Extreme Impact Contamination Events Sampling for Real-Sized Water Distribution Systems

Lina Perelman¹ and Avi Ostfeld, F.ASCE²

Abstract: Contamination warning systems are being designed to protect water distribution systems against deliberate contamination intrusions. To design a contamination warning system, contamination intrusion events need to be selected. Because contamination intrusions are random, even for a medium-size network the theoretical number of possible injection events is huge, and thus the number of contamination events which can be considered in the design process is limited. To effectively cope with the threat of contamination events there is a need to identify those critical instances. A straightforward approach of enumerating all possible contamination intrusions from which critical events can be selected is limited to small systems. As critical events are rare the probability of revealing them using common Monte Carlo randomized simulations is very small or requires an extensive impractical computational amount of trials. In this study a methodology utilizing importance sampling and cross entropy based on a recent published work of the authors is further tested on real-sized water distribution systems of increasing complexity. The results demonstrate the robustness of the methodology in terms of improved run times, suggesting computational feasibility for problems in which size prevents full enumeration or application of direct Monte Carlo simulation techniques. DOI: 10.1061/(ASCE)WR.1943-5452.0000206. © 2012 American Society of Civil Engineers.

CE Database subject headings: Water distribution systems; Sampling; Water pollution; Security.

Author keywords: Water distribution system; Water sampling; Rare events; Contamination; Security; Real-sized systems.

Introduction

Designing a sensor network capable of detecting all possible contamination events in a water supply system is unrealistic given limited budgets and technologies. It is, however, possible to design a sensor network that optimizes desired objectives (e.g., minimize expected volume of contaminated water delivered, or population infected prior to detection). In recent years a number of methods have been developed, formulating the optimal sensor placement problem as a single- or multi-objective problem using combinatorial, data-driven, and heuristic solution techniques (e.g., Kessler et al. 1998; Ostfeld and Salomons 2004; Berry et al. 2006; Krause et al. 2008; Preis and Ostfeld 2008; Ostfeld et al. 2008; Xu et al. 2008).

One of the primary obstacles facing optimal sensor placement modeling is the significant computational effort imposed by the hydraulic and water quality simulations required for modeling transport of a constituent in a distribution system. Some of the preceding studies addressed the importance of the contamination sampling stage and its influence on sensor network design (e.g., Berry et al. 2006; Krause et al. 2008; Ostfeld et al. 2008; Preis and Ostfeld 2008). Complete enumeration of all contamination events is impractical because the number of possible intrusion events and simulation run times grow substantially with system size. A heuristic scheme was previously suggested by Perelman and Ostfeld (2010) for establishing a representative sampling of critical contamination events which reduces the problem’s complexity for simulation-based techniques. The method proposes an importance sampling (IS) (Rubinstein 1997) approach combined with a cross entropy (CE) (Rubinstein and Kroese 2004) optimization algorithm for rare events simulations.

In this study the ability of the proposed approach suggested by Perelman and Ostfeld (2010) on real-sized water distribution systems and provides recommendations for selection of CE parameters.

Methodology

The basic idea of IS is to gain simulation speed-up by causing the occurrence of rare events to become more frequent through conducting simulations under a modified probability distribution. The difficulty in utilizing the IS approach lies in selecting the modified probability.

An approach for estimating the IS reference parameters was introduced by Rubinstein (1997) through minimizing the Kullback-Leibler distance measure (Kullback and Leibler 1951) or the CE between two probability distributions. The CE algorithm is a two-stage iterative procedure involving generating random samples and updating sampling probability. Starting with some initial probability density, the IS probability is updated each iteration such that the event of interest is rendered less rare. The primary CE algorithm employs the following stages:

1. Choose an initial probability \( \hat{\rho}_0 \). Set an iteration counter.
2. Generate a random sample \( X_1, X_2, \ldots, X_N \) from the density \( \hat{\rho}_{t-1} \) and evaluate each sample using some measure function (e.g., contamination level \( CL(X_i) \)). Sort their performances in descending order and select \( \rho \) percentage of samples with the
Applications

The performance of the proposed CE method for sampling rare contamination events is tested on three water distribution systems of increasing complexity: (1) Network 1 of the Battle of the Water Sensor Networks (BWSN1) (Ostfeld et al. 2008); (2) Richmond water distribution system, Yorkshire, UK (Centre 2001); and (3) Network 2 of the Battle of the Water Sensor Networks (BWSN2) (Ostfeld et al. 2008).

