METHODS FOR EVALUATING WATER DISTRIBUTION NETWORK EARLY WARNING SYSTEMS

Dan Kroll and Karl King

ABSTRACT

In recent years a wide variety of early warning systems for the continuous on-line detection of events relating to water security and quality have become available. Numerous methodologies and criteria have been suggested to determine the efficacy of these methods in real world scenarios. It is the authors' contention that many of these evaluation techniques have been deficient in evaluating all of the factors that would be crucial in defining a system for deployment for the protection of key installations, such as military bases. The following criteria along with a new method for determining Receiver Operating Characteristic Curves (ROC Curves) are suggested as a good way to compare various technologies. Similar considerations should come into play when evaluating systems for civilian uses.

INTRODUCTION:

The ability to detect and act upon changes in water quality is a critical component in the drive to protect our drinking water supplies from intentional or accidental contamination. The twin motivators of the terrorist threat to water along with consumer demands for safe and potable supplies has lead to a sea change in the drinking water industry. The distribution system represents the last analytical frontier in the water quality industry that has been, to date, overlooked and ignored since the inception of modern drinking water systems. The monitoring of source water and treatment plant processes has progressed to a level at which we can be confident that we are providing good quality water from the plant to the distribution system. Once the water reaches our aging distribution systems, however; our knowledge as to its continued integrity is limited by the quality and amount of available data. From a historical perspective, most monitoring in the distribution system has been relegated to the occasional snapshot provided by grab sampling for a few limited parameters or the infrequent regulatory testing required by mandates such as the Total Coliform Rule.

A number of studies conducted since 9/11 have shown that bulk monitoring of basic water quality parameters has the potential to indicate the presence of many harmful agents in water at the levels of interest. (EPA, 2005; Kroll, 2006; Hall et. al., 2007). The realization of the potential of bulk parameter monitoring as a practical tool to detect terrorism related events has lead to the development of a number of sensor packages designed for deployment in the distribution system.

Since 9/11, numerous communities have installed multi-parameter monitoring stations in various locations through out their distribution network as early warning systems to detect potential water security threats as well as providing operational data. These continuous on-line systems have recorded large streams of data (at some sites for a number of years) relevant to water quality in the distribution systems in which they have been deployed.

These data streams are quite complex and it becomes a Herculean task to differentiate what is normal background noise and fluctuations due to normal everyday events from changes that are indicative of a deviation in water quality deserving further attention. Unless a full-time team of statisticians is to be employed to make sense of this information, the need for computer aided data interpretation becomes obvious. Intelligent algorithms are a necessary adjunct to bulk parameter monitoring if useful decision trees are to be built from this type of monitoring. Intelligent algorithms are necessary to streamline the process of data interpretation. These algorithms should be capable of detecting the subtle changes in bulk parameter readings that are indicative of an incursion into the system without burdening the operators with a constant stream of false or trivial alarms. They should also be capable of differentiating the unique pattern of responses that are elicited by different classes of agent. These differences may be enough to narrow down the cause of events and, possibly, to fingerprint the class of disturbance that caused the event.

Over the past several years a number of such algorithms designed for this use have been under development by private industry, government programs, universities and national labs. The question then becomes how to evaluate the effectiveness of these potential early warning system (EWS) solutions. While a number of research studies (McKenna et al, 2006; Umberg, 2008) and programs such as the EPA's ETV program (ETV2005) have attempted to set criteria and means for evaluating such systems, many important functional criteria for such systems have been overlooked in past studies. A number of factors that are mandatory for the success of any such system are outlined below.

DUAL USE VALUE

The first priority for any EWS is that it detects contaminants so that people and infrastructure can be protected in some effective way. While the ability to detect contamination is critical, if a system is to be capable of offering a likely return on investment (ROI) and achieve widespread deployment, it would also be useful if the EWS can inherently provide information that would be useful on a day-to-day basis. It is presumed that an EWS spends very little of its total time in an actual alarm condition. What is it doing for the rest of the time? Justification of the cost can include:

1) Optimization of daily operations within the system monitored,

2) Providing alarms for operational events not related to contamination threats,
3) Replace grab sampling for compliance purposes with continuous monitoring,
4) Document system operation anomalies to assist with planning maintenance activities or planning and justifying major system upgrades (line replacement)
5) Building consumer confidence by continuously documenting system water quality

Such capabilities can reduce the cost of operations for the system being monitored, and possibly provide information that can be useful to those people who use the monitored system.

DETECTION CLASS REQUIREMENTS

When discussing early warning systems, there are a number of different classification levels relating to their effectiveness commonly utilized in categorizing systems for military use. The following Detection Classes are possible:

1) Detect to treat: very high confidence level, nearly 100%, which allows treatment to proceed on those exposed. Response time slow, cost high.
2) Detect to protect: high confidence level, > 99%, that allows for protection by limiting exposure without the need for confirmatory testing. Response time fast, cost medium.
3) Detect to warn: Presumptive confidence level, <99%, which would allow for protection by limiting exposure while confirmatory tests are run. Response time fast, cost low.

Confidence levels are usually a function of analysis time and cost, with higher confidence levels increasing both. An EWS will need to respond rapidly to be effective, and it will need to be relatively low in cost to provide wide coverage, so this paper will assume that the first deployed defense against contamination will be systems for Detect to Warn.

COVERAGE CHARACTERISTICS

1) Cost

Cost may not be significantly limiting when selecting an EWS for an icon or high-profile facility, but if the coverage is large, such as a major metropolitan area, budgets may be constrained, and the degree of coverage becomes a function of the cost per point monitored. Therefore, cost becomes an issue.

2) Area of protection

Note that effective coverage may be a function of the hydraulics of the distribution system within the geographical setting. Not all ideal deployment sites from a protection standpoint will be capable of being utilized due to system requirements, site “noise functions” and other logistics. While “protection of all” may be a laudable goal, reality may constrain the degree of implementation, forcing tradeoffs.

3) Communication

Monitoring of multiple points in a geographical area immediately raises the need for communication at least from the remote points to some centralized facility where the data can be interpreted, and actions taken. The EWS can accomplish highly sophisticated interpretation of the local data, but it cannot take actions in a complex situation affecting possibly millions of people. Human interaction and analysis will be required. The information from the multiple points must be communicated to an analysis and command center.

Thus, an EWS must be structured for secure communication. Both the instrument and the network must have tools in place to affect a high level of security so that information cannot be blocked, or false information transmitted on the network.

OPERATIONAL CHARACTERISTICS

There are a number of operational characteristics that need to be incorporated into the design of such a system.

