EPA/600/R-12/735 | December 2012 | www.epa.gov/ord

United States
Environmental Protection
Agency

Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE) Water Security Module

Office of Research and Development
National Homeland Security Research Center

Disclaimer

The U.S. Environmental Protection Agency (EPA) through its Office of Research and Development's National Homeland Security Research Center, funded and managed the research described herein under Contract EP-C-06-074 to John Hopkins University Applied Physics Laboratory. Although reviewed by the Agency, it does not necessarily reflect the Agency's views. Official endorsement should not be inferred. EPA does not endorse the purchase of sale of any commercial products or services.

Mention of trade names or commercial products in this document does not constitute endorsement or recommendation for use.

Address questions concerning this document or its application to:

Cynthia Yund, Ph.D.
National Homeland Security Research Center
Office of Research and Development (NG 16)
U.S. Environmental Protection Agency
26 West Martin Luther King Drive
Cincinnati, OH 45268
(513) 569-7779
yund.cynthia@epa.gov

Acknowledgments

Contributions of the following individuals and organizations to the development of the ESSENCE Water Security Module are gratefully acknowledged.

United States Environmental Protection Agency (EPA)

- **Office of Research and Development, National Homeland Security Research Center**
  Kathy Clayton (formerly with NHSRC)
  John Hall
  Terra Haxton
  Regan Murray
  Cynthia Yund
- **Office of Water, Water Security Division**
  Steve Allgeier
  Chrissy Dangel
  Dan Schmelling

Johns Hopkins University Applied Physics Laboratory

Steven Babin
Howard Burkom
Sheri Happel Lewis
Mohammed Hashemian
Charles Hodanics
Rekha Holtry
Wayne Loschen
Zaruhi Mnatsakanyan
Liane Ramac-Thomas
Michael Thompson
Joseph Skora
Svenson Taylor
Richard Wojcik

City of Milwaukee, Wisconsin

- **Milwaukee Water Works**
- **City of Milwaukee's Health Department**

City of Seattle, Washington

- **Seattle Public Utilities (SPU) Water System**
- **Public Health - Seattle & King County**

National Capitol Region (Washington, D.C. and surrounding areas)

- **Washington Suburban Sanitary Commission (WSSC)**
- **Montgomery County Department of Health and Human Services (Maryland)**
- **Prince George's County Health Department (Maryland)**

Contents

List of Figures

List of Tables

1 Abbreviations

API	Application Programming Interface
B(B)N	Bayesian (Belief) Network
CDC	Centers for Disease Control and Prevention
CPT	Conditional Probability Table
CVA	Cerebral Vascular Accident
DO	Dissolved Oxygen
ED	Emergency Department
EDS	Event Detection System
EPA	U.S. Environmental Protection Agency
ESSENCE	Electronic Surveillance System for the Early Notification of Community-based Epidemics
ftp	File Transfer Protocol
GI	Gastrointestinal
GUI	Graphical User Interface
HPC	Heterotrophic Plate Counts
HSPD	Homeland Security Presidential Directive
https	Secure Hypertext Transfer Protocol
IDSN	Intelligent Decision Support Network
IDSS	Intelligent Decision Support System
JHU/APL	The Johns Hopkins University Applied Physics Laboratory
MWW	Milwaukee Water Works
ORP	Oxidation-Reduction Potential
OTC	Over-the-Counter
OUA	Operational Utility Assessment
pD	Probability of Detection
pFA	False Alarm
ROC	Receiver Operating Characteristics
Sftp	Secured File Transfer Protocol
SPU	City 3 Public Utility
SVM	Support Vector Machine
TIA	Transient Ischemic Attack
TOC	Total Organic Carbon
WAN	Wide Area Network
WSSC	Washington Suburban Sanitary Commission
XML	Extensible Markup Language

EXECUTIVE SUMMARY

ESSENCE Water Security Module

Homeland Security Presidential Directive 9 (HSPD 9) directed the Environmental Protection Agency to develop robust, comprehensive, and fully coordinated monitoring and surveillance systems for the water sector. By its authority under section 300i-3 of the Safe Drinking Water Act (42 USC section 1434) and to address the monitoring and surveillance requirements of HSPD 9, EPA intends for the Water Security Initiative (WSi) to build on existing Agency and utility efforts to enhance the ability to detect and respond to contamination threats. WSi serves as a demonstration project, for designing and implementing an effective contamination warning system (CWS) in a drinking water distribution system. A CWS should encompass monitoring technologies and detection strategies, combined with enhanced public health surveillance to collect, integrate, analyze, and communicate information to provide a timely warning of potential water contamination incidents and initiate response actions to minimize public health and economic impacts. The success of a CWS depends on the ability to effectively integrate these components and analyze the resulting information in a timely manner to inform response actions that can substantially reduce the potential consequences of a contamination incident.

The task for John Hopkins University Applied Physics Laboratory (JHU/APL) was to build a module for the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) syndromic surveillance system to include water quality data with health indicator data for the early detection of a drinking water contamination event. ESSENCE is a web-based syndromic surveillance system designed for the early detection of disease outbreaks, suspicious patterns of illness, and public health emergencies. ESSENCE incorporates traditional and non-traditional health indicators from multiple data sources (emergency department chief complaints, and over-the-counter medication sales, and results of laboratory tests). Data are categorized into syndromes and sub syndromes to detect aberrations in the expected level of disease. Automated statistical algorithms are run on each (sub) syndrome and alerts are generated when the observed counts are higher than expected.

The purpose of the contract was to develop a prototype system for surveillance of certain water quality parameters and water distribution system operating conditions that may be correlated in space and/or time with public health events possibly related to drinking water contamination. Algorithms were developed and implemented to identify triggers that would initiate investigations by public health epidemiologists and/or water utility personnel. Typically, such investigations would begin by simply looking further within this surveillance system for indicators and data suggesting possible explanations of the anomaly that resulted in the trigger. If an anomaly of public health concern could not be ruled out, the appropriate public health and water utility officials would have health and water indicators and data on which to base a decision as to whether a more exhaustive investigation were warranted. Visual and analytical tools were developed to aid an investigation and foster collaboration between the health departments and the water utilities

This project produced a novel prototype warning system that employs water quality sensor clustering plus newly defined sub syndromes in existing hospital emergency room visit data in a Bayesian Network (BN) analyses. The techniques used address the challenges of synthesizing results from disparate data types with different data rates and complex environmental and operational responses into a warning system that can provide the user with a measure of the likelihood of occurrence of either an intentional or unintentional drinking water contamination event based on the particular combination of recent data anomalies. The surveillance tool developed for this project is a hybrid of BN implementations and outputs of alerting algorithms applied to both health indicator and water quality data.

