem from a different viewpoint, i.e. where and when to conduct manual quality sampling in the network, after a contamination has been detected at a fixed quality sensor, in order to evaluate the possible contamination impact on the system, as well as an area where the contaminant could have entered into the network, with as few samplings as possible. The motivation is that there may be a small number of specialized quality sensors installed within a water distribution network, and they may be installed at locations which facilitates detection, but not isolation. The proposed algorithm can be part of the expanded sampling strategy, as described by the EPA recommendations.

The paper is organized as follows: in Section 2 the problem formulation is described and key concepts and definitions are introduced; in Section 3 the solution methodology is presented. In Section 4 simulation results are demonstrated on an illustrative and a real network to verify the proposed algorithm. Section 5 concludes the paper with some final remarks and directions for future work.

2. PROBLEM FORMULATION

The topology of a water distribution network can be modelled as a directed graph $G = (V, E)$, where $V = \{v_1, v_2, ..., v_{N_v}\}$ is the set of N_v nodes (pipe junctions, water storage tanks) and $E = \{e_1, e_2, ..., e_{N_e}\}$ is the set of N_e edges (pipes) connecting two nodes, such that $e_j \in V \times V$, for $j \in \{1, ..., N_e\}$. Let $d_i(k)$ be the consumer outflow demand at node $v_i \in V$ and let $q_j(k)$ be the flow at edge $e_j \in E$ at discrete time k. The outflow demand vector is defined as $d(k) = [d_1(k), ..., d_{N_v}(k)]^\top$ and the flow vector $q(k) = [q_1(k), ..., q_{N_e}(k)]^\top$. In addition, let $p = [p_1, ..., p_{N_e}]^\top$ correspond to the pipe length vector, where p_j is the pipe length of the edge $e_j \in E$.

Let $f_q : \mathbb{R}^{N_v} \mapsto \mathbb{R}^{N_e}$ correspond to a hydraulic solver algorithm, e.g. as described in EPANET (Rossman, 2000), for computing the flow vector $q(k)$ according to the consumer outflow demand vector $d(k)$, such that $q(k) = f_q(d(k))$, for a given network topology and structural characteristics.

Let $V_s \subset V$ be the subset of nodes where fixed online quality sensors have been installed, after solving the sensor placement problem described in (Ostfeld et al., 2008). Let $V_m \subseteq V$ be the node subset where manual sampling can be performed; in general, some nodes may not be available for manual sampling, whereas some nodes with fixed sensors may require additional sampling. The objective is to develop an algorithm for selecting the most suitable nodes for manual sampling among V_m.

The partial differential equations describing the propagation of a contaminant substance through the water distribution dynamics, can be represented, using a suitable numerical solution, by a set of discrete-time state-space equations (Eliades and Polycarpou, 2010). For the state-space equation formulation, it is considered that the water distribution network is segmented into N virtual finite-volume cells, where $N > N_v$, assuming a certain time step Δt.

Let $x_i(k)$ be the average contaminant concentration within the i-th finite-volume cell at time k, for $i \in \{1, ..., N_v, ..., N\}$. In addition, for $i \in \{1, ..., N_v\}$, let $x_i(k)$ correspond to the node $v_i \in V$. The state-space equations which describe the propagation dynamics can be formulated as

$$x(k+1) = A(q(k))x(k) + B\phi(k - \tau_0), \tag{1}$$

where $x(k) \in \mathbb{R}^N$ is the average contaminant concentration vector at time k, such that $x(k) = [x_1(k), ..., x_N(k)]^\top$; $A : \mathbb{R}^{N_e} \mapsto \mathbb{R}^{N \times N}$ is a time-varying matrix describing the advection dynamics, based on consumer outflow demand $q(k)$ at time k, for a given network topology and structural characteristics.