1) Ease of use: A system that may be absolutely critical in a crisis cannot be difficult to use. User interface operation must be intuitive enough that minimally skilled operators can obtain necessary information without resorting to an operator’s manual.
2) Automated: The system must normally operate without requiring the presence of a human. Human intervention should only be needed for service or maintenance.
3) Continuous: The system must run without any long gaps in analysis that enable contamination of significant duration to slip by undiscovered. The maximum time for non-observability should be relatively short - on the order of 3 minutes. Longer response times unduly endanger a larger percent of the population by exposing them to the hazard before remedial actions can be taken.
4) Reliable: A non-working system is an opportunity for exploitation.
5) Cost Effective: Amortized cost per day comparable to or less than the labor, travel, equipment and reagents for an existing grab sample program. When evaluating this criterion, it is important to factor in the overall better picture of operations afforded by continuous versus discrete measurements.

PERFORMANCE CHARACTERISTICS

1) Detection of a broad spectrum of contaminant classes

   There are many contaminants that could cause serious harm if introduced into a drinking water distribution system. It should be noted that chemicals within a given category could have fundamentally different chemical characteristics. The commonality that one finds within a category does not preclude different responses of analytical sensors used in an EWS. The ability to detect with significant specificity is dependent upon seeing a chemical in many dimensions. Thus, single parameter systems have inherent limitations. There are a number of important categories of contaminant that such a system should detect including; Organo-phosphates, Carbamates, Other pesticides, Herbicides, Nematicides, Animal poisons, Petroleum products, Heavy metals, Infectious agents, Warfare agents (tested against real threat agents), Toxic Industrial Chemicals, Poisons, Cyanides, Toxins, Drugs and pharmaceuticals both legal and illicit

2) Rapid

The response time of any detection method that is used on a flowing stream is really a measure of the delivery rate of harmful contaminants. A simple way to view this is given by the equation:

$$\#\text{ of lethal doses delivered} = \frac{\text{stream flow rate X response time X conc.}}{\text{toxic dose}}$$

With this in mind, it becomes readily apparent that response times of a few minutes are required. Considering the flow rates of the pipes involved for most deployment scenarios, response times of hours or even 10’s of minutes are unacceptable.

3) Specific

Specificity of contaminant identification can be obtained in two ways:

a) Analyze for a specific molecule or organism.

b) Analyze generally across multiple orthogonal dimensions via mathematical analysis of the data from multiple sensors.

Detectors of both types exist, so the value of comparing the two approaches is evident. Given the immense number of possible contaminants that could be put into a water distribution system, looking for specific molecules would require an immense number of sensors. There are so many problems associated this approach that it quickly become untenable. The approach of using a manageable set of orthogonal sensors faces only the difficulty of obtaining sufficient information to apply pattern recognition methods that can differentiate between contaminants or classes of contaminants. One advantage of using an orthogonal set of different sensors is that it becomes nearly impossible to find some contaminant that goes by all sensors unnoticed.

4) Reproducible

An EWS must be reproducible to be trustworthy. Reproducibility can be demonstrated via testing with actual contaminants (for example: Aldicarb, Anthrax culture, Cyanide, Fluoroacetate, Nicotine, Ricin, Sarin, VX)

5) Low false negatives, low false positives

If a system can be blind to certain classes of chemicals (for example, not visible in the Ultraviolet 254 spectrum) then it represents an opportunity to contaminate the system without an alarm. Systems with multiple types of sensors are difficult if not impossible to fool, so that kind of system will normally produce some kind of alarm upon contamination, and false negatives will be very unlikely.

False positives can come from two sources: random system noise, and insufficient information during the analysis of event data. Given that systems are not noise-free, and there will always be some degree of insufficient information, false positives are bound to happen. The ultimate question is the determination of the number per time in a given installation (since noise is site-specific) and the acceptable frequency. False positives caused by noise can be reduced by the proper choice of alarm threshold according to ROC curves.

False positives caused by imperfect information will happen in an EWS using an inferential method. Given that a multi-parameter system implies an inferential method that can give false positives, and that an actual contamination event is a very rare occurrence, it should be recognized that there will be more false positives then actual positive events. Thus, any detection must be considered presumptive until follow-up testing verifies or denies the detection.

6) Qualitative

Contaminants can be named if the EWS is specific, but people using the information may not have sufficient training to recognize the nature of the contaminant found from its name. The system should assign a general classification category for simple clarification.

7) Quantitative

Ideally, an EWS would provide enough information to give some quantitative information about any contaminant presumably found in the water system. Such information could be useful in determining the threat level during a given episode, and also might be useful if treatment of people or infrastructure become necessary.

8) Sensitive

Clearly an EWS must be sensitive to contaminants present in harmful amounts, but quoting a simple minimum detection limit can be misleading. Contaminants change water chemistry and those changes can be analyzed as part of an alarm process that generates a Trigger signal. Changes are seen against a background of noise, or natural fluctuations in measured parameters. Accordingly, one cannot address sensitivity unless it is relative to a site's noise properties. Fortunately, military needs have produced a useful tool for simply stating a detector's capability: the Receiver Operating Characteristic (ROC) curve. Such curves allow an operator to select a detection alarm threshold according to local noise characteristics. Note that this allows an operator to increase sensitivity, accepting the higher probability of a false alarm when it is suspected that contamination is more likely. Figure 1 shows a representation of ROC space.

A ROC curve is a graphical representation of the trade off between the false negative and false positive rates for every possible cut off. Equivalently, the ROC curve is the representation of the tradeoffs between sensitivity (Sn) and specificity (Sp). By tradition, the plot shows the false positive rate on the X axis and 1 - the false negative rate on the Y axis. You could also describe this as a plot with 1-Sp on the X axis and Sn on the Y axis. A good diagnostic test is one that has small false positive and false negative rates across a reasonable range of cut off values. A bad diagnostic test is one where the only cut offs that make the false positive rate low have a high false negative rate and vice versa.

> An evaluator is usually satisfied when the ROC curve climbs rapidly towards upper left hand corner of the graph. This means that 1- the false negative rate is high and the false positive rate is low. We are less happy when the ROC curve follows a diagonal path from the lower left hand corner to the upper right hand corner. This means that every improvement in false positive rate is matched by a corresponding decline in the false negative rate. You can quantify how quickly the ROC curve rises to the upper left hand corner by measuring the area under the curve. The larger the area, the better the diagnostic test. If the area is 1.0, you have an ideal test, because it achieves both 100% sensitivity and 100% specificity. If the area is 0.5, then you have a test, which has effectively 50% sensitivity and 50% specificity. This is a test that is no better than flipping a coin. In practice, a diagnostic test is going to have an area somewhere between these two extremes. The closer the area is to 1.0, the better the test is, and the closer the area is to 0.5, the worse the test is. The remainder of this paper will focus on the development and real-world utilization of an alternative ROC curve method that is more suitable to continuous monitoring situations.