For health indicator data, existing ESSENCE algorithms are applied to syndromes and sub regions modified for waterborne disease detection (both microbial and chemical) and for the distribution system characteristics such as site locations. For water quality anomalies, the algorithms were adapted from the CANARY event detection software developed at Sandia National Laboratories in collaboration with EPA.

The BN is designed as a hierarchy of networks, with the top-level fusion analysis of outputs from 1) water quality data and 2) health indicator data designed to detect diseases potentially caused by contaminated drinking water, as depicted in Figure 1.

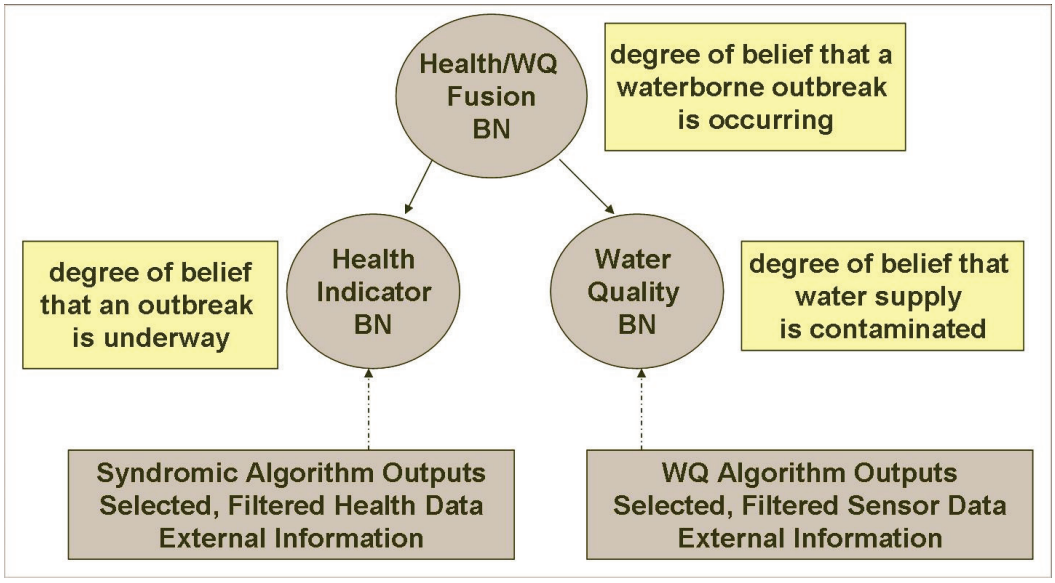


Figure 1 Top-Level Design of Bayesian Network to Detect Waterborne Disease

Visual and analytic tools specific to a water event were developed in the first two phases of this project, then refined and replicated in a separate major metropolitan city.
The benefits of having access to these data include:

- Data on specific water quality sensors and/or sources
- Summary visualization of a possible event for both water utility and health personnel
- Early recognition and warning of a possible drinking water contamination event
- Rapid confirmatory analysis
- Open lines of communication between public health and water utilities
- Prompt response for confirmatory testing and mitigation.

The ESSENCE enhancements were introduced in three pilot locations for both the water utilities and public health departments. Feedback from the users indicated a value in the tool, but also suggested areas of improvement.

The current system is a prototype developed for a proof-of-concept demonstration. The system has not undergone extensive testing on the software implementation side and the detection and fusion algorithms have only been tested on a limited set of simulated data.

Expansion to other cities is optional and feasible when users a) are willing to learn a new system, b) accept and trust an unfamiliar method for detecting anomalies in their data, and c) invest up-front time to develop appropriate region definitions for water quality and health indicators for their city.

Currently the ESSENCE public health software is used in states and cities across the nation: including a) states - Indiana and Missouri, b) counties - Miami-Dade, Broward, Hillsborgh, FL; Cook County, ILL; San Diego, Los Angeles, Santa Clara, CA; Tarrant County, TX, and King and Pierce, WA (Seattle) and c) the national capital region (Washington D.C., Virginia, Maryland) The United States Department of Defense uses a version of the surveillance software for military bases throughout the world.

Cost estimate for implementation in current ESSENCE cities is provided in the final report.

The project has demonstrated the value of collecting temporal and spatial water quality data for the water utilities. Modification of the existing ESSENCE software package can specify health events that may be caused by microbial and chemical contamination of drinking water.

This page intentionally left blank.

2 Background

Although drinking water-related disease outbreaks are relatively uncommon in the United States (U.S.), the Centers for Disease Control and Prevention (CDC) report 13–14 outbreaks per year, affecting an average of approximately 1000 people annually.[1] Throughout history, there have been numerous instances of the deliberate poisoning of drinking water supplies or denial of drinking water service to an enemy.[2] Deliberate contamination of a water system remains a feasible form of attack because it could easily be performed in a covert manner, and the resulting health effects, as well as potential panic, could be significant. As the threat of terrorism within the U.S. persists, it is therefore prudent for water systems to evaluate their infrastructure vulnerabilities, to mitigate risks where possible, and to prepare to respond in the event of an incident. In December 2003, the President of the U.S. issued Homeland Security Presidential Directive 7 (HSPD-7)[3], which designated the U.S. Environmental Protection Agency (EPA) as the agency responsible for protecting the nation's drinking water infrastructure. On January 30, 2004, the President issued HSPD-9[4], which directed the EPA to develop robust, comprehensive, and fully coordinated monitoring and surveillance systems for water quality. The EPA Water Security Initiative is piloting contamination warning systems at several U.S. water utilities to test their feasibility. These pilot programs are integrating online water quality monitoring, sampling and analysis, automated public health surveillance systems, consumer complaint monitoring, and enhanced physical security monitoring.