The term $B\phi(k - \tau_0)$ denotes the deviation in the system dynamics at discrete time k, due to a contamination fault affecting some states, with initialization time τ_0; in general, τ_0 is considered unknown. In this work, for notational simplicity it is assumed that contamination faults may be initialized (triggered) only at the network nodes; let $B \in \{0, 1\}^{N \times N_v}$ be the diagonal matrix for which its (i, i)-th element is $B_{(i,i)} = 1$ for $i \in \{1, ..., N_v\}$ and $B_{(i,i)} = 0$ otherwise.

In addition, let $\phi : \mathbb{N} \mapsto \mathbb{R}^{N_v}$ be the contamination fault function affecting the states, such that $\phi(k - \tau_0)$ corresponds to the contamination fault injection signal which initiates at time τ_0. Single-source contamination faults are considered in this work.

It is considered that the contamination fault function $\phi(k - \tau_0)$ belongs to the contamination fault function set $\Phi(k - \tau_0) = \{\phi^1(k - \tau_0), ..., \phi^{N_v}(k - \tau_0)\}$, where $\phi^i(k - \tau_0)$ is a vector whose elements are all zero except for the i-th element which corresponds to the contaminant injection signal initiating at time τ_0. A special case of a

contamination fault function $\phi^i(k-\tau_0)$ is the unit-step function, such that the i-th element of $\phi^i(k-\tau_0)=0$, for $k<\tau_0$ and $\phi^i(k-\tau_0)=1$, for $k\geq\tau_0$.

Let S^* be the contamination fault scenario set, comprised of all contamination fault functions which initiate at certain time instances, such that

$$S^*=\{\phi(k-\tau_0)\in\Phi(k-\tau_0)\mid\tau_0\in\{0,1,...,\tau\},k\geq\tau\} \quad (2)$$

It is important to also consider the impact dynamics, which characterize the amount of "damage" caused by a contamination fault scenario $s\in S^*$ affecting a demand node at each time step, measured with a certain impact metric (e.g. number of people affected), as discussed in (Eliades and Polycarpou, 2010). Let $\xi_i(k;s)$ be a state describing the impact damage caused on node v_i at discrete time k due to contamination fault scenario $s\in S^*$, assuming zero initial conditions. The state-space equations describing these dynamics can be formulated, for $i\in\{1,...,N_v\}$ as

$$\xi_i(k+1;s)=\xi_i(k;s)+f_\xi(x_i(k),d_i(k)) \quad (3)$$

$$\psi(k;s)=f_\psi(\xi(k;s)), \quad (4)$$

where $\xi(k;s)\in\mathbb{R}^{N_v}$ is the nodal impact vector for $s\in S^*$, $f_\xi:\mathbb{R}\times\mathbb{R}\mapsto\mathbb{R}$ is the impact function which depends on the nodal water consumption $d_i(k)$ and contaminant concentration state $x_i(k)$ at node v_i, at discrete time k. In addition, $\psi(k;s)$ is a measure of the overall impact at discrete time k for the contamination fault scenario $s\in S^*$, and $f_\psi:\mathbb{R}^{N_v}\mapsto\mathbb{R}$ is the overall-impact function which depends on the nodal impact vector. For example, $\xi_i(k;s)$ may correspond to the number of people affected due to water consumption at the i-th node, and $\psi(k;s)$ to the total population affected.

A multi-objective problem is formulated for selecting a node from the node set V_m to conduct manual quality sampling. The solution algorithm is repeated recursively, to evaluate the contamination impact more accurately and to reduce the region of the possible source-area. In this work the problem considering the availability of one response team is formulated.

Let $y^j\in V_m$ correspond to the node where manual sampling should be conducted for the j-th sampling event after the fault has been detected, for $j\geq1$; let $\delta^j\in\{0,1\}$ be a contamination flag, such that $\delta^j=1$ if node y^j is contaminated, and $\delta^j=0$ otherwise. The problem is formulated as

$$y^j=f(y^{j-1},\delta^{j-1};S^{j-1}) \quad (5)$$

where y^0 is the node where contamination was first detected, $\delta^0=1$ and $S^0\equiv S^*$. The contamination fault scenario set $S^j\subseteq S^*$ is computed after each sampling event and is comprised of all the feasible faults which may have caused the contamination fault. Function $f:V_m\times\{0,1\}\mapsto V_m$ is used to compute the node where manual sampling should be conducted in the next sampling event. This function depends on the network topology and the state-space equations describing the advection and impact dynamics. Function $f(\cdot)$ may be considered as an optimization algorithm to find the best location for manual sampling, with respect to multiple objectives: a) the number of possible contamination fault scenarios which may have caused the detected fault and b) the contamination fault impact damage. These objectives may be conflicting, and it is possible that no single solution is optimal for both objectives; instead, a Pareto front of solutions may be computed, and from that, a single solution is selected.

