SAND2008-0383C

Distributed Network Fusion for Water Quality[+]

Mark W. Koch[1] , Sean A. McKenna[2]

[1] Sandia National Laboratories[*], P.O. Box 5800, MS 1163, Sensor Exploitation Applications Department, Albuquerque, NM; mwkoch@sandia.gov

[2] Sandia National Laboratories, P.O. Box 5800, MS 0735, Geohydrology Department, Albuquerque, NM; samcken@sandia.gov

Abstract

To protect drinking water systems, a contamination warning system can use in-line sensors to detect accidental and deliberate contamination. Currently, detection of an incident occurs when data from a single station detects an anomaly. This paper considers the possibility of combining data from multiple locations to reduce false alarms and help determine the contaminant's injection source and time. If we consider the location and time of individual detections as points resulting from a random space-time point process, we can use Kulldorff's scan test to find statistically significant clusters of detections. Using EPANET, we simulate a contaminant moving through a water network and detect significant clusters of events. We show these significant clusters can distinguish true events from random false alarms and the clusters help identify the time and source of the contaminant. Fusion results show reduced errors with only 25% more sensors needed over a nonfusion approach.

1. Introduction

To maintain the safety and security of drinking water, water utilities need innovative technologies to detect deliberate or accidental contamination in water distribution systems. One approach uses water quality sensors in the water distribution system and measures attributes of the water such as free chlorine, total organic carbon, pH, temperature, and electrical conductivity. While these measurements do not necessarily measure contaminant levels directly, a sudden change in their readings can indicate contamination or an abnormal operation of the water distribution system. One approach uses change detection algorithms to compare the current measurements with models of the background. We call each location with sensors and algorithms a *sensing-node* and a change in water quality detected by the algorithms an *event*.

In conjunction with the National Homeland Security Research Center we are extending research from detection at a single sensing-node to detection at multiple nodes distributed throughout the water distribution network. Here, we want to use the topology of the water distribution network and sensor fusion to combine multiple

[+] This work was funded by the U.S. EPA National Homeland Security Research Center (NHSRC)

[*] 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-AC04-94AL85000

detected events together. This approach allows the reduction of errors for the entire system and the determination of the source of the contaminant.

In detection problems, the two types of errors are false alarm (FA) errors and missed detection (MD) errors. FA errors occur when the change detection algorithm alarms on a nonexistent event and MD errors occur when the change detection algorithm misses an actual event. The two errors are linked, so that decreasing one of the errors increases the other. If the FA errors are too large then operations personnel will begin to mistrust the system, but if there are too many MD's then the system becomes ineffective. These FA problems multiply when we add more sensing-nodes to determine the source location and extent of the contaminant. Suppose a single sensing-node has one FA per day then a network of 100 sensing-nodes would have four FA errors per hour! This would be unacceptable in a working system.

Figure 1 shows a space-time cube for a water distribution system called Anycity with 100 randomly placed sensing-nodes. The width and depth dimensions of the cube show the spatial dimension with the water distribution network shown at the top and bottom of the cube. The time dimension is along the height of the cube with time increasing from bottom to top. The circles represent simulated detections from a change detection algorithm. The open circles show randomly generated FA's assuming a sensor at every junction and a single sensing-node FA rate of once per day. The entire cube represents sensor activity over 25 hours. For 1 FA per day we expect an average of 104 FA's within the cube. Using EPANET (Rossman 2000), we simulate a tracer injected into the network. This tracer represents the contaminant and the solid circles show the detections of this tracer. In actuality, we would not know if the detections were real or FA's (solid or open circles). Using sensor fusion, we would like to separate the real detections from the FA's and reduce the errors of the entire system.

We consider the location and time of an event as a point resulting from a random space-time point process. Statistically significant clusters of events indicate a set of true detections, whereas a set of purely random events would indicate false detections. We use Kulldorff's scan statistic to fuse the detections of the sensing-nodes in this distributed network. Here, Kulldorff's scan statistic can identify statistically significant clusters of events in space and time. The location and size of the significant clusters indicates the location and the extent of the contamination. The scan test uses sliding windows of different sizes in space and time to search for clusters. We use the distribution network's topology to define the space dimension.

