

Source Location Inversion and the Effect of Stochastically Varying Demand

Sean A. McKenna¹, Bart Van Bloemen Waanders¹, Carl D. Laird², Steven G. Buchberger³, Zhiwei Li³, Rob Janke⁴

¹Sandia National Laboratories, Albuquerque, New Mexico, *samcken@sandia.gov*

²Department of Chemical Engineering, Carnegie, Mellon University, Pittsburgh, PA

³Department of Civil and Environmental Engineering, University of Cincinnati

⁴U.S. EPA National Homeland Security Research Center, Cincinnati, Ohio

Abstract

In recent years, significant advances have been made in the development of gradient-based optimization algorithms and their application to inverse problems in water distribution systems. We apply a gradient-based optimization procedure to the problem of identifying the location of a contaminant injected into a distribution system based on data collected at a finite number of sensors. The solution of this problem is complicated by uncertainty in the instantaneous water demands occurring at nodes throughout the distribution system. We characterize the effect of this demand uncertainty on the ability of the inversion algorithm to accurately and precisely identify the correct source location by varying the time step at which the variable demands are aggregated from 30 minutes to 24 hours. These calculations determine the effect of demand aggregation on the inversion results by comparing the results across time step sizes to the results achieved at the smallest time scale (30 minutes). In a distribution system the true water demands at any time step are unknown and represent irreducible uncertainty. We show how large of an effect this irreducible uncertainty has on our ability to locate the source location of contaminants within a distribution system. The calculations are done on a moderately sized distribution system network and the stochastic demands are generated using a recently developed Poisson Rectangular Pulse (PRP) demand generator. The contaminant is simulated with tracer transport using EPANET. Results for the example problem examined herein using 100 sensors show that the inverse approach is capable of identifying the correct source node at all time step aggregations.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94-AL-85000

Introduction

The threat of accidental contamination of water distribution systems is not new. However, in the past few years, concern over malevolent contamination of municipal water networks has increased consideration of novel protection measures for distribution systems. Water distribution networks are especially vulnerable to biological and chemical attack due to the distributed nature of the network over large spatial areas and the large number of access points within a network. Any water outlet, such as a hydrant or even a household water faucet, can be an access point for backflow contamination into the network. Physical security can only provide a limited amount of protection. As an alternative to physical security alone, sensors can be installed in the network to detect contamination and initiate a containment and restoration strategy.

Given the necessity of applying a limited number of sensors within a distribution system to detect a contamination event, the optimal location of sensors within networks has been an area of active research. Much of this research effort has been focused on development of algorithms for both the optimal placement of water quality and, eventually, contaminant specific, sensors within water distribution systems as well as inversion algorithms for using information from sensor networks to identify the source location of the contaminants. The sensor placement algorithms include both heuristic sensor location optimization schemes (e.g., Ostfeld and Salomons, 2004; Uber et al, 2004) and exact solution techniques including integer programming approaches (e.g., Berry et al., in press). Relatively less attention has been paid to the source location inversion problem, but in recent years advances in the application of gradient-based optimization to solve this problem have been realized (Laird, et al., 2005; van Bloemen Waanders et al., 2003).

An aspect of water distribution systems that has not yet been considered as part of the source location inversion process is that demands within the system are unknown and typically of a much shorter time scale than the length of the hydraulic time step used in distribution system models. Over the past decade, considerable effort has gone into characterizing small-scale demands within networks. Buchberger *et al* (2003) monitored demands in a small neighborhood at a one-second time interval for a period of nine months. This extensive data set was used to corroborate the Poisson Rectangular Pulse (PRP) model of residential water use, first proposed by Buchberger and Wu (1995). This model has been used to examine the effect of fine scale demand patterns on the transport of tracers within distribution systems (McKenna, et al., 2004). Errors in the prediction of tracer concentrations can occur when the true fine time scale demands are averaged over larger and larger time steps.