In a 2003 U.S. General Accounting Office report[5] documented the results of interviews with water experts. The water distribution network was identified as the most vulnerable component of the U.S. drinking water infrastructure. Recognizing this vulnerability, water utilities are increasingly monitoring water quality parameters in their distribution systems. To improve assessment that data anomalies might be related to the occurrence of a waterborne disease outbreak, The Johns Hopkins University Applied Physics Laboratory (JHU/APL), in a collaborative project with EPA, tested the feasibility of a warning system prototype that integrates disparate types of data from a number of diverse sources. Traditional water quality parameters can be measured by online water quality sensors and by analysts evaluating routinely collected grab samples (to be described subsequently). Community health data identified through automated public health surveillance systems may include early signs and symptoms of diseases of water origin. Such public health syndromic surveillance systems use automated feeds of pre-diagnostic health data in near real-time to identify changes in community health status, facilitating notification to those charged with investigation and follow-up of potential public health crisis. JHU/APL has previously developed such a system called the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE).

Because ESSENCE is capable of analyzing, fusing, and displaying disparate types of data from diverse sources, the EPA established a contract with JHU/APL to

combine this system with water quality data provided by participating public water utilities. The enhanced syndromic surveillance system is designed to provide an additional indicator facilitating detection and response to drinking water contamination incidents. Additionally, coordination and communication between drinking water utilities and local public health officials has been and continues to be a critical factor in the design, implementation, and maintenance of this contamination warning system.

The purpose of the original contract was to develop a prototype system for surveillance of certain water quality parameters and water distribution system operating conditions that may be correlated in space and/or time with public health events possibly related to drinking water contamination. Algorithms were developed and implemented to identify triggers that would initiate investigations by public health epidemiologists and/or water utility personnel. Typically, investigators would begin by simply looking further within this surveillance system for indicators and data suggesting possible explanations of the anomaly that resulted in the trigger. Visual and analytical tools have been developed that would aid this type of investigation and foster collaboration between the health departments and the water utilities. If a public health concern could not be ruled out, the appropriate public health and water utility officials would have indicators and data for public health status and water quality on which to base a decision as to whether a more exhaustive investigation was warranted to determine the possible relationship between drinking water quality and the disease/illness present in the community.

During the original contract period, water quality data and health data from two locations were analyzed, and detection and fusion algorithms were developed to detect water contamination-related health events. Extensive analysis of the different water quality data sources was performed and Bayesian Networks (BNs) were built for finding gastrointestinal (GI)-related health events and possible water contamination, and for fusing information from the different data sources.[19] Additional tasks covered in this report include the development of an additional BN to detect chemical or neurological health events, analysis of water data from a new location, and improvements to the user interface.

3 Algorithms and Evaluation

The objectives of this task included analyzing historical data from a third independent water utility, developing one new health BN for finding health events related to chemical contamination of the drinking water supply, testing the algorithms and BNs with the historical data, and proposing alert levels for both the water utilities and public heath epidemiologists.

3.1 Algorithms

This section describes the methods used to detect anomalies in the water sensor and public health data time series and to fuse those detection results to find possible drinking water contamination events. The results of the time series detection algorithms were the inputs to the BNs that were used to fuse the disparate data sources.

3.1.1 Anomaly Detection Approach

The most difficult analytic challenge facing modern public health surveillance has been the automated recognition of outbreak scenarios from nonspecific data. Many algorithms have been applied to streams of data recorded during population care-seeking. Multiple authors have estimated the sensitivity of these algorithms, i.e., the probability of an alert given the occurrence of an outbreak with an estimated data effect.[6,7,8] However, the health monitor, the professional using an automated system, routinely faces the opposite question: given an alert or a combination of alerts, what is the probability that an outbreak has begun? The monitor needs a system sensitive to outbreak indications to corroborate early clinical evidence, to help judge the nature and extent of an outbreak, and, thus, to inform investigation and response decisions. The monitor's routine decision requires a prior outbreak probability, just as the estimation of the positive predictive value of a clinical test requires the disease prevalence along with test sensitivity and specificity.[9] However, for many surveillance objectives, there is no reliable estimate of the prior outbreak probability. Historical datasets labeled with outbreak signals are generally unavailable, especially for multiple data types. Data simulations can help with sensitivity estimation, but simulations that can estimate a prior outbreak probability in a population were unavailable for this study, if they exist at all.

The objective of this project was to provide an automated early detection capability, using pre-diagnostic health care data together with water quality data, for a disease outbreak caused by naturally-occurring or deliberate contaminated drinking water, whether they are contamination. (These data types were described above in Section 3.1.1.)

There are some particular challenges for the current objective of waterborne disease detection. Waterborne outbreaks are rare, but as noted in Section 1, they do occur with a significant disease burden. However, applying detection algorithms to multiple types of water quality and health data can lead to nuisance alerting that can render a surveillance system useless. The challenge is to combine the available information types to gain sensitivity to these rare outbreak events without excessive warning flags. To achieve this capability, indicators of abnormal data must be weighted with knowledge of the likely effects of contamination and resulting disease.

Another challenge was the fusion of data types. One EPA project requirement was to design a capability that could be

transferred to different geographic regions. However, the usefulness of water quality data depends on the operations and data-sharing policies of the local water utility. Similarly, the value of the public health data feeds depends on the detail and reliability of information available from care-providing institutions and on the population coverage of these institutions. Accurate fusion of these data types for a coherent picture of population health depends on management of these data issues.

