Water Quality Sensor Placement in Water Networks with Budget Constraints

Jonathan W. Berry* William E. Hart* Cynthia A. Phillips*
James G. Uber† Thomas M. Walski‡

Abstract

In recent years, several integer programming models have been proposed to place sensors in municipal water networks in order to detect intentional or accidental contamination. Although these initial models assumed that it is equally costly to place a sensor at any place in the network, there clearly are practical cost constraints that would impact a sensor placement decision. Such constraints include not only labor costs but also the general accessibility of a sensor placement location. In this paper, we extend our integer program to explicitly model the cost of sensor placement. We partition network locations into groups of varying placement cost, and we consider the public health impacts of contamination events under varying budget constraints. Thus our models permit cost/benefit analyses for differing sensor placement designs. As a control for our optimization experiments, we compare the set of sensor locations selected by the optimimzation models to a set of manually-selected sensor locations.

1 Introduction

In this paper we improve, extend, and experiment with the integer-programming model introduced by Berry et. al. [1]. Those authors used a discrete event simulation to track "balls" of contaminant flowing through a municipal water network, recording consumption of contaminant with each visit of a ball to a demand source. The integer program then selected locations for k sensors to minimize the expected consumption of contaminant across the set of simulated scenarios.

We improve this basic model of [1] in several ways:

*Algorithms and Discrete Math Dept, Sandia National Laboratories, Albuquerque, NM; PH (505)284-4021,(505)844-2217,(505)845-7296 {jberry,wehart,caphill}@sandia.gov. Sandia is a multipurpose laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

†USEPA, Cincinnati, OH and Dept. of Civil Engineering, University of Cincinnati, Cincinnati, OH; PH (513)569-7974 uber.jim@epa.gov.

‡Bentley Systems, Incorporated – Haestad Solutions Center, Waterbury, CT; PH (570) 735-1368 tom.walski@bentley.com

- We replace the discrete event simulation with a direct EPANET simulation of each attack scenario.
- We associate placement costs with each potential sensor location.
- We more accurately model an attack. In particular, we remove the unreasonable assumption made in [1] that contaminant completely replaces outflowing water at the attack site. We now model a reasonable EPANET "mass" injection over a given period of time.
- We added an evaluation phase to ensure that post-solve simulation using EPANET reproduces the predictions made by the model.

We explain each of these improvements in detail in Section 4.

2 Background

Following Berry et al. [1], we model a municipal water network as a graph $G = (V, E)$, where the vertex set V is a set of junctions, tanks, or locations of water consumption and the edge set E is a set of connections (pipes, pumps, and valves). Networks are typically skeletonized, so each node may represent an entire neighborhood or set of facilities.

We assume an attacker contaminates the network at precisely one point through a single continuous, constant-rate injection of a given duration, beginning at a given time. An attack scenario is thus an 4-tuple (l, t, r, d), where l is a location, t is a time, measured from the beginning of the simulation, r is an injection rate in contaminant units per minute and d is the attack duration in minutes. For this paper we always set the injection rate $r = 100$ contaminant units per minute and attack duration $d = 450$ minutes. Our model can incorporate a risk profile, but we report results only for the case in which all possible attack scenarios are equally likely.

We wish to place a limited number of sensors in a water network to most effectively protect the city's population. In this paper, we limit the potential sensor locations to the set of nodes V. We assume that the sensors are perfect and that they raise a general alarm precisely when passed by contamination of sufficient magnitude. The general alarm halts any consumption of contaminant, so we assume the health-risk damage stops at the moment of detection.

Constructive criticism of Berry et al. [1] from the water community suggests that one of the most important shortfalls of this model is that it does not account for sensor placement costs and budgetary constraints. A human expert, using intimate knowledge of the utility, can eliminate all sites from consideration that are too expensive or too difficult to use for whatever reason. If there are few remaining sites, does this make the problem so easy that that a local expert can effectively solve the sensor placement problem by inspection? In this paper we explore this question by comparing placements generated by our model with common-sense sensor placements by a human. The goal of our study is not to declare a winning strategy, or to discredit either modelling or the application of human expertise.