Here we examine the ability of a contaminant source location inversion approach to correctly identify the source location using a network model with larger hydraulic time steps than those used in the true contaminant transport. This work combines recently developed source location inversion algorithms for water distribution networks with a new PRP-based demand simulator. The goal of this work is to

identify how well the inversion algorithm can identify the true source location as the variable demands are aggregated over larger and larger time steps.

Simulation Approach

This paper brings together two distinct lines of research: characterization of instantaneous water demands and non-linear inversion schemes for source location identification. Stochastic simulation is used to generate water demands at different scales of temporal discretization and dynamic optimization techniques are used to solve the source location problem in real time. A brief background on each of these simulation approaches is given below along with references where more detail can be obtained.

Demand Simulation

Residential water demands at single family homes are assumed to behave as a nonstationary PRP process (Buchberger and Wu, 1995). Under the PRP hypothesis, the frequency of residential water use follows a Poisson arrival process with a time dependent rate parameter. This process produces an exponential distribution of arrival times. When a water use occurs, it is represented as a single rectangular pulse of random duration and random but steady intensity, or consumption rate. When a home draws a pulse of water from the supply network, it is considered to be "busy"; otherwise, the home is considered "idle".

Buchberger and Wells (1996) found that over 80 percent of indoor residential water demands occur as single pulses and that complex demand patterns are easily converted to an equivalent single pulse. By virtue of the Poisson assumption, it is unlikely that more than one pulse will start at the same instant. Owing to the finite duration of each water pulse, however, it is possible that two or more pulses with different starting times will overlap for a limited period. When this occurs, the total water use at the residence is the sum of the joint intensities from the coincident pulses. Owing to its simplicity and versatility, the PRP approach offers an effective new way to model the temporal and spatial variability of residential water demands across a municipal distribution system.

A computer program **PRPsym** was developed to simulate residential water demands at various time averaging intervals for each node in a municipal distribution system. The simulation process involves three steps:

Step 1: Generate instantaneous PRP water demands at a 1-second resolution.

- a) Draw the time, T , to the next demand from the exponential distribution of times between demand, $\phi_T(t)$, consistent with the specified Poisson arrival process.
- b) For each new demand pulse draw the duration, D [T], and the intensity, Q [L^3/T], from their respective log-normal distributions, $\phi_D(d)$ and $\phi_Q(q)$.
- c) Check for end of simulation time, if not then return to step 1a.

Step 2: Integrate the instantaneous demands. Some water demand pulses at a node will overlap. Add the coincident pulses to get the total instantaneous demand at a node.

Step 3: Convert the total instantaneous water demands to an equivalent series of time-averaged intensities. Divide the total volume of water use during a specified time step by the duration of the time step to get the time-averaged water demand.

Scaling of the Poisson arrival parameter allows for demand nodes to be configured to represent from 1 to over 1,000 homes. Each node requires a PRP pulse template to specify statistical properties of indoor and outdoor water demands and a multiplier pattern to define the hourly variation in arrivals for indoor and outdoor water use that occur throughout the day.

Non-Linear Optimization for Source Location

Details on the application of non-linear optimization approaches to the problem of identifying the location of a contaminant source within a water distribution network can be found in Laird et al (2005). The approach is outlined briefly below.

With known flow rates and velocities as inputs, the water quality model for the network, using P , J and S to refer to the complete sets of all pipes, junctions, and storage tanks, respectively, is developed. The concentration in the pipes is denoted as $c^p(x; t)$; $i \in P$ and $c^n_k(t)$; $k \in N$ represents the concentration at the nodes, where $N = J \cup S$ is the complete set of all nodes, including junctions and storage tanks. Here, $t \in [0::t_f]$ is time, and $x \geq 0$ is the displacement of flow along a pipe. In developing the model, we refer to connections and concentrations at pipe boundaries. Note that these designations are time dependent and change with the flow direction. Pumps and valves are modeled as zero length pipes, and reservoirs are modeled as junctions with known external sources. We assume there is no decay reaction for the contaminant, although first order decay can easily be included in the formulation. Consistent with the transport formulation within EPANET, plug flow and perfectly mixed concentrations, no dispersion, are assumed.