To test our distributed detection algorithms, we use EPANET to simulate a city's water distribution system. Combining EPANET's simulation of the transport of a tracer and the performance models of the change detection algorithms, we show how multiple sensing-nodes improve the event detection performance over a single sensing-node. We also show how the system's performance changes with number of sensing-nodes and how well the scan test determines the injection location and time of the contamination.

The rest of this paper first discusses related work in Section 2, and then describes our approach and implementation in Section 3. Section 4 describes how we evaluated our system and presents the results. Section 5 presents our conclusions and describes future work.

2. Related Work

Recent research on using water quality measurements to identify periods of anomalous water quality has focused on data obtained at a single monitoring location. Various algorithms have been applied to these data sets to extract anomalous signals from the often noisy water quality background (e.g., Cook et al., 2006; Jarrett et al., 2006; Kroll and King, 2006). Research at Sandia National Laboratories has involved development and testing of multiple robust multivariate statistical algorithms (Klise, 2006; McKenna, 2006 and 2007) and these are embedded in the CANARY software (Hart et al., 2007). The algorithms provide a means of automatically detecting changes in water quality sensor measurements by comparing the current measurements to their predicted values based on their previous history. Essentially, the algorithms create a current measurement vector from all the available sensors. This measurement vector is compared to a prediction vector based on previous sensor data.

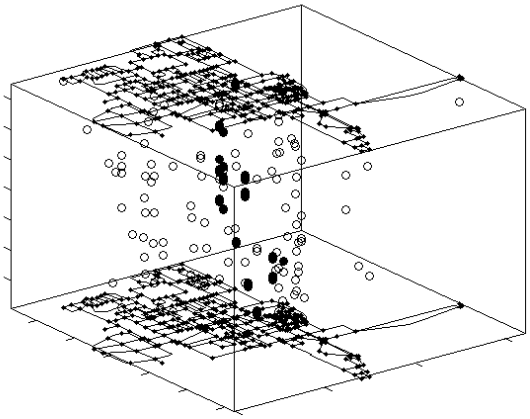

Figure 1. Space-time cube of Anycity with simulated sensors at every junction. The water distribution network is shown in the space dimensions (width and depth) and time is the height dimension from bottom to top. The circles represent detected events with open circles shown as false alarms and filled circles as correct detections.

The concepts of distributed detection, where sensor responses from multiple locations across a network are fused to provide a "network-wide" detection capability, have not been fully applied to water distribution networks. Initial work towards integrating responses from more than one sensor location has recently been reported (Yang 2007). In this work, the authors use water quality sensors at two locations to improve the water quality signal. One of the locations acts as a reference that allows for adaptive compensation at the second location to account for calibration errors and background noise in that second sensor. The authors show how improved signals could allow classification of the contaminant.

3. Approach and Implementation

If m sensor nodes are randomly placed in a network of M junctions then the probability of having at least x detections in a contaminant plume with a size of X junctions is:

$$\Pr(x \mid m) = \sum_{i=x}^{X} \binom{X}{i} \binom{M-m}{X-i} \Big/ \binom{M}{m} \qquad \textbf{(1)}$$

where $\binom{n}{k}$ represents the combination of n things taken k at a time. For a nonfusion approach and perfect sensors in the Anycity network (M=396) and a contaminant plume size of 20 junctions we would need 80 sensors to have a 0.99 probability of

having one sensor in the contaminant plume. For many imperfect sensors we have the potential for a large increase in false alarms, as discussed in Section 1. To remedy this problem, we propose a modest increase in the number of sensing-nodes and a distributed fusion approach to combine results at multiple nodes to reduce the false alarms errors.