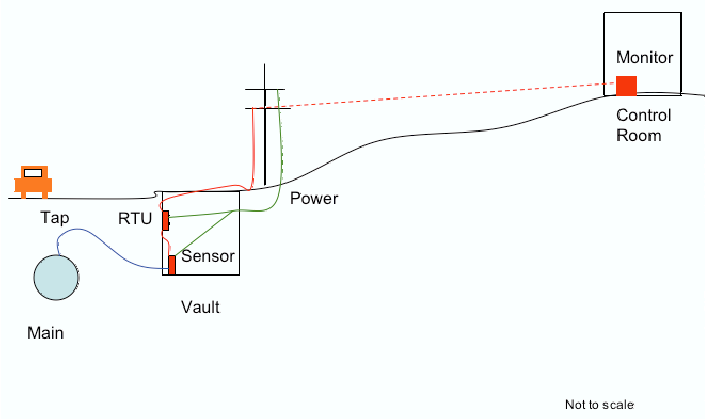


Figure 1: The components of a typical water sensor.

Rather, our results suggest that a collaboration between modellers and those with practical water system expertise can improve the effectiveness of sensor placement decisions.

The remainder of the paper is organized as follows. In Section 3, we discuss the determining factors of sensor placement costs. Section 4 details the water transportation model and the EPANET simulation phase that generates model data. In Sections 5, 6 and 7 we describe the real-world data set we use, our experimental design, and our results, respectively.

3 Sensor Placement Costs

The cost of placing water quality sensors can vary widely depending on the features of the site. This involves much more than simply the cost of the sensor. The lowest-cost locations are those that already have SCADA (Supervisory Control and Data Acquisition) equipment and an exposed water line. The highest-cost sites require purchase of land, installation of a building or vault, installation of an RTU (Remote Telemetry Unit), bringing power to the site, installing communication equipment and upgrading/installing equipment at a central control room. See Figure 1 for a schematic of a typical sensor location.

Our new sensor placement model accounts for these site-specific costs. Planners should estimate these costs only to the fidelity necessary for planning and only to a level consistent with the other model data. Given the other assumptions in the model (such as assuming water transport is driven by average demands over a small number of repeating patterns), we estimated the costs detailed below only to within an order of magnitude. Real costs for specific projects will vary widely.

Land. Ideally, the utility can place a sensor in a building, vault or other facility it (or the municipality) owns, so there will be no cost to purchase land or easement. There is no land cost for placing a vault in a public right-of-way, but structure costs will be higher for construction and repaving. If the land is leased then one should use the present worth of the lease for this cost.

Structure. Sensors must be housed. Ideally, the utility already has a building (e.g. pump station) or underground vault (e.g. valve vault) in a good location. If

not, the utility must construct a building or some other above-ground structure. In congested areas, the utility must install a buried vault to house the equipment. The architectural design depends on the neighborhood and may be little more than a pre-engineered enclosure the size of a dog house.

Sensors. The cost of the sensor depends on the water quality constituent being detected. Prices vary widely and are changing as new technologies become available. The cost of sensors should be independent of the location.

Link to Communications. Most sensors produce some type of analog signal (usually 4-20 mA) which must be converted to a digital signal and then put in the appropriate format for the SCADA system. This is usually accomplished in an RTU. An existing pump station, tank or valve vault may already have an RTU. Adding a signal to an existing RTU is much less expensive than purchasing and installing a new unit. A utility can place an RTU on a utility pole outside an underground vault to avoid flooding and minimize the need for confined space entry.

Power. While some units may run off battery or solar power, most remote sensors require standard electrical power. In most areas the cost for a power hookup is nominal (if any). However, in some remote locations there may be a substantial cost to run power. Some buried vaults in urban settings may require excavation and repaving to supply power. Some situations may require battery backup.

Communications. A means must be established to deliver a signal from the RTU to the main control room. The cost is usually nominal for a dial-up phone connection. Cellular phones are only slightly more expensive. A dedicated phone line has a higher operating cost. Where phone service is not available, the utility may need radio (or in some cases satellite) communication. This has a higher initial cost for the transmitter and antenna.