3. SOLUTION METHODOLOGY

In the next paragraphs a methodology is presented for solving the optimization problem defined in (5), based on the classification technique of decision trees.

As standard practice, water utilities install hydraulic sensors at selected locations in the network to monitor the water demands or the pipe flows. Computing an approximation of the system hydraulics is a challenging research task, which may be solved by considering historical data of water demand and pipe flow. By assuming approximately periodic hydraulic dynamics, with a 24-hour period, it is possible to compute approximations of the outflow demand vector $d(k)$ and of the flow $q(k)$. From a practical viewpoint this is a reasonable assumption which can be verified by statistical analysis of the monitored flow dynamics.

The set S^* of all contamination functions is infinite and this is a significant difficulty in solving the isolation problem; it is desirable to compute a finite subset, taking into consideration the periodicity of the hydraulic dynamics, to reduce the contamination function space. Contamination fault scenarios with starting time-step within the first day of the simulation are considered. In this work, uncertainties in the system (e.g. due to consumer demands) do not affect significantly the nominal periodic dynamics considered. Let T_h correspond to one period of 24 hours, such that $T_h=\frac{24\ hr}{\Delta t}$, where Δt is the time step. The set $S\subset S^*$ of all single-period contamination fault scenario is given by

$$S=\{\phi(k-\tau_0)\in\Phi(k-\tau_0)\mid\tau_0\in\{0,...,T_h-1\}\}. \quad (6)$$

3.1 Decision Tree Algorithm

In the following paragraphs, an algorithm for function $f:V_m\times\{0,1\}\mapsto V_m$ described in (5) is introduced. The algorithm initiates when a fault alarm is activated and a quality fault is confirmed at the node $v_i\in V_s$. Expanded sampling nodes are computed iteratively through the function $y^j=f(y^{j-1},\delta^{j-1};S^{j-1})$, for $j\geq1$, where $y^j\in V_m$ is the sampled node at the j-th quality sampling event; $\delta^j\in\{0,1\}$ is the contamination flag and $S^j\subseteq S$ feasible fault scenario set computed for the j-th manual sampling event. As initial conditions, it is considered that $y^0\equiv v_i$, where v_i is the detected and confirmed contaminated node, $\delta^0=1$ and $S^0\equiv S$. The feasible fault scenario set $S^j\subseteq S$ for the j-th sampling event, given by $S^j=f_s(y^{j-1},\delta^{j-1};S^{j-1})$; $f_s(\cdot)$ corresponds to the *backtracking algorithm* (Shang et al., 2002), or other equivalent algorithms.

The normalized impact damage $W^j\in[0,1]^{|S^j|}$ measures the severity of each feasible contamination fault in S^j for the j-th sampling event, given by $W^j=f_w(S^j)$; $f_w(\cdot)$ is a function which calculates the impact increase between two

time instances, e.g. from the moment the fault occurred until the moment the fault was accommodated.

Linguistic labels are assigned, which describe the severity of a certain contamination fault, to assist in the decision-making process. Let Λ be the set of all impact labels considered; one such set with three labels may correspond to $\Lambda = \{$'High', 'Moderate', 'Low'$\}$. Let $f_l : [0,1]^{|S^j|} \mapsto \Lambda^{|S^j|}$ be the function which maps the normalized impact damage vector W^j, for the j-th sampling event, to the impact label vector $L^j = f_l(W^j)$. In the decision tree algorithm, a label is computed at each iteration to characterize the worst-case fault scenario which may have caused the detected contamination fault.