To combine detections from multiple sensing-nodes, we use Kulldorff's scan statistic (Kulldorff 1997). Scan statistics are used to determine whether a set of points are randomly distributed or show signs of clustering. Scan tests count events in sliding windows over an area A and use the counts to determine if there is a cluster of significant events. Kulldorff calls this set of windows *zones*.

Although computationally intensive for estimating the null distribution, Kulldorff's approach can handle multiple dimensions, overlapping zones of different sizes and shapes, and it directly determines the locations of the clusters. By using a likelihood ratio and a clearly defined alternative hypothesis it avoids the multiple and dependent testing problem. It is also a unique test making it unnecessary to perform a separate test for each cluster size and location. We use the binomial version of the test (Kulldorff 1995), and assume the single sensing-node produces a yes/no or 1/0 decision on the presence/absence of an event.

Kulldorff's scan test is conditioned on knowledge of the total number of events C. Here we need to know the geographic area and time interval of interest A, and how the region is covered with the set of all zones Z. Kulldorff's test has two hypotheses:

1. Null hypothesis H_0: For all the zones, the probability of an event inside the zone p is the same as outside the zone, q, i.e. $p=q$.
2. Alternative hypothesis H_1: There is at least one zone where the probability of an event inside the zone is greater than the probability outside, i.e. $\exists z \in Z \mid p > q$.

The likelihood function $L(z,p,q)$ for the scan test is:

$$L(z, p, q) = p^{c_z}(1-p)^{n_z-c_z} q^{C-c_z}(1-q)^{(N-n_z)-(C-c_z)} \quad \textbf{(2)}$$

and represents the likelihood that the number of events inside zone z is c_z and the number of events outside zone z is $C-c_z$. Here N represents the total number of possible events in A and n_z represents the number of possible events in z. Using (2) the likelihood ratio becomes:

$$\frac{L(z)}{L_0} = \frac{\sup_{z \in Z, p>q} L(z,p,q)}{\sup_{p=q} L(z,p,q)} \quad \textbf{(3)}$$

Thus the scan test uses the largest likelihood ratio to combine results from multiple zones. The scan test statistic λ is:

$$\lambda = \frac{\max_z L(z)}{L_0} \quad \textbf{(4)}$$

In general the distribution of λ has no simple analytical form. To determine the distribution of λ for the null hypothesis, Kulldorff suggests using Monte Carlo randomization. Since the test is conditioned on the number of cases C, we can generate random examples using $p = C/N$ (sensing-node FA error estimate) and compute the scan test for each. As long as the number and performance of the

sensing-nodes stays the same then estimation of the null distribution can be accomplished offline and prior to application.

4. Implementation

To fuse information from multiple sensing-nodes in a water distribution system we use Kulldorff's scan test. Figure 2a shows a hypothetical time series for three sensing-nodes. Assume sensing-nodes *B* and *C* are downstream from sensing-node *A*, and the sensing-node produces a 1/0 decision for anomaly and no anomaly, respectively. In the figure, a dot with a stem represents a 1 and just a dot represents a 0. For our problem we have 2 dimensions: space and time. The network of pipes and junctions represents the space dimension. We represent the space dimension as the travel time between sensing-nodes. We use EPANET simulations to find the median velocity over a 24 hour period for each pipe in the network and use these median values to compute an estimate for the travel time between any two nodes.

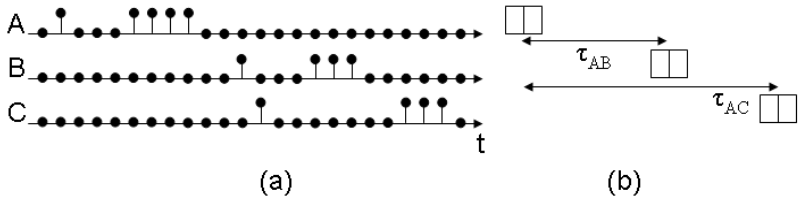

Figure 2. (a) Example hypothetical time series for 3 sensing-nodes A, B, and C. Here, a dot with a stem represents a 1 or a detection and a dot with no stem represents a zero or no detection. (b) A 3×2 space-time template for a space-time cluster centered at node A.