To meet these challenges, JHU/APL developers chose a BN analytic approach for the capability to detect a waterborne disease outbreak using the locally available, diverse set of data streams of varying reliability and relevance. This approach had been applied in a previous project for the D.C. Health Department to analyze asthma exacerbations using a combination of clinical patient record data and air quality data.[10]

3.1.2 Data Fusion Approach

The BN is often represented as a directed acyclic graph, a diagram with nodes and directed edges. Five types of nodes are referred to in this report:

- input nodes (representing the measured data),
- output nodes (any nodes whose values are of interest to the user),
- intermediate nodes (connect input to output nodes),
- parent nodes (dependent on child node values, arrows point away from these nodes in the graphs)
- child nodes (arrows point toward these nodes in the graph)

The output nodes represent hypotheses that can be true or false (or some probability of being true or false) based on conditional probability tables (CPTs) (described below) and the findings at the input nodes (Figure 1). The BN implementation used in this study adopted the convention of setting findings at the input nodes as either true or false and probabilities at the output nodes as continuous values. The connected nodes are linked by conditional dependencies that can be based on expert reasoning and/or data-derived inference. BNs thus incorporate data history and expert knowledge. They have been applied to use disparate types of evidence to determine likelihoods of significant data anomalies.[11,12]

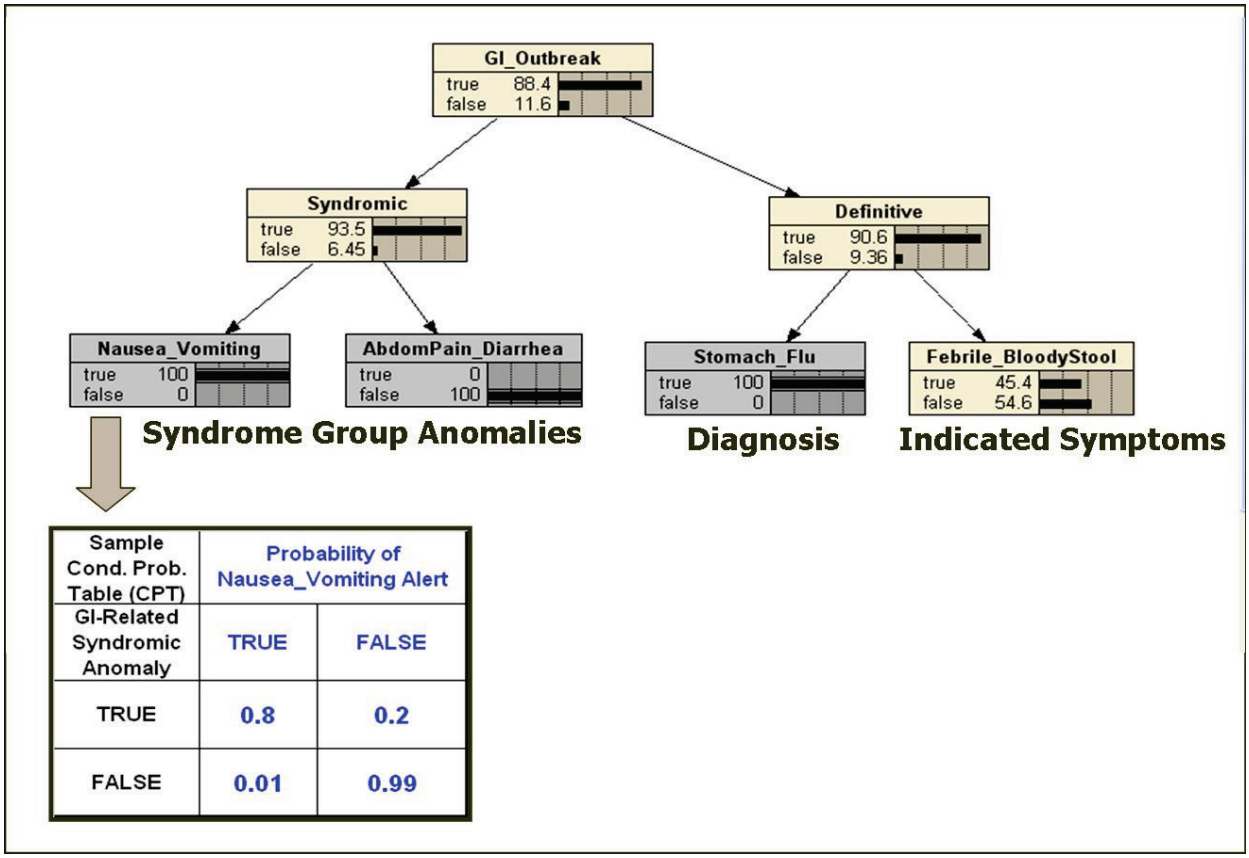


Figure 1 Example of How a Bayesian Network Can be Used in Probabilistic Decision-Making

The BN approach was chosen for the following reasons:

1. The BN paradigm provides an umbrella for managing multiple data types and rates, and for combining statistical and non statistical evidence.
2. BN outputs are not restricted to "black box" alerts. Probability values at any node within the BN structure can be displayed so that a user can see which nodes contribute most to an alert.
3. BNs can include nodes that encapsulate the effects of unstructured evidence such as reports of waterborne outbreaks in areas neighboring the monitored system or intelligence reports of terrorist activity.
4. The BN probabilistic structure can accommodate both continuous and discrete data, multiple data rates, and missing or sparse values.

A common criticism of BNs is the computational demands of large networks. Scaling issues are avoided in the current design by summarizing raw data information with statistical algorithm calls. A BN used to assess a given threat obtains processed inputs from a suite of alerting algorithms applied to the collection of available data and possibly from additional non-statistical information. This additional information may be "hard" evidence such as a sensor detection of a specific pathogen or "soft" evidence such as increased internet activity or an intelligence report of a suspected attack on a region. BN modeling is amenable to both evidence types and can

weight them in a transparent way according to their credibility.

Functional delegation drastically reduces the number of fused random variables and, thus, the BN size. It also avoids the need for computationally expensive agent-based fusion (based on characteristics and behaviors of individual people) and management of raw data streams. This approach also allows modular efficiency gains, such as decomposing a model into smaller BNs with assumptions of independence where very weak dependence is suspected.

BNs are one possible choice from an array of data fusion approaches. While rule-based approaches concentrate on expert knowledge and newer approaches, such as support vector machines (SVMs), are powerful tools for multidimensional data inference, the BN approach readily employs both evidence types. In addition, by fusing degrees of belief, BNs can fuse results from arbitrary algorithms, SVMs, or expert systems that may solve a portion of the problem.