Control Room. If the utility has an existing SCADA system and wants to tie the sensors into it, the cost is nominal. However, if a large number of new signals are being brought into a system, there may be a cost to upgrade the system. Setting up an independent SCADA system for water quality signals incurs a significant cost for components in the control room.

Operation and Maintenance. Utilities should consider long-term costs after sensor installation. Most operation and maintenance costs are independent of location and therefore could be excluded in an analysis of location. These costs include power cost, phone (or other communication) costs and visits from technicians to check and calibrate the sensors. One exception is facilities defined as confined spaces. These facilities require at least two technicians per visit and these technicians may require special confined space entry equipment.

Application. Most nodes in a water distribution system are expensive locations for placing sensors because they require a structure and RTU in addition to the sensor itself. Thus we expect that most nodes in a sensor placement formulation will have a "typical node" cost. Sensor placement locations that coincide with existing facilities will have lower costs. Those that have accessibility or ownership difficulties will have higher costs.

4 The Dynamic Model With Placement Costs

The sensor placement model presented in Berry et al. [1] consists of two phases: a simulation phase and an integer programming (IP) phase. We call this the *dynamic model* to distinguish it from earlier *static* IP models that did not explicitly represent time. IP is a well-understood optimization technology that is useful for making concrete decisions such as whether to place a sensor at a given location or not.

4.1 The Simulation Phase

In the first phase of the dynamic model, we simulate a set of *attack scenarios* one by one using a custom application based on the EPANET toolkit [3]. An attack scenario is an ordered pair consisting of a vertex and an attack time: (v, t), where v is a vertex (a junction, tank, well, or reservoir), and t is the number of minutes since the beginning of the simulation (midnight in our experiments). At the starting time for attack a, we assume an attacker initiates an injection of a constant rate a_r of mass units per minute. They maintain the injection for a duration a_d. During the simulation, we record each first visit of a nonzero concentration of contaminant to a vertex. We also record, at that instant, the total consumption of contaminant, in mass units, since the beginning of the simulation. The result of one simulation is thus a sequence of ordered pairs $\{(v_1, m_1), (v_2, m_2) \ldots (v_k, m_k), (l_d, m_{\text{tot}})\}$, where m_i is the total consumption of contaminant at the instant v_i first experiences contaminant. Each sequence ends with an exposure at the *dummy location*, paired with the total consumption of contamination through the entire simulation horizon.

4.2 The Integer Program

The variant of the dynamic model described below places sensors to minimize the expected consumption of contaminant over all simulated attack scenarios. This is just one of many plausible models. One could consider other objectives, alter the model to explicitly address data uncertainties, or add other constraints to more accurately model a real water system.

In our IP model, every attack a has a *witness*. This is the first (placed) sensor hit by contamination from attack a. Thus, the witness signals the alarm. Every attack has a witness, with the costless dummy sensor "witnessing" all attacks that aren't detected by any sensor in the placement.

The input data to our dynamic IP model are as follows:

- $G = (V, E)$, the network. $V = v_1, \ldots v_n$ and E $= e_1, \ldots e_m$.
- α_{it}, the probability of an attack at node v_i at time t.
  Assuming exactly one location is attacked sometime during the day, we have $\sum_{(i,t)\in\mathcal{A}} \alpha_{it} = 1$, where $\mathcal{A} \subseteq V \times \tau$ is the set of attacks and τ is the set of possible attack times.

- B, the sensor placement budget in dollars.

- T, the time horizon of the simulation.

- $L \subseteq V$, the set of possible sensor locations.

- C_l, where $l \in V$, the cost of placing a sensor at location v.

- $L_a \subseteq V$, the set of network locations (nodes) contaminated by attack a.

- w_{aj}, for $a = 1, \ldots, |\mathcal{A}|$ and $j \in L \cap L_a \cup \{q\}$, where q is a dummy location; weights from the EPANET simulation output equal to the amount of contaminant consumed if a sensor at location j witnesses attack a.
  We compute the w_{aj} as follows. For each $v \in L_a$, let t_v be the time node v is contaminated by attack a. Let $d_v(t_1, k)$ be the total demand at node v from time of day t_1 through the next k time units ($d_v(t_1, k) = 0$ if $k \leq 0$). Recall that k can be longer than a day. The weight w_{aj} is the amount of contaminant consumed by the network before detection if a sensor at location j raises the alarm. That is, all potential sensor locations contaminated before t_j are not given a sensor. Thus we have $w_{aj} = \sum_{v \in L_a} d_v(t_v, t_j - t_v)$ for all $j \in L \cap L_a$. w_{aq} is the amount consumed if no sensor raises an alarm: $w_{aq} = \sum_{v \in L_a} d_v(t_v, T - t_v)$.