To search for clusters of detections we specify the zone sizes in space and time as $s \times \tau$. Here s represents the size of the zone in space and τ represents the size of the zone in time. The space size s represents the number of sensing-nodes closest to and including the center of the zone (in travel times). For example, Figure 2b shows a template for a 3×2 space-time zone centered at node *A*. The number of adjacent boxes represents the zone size in time ($\tau = 2$) and the number of sets represents the zone size in space ($s = 3$). The horizontal distance τ_{AB} represents the estimated travel time between nodes *A* and *B* and τ_{AC} represents the estimated travel time between nodes *A* and *C*. This zone template searches for clusters that have a contaminant source at node *A*.

Figure 3 shows the 3×2 template aligned with the time series in space and time for the current time t_c. Here, the leading edge of the template is aligned with the current time for sensing-node *C*. The total number of detections in this template is one. This becomes c_z equation (2). Even though an event was detected at node *A* there is no correlating evidence at the other nodes, so the scan

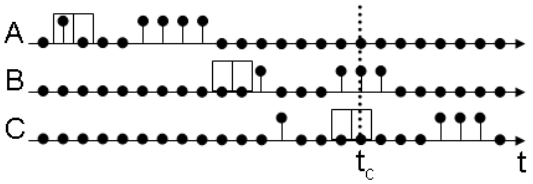

Figure 3. The 3×2 template aligned with the time series in space and time at current time t_c.

test would not detect a significant cluster for this location and time.

Figure 4 shows the template advancing to a new time. Here there are six detections in this template. If we assume a significant cluster is detected and the contaminant was introduced at time step t_i, then the detection delay is given by $t_c - t_i$.

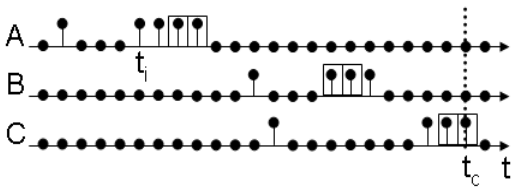

Figure 4. The 3×2 template aligned with the time series in space and time at a new time t_c.

Since we do not know the actual size of the space-time cluster, we need to test with multiple zone templates of different sizes. We also do not know the source location of the contaminant, so we need to test with zone templates that assume a source at the other sensing locations. At each point in space and time these zones are combined by taking the one that produces the largest scan test score (3). Note, for $\tau > 1$, counts in the zone template at one time may be used for counts for a zone template at neighboring times. Because of this and the different zone sizes, the random variables representing the counts are not independent. This makes it difficult to determine the null hypothesis analytically.

5. Evaluation and Results

To evaluate our distributed fusion approach we built an event simulator called DetectNet using Matlab and the EPANET toolkit (Rossman 1999). The objective of DetectNet is to simulate sensing-node detections from an algorithm like CANARY (Hart 2007) in a water distribution system.

The heart of DetectNet is EPANET. EPANET takes a description of a water distribution system including stochastic demands and a chemical tracer and determines the concentration of the tracer throughout the network at different time steps. This tracer serves as a proxy for a contaminant introduced into the water distribution system. The parameters specified for the tracer are initial concentration, start time, and length of the tracer injection. DetectNet takes the tracer simulation results and produces a set of detections based on the performance characteristics of a suite of sensors and the associated sensing algorithm (e.g., CANARY). The performance characteristics are based on FA and MD errors. Here, we use a pseudo random number generator to add extra detections based on the FA error and remove detections based on MD errors. We also use EPANET to extract the network constraints for the senor fusion algorithm. These constraints are the connectivity of the network and median travel times between junctions.