For the proposed algorithm, it is useful to construct a binary matrix M^j for the j-th sampling event describing the fault propagation. The binary matrix M^j is of size $|S^j| \times |V_m|$, and is computed with $M^j = f_m(S^j)$, where $f_m(\cdot)$ is the fault propagation function, defined as follows: for the i-th contamination fault scenario, simulate the contaminant propagation by using the mathematical model of the system; if the contaminant reaches the l-th node in the set V_m, some time after the contamination fault has been detected, then $M_{(i,l)} = 1$, otherwise if it does not reach the l-th node, $M_{(i,l)} = 0$.

A modified version of the classical decision tree splitting algorithm, described in (Alpaydin, 2004), is presented. In the following paragraphs it is demonstrated how the algorithm computes the information gain as well as the maximum impact, constructs a Pareto front of solutions and returns one node where manual sampling should be conducted in the next sampling event. Two objective metrics are considered: a) the information gain, and b) the maximum worst-case impact.

The information entropy metric is considered as a measure of the homogeneity of a set. Let $f_e(\cdot)$ be the function which computes the information entropy; $f_e(L^j)$ is the entropy metric corresponding to the label set L^j, at the j-th sampling event. If $f_e(L^j) = 0$, then all elements in L^j have the same label value; this terminates the algorithm and returns to the operator: a) the set S^j which corresponds to an area of possible source nodes and b) the common label in L^j as the impact evaluation. The information gain $Z_i \in \mathbb{R}^{|V_m|}$ describes the change in the entropy when sampling at a certain node for the i-th node. In the case where the maximum information gain is zero, i.e. $\max_i\{Z_i\} = 0$, then the solutions cannot be improved; this terminates the algorithm and returns to the operator: a) the set S^j which corresponds to an area of possible source nodes and b) the worst-case impact label in L^j as the impact evaluation.

It may be preferable to give a greater weight on high-impact contamination faults. In addition, it may be preferable to take into consideration True-Positive quality measurements (i.e. select a node to conduct manual sampling such that if it is contaminated, lower-impact faults are excluded, rather than exclude higher-impact faults if no contamination is detected). Let Ω_i be the maximum impact metric computed for the i-th node; this is given by

$$\Omega_i = \max\{W^j_\kappa \mid M_{(\kappa,i)} = 1, \kappa \in \{1, ..., |S^j|\}\}. \quad (7)$$

Both the vectors $Z \in \mathbb{R}^{|V_m|}$ and $\Omega \in \mathbb{R}^{|V_m|}$ are used in computing the node where manual sampling should be performed. Let $f_p : \mathbb{R}^{|V_m|} \times \mathbb{R}^{|V_m|} \mapsto V_m$ be the function which computes the next sampling node, such that $y^j = f_p([Z,\ \Omega]^\top)$. In general, the i-th node is searched, which corresponds to the maximum value in both Z and Ω. Sometimes a single optimal solution for the two objective metrics cannot be found; therefore a multi-objective algorithm is required to compute the Pareto front, i.e. the subset of node attributes which are non-dominant to each-other. From the set of Pareto solutions, a single solution is selected using some heuristic (e.g. the smallest Euclidean distance from the upper-right coordinate $[\max\{Z\},\ \max\{\Omega\}]^\top$).

The response team is send to examine whether the selected node is contaminated; if this is true, then $\delta^j = 1$, otherwise, $\delta^j = 0$. This process iterates by repeating the computation of (5), until the source-area is isolated and the fault impact evaluated.