The goal of the optimization algorithm is to minimize an objective function, Ψ , that is the weighted sum of squared errors between predicted and observed concentrations at the nodes:

$$\min_{m(t), c^p(x, t), c^n(t)} \Psi = \sum_{r \in \Theta_s} \sum_{k \in N_s} \frac{1}{2} \int_0^{t_f} w_k(t) [c_k^n(t) - c_k^{n*}(t)]^2 \delta(t - t_r) dt \quad (1)$$

Where $m(t)$ is the unknown contaminant mass flow rate, $w_k(t)$ is a time dependent flow based weighting scheme for each node – this weighting shifts the errors from being between measured and predicted concentration to mass, $c_k^{n*}(t)$ is the measured concentration at node k , $\delta(t)$ is the Dirac delta function, t_f is the final simulation time and r indicates the time step at which concentrations at sensors are evaluated. This

function is minimized subject to physical constraints on transport of the contaminant within the pipes:

$$\left. \begin{aligned} \frac{\partial c_i^P(x,t)}{\partial t} + u_i(t) \frac{\partial c_i^P(x,t)}{\partial x} &= 0 \\ c_i^P[x = I_i(t), t] &= c_{k_i(t)}(t) \\ c_i^P(x, t = 0) &= 0 \end{aligned} \right\} \forall i \in P \quad (2)$$

As well as conservation of mass within the tanks and nodes. Additionally, a regularization term is added to the overall objective function to force a unique solution to the inverse solution. The discrete forms of the objective function and the physical constraints on mass transport as well as the equations for conservation of mass are solved using a direct simultaneous method (see Laird et al., 2005).

Example Problem

The problem of inverse source location using different hydraulic time step length under conditions of variable demand is illustrated with a set of calculations on an example distribution network. The distribution network is shown in Figure 1 and has 1 tank, 394 nodes and 534 pipes. The base demands are shown in Figure 1. Base demand at any one node ranges from zero to over 150 gpm. The base demand at each node has a periodic trend throughout the day. The **PRPsym** code is used to generate demand multipliers that are then applied to these base demand values.

The concentration is injected at node # 435 (Figure 1) as a square wave at hour 6 with a duration of 1 hour. **EPANET** is used to solve the hydraulic and transport equations. A 5 minute water quality time step is used in all simulations and the hydraulic time step is varied as discussed below. A total of 100 sensors are distributed randomly throughout nodes in the system with the probability of selecting any given node as a sensor location being approximately proportional to the amount of flow through that node. The sensors are assumed to provide error and noise free readings of the actual concentrations.

The parameter values for residential water demand simulations done herein are listed in Table 1. Two types of output are produced by the **PRPsym** software for each node: (i) water demand time series and (ii) water demand simulation statistics. The **PRPsym** code is used to generate demands for 0.5, 1, 2, 4, 6, 12 and 24 hour time step lengths. The base case simulation used as the ground truth in this work simulates instantaneous demands using the parameters in Table 1 and then aggregates these demands over 30 minute time steps. These aggregated demands are used in **EPANET** for simulation of the transport from node 435. The concentrations observed at the nodes with sensors are then recorded as the observed concentrations.

The same demand generation parameters, Table 1, are then used to generate demands aggregated over 1, 2, 4, 6, 12 and 24 hour time steps. The base demand at each node is the same for all simulations and the total demand across the entire network is

preserved across all aggregation levels. The random number seed used to generate the stochastic demands is also kept the same across all time step sizes.