Figure 1 showed the network we use for the simulation. The network has 396 junctions, 534 pipes, 2 tanks, 4 valves, and no pumps (gravity fed). The simulation runs for 24 hours with one minute time steps. For randomly selected locations with no demand, a 30 minute tracer injection with a concentration of 50 mg/L is simulated. The tracer's concentration decreases as it moves through the network. Assuming there is a sensor at a junction, if the concentration at that junction is greater than 5 mg/L then we allow a possible true detection by the sensing-node otherwise we allow only false detections. We selected runs that gave an average plume size, at

concentrations above the detection limit, equivalent to a portion of the network that would contain 20 nodes.

For sensing-node performance we assume a 10 minute sample interval, a FA rate of 1/144 (once per day) and a 0.01 MD error. This FA error was selected as a plausible worst case performance that demonstrated the fusion algorithm's abilities to identify a contaminant in background clutter. The FA error does not necessarily reflect current or projected sensor node performance. We tested Kulldorff's scan test with varying numbers of sensor nodes whose locations were randomly selected. We investigated 396 (sensors at every junction), 200, 150, 100, 50, and 20 sensing-nodes. Using equation (1) and for 20 sensing-nodes, we do not expect very good results, since there is less than a 20% chance that at least two nodes will randomly be placed in the contaminant plume for the contaminant injection characteristics used here. For each set of sensing-nodes we generate 100 different days of background clutter data using the FA error rate and compute the scan statistics for each time step and cluster location. The exact location of the sensing-nodes does not change the null distribution, since the space dimension is based on the s closest nodes. For all sensor configurations we use all combinations of clusters sizes of (1, 3, 6, 12) in space combined with and (1, 3, 6) in time, except $s \times \tau = 1 \times 1$.

Using EPANET we simulate the introduction of a contaminant at 5 different locations. For each separate injection location we generate 100, 1-day simulations with different randomly selected sensing locations and different background FA's. Figure 5 shows a portion of the Anycity network overlaid with sensing-node detections and significant clusters. Here we have a sensor at every junction indicated by the black dots. Circles represent all the detections up to and including the time stamped in lower left corner. Circles filled with white are true detections and circles not filled (shows junction and links) represent false alarms. In actuality, we do not know the truth of the detections, but this labeling makes it easy to see how the network is performing. The contamination is introduced at 12:00 AM.

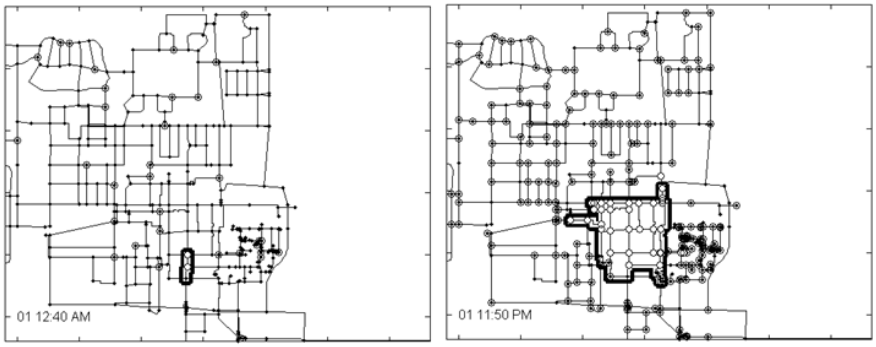

Figure 5. Example results for scan test with 396 sensors. Circles represent detections. Circles filled with white represent true detections and circles with no fill (show junctions and links) are false detections. The solid black line indicates the extent of the significant cluster at this time step. (a) First significant cluster detected. (b) Intersection of significant scan clusters after 24 hours.

In Figure 5a, the heavy black line shows the first significant cluster detected by the scan test at 12:40 AM. Thus it took 40 minutes after the introduction of the

contaminant to detect a significant cluster. Figure 5b shows the intersection of all the significant clusters for that day. Here the scan test accurately reflects the extent of the contamination plume though it does include some FA's near the bottom. This is because we do not make use of any knowledge of the flow direction in any one link. Note the contaminant did not spread very much in the southerly direction.