The surveillance tool developed for this project is a hybrid of BN implementations and outputs of alerting algorithms applied to both health indicator and water quality data. For health indicator data, existing ESSENCE algorithms are applied to syndromes and sub regions modified for waterborne disease detection and for the distribution system characteristics such as site locations. For water quality anomalies, the algorithms were adapted from the CANARY event detection software[17] developed at Sandia National Laboratories in collaboration with EPA.[13] (CANARY is an event detection tool that integrates data from water quality sensors in real-time and predicts whether the recorded water quality changes are anomalous or otherwise unexpected.) Adaptations were needed to manage data rate and reliability problems. For example, output from a substantial number of sensors was intermittent, so to use whatever current data were available, stable baselines were formed by pooling baseline data from sensors whose outputs were similar in phase and magnitude and whose locations reflected similar water source characteristics.

As depicted in Figure 2, the BN is designed as a hierarchy of networks, with a top-level fusion network whose inputs are the outputs from two detection networks, a water contamination detection network and a health indicator network which is designed to detect diseases potentially caused by contaminated drinking water. The water contamination detection network is a BN implementation of a contamination event detection system (EDS) that can include and combine multiple EDS indicators.

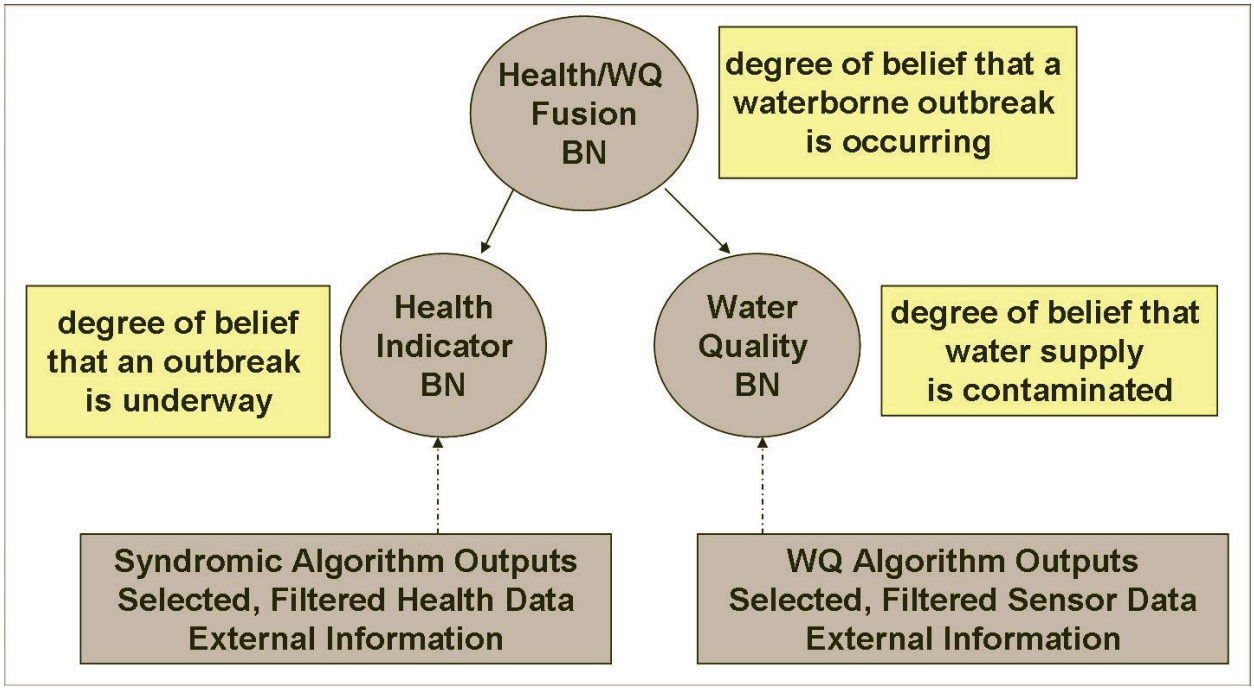


Figure 2 Top-Level Design of Bayesian Network to Detect Waterborne Disease

WQ = Water Quality, BN = Bayesian Network

Each component network is in turn a probabilistic hierarchy whose basic inputs are the results of statistical alerting algorithms applied to individual data streams, either counts of syndromic patient records on the health indicator side, or water parameter measurements on the water quality side.

For the health BN, algorithm inputs are daily counts of records of patient visits to emergency departments filtered by age group and chief complaint category. The chief complaint filtering is based on categories of health database queries chosen for the indication of waterborne disease. JHU/APL follows the practice of referring to these categories as syndromes and subsyndromes.[14]

Inputs to the water quality BN are outputs of EDS contamination indicator algorithms. In turn, inputs to these algorithms are time series of normalized measurements of water quality parameters such as free chlorine concentration. Because of relatively noisy variations, these water quality measurements are normalized using the time series mean and standard deviation. The choice of these measurements and their normalization was guided by informal discussions with EPA scientists and water quality engineers who participated in this project and by analysis of available data.

These BN tools were designed for operation on a daily basis, though more frequent monitoring is possible depending on the input data rate. The top-level node (Figure 2) gives the likelihood of a waterborne outbreak based on recent data, and layered visualizations provide transparency so that users at the local public health department `or water utility can see the basis of the BN indication. Details for the water quality, health, and integrated networks are supplied in Sections 3.1.3, 3.1.4, 3.1.5, and 3.1.6, respectively.

3.1.3 Water Quality Bayesian Network

The structure of the Water Quality BN is shown in Figure 3. Detection results of the water-quality data-detection algorithms are fed to the input nodes of this BN on a site-by-site basis. Input nodes were included in the BN based on typical measurements taken at both grab sample and continuous sites. The same BN structure can be used for both types of data. Unavailable information related to a particular input node can remain unknown and will not affect the ability of the BN to infer probabilities associated with the output nodes. Anomalies are found based on individual measurements and/or multiple anomalies across all input measurement types.