The integer program (IP) uses the following variables:

- decision variable s_i for each potential sensor location $i \in L$. This variable is 1 if we place a sensor at location i and is 0 otherwise.

- derived variables b_{ai} for $a \in \mathcal{A}$ and $i \in L \cap L_a \cup \{q\}$, where q is a dummy location. Variable b_{ai} is 1 if location i witnesses attack a. These variables need not have formal integrality constraints. They will always be binary in any optimal solution provided the s_i variables are binary. Omitting unnecessary integrality constraints can improve the practical performance of IP solvers.

For ease of presentation, we assume that $L = L_a$. Let $\mathcal{L} = L \cup \{q\}$, the set of useful sensor locations for attack a plus the dummy location. The IP to minimize consumption is:

$$\text{(MC)} \quad \text{minimize} \quad \sum_{a \in \mathcal{A}} \sum_{i \in \mathcal{L}} \alpha_a w_{ai} b_{ai}$$

$$\text{where} \quad \begin{cases} \sum_{i \in \mathcal{L}} b_{ai} = 1 & \forall a \in \mathcal{A} \\ b_{ai} \leq s_i & \forall a \in \mathcal{A}, i \in \mathcal{L} \\ \sum_{i \in \mathcal{L}} s_i C_i \leq B \end{cases}$$

The first type of constraint assures that there is exactly one witness for each attack scenario. The second set enforces that a sensor cannot witness any attack if it is never installed. The objective-function pressure then assures that the first eligible sensor in the list for attack a is chosen as witness (barring zero-demand nodes or very small probability attacks). The last constraint enforces the sensor placement budget. The objective minimizes the total consumption over all attacks (weighted by risk).

4.3 The Evaluation Phase

Given a set of sensor placements, we compute the expected health effects using the EPANET toolkit attack simulation code. We simulate all attacks, and for each attack the code records the first visit of contaminant to a sensor location and the total consumption at that instant. We could easily modify the code to continue the attack after first detection in order to simulate delays in response to any attack. However, the data we present were generated under the assumption that the attack is over at the instant of detection. For the IP solution, this value exactly matches the IP objective as long as the attack scenario set is precisely the set represented in the IP constraints.

5 Data

For our experimental study, we consider the SNL-5 data, which has approximately 3600 nodes, 3800 pipes and 90 pump stations. This is a skeletonized model of a large U.S. city, and many water sources feed this system. The EPANET model of this network had calibration problems relating to pump curves and well output, but the study proceeded anyway, as a real-life, time-constrained sensor placement effort would. We were able to simulate the network model as it was calibrated for 24 hours, but the system became unbalanced after roughly 40 hours (most likely because we did not have complete information on the manner in which sources were controlled).

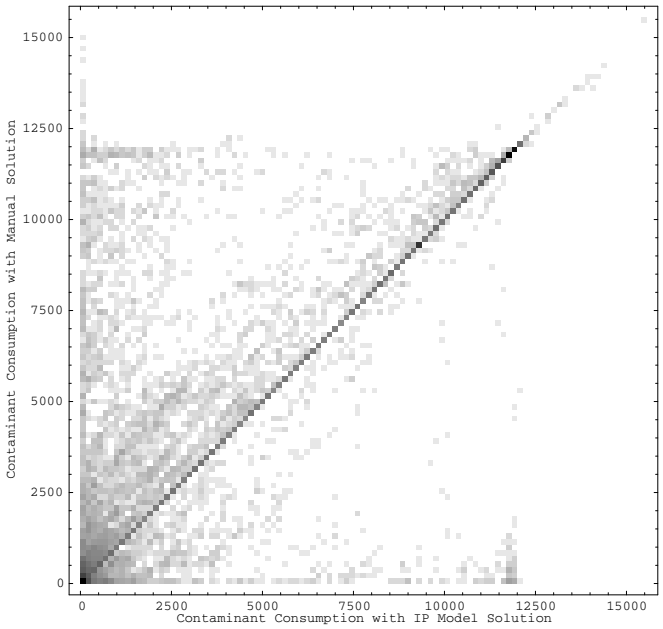
The water utility provided Sandia National Laboratories with the network model and a list of SCADA-capable network locations.[1] Using the cost guidelines of Section 3, we set the sensor placement costs at various classes of network locations as follows:

Pump stations with SCADA	\$20,000
Pump stations without SCADA	\$30,000
Other locations with SCADA	\$30,000
Other locations	\$70,000

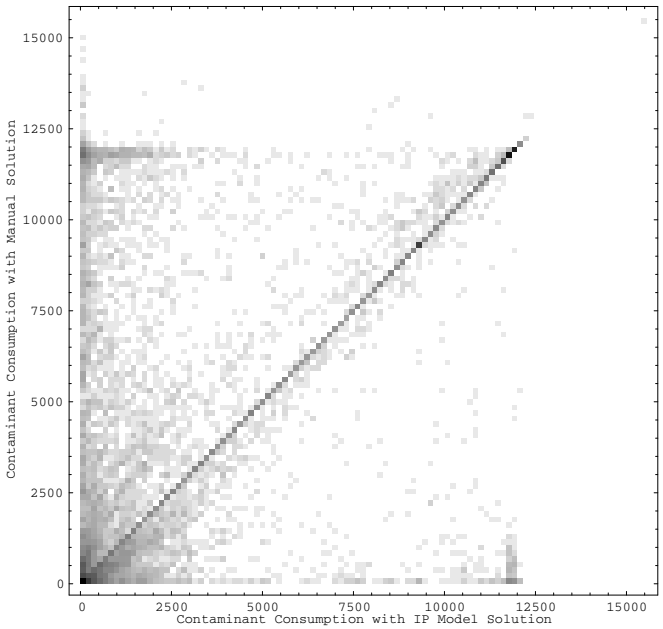
6 Methods

We chose a simple experimental design due to limited resources and time. The 5th author manually generated sensor placements to serve as a control for our

[1]The utility considers this data sensitive. Even non-Sandia co-authors of this paper received only an anonymized version of the data to protect the identity of the utility.



(a)



(b)

Figure 2: Comparative results for budgets of (a) \$250,000 and (b) \$1,000,000. Darker regions indicate greater densities of data points. Data points above the $y = x$ line indicate better values for the IP model sensor placements, while data poitns below the line indicate better values for the manual sensor placements.

optimization experiments. He had no experience with this particular utility, so he could not able to employ the local expertise that a utility manager would have. Furthermore, he had limited time to make his decisions. However, he intentionally employed the anticipated strategy of many water utility managers: limiting the placement choices to those cheapest, most convenient ones that would yield a common-sense coverage of the network.

We considered sensor placements budgets of $250,000 and $1,000,000. For each attack scenario in the simulation phase of the dynamic model, we ran a 24-hour EPANET water quality simulation to gather our data. The set of attack scenarios was the cross product of the vertex set V and four attack times evenly spaced throughout the 24 hours. Each attack had a contaminant injection strength of $a_r = 100$ mass units per minute, where the units could be, for example, grams. The duration of each attack, a_d was 450 minutes, a plausible value for an attacker's competing desires to do great damage, yet leave the scene before detection.

We used the AMPL modeling language [2] to formulate the integer program (MC). In all cases, we solved this IP using AMPL 9.0.2, which applied the CPLEX 9.0.2 integer programming solver. These tools ran on a dual-processor 64-bit Linux workstation with 20Gb of RAM. The combined cost of hardware and software used to obtain our results was roughly $50,000. As we will see, this is a cost that may be justified in a real effort to place monitoring stations in a moderate-sized city. CPLEX has an option to relax the requirement that an optimal solution be found. We employed this option to save time; our results are guaranteed to be within 1% of optimal. With this option, each integer programming solve for this paper ran for a matter of hours, not days, on the 64-bit workstation. The simulation phase required similar amounts of time.