4. SIMULATION EXAMPLE

In this example, the solution methodology is demonstrated on a small-scale real water distribution network. Figure 1 depicts one of the benchmark networks in the "Battle of the Water Sensor Networks" design competition (Ostfeld et al., 2008). This network is composed of 178 pipes connected to 129 nodes (126 junctions, two tanks and one reservoir). The structural characteristics (e.g. pipe lengths and diameters) are considered to be known. Each junction node is assigned with a daily average consumption volume as well as a discrete signal describing the rate of water consumption within 48 hours, with a 30-minute time step. These are assumed to describe the normal operation. The hydraulic dynamics are computed using the EPANET solver (Rossman, 2000); a daily period initiates at 8 *am* and terminates in 24 hours. It is assumed that the contamination substance does not react with other substances flowing in the water distribution network. Five nodes in the network are monitored using fixed on-line quality sensors, and are indicated in Fig. 1 as $V_s = \{$'17', '31', '45', '83', '122'$\}$; this is an optimal solution for the multiple objectives described in (Ostfeld et al., 2008; Krause et al., 2008). It is further considered that the water flows are approximately periodic and known, the time is discretized with $\Delta t = 5\ min$ and the daily period is $T_h = 288$. For the proposed sensor placement scheme, the contamination scenarios set S is constructed, which is comprised of 28 076 fault scenarios which initiate within the first one-day period. A constant contaminant injection rate is considered for each fault scenario.

Let $\Lambda = \{$'High','Moderate','Low'$\}$ be the set of all impact labels; for the i-th fault scenario, it is considered that if $0 \leq W^j_i < 0.33$, a 'Low' impact label is assigned; for $0.33 \leq W^j_i < 0.66$ and $0.66 \leq W^j_i \leq 1$, a 'Moderate' and a 'High' impact label are assigned respectively, for the j-th sampling event. As impact metric, the volume of contaminant mass consumed is considered. From the set S, 65% of the fault scenarios have 'Low' impact, 21.4% have 'Moderate' impact, and 13.6% have 'High' impact.

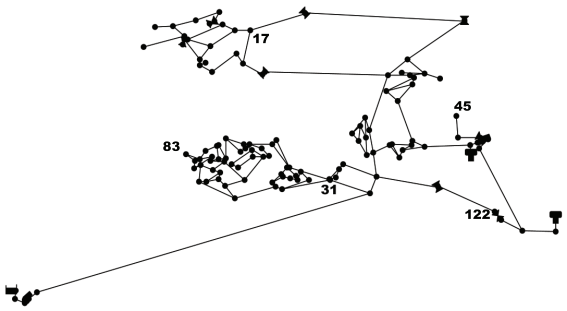

Fig. 1. The real water distribution network. The indices correspond to the locations of online quality sensors.

Table 1. Confusion Matrix for 10 000 Fault Scenarios

High	Moderate	Low	Classification
1372	36	154	as High
0	2117	275	as Moderate
0	0	6046	as Low

To evaluate the effectiveness of the proposed solution methodology, an experiment is conducted in which 10 000 fault scenarios are simulated, selected at random from the contamination fault scenario set S. For each fault scenario, the proposed algorithm is applied, to evaluate the impact damage and isolate the contamination source area.

The impact labels were correctly evaluated for 95.35% of all the fault scenarios considered, and incorrectly for 4.65%. Note that because the highest impact of all fault scenarios candidates is returned when there is no information gain, the algorithm essentially has a bias towards the worst-case impacts, as it is preferable to misclassify a lower-impact fault as a higher-impact fault, instead of the opposite. The Confusion Matrix in Table 1, summarizes the frequencies of misclassifications across each label. For example, when the actual fault impact was 'High', all instances were correctly classified as 'High', whereas when the actual fault impact was 'Low', 2.38% of these fault scenarios were misclassified as 'High' and 4.28% as 'Moderate'. This is in accordance to our solution methodology.

Detailed accuracy metrics for each label are presented in Table 2. The "True Positive (TP) Rate" (also referred to as "Recall") for the i-th label, is the number of fault scenarios which have been correctly classified, over the number of all fault scenarios which have been classified as the i-th label.The 'High' label has the maximum TP Rate, which is equal to "one". The "False Positive (FP) Rate", for the i-th label, is the number of fault scenarios misclassified, over the number of all fault scenarios which do not been classified as that label. There were no misclassifications as 'Low' labels, thus the lowest FP rate; in fact most of the misclassifications were towards the 'Moderate' and 'High' labels. The "Precision" metric for the i-th label is the number of fault scenarios which have been correctly classified in that label, over the number of all fault scenarios which have been classified in that label, correctly or not. Similar to the previous metric, the 'Low' label has the highest Precision. The final metric considered is the "F-measure", which is a statistical measure of the harmonic mean of Precision and Recall (TP Rate), describing the accuracy