Figure 6 shows the results for 100 sensing-nodes or 25% coverage. Here, we introduce two contaminant source locations. Figure 6a shows that it takes 4.5 hours to detect both contaminant sources. Figure 6b shows the results after 24 hours.

Figure 6. Scan test results for 100 sensing-nodes and 2 contaminant sources. (a) Results when both sources are first detected. (b) Results after 24 hours.

Figure 7 shows the operating characteristics for different numbers of sensors. We call a correct detection if the scan test finds a significant cluster intersecting the contaminant plume. We call a false detection if the scan test finds a significant cluster during a day of background clutter. The scan test has excellent performance until the number of sensors drop to 50 and below. At this point the chances that at least 2 sensors will be within the contaminant plume start to drop rapidly. For greater than 50 sensors the results are excellent considering the high numbers of individual sensing-node FA's. At 100 sensors we have very low errors. Recall, 80 sensors are needed for the nonfusion single-detection approach assuming perfect detection and 99% detection of plumes with a size of at least 20 junctions. Thus the distributed fusion approach requires a 25% increase in numbers of sensing-nodes.

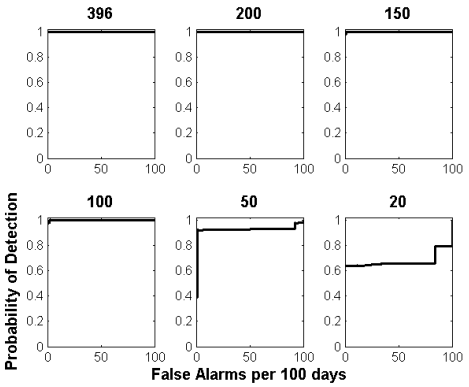

Figure 7. Operating characteristics for varying numbers of sensors.

Figure 8 shows histograms for the time to detection (hours) and distance to detection (links) for the 500 simulated cases. These statistics are based on the time and location of the first significant cluster to be detected (assuming there is a detection). The threshold for distributed detection is selected to achieve one FA per 100 days. As expected the time to detection and distance to detection increases as the

number of sensors decrease. If there is a sensor at every node then we detect the contaminant within one hour and the correct starting location more than 50% of the time. For 100 sensors, those results reduce to 3.5 hours within 2 network links.

Note that the time to detection histograms become bimodal as the number of sensors decrease. Here, we hypothesize that as the number sensors decrease it is more likely that the first sensor to encounter the contaminant will be on the edge of a plume than the center, since the plume expands as time increases. Sensors on the plume edge eventually detect the contaminant, but the delay to detection increases.

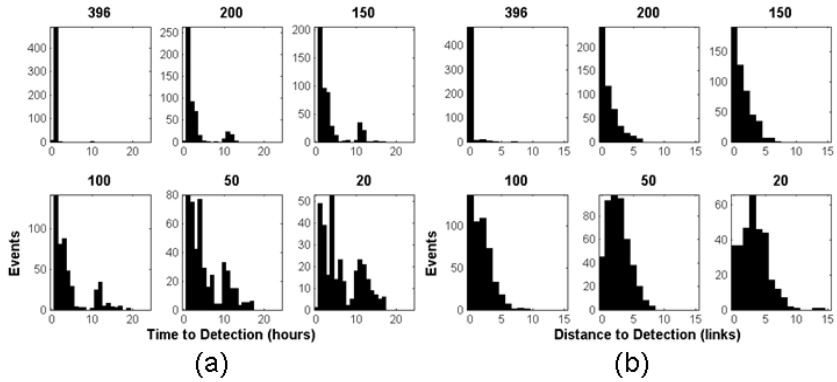

Figure 8. Detection statistics. (a) The time to detection (hours) for different numbers of sensors. (b) The distance to detection (links) in terms of network links. Both are based on the center of the first cluster to be detected.