DISCUSSION OF A NEW METHOD FOR DETERMINING ROC CURVES

The classical ROC curve is plotted parametrically. That is, from False Alarm Rate as a function of Trigger Level threshold versus Hit Rate as a function of Trigger Level threshold. An example is shown in Figure 2.

This format works quite well for discrete measurements, such as in test kits but may not be optimized for continuous monitoring technologies. With this in mind, the authors, in collaboration with US Military Personnel at the Army Corps of Engineers Research Laboratory and Edgewood Chemical and Biological Command, have developed an alternate presentation that is more useful when the operational analysis is continuous in time. This alternate presentation makes it easy for the system operator to select the Trigger Threshold to balance trigger sensitivity against the frequency of triggers caused by system noise.

In the alternate presentation, the False Alarm Rate function is expressed as Mean Time Between False Alarms versus Trigger Threshold. The Hit Rate function is translated into the amount of agent, expressed as the Percent of LD50 (1 liter of water for a 70 kg person), required to give a 100% hit rate, versus Trigger Threshold. See figure 3.

This format allows the user to select the Trigger Threshold based on desired trigger sensitivity versus the acceptable time between false alarms due to process noise. The blue markers are points of different Trigger Threshold value (increasing to the right).

To plot the classical ROC curves, we need the curves for:

False Alarm Rate versus Trigger Threshold

Hit Rate versus Trigger Threshold

To plot the alternate form of ROC curve, we need the curves for:

Mean Time Between False Alarm versus Trigger Threshold

Trigger Amount (%LD50) versus Trigger Threshold

HIT RATE VERSUS TRIGGER THRESHOLD CURVE

The HR curve for the classical ROC curve format is derived from a Monte Carlo simulation with a run of 1000 trials for each point on the HR curve. An agent is selected, which also defines the LD50 concentration for the agent. The lab test data for that agent at a known concentration are entered into the spreadsheet. These data are the five deviations from baseline (one per water quality parameter) produced when that concentration of agent is introduced into drinking water of the same type as used to define the FAR curve.

The statistics for the process water + measurement noise levels for each parameter are defined. Rather than use the standard deviation for the entire 5000 point population, 4970 values for the standard deviation over 30 minutes run time were calculated for each parameter and the typical standard deviation values used to define the measurement + noise statistics.

The resolution of the measurement devices is defined.

A Trigger Threshold Level, and a Dose value are entered.

The spreadsheet then calculates 1000 independent points where the process has been dosed, and noise added via independent random number generators. The noise figures are max values observed during 30-minute intervals of reference data (with no agents or upsets present). The probability density function for each parameter noise component was selected pessimistically to be a flat distribution between the extreme values of the noise component. The flat distributions used could be considered worst-case representations of the noise.

The measurement values are then rounded to represent the quantization noise found in the analog to digital conversion electronics of the system. Pessimistic values for quantization are used.

Those vector values are then processed by the Trigger Algorithm to produce a Trigger Signal for each vector. The values are then compared to the given Trigger Threshold values in the Monte Carlo simulation to obtain the number of hits per 1000 trials. The number per 1000 is taken as the Hit Rate percent.

Hit Rate (%) = 100*Number of hits/1000

Monte Carlo simulations are done for a number of Trigger Levels, generating a HR curve.

A sample Hit Rate versus Trigger Threshold curve is shown in figure 4.

The FAR values at those selected Trigger Threshold levels are calculated according to equation {1}.

With the Hit Rate data, and the FAR data, a ROC curve for the selected agent concentration can then be plotted.

Other concentrations can be selected and the process repeated to obtain a family of ROC curves at different agent concentrations. Table 1 shows typical data with this approach.

The ROC curves can then be plotted for the agent at the given concentrations. Figure 5 is an example for the agent Sodium Fluoroacetate, at three concentrations. Note that most of the markers for the 1 percent curve are covered by those for the 0.5 percent curve.

TRIGGER AMOUNT (%LD50) VERSUS TRIGGER THRESHOLD CURVE

For the modified format ROC curve, the Trigger Threshold is set at a selected value. The dose is then adjusted to find the dose value where the hit rate becomes 100%. A new Trigger Threshold value is selected and process repeated. This process defines the Dose versus Trigger Threshold curve. See figure 6. This is plotted parametrically with the MTBFA versus Threshold curve to generate the modified format ROC curve. See Figure 7.

DETERMINING ROC CURVES FOR DATA FROM SITES ON MILITARY BASES

The US Army Corps of Engineers Research and Engineering Laboratory (CERL) under a Cooperative Research and Development Agreement (CRADA) with Hach Homeland Security Technologies have worked for the past several years to evaluate early warning systems in real world deployment scenarios on military bases. As well as helping to develop theories of operation and response, these deployments have resulted in a large body of data from a variety of sites for which ROC curves have been calculated.

In the system evaluated in this study, once each minute, the EDS algorithm processes five water quality parameter readings: pH, conductivity, chlorine, turbidity, and Total Organic Carbon. The algorithm calculates the deviations of those parameters from the process baseline, numerically scales those five deviations and then calculates a single Trigger Signal via a distance measure. The Trigger Signal is compared to a user selectable Threshold level. If the Trigger Signal is greater than the Threshold, the system is said to be in alarm, indicating significant deviation from baseline conditions in the water. U.S. Patent 6,999,898

EDS systems with the above described algorithm were tested at two military installations (a BASE in the southeast and a FORT on the east coast) and at a municipal site in a large Midwestern city where data were collected for many months. In addition to providing the on-line monitoring function, the EDS at each site provided data that could be used for two parallel analyses: 1) Receiver Operating Characteristic (ROC) curve analysis via a Monte Carlo simulation in a spreadsheet, and 2) simulation of agent additions to the water at each site via superposition of agent deviation data on the site data, followed by playing the resultant data files through the EDS algorithm. The results from the Monte Carlo simulation were compared to the simulation with actual site data over time.

MONITORING SITES AND THEIR CHARACTERISTICS

The water quality at the municipal site is very consistent, with low variance in the water quality, hence the data from the site are very easy to analyze. This site was included in the study because it provides data that should provide the best EDS performance. Analysis results from these data demonstrate the upper limit of performance that can be expected from the algorithm.

The water quality at the FORT is more variable, but still controlled. This is a more typical site.

The water quality at the BASE is much more variable than typical, with pH and chlorine levels being uncontrolled. There is also some blending of waters from two sources. These changing conditions make for a more challenging analysis. Analysis results from these data demonstrate lower performance, as expected.