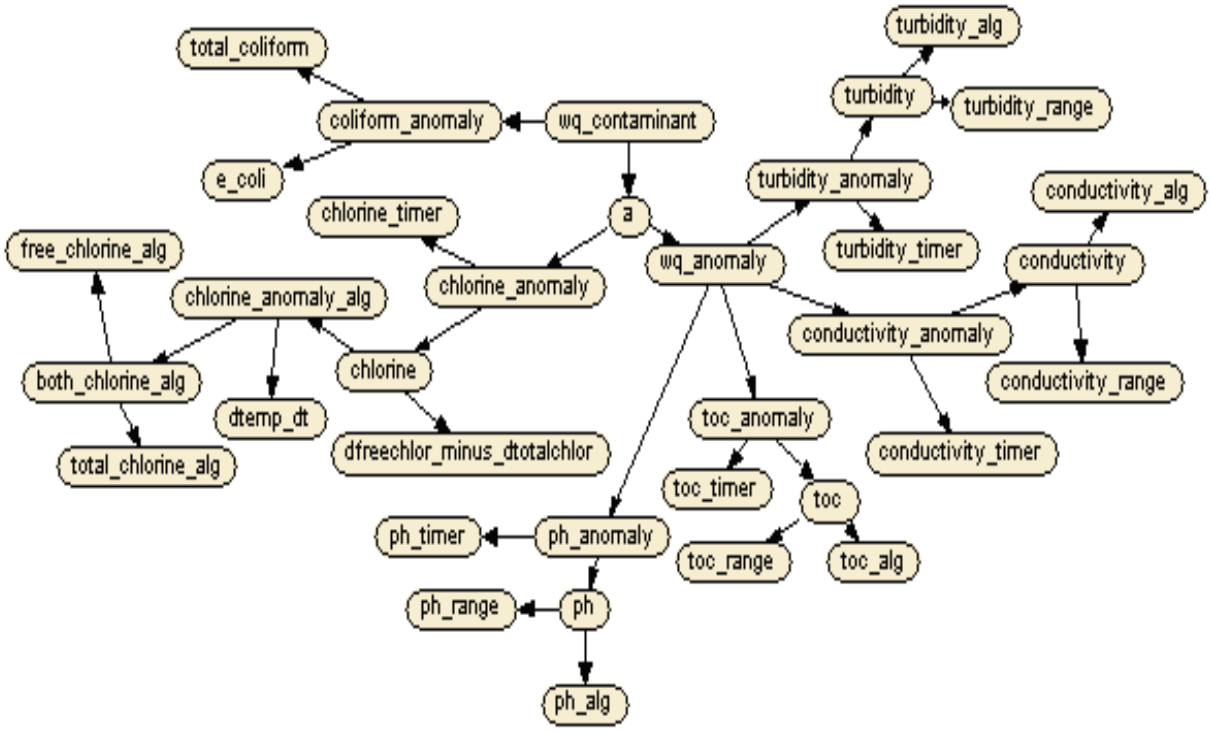


Figure 3 Water Quality Bayesian Network

The input measures used for this BN include *Escherichia coli* (*E. coli*) and total coliform (both are binary values that indicate absence or presence), free chlorine, total chlorine, pH, total organic carbon (TOC), conductivity, and turbidity. Anomalies are found in the non-binary value measurements (e.g., free chlorine) in two different ways. A detection algorithm is run on the time series data to look for unusual drops or increases; the measured time series values are also compared to typical ranges to determine if the current measurement is within range.

The goal of this design was to provide increasing probabilities of water contamination (at the 'wq_contaminant' node) when more than one type of water quality measurement has a high probability of an anomaly being present. The output and intermediate nodes that feed into the fusion BN include: 'water

quality_contaminant', 'coliform_anomaly', 'chlorine_anomaly', 'ph_anomaly', 'toc_anomaly', 'conductivity_anomaly' and 'turbidity_ anomaly.' Where the '(measurement type)_anomaly' nodes have higher probabilities of anomaly when something is detected in the measurement or when the measurement is outside of an acceptable range.

Outputs are dependent on performance of the anomaly detection algorithms. Therefore, the BN will not find an anomaly in the water data unless detections are passed in through at least one of the input nodes. A discussion of the Water Quality BN performance against a set of defined scenarios is given in Section 3.4.3.2.

3.1.4 Gastrointestinal (GI) Bayesian Network

GI-related illness is one of the syndromes monitored daily by public health departments. For automated GI-related illness monitoring, several public health departments use electronic syndromic surveillance systems. Such systems collect data from hospitals, other medical facilities, and over-the-counter (OTC) drug retailers. Hospital data, in most cases, are limited to chief complaints. Statistical anomaly detection algorithms will generate GI-related illness alerts if the number of visits with GI-related chief complaints exceeds the expected number. This causes large numbers of epidemiologically insignificant alerts because the definition of GI-related chief complaints is too broad.

The JHU/APL approach is to build a BN-based model that will estimate detection algorithm outputs and distinguish epidemiologically significant alerts from the ones that are just mathematical anomalies in the GI-related illness data. Chief complaint data for different age groups (Table 1) were queried and processed with statistical anomaly detection algorithms. The rationale for choosing these specific queries can be found in EPA Task Deliverable 1(b).[19] Age groups were selected based on default groupings used in ESSENCE. Naming conventions for the BN node age groups were chosen to provide unique names to each year grouping. Mappings of the named age groups to age groups in years for both the GI BN and Chem-Like/Neurological BN are as follows: infants/i (0-4 years), children/c (5-17 years), adults1/a1 (18-44 years), adults2/a2 (45-64 years), and elderly/e (65+ years). Each query was paired (mapped) to the specific node within the GI BN (Figure 4).

The node names and descriptions listed in Table 1 directly correspond to the node names in Figure 4. The first column, 'Chief Complaint Based ESSENCE Query,' contains the specific chief complaint terms that were queried to as inputs to the BN. A more detailed description of chief complaint queries is provided in Section 3.1.12. Note that sub-syndromes were not used in the design of the GI BN. The third column of Table 1 contains node names that correspond to the BN input nodes in Figure 4. For each day, the BN node will receive a "true" state if the anomaly was detected within its pair query during the past seven days. Nodes with the "true" states propagate the GI BN to recognize epidemiologically relevant patterns. The more complete the pattern, the higher probability of a true GI event. Additionally, the GI BN has intermediate nodes that provide the probability of the GI event within each of the five age groups.