The complete data and EPANET (EPANET 2.0) input files of all three networks are available at the Centre for Water Systems (CWS) Benchmarks (2001). The following assumptions have been made for the applications: (1) the impact of a deliberate or accidental contamination event is denoted as the contamination level (CL) and is quantified as the expected population infected prior to detection as defined in Ostfeld et al. (2008); (2) random variables—location and starting time; (3) event parameters—single injection location with mass injection rate of 100 (g/min); expected detection time of 8 h; and a minimum hazardous concentration level of 5 (mg/L); and (4) an evaluation time step of 1 min for both the hydraulic and water quality extended period simulations.

The BWSN1 example was also explored in Perelman and Ostfeld (2010). The BWSN1 example in this research is utilized for a different contamination level criterion, and for comparing to the other two example applications of Richmond and BWSN2.

Example 1—Battle of the Water Sensor Networks 1

BWSN1 is example 1 of the Battle of the Water Sensor Networks (Ostfeld et al. 2008). The system consists of 126 nodes, one constant head source, two tanks, 168 pipes, two pumps, and eight valves (Fig. 1). It is subjected to a varying demand pattern of 48 h. It is assumed that an intrusion lasting 5 min could occur at any of the 126 nodes every 5 min during 48 h of the simulation. The total number of decision variables (denoted ndv) is thus 126 × (60/5) × 48 = 72,576, and the total number of possible contamination events is 126 × (60/5) × 48 = 72,576.

All 72,576 contamination events were generated revealing high diversity among the events with an average of 99 people infected and a standard deviation of 125. The top ten most extreme events (0.01% of the entire solution space) were declared as critical contamination events having more than 630 infected people (CL > 630) (Table 1).

The CE method was executed multiple times to estimate the IS probability of the extreme contamination events with a sample size of N = 5 × ndv = 3,510, ρ = 0.02, and α = 0.7. The CE algorithm converged each time after three iterations revealing on average five extreme solutions and the estimated IS probability, and requiring 15% of the computational effort of the full enumeration. The ten most critical events are composed of different combinations of four locations (Fig. 1) and four timings of contaminant injections. The obtained results revealed similar critical injection locations to the locations reported in Perelman and Ostfeld (2010) for a different contamination level criterion, all in the vicinity of the source.

Without any prior knowledge of the events, a uniform distribution assigns 10 × 1/(126 × 576) = 0.000138 probability to sample the extreme events, whereas the estimated IS probability assigns much more weight to the extreme contamination events (on average 0.6117). Increasing the sample size to N = 6 × ndv = 4,230 and setting ρ = 0.015 (α = 0.7) revealed seven events on average with 17% computational effort of a full enumeration. Table 1 summarizes the network data, the impact of contamination events on

![Fig. 1. Results for example 1—Battle of the Water Sensor Networks 1](image-url)
### Example 2—Richmond Water Distribution System

This example incorporates the model application to the Richmond water distribution system, Yorkshire, UK. The system consists of one source, 949 pipes, 865 nodes, seven pumping stations, and six tanks (Fig. 2). It is subjected to a demand cycle of 24 h. An intrusion lasting 5 min could occur at any of the 865 nodes every 10 min during the 24 h of simulation. The total number of random variables ($ndv$) is thus $865 \times (60/10) = 1,009$, and the total number of possible contamination events is $865 \times (60/10) = 126,560$. Enumerating all contamination events revealed 100 events causing more than 33 people to become infected (CL $> 33$). The average CL, and CL standard deviation for Richmond, are three and five people, respectively (Table 1). This is in light of the total population estimated as 556 people for Richmond, 11,130 for BWSN1, and 356,105 for BWSN2, following the CL matrix in Ostfeld et al. (2008).