Analyzing data from sites that could be characterized as easy, medium and difficult gives a picture of the range of performance that the EDS can deliver. Analyzing only easy or difficult data would not present a broad picture of capability.

SITE DATA STATISTICS

The ROC curve analysis used in this study requires statistical analysis of the noise content of the parameter signals. Months of data were available from each site, from which, nine continuous days of representative data were selected from each site. It was important to select data sets that did not contain any abnormal events because these data sets were used to derive ROC curves. The background data set for ROC curve derivation must not contain abnormal, maintenance or calibration events, otherwise the statistics will be distorted and an incorrect presentation made.

The parameter signals for all the sensors at a site were statistically characterized over the selected data set, and those statistical values were input to the ROC curve analysis spreadsheet. The parameter values included not only the variation in the water, but also any noise or errors contributed by the sensors in normal operation.

FALSE ALARM RATE ANALYSIS OF THE TRIGGER SIGNALS

A statistical characterization of the Trigger Signal values for the individual baseline data sets is needed for a ROC curve analysis. The data collected from the sites provided the Trigger Signal values associated with the raw parameter data.

Having the Trigger Signal data, it was possible to produce curves of False Alarm Rate versus Threshold setting. Note that this is the alarm rate attributed to process and instrument noise with no agents or abnormal events present. For the non-classical ROC method used in this study, the Alarm Rate was converted to a more useful measure: Mean Time Between Alarms. For example, if there were 6 alarms in 12 months, the MTBA would be 2 months per alarm. MTBA curves for the BASE and FORT sites are shown in the graph in figure 8. The BASE shows a much lower MTBA than the FORT because of the BASE's higher parameter noise content, and more variable parameter signals.

MONTE CARLO ANALYSIS

A Monte Carlo analysis was done in a spreadsheet program to generate ROC curves. This method of analysis takes much less time to do than simulation runs with site data. It also permits "what if" analysis such as exploring the benefits of tighter control at a site (reduced variation) for parameters such as pH and chlorine.

In the Monte Carlo analysis, the deviations of the process parameters caused by an agent addition were combined with random noise from the process + sensors. In the analysis, the amount of added agent can be varied. The resultant Trigger Signal is calculated and compared to the selected Threshold value to see if the system would be in alarm.

In this study, three plausible threat agents: the pesticides Aldicarb and Oxamyl, and Cyanide. Dose concentrations in the study were expressed as a percent of LD50, the dose that would kill 50% of a typical population. The LD50 values, mg/kg in a liter were:

Aldicarb 1 mg/kg/L, Cyanide 6.4 mg/kg/L, Oxamyl 5.4 mg/kg/L

The actual dose was based on that for a 70 kg person.

To simulate a statistically large sample, 10,000 samples of random noise according to site statistics were added to the parameter deviations, and the Trigger Signals calculated.

Suppose the amount of agent added brings the Trigger Signal close to the Threshold value. It may occur that because of the noise, the EDS alarms 4,391 times, and does not alarm 5,609 times at the given dose and Threshold setting. The agent dose for the simulation could be increased until 10,000 alarms occur. It can then be stated that the probability of detection at that dose is at least 99.99% - the worst case being 10,000 alarms out of 10,001 trials. In this study, detection is considered a failure if the detection rate is not 100%. For example, detection rates of 70% are considered ethically unacceptable. Thus, the dose value that overcomes all of the noise in the system to generate an alarm 100% of the time can be determined.

By doing this analysis at various Threshold values, the values for a curve Detection Concentration versus Threshold can be obtained. The graph in figure 9 is for the case of Cyanide at the BASE. The ROC curve is then plotted parametrically from that curve and the curve of MTBA versus Threshold. The ROC curve for cyanide at the BASE is shown in figure 10.

ROC curves of this type are practically more useful than the classical probability-based curves. With this form of the ROC curve, the Threshold setting of the EDS can be set according to easily understood parameters: concentration represented as %LD50 and desired mean time between alarms. The ROC curve defines the system sensitivity at a given MTBA. Managers can opt for a long MTBA, or increased sensitivity. The corresponding MDL can be read from the ROC curve. The curve of MDL concentration versus Threshold setting then gives the associated Threshold setting for the EDS. Figure 11 shows the Aldicarb ROC curves for the BASE and FORT. This graph shows that the larger signal variation at the BASE pushes up the Minimum Detection Level for the same agent, compared to the MDL for that agent at the FORT.

The Monte Carlo analysis can also give the curve for Detection rate versus Concentration at a constant Threshold. The graph in figure12 shows the curve for Aldicarb at two sites. The

Threshold in this case is 1. The difference in slope is an indication of the difference of the noise content of the parameter signals at the two sites. The slope of the curve from the BASE data is lower because that site has more noise.

PLAYING A DOSED SIMULATION FILE THROUGH THE EDS ALGORITHM

Agent doses can be arithmetically superimposed across the site data files (157 doses per file, 20 minute dose duration) and those files can be played through the EDS to see if the results obtained from the EDS algorithm match those predicted by the Monte Carlo analysis.

Before playing the files through the EDS algorithm, the EDS can be “tuned” to the noise content of a site’s water and sensors by playing through the raw data set from the site without dosing. The data were played through in the simulation mode that runs 60 times faster than real time to speed the process. In actual on-line operation, the EDS analyzes data as it arrives in real time to do the tuning.

ANALYSIS COMPARISON

The Monte Carlo analysis predicts a Trigger Signal peak value for a given agent dose at a given site. The same dose applied to the site data also gives a mean value for the peak Trigger Signal over multiple injections. Comparison of the two methods shows that both give the same, expected result. The graph in figure 13 shows the percent difference between the predicted Trigger Signal peak from the Monte Carlo study, and the mean value of the Trigger Signal peaks found in the injection study. All difference values were less than 1.5%. (Ald => Aldicarb, CN => Cyanide, Oxamyl).

Table 2 shows the comparison of the Minimum Detection Limit (MDL) from the Monte Carlo study to the MDL found in the CSV dosing study. Units are % LD50 for the agent. The Monte Carlo analysis is conservative, predicting slightly higher MDL values than those found in the dosing study.

DETECTION TIME ANALYSIS

The nature of the EDS detection algorithm, and the sensors used for parameter measurement in this study allow for calculation of the Detection Time after contaminated water arrives at the sensor package. The Detection Time is a function of the response time of the sensor package, the time for the dose curve to reach its maximum value, and the peak Trigger Value of the dose. The graph in figure 14 shows results for the sensor set used in the studies, and two reasonable dose response times of 20 and 40 minutes. The Detection Time is then shown as a function of the peak Trigger Value.

For any agent contamination of consequence (Trigger Signal > 4), the time to detect will be 7 to 9 minutes. Detect times over 15 minutes are considered ethically unacceptable.