Table 1 Selected Queries

Chief Complaint Based ESSENCE Query	Query Description	Input Node Name	Age Groups
diarrhea and vomiting	Note that the queries (in the first column) used for the input nodes of the GI BN are specific chief complaint queries and not sub syndromes. The input node names (in the third column) are simply variable names used in the BN to represent the queries.	dv	0-4, 5-17, 18-44, 45-64, 65+
diarrhea, vomiting, and fever		dvf	0-4, 5-17, 18-44, 45-64, 65+
abdominal pain and diarrhea		abd	0-4, 5-17, 18-44, 45-64, 65+
enteritis or gastroenteritis or stomach flu		enteritis	0-4, 5-17, 18-44, 45-64, 65+
bloody diarrhea and fever		bdf_all	0-65+
Specific diseases (ecoli, salmonella, etc.)		reportable_list	0-65+
fever and cough or fever and sore throat (influenza like illnesses)		ili	0-4, 0-65+

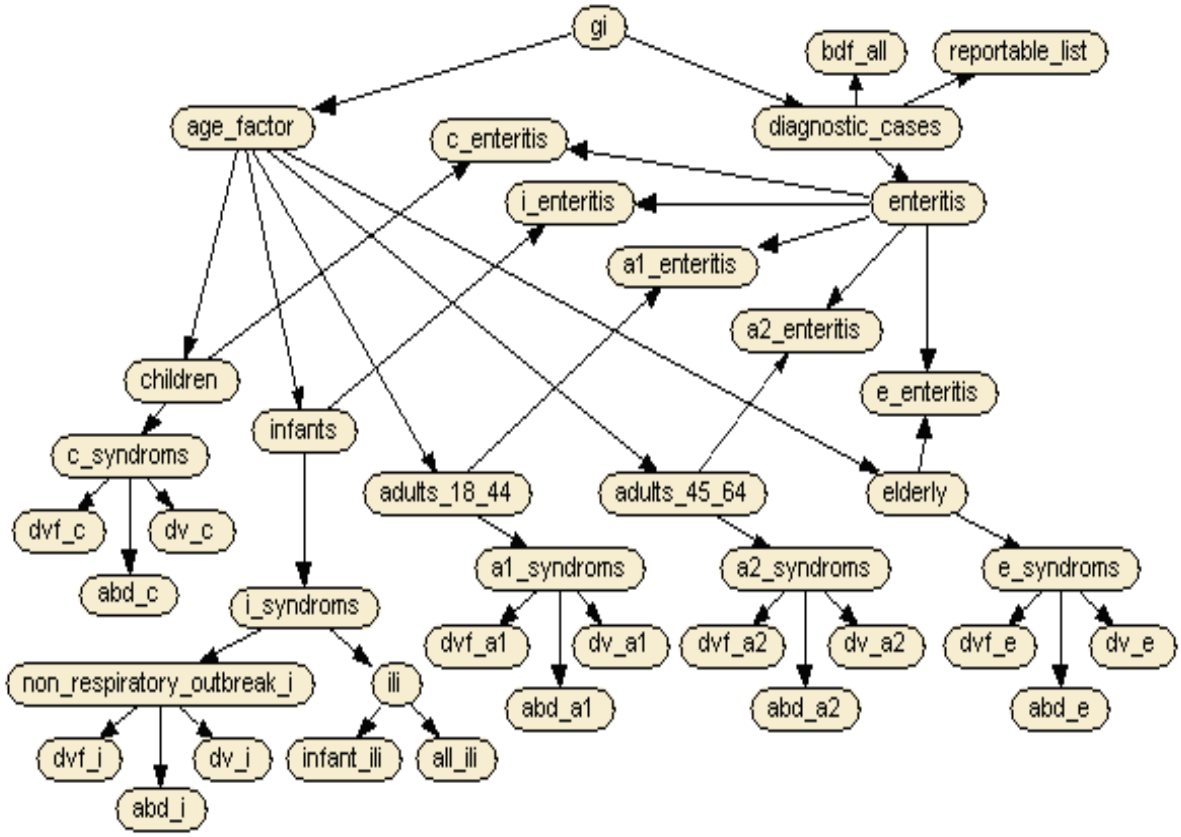


Figure 4 Gastrointestinal Bayesian Network Structure

Eleven GI outbreaks were simulated. Each outbreak involved 14–23 cases within a 2–4 day time period. Each case was represented by a randomly selected chief complaint. Content for the chief complaint was selected based on CDC's case definitions for GI-related diseases.[23] Artificial cases ("injects") were injected into the background data to create an outbreak. The background data were created using one year of archived real data that was free of outbreaks and the detection threshold at the 'gi' output node was set to a value of 0.25 or above. The model detected ten of the eleven injected outbreaks, with two false positive detections (alerts where no known events had occurred) during a two-year period.

3.1.5 Chemical Contamination/Neurological Bayesian Network

It is more difficult to design a BN to detect health events related to chemical contamination in a drinking water distribution system because there are no known events in the historical data to test against. Even so, looking for health events that are related to chemical contamination is an important piece of a water security alert system. Instead of trying to detect health events with specific chief complaint queries (as for the GI BN), special sub syndromes related to chemical contamination in drinking water were developed. The Chemical Contamination/Neurological BN was built taking into consideration the list of symptoms associated with exposure to

different contaminant classes provided by the EPA.[13]

The structure of the Chemical Contamination/Neurological BN is shown in Figure 5. This BN attempts to cover a wide range of chief complaints associated with chemical contamination. Three types of patterns pulled out of the chief complaint data include 1) specific neurological complaints ('neuro' nodes), 2) specific references to contaminated water ingestion ('reportable_list' node), and 3) patterns of chief complaints that may suggest ingestion of contaminated water ('gi,' 'uncommon,' and 'common_non_gi' nodes in Figure 5). Patterns in the third group must find a correlation between detections in either the GI complaints and other common non-GI symptoms or the GI complaints and uncommon non-GI complaints. A higher probability is assigned to the pattern of symptoms that includes the group of uncommon non-GI symptoms. Descriptions of the sub syndrome based queries used for the Chem-Like/Neurological BN are listed in Table 2. As with the GI BN, the queries were divided into five age groups. The node names and descriptions listed in Table 2 directly correspond to the node names in Figure 5. More details on how sub syndromes were chosen for this BN are given in Section 3.1.13.