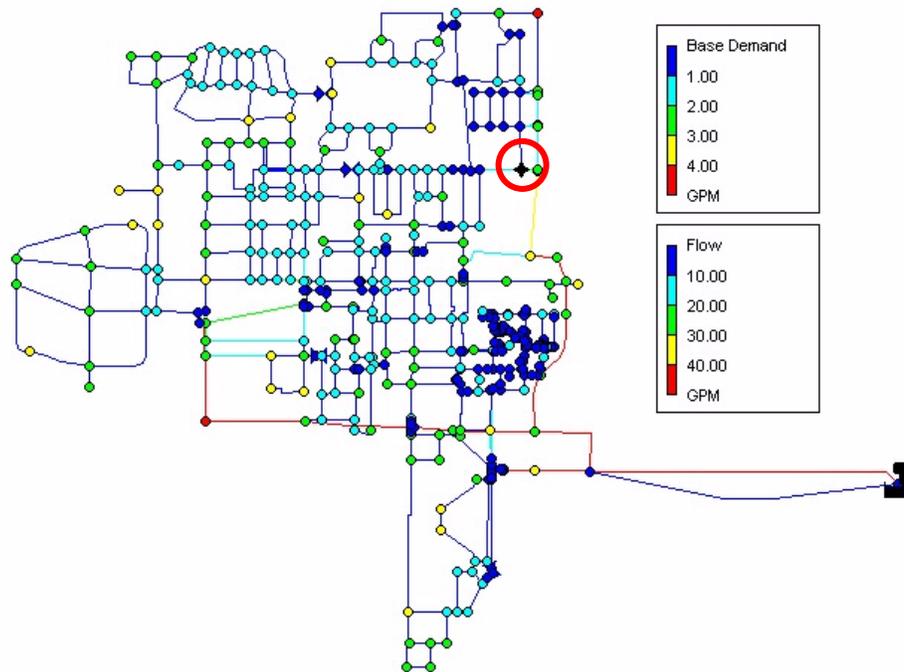


Figure 1. Example network used in simulations. The color scales show the base demand (gpm) at the nodes and the flow (gpm) in the pipes. The red circle shows the location of node 435, the contaminant source.

Table 1. Input parameters for the PRP demand simulator, **PRPsym**

Demand Characteristic		Indoor Use
Intensity:	mean	2.00 gpm
	variance	1.56 gpm ²
	coef of var	0.62
	distribution	log-normal
Duration:	mean	1.00 min
	variance	4.00 min ²
	coef of var	2.00
	distribution	log-normal

(*) Values are derived from Buchberger and Wells [1996] and Buchberger *et al* [2003].

The inverse source location approach minimizes Equation 1 by matching the data observed in the base case simulation with 30 minute time steps with each of the models having larger time steps. The inverse approach optimizes the source location to achieve this minimization. Only a single **EPANET** simulation is necessary to provide the hydraulic information for each time step size. For each time step length,

the ability of the inverse approach to identify the correct source node is quantified as discussed below.

Results and Discussion

For each inverse estimation, the fraction of the total contaminant mass in the source term is partitioned to different nodes within the network. An exact solution would assign a fraction of 1.00 to the true source node. The inverse approach with regularization results in a very small fraction of mass being assigned at all nodes in the network; however, here we only identify the nodes in the solution that are assigned a mass fraction of at least one percent of the maximum mass fraction assigned to any node. For example,

Figure 2 shows the scaled mass fraction for an inverse solution with 2 hour time steps. Four nodes are assigned a significant fraction of the source mass, nodes 435, 2301, 431 and 2401, with the majority of the mass assigned to the true source node (435).

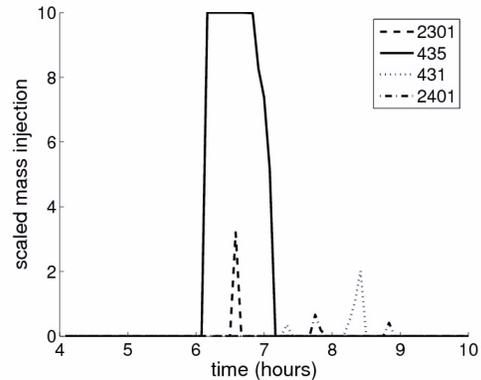


Figure 2. Example output showing amount of source mass assigned to different nodes.