Our evaluation runs verified the objective of the IP and computed the equivalent objective for the manual placement. We could use the same system to test the effectiveness of the IP sensor placement for attacks that are "close" to the ones explicitly modeled. For example, we could double the number of attack times. We could then compare the performance of the sensor placement with lower-fidelity information to the objective value of a larger IP that considers all the attacks. We are performing these tests now and will report the results when the tests are complete.

7 Results

Table 1 shows the evaluation of the sensor placement solutions generated by hand and with the IP model, for budgets of $250,000 and $1,000,000. These results show the expected contaminant consumptions, in mass units, for the evaluation runs. Figure 2 shows more details of the evaluation runs; each data point (cross) in these figures represents a comparison of the IP vs. the manual model for a single attack (a single location, time pair). Thus for crosses above the diagonal, the IP solution's value is better, and for those below the diagonal, the human did better.

For each budget, the IP model's solution selected fewer sensors than our human expert, targeting some expensive locations for sensor placement. In particular, for the $250,000 budget the IP model selected 7 sensors, compared to 10 for our

Expected Mass Units Consumed		
Budget	IP Model	Human
$250,000	5053	5698
$1,000,000	3919	4883

Table 1: Results of the evaluation phase: expected consumptions

human expert. For the $1,000,000 budget, the IP model selected 24, while the human expert selected 38.

As stated Section 6, the startup hardware and software costs for employing the dynamic model are roughly $50,000. We reconsidered the performance of the IP model to directly account for these costs. The logic here is that the use of IP solver technologies may require a water utility to invest in an AMPL/CPLEX solver licenses, as well as a computation workstation with which to do these analysis. But is this investment worthwhile compared to simply doing a manual sensor placement. For this analys, we ignoring systems administration and maintenance costs for this hardware, but simply subtract these investment costs from the initial budget use for IP-based sensor placement. We re-solve the problems and present the results in Table 2. The expected consumption with the slightly smaller budget is not affected significantly, and maintains a general advantage over the manual solutions.

Expected Mass Units Consumed	
Budget	IP Model
$200,000	5141
$950,000	3970

Table 2: Reduced-budget model results, assuming that some of the sensor placement budget is devoted to computing resources

8 Discussion

This paper explores the role of mathematical models in real-world sensor placement problems. Optimization models should not be used as definitive artibers of monitoring station placements. Decisions ultimately must be made by humans, who are better able to recognize and weigh different objectives. These may range from trading-off early warning effectiveness for better dual-use water quality monitoring ability to not upsetting the powerful city politician whose nephew is in charge of the fire station on Central Avenue.

We find that solving a mathematical model provides a value-added service to decision makers who will ultimately decide upon sensor locations. For example, a human decision maker will likely be curious to know where his or her placements differ from the model's. The exercise of comparing these placements might generate new insights or bring to light new considerations.

In fact, our human expert generated an interesting set of questions after seeing the results. Which locations detected the most attacks during evaluation? Which detected the consumption-weighted most significant attacks? Would ranking sites by attack detection value yield a reasonable heuristic? All of these are interesting questions, and, we conjecture, typical of those that will arise from a real sensor location scenario.

We plan to consider a simple modification to our experiments: allowing the contaminant to continue to propagate for some period of time after detection. This will simulate the inevitable delay before a general alarm is raised, as the utility assesses the situation. This modification may tend to blur the distinctions in effectiveness between sensor placements, but the magnitude of this effect is an open question.

References

[1] J. Berry, W. E. Hart, C. A. Phillips, and J. Uber. A general integer-programming-based framework for sensor placement in municipal water networks. In *Proceedings of the World Water and Environment Resources Conference*, 2004.

[2] R. Fourer, D. M. Gay, and B. W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming*. Brooks/Cole, Pacific Grove, CA, second edition, 2002.

[3] L. A. Rossman. The EPANET programmer's toolkit for analysis of water distribution systems. In *Proceedings of the Annual Water Resources Planning and Management Conference*, 1999. Available at `http://www.epanet.gov/ORD/NRMRL/wswrd/epanet.html`.