Next, the CE algorithm was executed converging each time after three iterations and revealing the estimated IS probability. The results demonstrate that 90–96% of the critical solutions were sampled using CE requiring only 12–15% of the entire computational effort. Comparing the performance of IS versus Monte Carlo sampling, for the same simulation effort only approximately 10% of the critical events were sampled using Monte Carlo simulations, versus more than 90% using IS. The critical events consist of combinations of six locations (nodes number 2, 3, 4, 298, 318, and 774) located at the vicinity of Tank A (Fig. 2), and 32 timings of contaminant injections. Without any prior knowledge of the events, a uniform distribution assigns 0.0008 probability weights to sample the extreme events, whereas the estimated IS probability assigns 0.707 probabilities on average to the extreme contamination events.

Table 1 summarizes the network data, the impact of contamination on the system, the CE parameters, and the results of this application. Comparing the performance of IS versus Monte Carlo (MC) sampling revealed that for the same simulation effort at most one critical event was sampled using MC.

### Example 3—Battle of the Water Sensor Networks 2

BWSN2 is example 2 of the Battle of the Water Sensor Networks (Ostfeld et al. 2008). The system consists of 12,523 nodes, 14,822 pipes, two constant head sources, two tanks, four pumps, five valves, and is subject to five variable demand patterns (Fig. 3). The system is simulated for a total extended period duration of 24 h. A pollutant can be injected at any one of the 12,523 nodes of the system every hour, with an injection duration of 1 h as well. The total number of random variables is thus $12,523 + 24 = 12,547$, and the total number of possible contamination events is $12,523 \times 24 = 300,552$.

The primary challenge related to this example application is the computational effort of each simulation which lasts approximately 3–3.5 min on a 2 Core 1.96 GB 778 MHz personal computer. Hence, to simulate the entire solution space, $300,552 \times 3(\text{min}) = 626(\text{days}) \approx 1.72(\text{years})$ are required. To test the potential of the CE method for real-sized water distribution systems it was, again, desired to perform a full enumeration. To cope with computational limitations, the code was parallelized and executed on NANCO (128 dual-core processor 2.2 GHz Advanced Micro Devices [AMD] opteron Linux cluster) located at the Technion computer center using up to 32 nodes simultaneously. The simulation took approximately one week to complete.
Fig. 2. Results for example 2—Richmond water distribution system

Fig. 3. Results for example 3—Battle of the Water Sensor Networks 2
A contamination event effecting more than 2,000 people was declared hazardous, 763 such events were identified concentrating in 90 nodes of the system located on the main supply line (Fig. 3). Table 1 summarizes the data and the results for this application. The results demonstrate that 78–87% of the critical solutions were sampled using CE requiring 13–25% of the entire computational effort relative to 4–8% critical solutions sampled using ordinary MC simulations. The estimated IS assigned 0.333 probabilities on average to the extreme contamination events versus 0.0025 assigned by a direct MC simulation.

Conclusions

The proposed methodology was demonstrated using three water distribution systems of increasing complexity. The impact of a contamination event on a water distribution system was quantified using the expected population infected prior to detection. Table 1 summarizes the data of the three networks and the CE parameters used. The impact of contamination events was explored for all networks, demonstrating diverse affects on the population. For all systems, the majority of extreme contamination events, having a small probability to occur but severe impact on the system (five to six times the average event), was identified and the IS probability was estimated. The resulting IS probability substantially increased the likelihood of rare contamination events, allowing them to be efficiently sampled.

Although CE is a heuristic algorithm requiring parameter tuning, the application results can provide some guidelines for their selection. The results demonstrate that the required sample size is in the range of 4–5% of the solution space, the elite sample size is up to 8% of the sample size, and the smoothing parameter was constant for all applications with $\alpha = 0.7$.

The results of the CE algorithm present its robustness in terms of improvement in running times, which is especially appealing for large complex water distribution systems, requiring a relatively small computational effort and capable of sampling the majority of the extreme contamination events. The proposed approach can offer computational feasibility in problems for which size prevents full enumeration or application of direct MC simulation techniques.

Efficiently revealing critical nodes location is highly important for further enhancing the physical security of a water distribution system through strengthening guarding, fencing, and surveillance equipment at those places.

Acknowledgment

This work was supported by the Technion Funds for Security research. The valuable parallel coding work of Ariel Tubaltzev for running the BWSN2 system on the NANCO machine is highly acknowledged.

References


EPANET 2.0 [Computer software]. US EPA, Washington, DC.