The results of the inverse source location calculations are summarized using three different quantities. The final value of the objective function (Equation 1) indicates the ability of the inverse model to match the concentrations observed at the sensor locations through adjusting the location of the source. These final objective function values are shown in the left image of Figure 3 and show that the sum of the squared errors between the observed and modeled concentrations increases linearly with increasing time step size up to time steps of six hour length.

The ability of the inverse approach to find the correct source node is quantified by the fraction of the total mass that is assigned to the true source node (middle image, Figure 3). For time step lengths of 2 hours or less, over 90 percent of the total mass is assigned to the correct source node. For longer time steps, the mass fraction correctly assigned to the source node decreases to less than 50 percent for all time step lengths greater than 6 hours. The assigned mass fraction results in the middle image of Figure 3 do not directly address the number of nodes to which a significant amount of mass is assigned. For example, in cases where the mass fraction correctly assigned to the source node is less than 50 percent it is not clear whether the majority of the mass was incorrectly assigned to a different node, or if the largest proportion of the mass was still assigned to the correct node, but the remaining amount of mass was spread over a greater number of incorrect nodes. All results were checked and in each case, the true source node was the node assigned the largest amount of mass. In general, the larger the time step length, the larger the number of nodes that were assigned a significant fraction of the mass.

Ideally, the amount of mass assigned to the correct node and the number of nodes over which the mass is spread could be summarized in a single term. Toward this goal, a measure of entropy is introduced as a way to quantify the uncertainty in assignment of a single node as the source node. The entropy is calculated as (after Harr, 1987):

$$H = \frac{\left(- \sum_{k=1}^g m_k \ln(m_k) \right)}{\ln g}$$

where m_k is the mass fraction assigned to the k th node out of a total of g nodes assigned a significant fraction of the mass. As $m_k \rightarrow 1.0$ for any k , $H \rightarrow 0$. Conversely, H reaches a maximum value of 1.0 when all m_k are equal and the determination of the most likely source node is highly uncertain. The entropy of mass fraction provides a convenient single measure of uncertainty and is adopted from similar approaches with classification probabilities; however, it is noted that the mass fraction is not directly equal to the probability of the node being the true source location. This change in entropy definition will be examined further in future work. Here, g is set to equal the total number of nodes with a mass fraction that is at least one percent of the maximum mass fraction assigned to any node. Therefore, g varies in these calculations from 3 to 14. The calculated entropy value increases to nearly 0.80 with increasing time step size up to a time step size of 6 hours after which it decreases slightly (Figure 3, right image). An alternative calculation for entropy would be to set g to the total number of nodes in the distribution system for all calculations. This approach will be also be explored in future work.

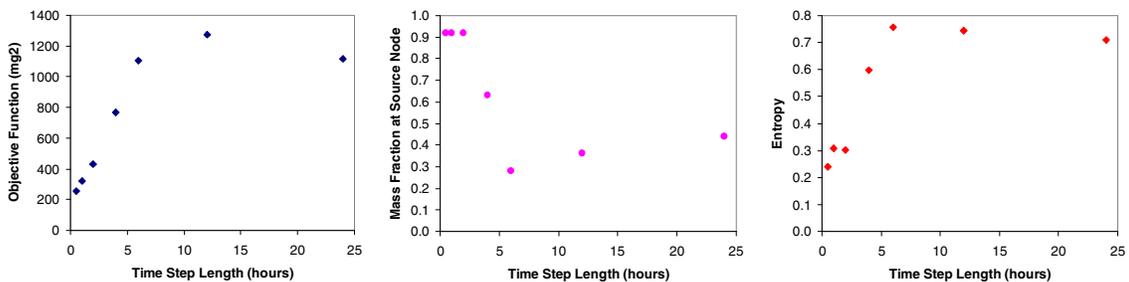


Figure 3. Summary of results showing the objective function value (left), the mass fraction assigned to the correct source node (middle) and the entropy (right) all as a function of time step size.