Table 2. Detailed Accuracy by Impact Label

TP Rate	FP Rate	Precision	F-Measure	Label
1.000	0.022	0.878	0.935	High
0.983	0.035	0.885	0.932	Moderate
0.934	0.000	1.000	0.966	Low

of the classifier, with the best value approaching "one". On average, 1.84 samples were necessary for evaluating the impact of a fault using decision trees. For 80% of the fault scenarios considered, no more than two manual samplings where necessary in evaluating the impact and for 90%, no more than four manual samplings.

Finally, Fig. 2 depicts the histograms of the number of possible source nodes for each fault scenario, before and after the expanded sampling methodology is applied. For this example, the average number of possible source nodes is reduced almost in half, from 29 source nodes to 15, before and after the expanded sampling respectively. In specific, it is observed that in the first histogram, more than 50% of the fault scenarios have more than 30 source node candidates (out of 129). By applying the proposed methodology, the histogram distribution is skewed towards zero, indicating that the number of possible source for a large number of fault scenarios is reduced dramatically.

5. CONCLUSIONS

In this work, a computational approach is proposed for choosing a sequence of nodes in the distribution network to perform expanded sampling, so that the water contamination impact is evaluated and the source-area is isolated, with as few quality samples taken as possible. The proposed method is based on constructing a decision tree using multiple objectives. To illustrate the solution methodology, results are presented based on a simplified and a real water distribution system. To examine the effectiveness of the proposed algorithm, a large number of contamination faults were simulated and the corresponding decision trees were constructed to evaluate the impact-level and isolate the area of the contamination. The results are not directly comparable to other approaches presented in the literature; however, in future work it will be investigated how the proposed algorithm compares with baseline approaches, such as sampling at locations selected using expert knowledge.

REFERENCES

Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press, Cambridge, MA, USA.

Brdys, M.A. and Ulanicki, B. (1994). *Operational control of water systems: structures, algorithms, and applications*. Prentice Hall, New York, USA.

Cristo, C.D. and Leopardi, A. (2008). Pollution source identification of accidental contamination in water distribution networks. *ASCE Journal of Water Resources Planning and Management*, 134(2), 197–202.

De Sanctis, A.E., Shang, F., and Uber, J.G. (2010). Real-time identification of possible contamination sources using network backtracking methods. *ASCE Journal of Water Resources Planning and Management*, 136(4), 444–453.

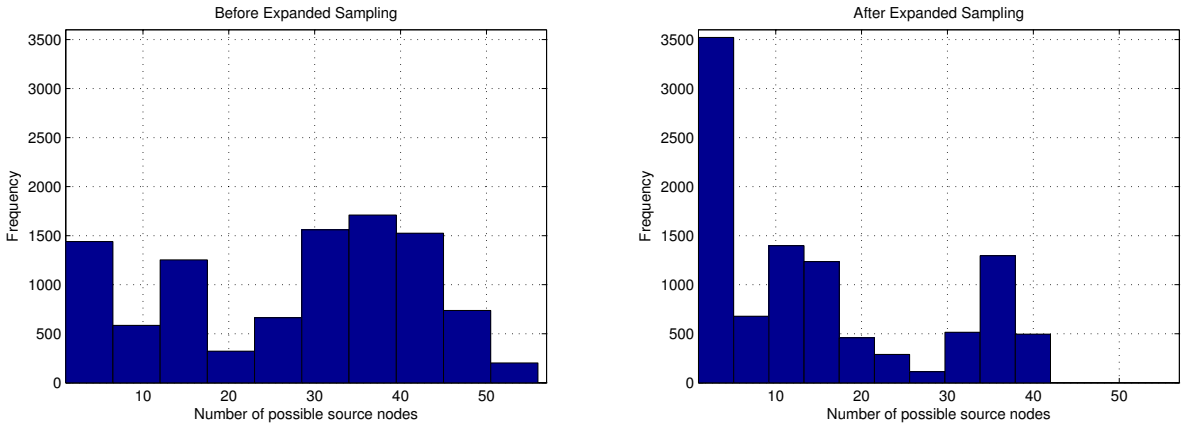

Fig. 2. Histograms of the number of source nodes possible for all fault scenarios, before and after expanded sampling.