6. Conclusions and Future Work

We have applied Kulldorff's scan test to the problem of detecting contamination using multiple sensors in a water distribution network. Kulldorff's test identifies significant clusters in space and time and can distinguish between clusters of true events from random background alarms. As the number of sensors in the water distribution network increases, the chance of a FA increases too. This makes it difficult to separate false detections from true detections. The approach developed here is general enough to handle improvements in change detection algorithms such as potential contaminant identification (Yang 2007) and real-time estimation of flow rates and directions from the network model. For a 25% increase in sensing-nodes from the nonfusion single-detection approach, distributed fusion results in very low error rates.

Currently we use Monte Carlo simulation to estimate the null distribution. Reestimination is required if the number of sensors change or the sensor characteristics change. Another approach would use a Bayesian scan test that would make more assumptions about the characteristics of the null distributions. The Bayesian approach would not require the time-consuming Monte Carlo techniques to estimate the null distribution. It is noted that, while time consuming, the current Monte Carlo calculation of the null distributions is done off-line using the assumed FA rate prior to the detection data becoming available. This makes the distributed detection approach developed here capable of functioning in a real-time mode.

Tracking the detections through the network and improving the extent determination is important for knowing how to respond to an event. Tracking

involves determination of which clusters are associated at different time steps and which belong to different contaminant plumes.

In our present approach, we project the detections back in time to determine the best estimate of start location and time. To improve extent determination we could also project detections forward in time. This would give more support to the detections at the edge of the plume and may guide location of portable sampling units to further identify and characterize the contamination event.

7. References

1) Cook, J.B., J.F. Byrne, Daamen, R.C. and E.A. Roehl (2006), "Distribution System Monitoring Research at Charleston Water System", *Proceedings of the 8th Annual Water Distribution System Analysis Symposium*, Cincinnati, OH.
2) Hart D, S.A. McKenna, K. Klise, V. Cruz, and M. Wilson (2007). "CANARY: A water quality event detection algorithm development tool," *World Environmental and Resources Congress*, Tampa FL, 1-9.
3) Jarrett, R., G. Robinson and R. O'Halloran (2006), "On-line monitoring of water distribution systems: Data processing and anomaly detection", *Proceedings of the 8th Annual Water Distribution System Analysis Symposium*, Cincinnati, OH.
4) Klise, K.A. and S.A. McKenna (2006). "Water quality change detection: multivariate algorithms," *Proceedings of SPIE, Defense and Security Symposium 2006*, Orlando, FL, J1-J9.
5) Kroll, D. and K. King (2006), "Laboratory and flow loop validation and testing of the operational effectiveness of an on-line security platform for the water distribution system", *Proceedings of the 8th Annual Water Distribution System Analysis Symposium*, Cincinnati, OH.
6) Kulldorff, M. and N. Nargarwalla (1995), "Spatial disease clusters: Detection and inference, *Statistics in Medicine*, 14:799-810.
7) Kulldorff, M. (1997). "A spatial scan statistic," *Communications in Statistics: Theory and Methods*, 26:1481-1496.
8) McKenna, S.A., K.A. Klise and M.P. Wilson (2006). "Testing water quality change detection algorithms," *Proceedings of the 8th Annual Water Distribution System Analysis Symposium*, Cincinnati, OH.
9) McKenna, S.A., D.B. Hart, K.A. Klise, V.A. Cruz and M.P. Wilson (2007). "Event detection from water quality time series," *Proceedings of: ASCE World Environmental and Water Resources Congress*, Tampa, FL.
10) Rossman, L.A. (1999). "The EPANET programmer's toolkit for analysis of water distribution systems." *Proceedings of the 26th Annual Water Resources Planning and Management Conference*, June 6-9, Tempe, AZ. Vol. 43: 1-10.
11) Rossman, L.A. (2000). *EPANET Users Manual*, EPA/600/R-00/057, United States Environmental Protection Agency, Version 2.
12) Yang, J.Y, R.C. Haught, J. Hall, J. Szabo, R.M. Clark, and G. Meiners (2007), "Adaptive water sensor signal processing: experimental results and implications for online contaminant warning systems," *World Environmental and Water Resources Congress*, Tampa, FL.