Conclusions and Future Work

The results above show that the inversion approach is capable of identifying the correct source node over all time step sizes investigated. These results indicate that an actual contamination event that is responding to the very fine-scale demands within a distribution system could be identified using a simulation model with much larger time steps. The ability of the inverse source location identification approach to

find the true source node becomes less certain for larger time steps, see the measures plotted in Figure 3, but for all of the cases examined here, the identification was accurate – the approach assigned the majority of the source mass to the correct source node across all time step sizes.

The results show that, in general, the quality of the source identification using the inverse approach degrades with increasing aggregation of fine-scale demands up to a time step size of approximately 6 hours. The quality of the source identification, whether measured in terms of the objective function, the fraction of the total mass assigned to the true source or the entropy, is roughly constant for the time step sizes beyond 6 hours (Figure 3). The reasons for this asymptotic behavior in the results will be examined in future work. The results in Figure 3 suggest that the solutions become less unique with larger time steps. One way to examine the uniqueness of these results with respect to the aggregated demands will be to generate multiple stochastic demand patterns for each time step size and characterize the range of objective function, mass fraction and entropy values across all simulated demands. Other simulations that can be done to assess the sensitivity of the results presented here include varying the parameters used in **PRPsym** (Table 1) to generate the demands and to vary the number of sensors within the network. Rapid assessment of these additional simulations is feasible due to the minimal computational burden that can be achieved by the inversion approach used herein.

References

- Berry, J., L. Fleischer, W.E. Hart, C.A. Phillips and J.-P. Watson, (in press), Sensor Placement in Municipal Water Networks, accepted for publication in: *Journal of Water Resources Planning and Management*, Special Issue on Water Distribution Network Simulation.
- Buchberger, S.G., Carter, J.T., Lee, Y.H., and Schade, T.G. (2003) “Random demands, travel times, and water quality in deadends.” *American Water Works Association Research Foundation*, Report 90963F, Denver, Colorado, 470 pages.f
- Buchberger, S.G. and G.J. Wells, 1996, Intensity, duration and frequency of residential water demands, *Journal of Water Resources Planning and Management*, ASCE, 122(1):11-19.
- Buchberger, S.G. and L. Wu, 1995, A model for instantaneous residential water demands, *Journal of Hydraulic Engineering*, ASCE, 121(3):232-246.
- Harr, M.E., 1987, *Reliability-Based Design in Civil Engineering*, Dover, Publications, Inc., Mineola, New York, 291 pp.
- Laird, C.D., L.T. Biegler, B.G. van Bloemen Waanders and R.A. Bartlett, 2005, Contamination Source Determination for Water Networks, *Journal of Water Resources and Planning*, March/April.

- McKenna, S.A., S. Buchberger, V.C. Tidwell, 2003, Examining the Effects of Variability in Short Time Scale Demands on Solute Transport, in proceedings of: ASCE 2003 World Water and Environmental Resource Congress, June 22-26, Philadelphia, Pennsylvania, January, June 23-26.
- Ostfeld, A. and E. Salomons, 2004, Optimal Layout of Early Warning Detection Stations for Water Distributions System Security, *Journal of Water Resources Planning and Management*, 130 (5), pp. 377-385.
- Uber, J. R. Janke, R. Murray and P. Meyer, 2004, Greedy Hueristic Methods for Locating Water Quality Sensors, ASCE World Water & Environmental Resources Congress 2004, Salt Lake City.
- van Bloemen Waanders, B.G., R. A. Bartlett, L.T. Biegler, C.D. Laird, 2003, Nonlinear Programming Strategies for Source Detection of Municipal Water Networks. ASCE World Water and Environmental Congress, Philadelphia, Pennsylvania, June 23-26.