CONCLUSION

A number of projects have been initiated to develop and/or evaluate EWS. Earlier studies focused on ability of common sensors to detect noticeable changes in water quality when a contaminant was present. (:ETV 2005; EPA 2005; Hall et. al. 2005, Kroll 2006) After these early studies verified the efficacy of bulk parameter monitoring, later studies focused on the development and testing of event detection algorithms and their utilization in interpreting the generated data streams. (Yang et. al. 2006; Kroll and King, 2007; McKenna et. al. 2007; Umberg, 2008). While these studies utilized some of the criteria for evaluation listed above including ROC curves none have concentrated on all of the criteria. Even those that did take into account such factors as ROC false alarm rates and time to response failed to establish meaningful goals that would indicate that these systems are ready for real world deployment. Factors such as concentration needed to detect, time to alarm, detection of the full contaminate range, etc. when evaluated were not held to specifications that would be protective of human health.

The systems that were deployed in the CERL study were found to perform acceptably under all of the evaluation criteria set forth in this paper. They demonstrated dual use by detecting a number of operational and non-security events including, dead ends, contamination by aviation fuel, pumping schedule problems and others. These results show that the EDS under examination can detect low levels of threat agents in a few minutes at 100% probability of alarm. These systems were detect to warn type systems and lived up to that definition in speed (time to respond was a little as 3 minutes) and cost criteria (When compared to grab sampling protocols). Coverage characteristics were found to be adequate and communications were determined to be simple and secure.

All operational criteria including continuous operation and reliability were verified. The systems were under a service contract with the manufacturer whom preformed routine maintenance and calibrations. Any required service beyond what was covered in the service contracts was minimal.

The operational characteristics of the systems were deemed to meet the criteria. The calculation of ROC curves using the new methodology detailed in this paper for the sites indicate that this equipment is capable of detecting likely threats at levels and with a time to alarm needed by an effective early warning system The new and improved ROC curve method also can be used by operators to set threshold alarm levels to minimize unknown alarms while maintaining the desired level of sensitivity. It was also demonstrated that the Monte Carlo analysis for the determination of ROC curves and Minimum Detection Limits at a site, closely matches the results of simulations with agent-superimposed data sets.

Water distribution systems early warning systems are shown to be an effective means of enhancing the security of base water supplies and the new ROC curve method is shown to be an effective way of validating these systems with a tool that can be useful to operators in running and tuning these systems.

AUTHORS

Karl King. Chief Technologist. Hach Homeland Security Technologies. 5600 Lindberg Drive. Loveland, CO. 80539. KKING@hach.com

Dan Kroll. Chief Scientist. Hach Homeland Security Technologies. 5600 Lindbergh Drive. Loveland, CO. 80539. DKROLL@hach.com

REFERENCES

EPA. 2005. “Technologies and Techniques for early Warning Systems to monitor and Evaluate Drinking Water Quality: A State-of-the-Art Review.” US Environmental Protection Agency Office of research and Development National Homeland Security research center Research report.

ETV. 2005. “Multi-parameter Water Monitors for Dsitribution Systems”. A series of EPA Environmental Technology Verification reports performed by Batelle. http://www.battelle.org/PRODUCTSCONTRACTS/etv/verifications.aspx#W12

Hall, John et al. 2007. "On-line Water Quality Parameters as Indicators of distribution System Contamination." *Journal of the American Water Works Association*. Vol. 99 Issue 1.

Kroll, Dan. 2006. "Safeguarding the Distribution System: On-Line Monitoring for Security and Enhancing Operational Performance" *Journal of the New England Water Works Association*. June 2006.

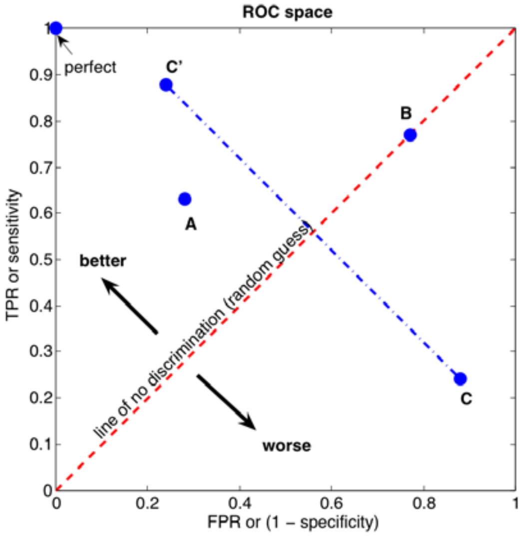
Kroll, Dan and King, Karl. 2007. "Operational and Laboratory Verification Testing of a Heuristic On-Line Water Monitoring System for Security" International Journal of High Speed Electronics and Systems. Vol. 17, No. 4 December.

McKenna, S. A. et al. 2006. "Testing Water Quality Change Detection Algorithms" in Proceedings of the 8th Annual Water distribution Analysis Symposium, Cincinnati, OH, August 27-30th.

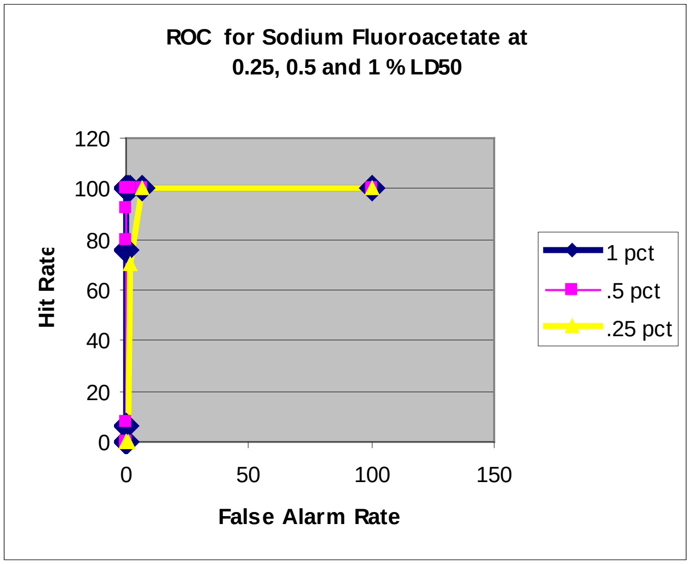
McKenna, S. A. et al. 2007. "Event Detection From Water Quality Time Series" In Proceedings of The World Environmental and Water resources Congress 2007: Restoring Our Natural Habitat. American Society of Civil Engineers.

Umberg, Katie. 2008. "Evaluation of Water Quality Event Detection Systems Deployed at the First Water Security Initiative Pilot Utility" Presentation at 2008 AWWA Water Security Congress. Cincinnati, OH April 6-8th.