Eliades, D. and Polycarpou, M. (2010). A fault diagnosis and security framework for water systems. *IEEE Transactions on Control Systems Technology*, 18(6), 1254 – 1265.

Guan, J., Aral, M.M., Maslia, M.L., and Grayman, W.M. (2006). Identification of contaminant sources in water distribution systems using simulation–optimization method: Case study. *ASCE Journal of Water Resources Planning and Management*, 132(4), 252–262.

Huang, J.J. and McBean, E.A. (2009). Data mining to identify contaminant event locations in water distribution systems. *ASCE Journal of Water Resources Planning and Management*, 135(6), 466–474.

Krause, A., Leskovec, J., Guestrin, C., VanBriesen, J., and Faloutsos, C. (2008). Efficient sensor placement optimization for securing large water distribution networks. *ASCE Journal of Water Resources Planning and Management*, 134(6), 516–526.

Kumar, J., Zechman, E.M., Brill, E.D., Mahinthakumar, G., Ranjithan, S., and Uber, J. (2007). Evaluation of non-uniqueness in contaminant source characterization based on sensors with event detection methods. In *Proc. ASCE World Environmental and Water Resources*, 513–520.

Laird, C.D., Biegler, L.T., and van Bloemen Waanders, B.G. (2006). Mixed-integer approach for obtaining unique solutions in source inversion of water networks. *ASCE Journal of Water Resources Planning and Management*, 132(4), 242–251.

Laird, C., Biegler, L., van Bloemen Waanders, B., and Bartlett, R. (2005). Contamination source determination for water networks. *ASCE Journal of Water Resources Planning and Management*, 131(2), 125–134.

Liu, L. (2009). *Real-time Contaminant Source Characterization in Water Distribution Systems*. Ph.D. thesis, North Carolina State University, Raleigh, North Carolina.

Ostfeld, A. and Salomons, E. (2005). Solving the inverse problem of deliberate contaminants intrusions into water distribution systems. In *Proc. ASCE World Water and Environmental Resources*, 12.

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., Khu, 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., and Walski, T. (2008). The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. *ASCE Journal of Water Resources Planning and Management*, 134(6), 556–568.

Preis, A. and Ostfeld, A. (2007). A contamination source identification model for water distribution system security. *Engineering Optimization*, 38(8), 941 – 947.

Rossman, L.A. (2000). *EPANET 2 Users manual*. National Risk Management Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, OH.

Shang, F., Uber, J., and Polycarpou, M. (2002). Particle backtracking algorithm for water distribution system analysis. *ASCE Journal of Environmental Engineering*, 128(5), 441–450.

Uber, J.G. (2005). Identifiability of contaminant source characteristics in steady-state and time-varying network flows. In *Proc. ASCE World Water and Environmental Resources*, 25–28.

U.S. Environmental Protection Agency (2008a). Water security initiative: Interim guidance on developing an operational strategy for contamination warning systems.

U.S. Environmental Protection Agency (2008b). Water security initiative: Interim guidance on developing consequence management plans for drinking water utilities.

Vankayala, P., Sankarasubramanian, A., Ranjithan, S.R., and Mahinthakumar, G. (2009). Contaminant source identification in water distribution networks under conditions of demand uncertainty. *Environmental Forensics*, 10(3), 253 – 263.

Zechman, E.M. and Ranjithan, S.R. (2009). Evolutionary computation-based methods for characterizing contaminant sources in a water distribution system. *ASCE Journal of Water Resources Planning and Management*, 135(5), 334–343.