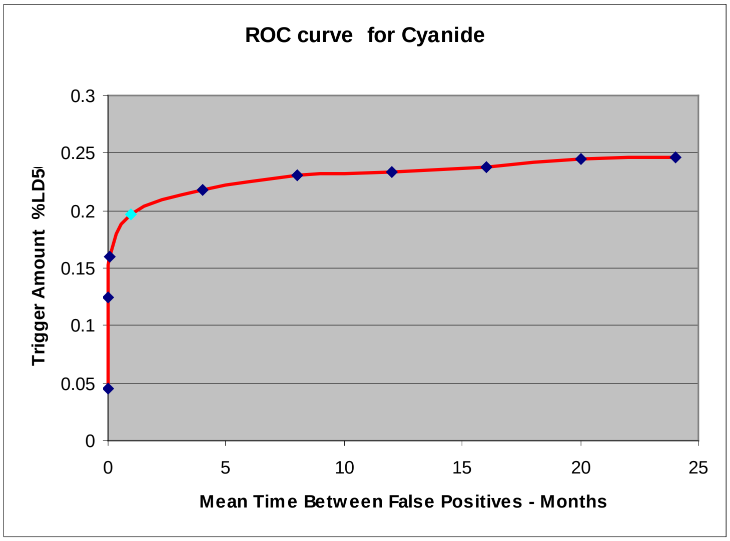
Yang, Jeffrey, et al. 2006. "Adaptive Monitoring to Enhance Water Sensor Capabilities for Chemical and Biological Contaminant Detection in Drinking Water Systems". In Proceedings of Optics and Photonics in Global Homeland Security II. Theodore Saito and Daniel Lehrfeld editors. SPIE Publications.



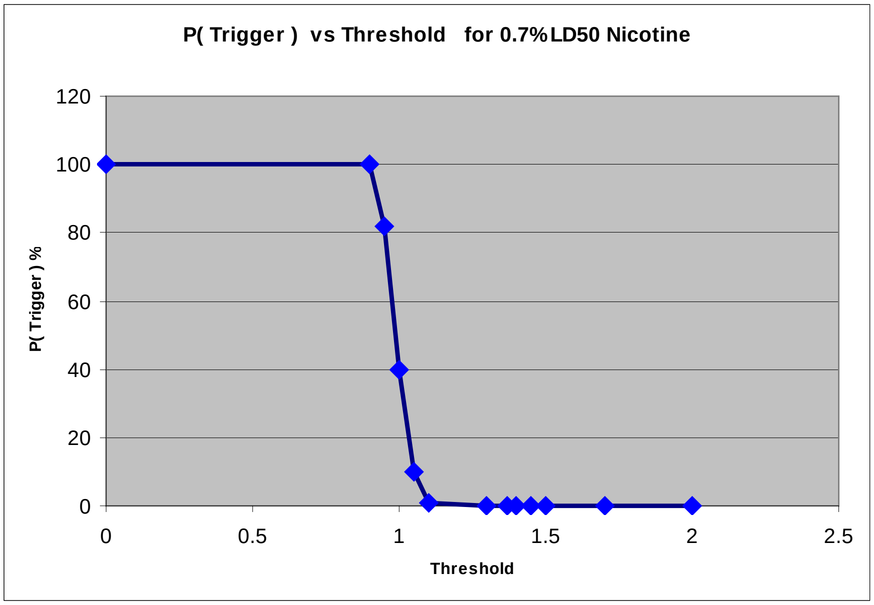
1. Figure 1. Classic ROC curve space.



2. Figure 2. Classic ROC curve format.



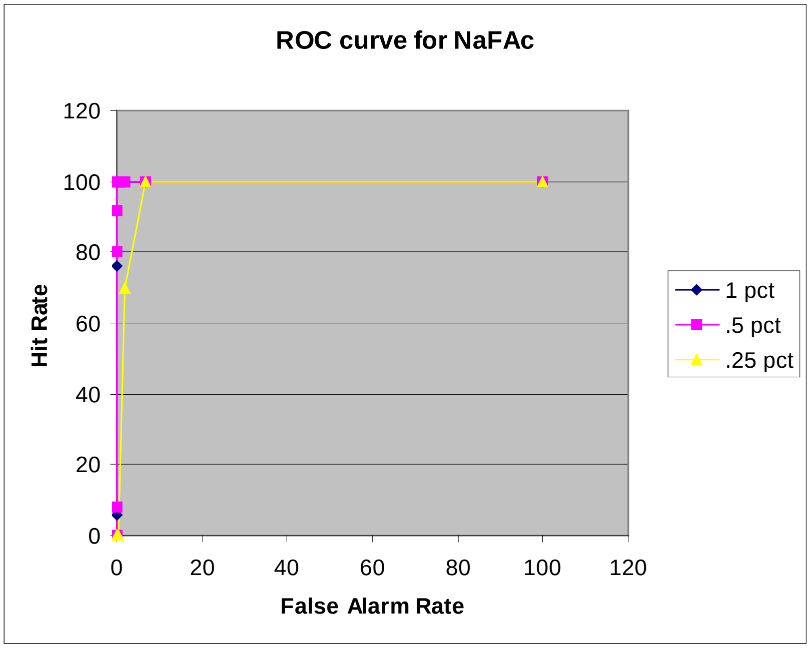
3. Figure 3. ROC Curve format for a continuous system.



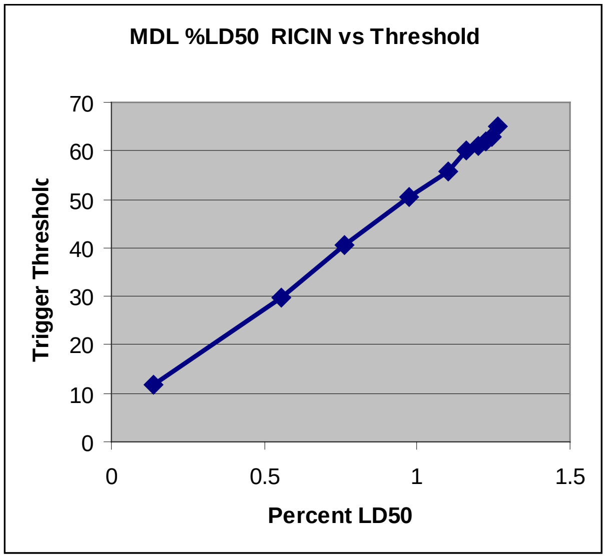
4. Figure 4. Hit Rate versus Trigger Threshold example

5. Table 1

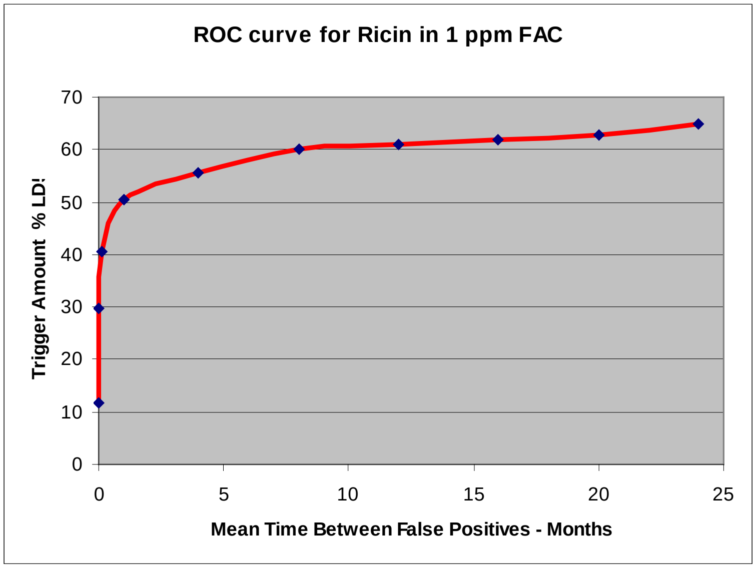
6. Trigger	7.	8. Hit Rate at X % LD50		
9. Level	10. FAR	11. 1 pct	12. .5 pct	13. .25 pct
14. 0	15. 100	16. 100	17. 100	18. 100
19. 0.2	20. 6.727948	21. 100	22. 100	23. 100
24. 0.3	25. 1.745115	26. 100	27. 100	28. 70
29. 0.4	30. 0.452653	31. 100	32. 100	33. 0
34. 0.5	35. 0.11741	36. 100	37. 100	38. 0
39. 0.55	40. 0.059797	41. 100	42. 92	43. 0
44. 0.6	45. 0.030454	46. 100	47. 80	48. 0
49. 0.7	50. 0.007899	51. 100	52. 8	53. 0
54. 0.8	55. 0.002049	56. 100	57. 0	58. 0
59. 1	60. 0.000138	61. 100	62. 0	63. 0
64. 1.2	65. 9.27E-06	66. 76	67. 0	68. 0
69. 1.3	70. 2.41E-06	71. 6	72. 0	73. 0
74. 1.4	75. 6.24E-07	76. 0	77. 0	78. 0
79. 1.5	80. 1.62E-07	81. 0	82. 0	83. 0
84. 2	85. 1.9E-10	86. 0	87. 0	88. 0
89. 3	90. 2.62E-16	91. 0	92. 0	93. 0



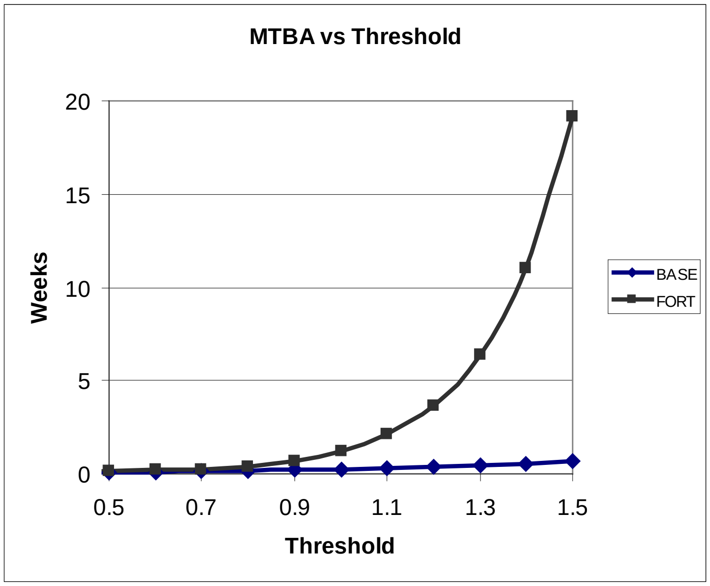
94. Figure 5. ROC curve for Sodium Fluoroacetate



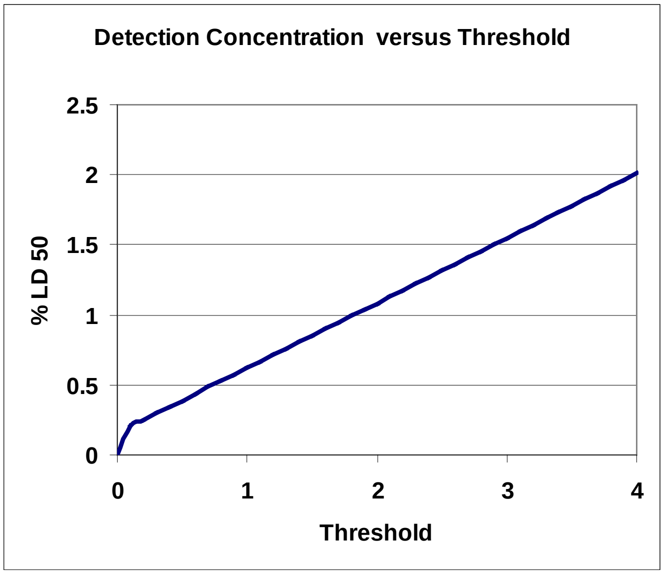
95. Figure 6. MDL versus concentration



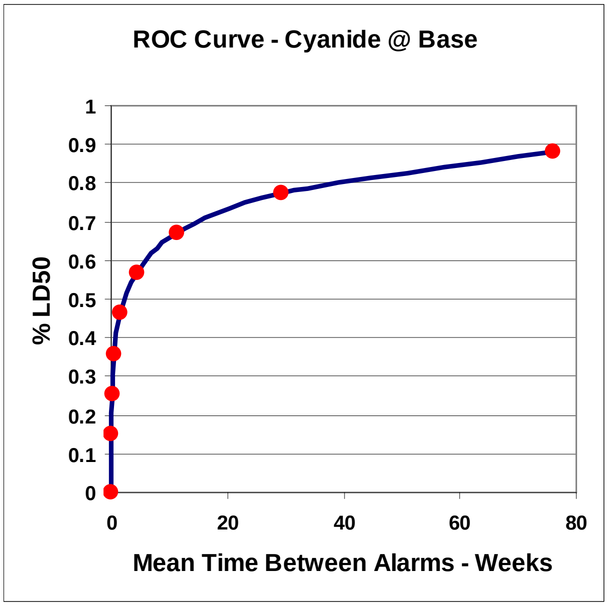
96. Figure 7. ROC curve for Ricin



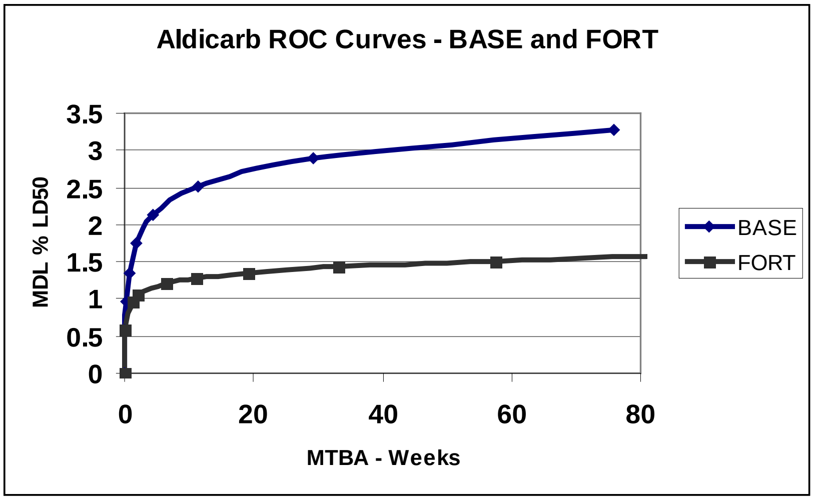
97. Figure 8. MTBA curves for the BASE and FORT.



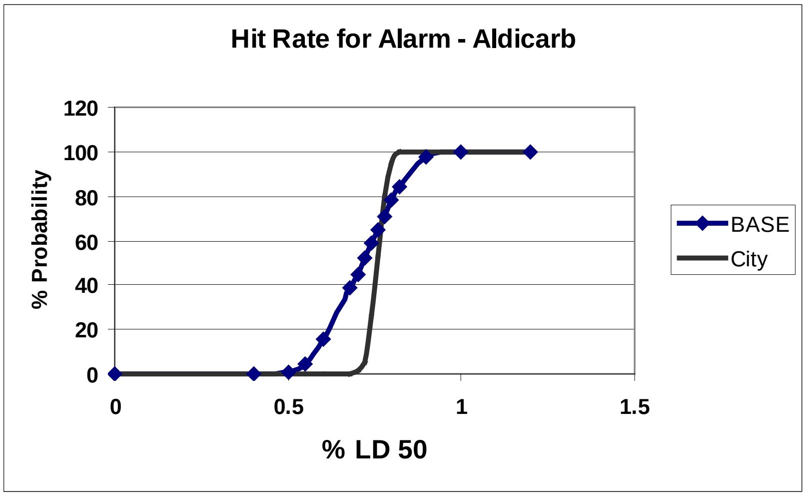
98. Figure 9. Detection concentration versus threshold for cyanide at the BASE.



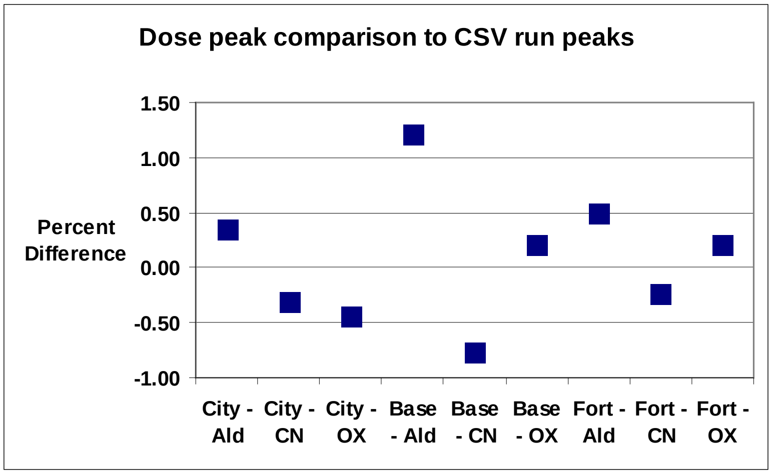
99. Figure 10. ROC curve for cyanide at the BASE.



100. Figure 11. Aldicarb ROC curves for the BASE and FORT.



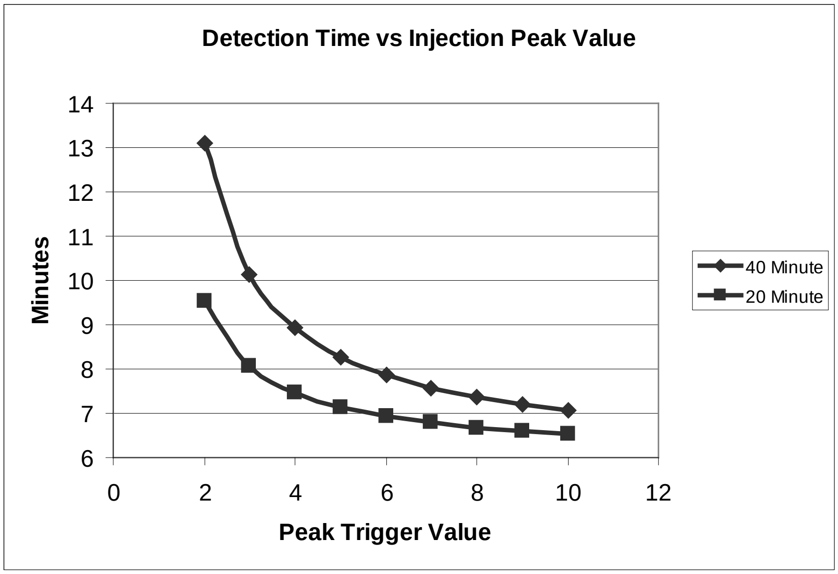
101. Figure 12. Detection rate versus Concentration for Aldicarb at 2 sites.



102. Figure 13. The percent difference between the predicted Trigger Signal peak from the Monte Carlo study, and the mean value of the Trigger Signal peaks found in the injection study.

103. Table 2. Comparing Minimum Detection Limit (MDL) from the Monte Carlo study to the MDL found in the CSV dosing study.

CASE	ROC MDL	CSV MDL	Difference
City - Aldicarb	0.81	0.779	0.031
City - Cyanide	0.523	0.499	0.024
City - Oxamyl	0.219	0.217	0.002
Base - Aldicarb	0.974	0.971	0.003
Base - Cyanide	0.609	0.498	0.111
Base - Oxamyl	0.258	0.229	0.029
Fort - Aldicarb	0.972	0.959	0.013
Fort - Cyanide	0.6	0.527	0.073
Fort - Oxamyl	0.257	0.228	0.029



104. Figure 14. Detection times.