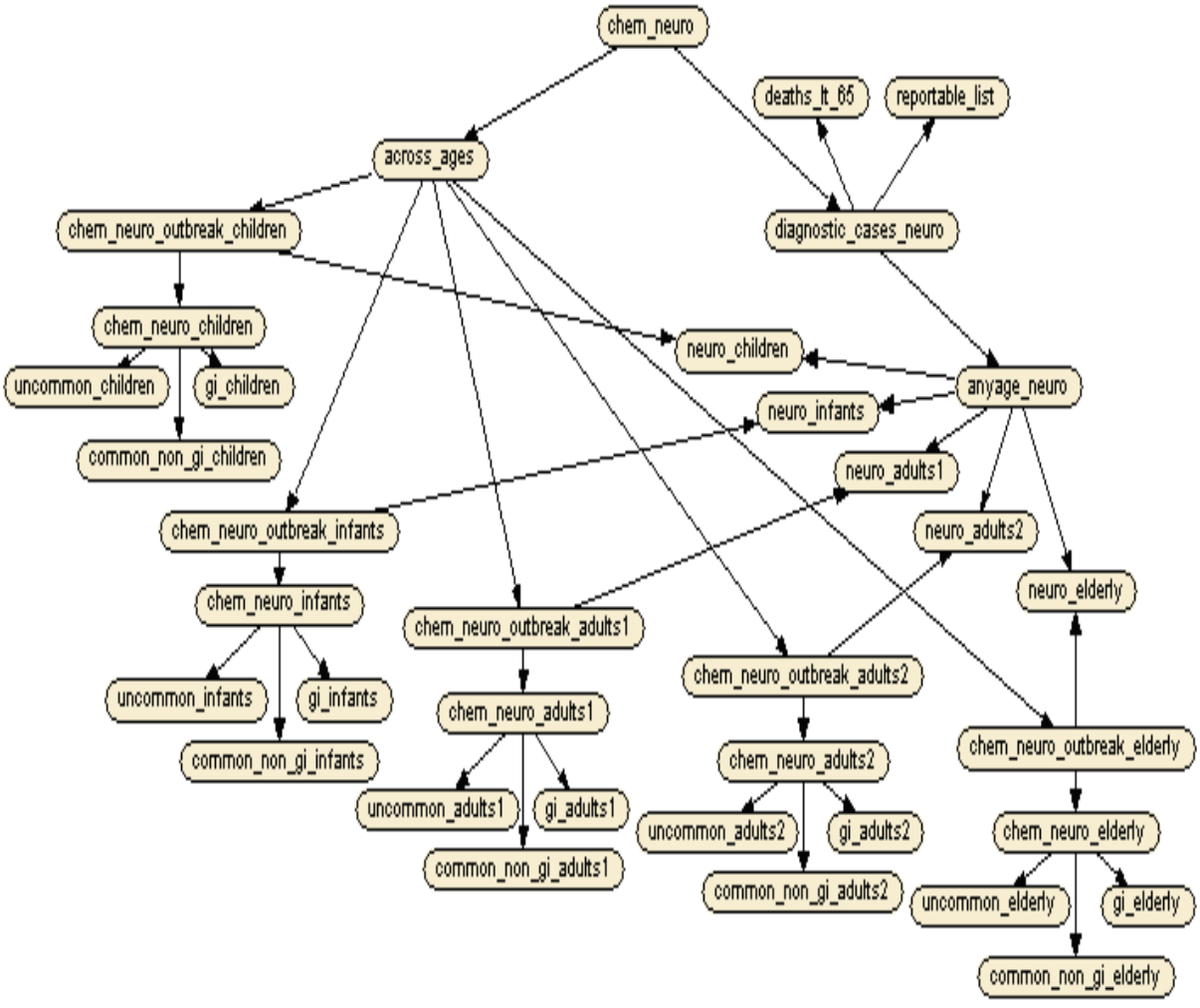


Figure 5 Chemical Contamination/Neurological Bayesian Network Structure

Table 2 Chemical Contamination/Neurological Bayesian Network Queries

Subsyndrome-Based ESSENCE Query Associated with input Node	Query Description	Input Node Name	Age Groups
NeurologicalEffects OR Confusion	Select neurological complaints. Excludes those related to stroke/ cerebral vascular accident (CVA)/transient ischemic attack (TIA). Excludes chronic conditions per symptoms from EPA contaminant classes; mostly cognitive deficits, loss of consciousness and seizure-related. Excludes impaired speech/swallowing.	neuro	0-4, 5-17, 18-44, 45-64, 65+
Dry mouth OR Dysphasia OR PinPointPupils OR Chloracne OR MucousMembraneErosion OR Hyperpigmentation OR IncreasedSaliva OR MetallicTaste OR SoreMouthGums OR SuscepToDiseases OR WateryEyes	Rare subsyndromes associated with EPA contaminant classes.	uncommon	0-4, 5-17, 18-44, 45-64, 65+
Chills OR Cough OR Dermatitis OR DifficultyBreathing OR Fever OR Headache OR RunnyNose OR SoreThroat OR Sweating OR SwollenThroat	Noisy non-GI (gastrointestinal) subsyndromes associated with EPA contaminant classes.	common	0-4, 5-17, 18-44, 45-64, 65+
Nausea OR Vomiting OR Diarrhea OR AbdominalPain	These are noisy GI subsyndromes with high false positives, but common to all EPA contaminant classes.	gi	0-4, 5-17, 18-44, 45-64, 65+
Death LT 65	Number of deaths for age groups less than (LT) 65 years.	deaths_lt_65	0-64
ContaminatedFluidExposure	These are atypical complaints associated with only contaminated fluid/water ingestion. Should be very rare.	reportable_list	0-65+

3.1.6 Fusion Bayesian Network

The Water Health Fusion BN structure is shown in Figure 6. Because the Fusion BN searches for patterns in the data that indicate a drinking water contamination health event, the highest level or fusion alert node of the BN should only have a high probability when anomalies are found in 1) the water data and 2) either the GI queries or the chemical/neurological queries. An anomaly in the water data without an anomaly in the health data or vice versa should not result in a fusion alert. The definition of high probability in this case is a debatable matter best left to experts who would use the system. The design of the BNs can accommodate adjustments to output node levels because the probabilities associated with each node can easily be modified to incorporate expert knowledge.

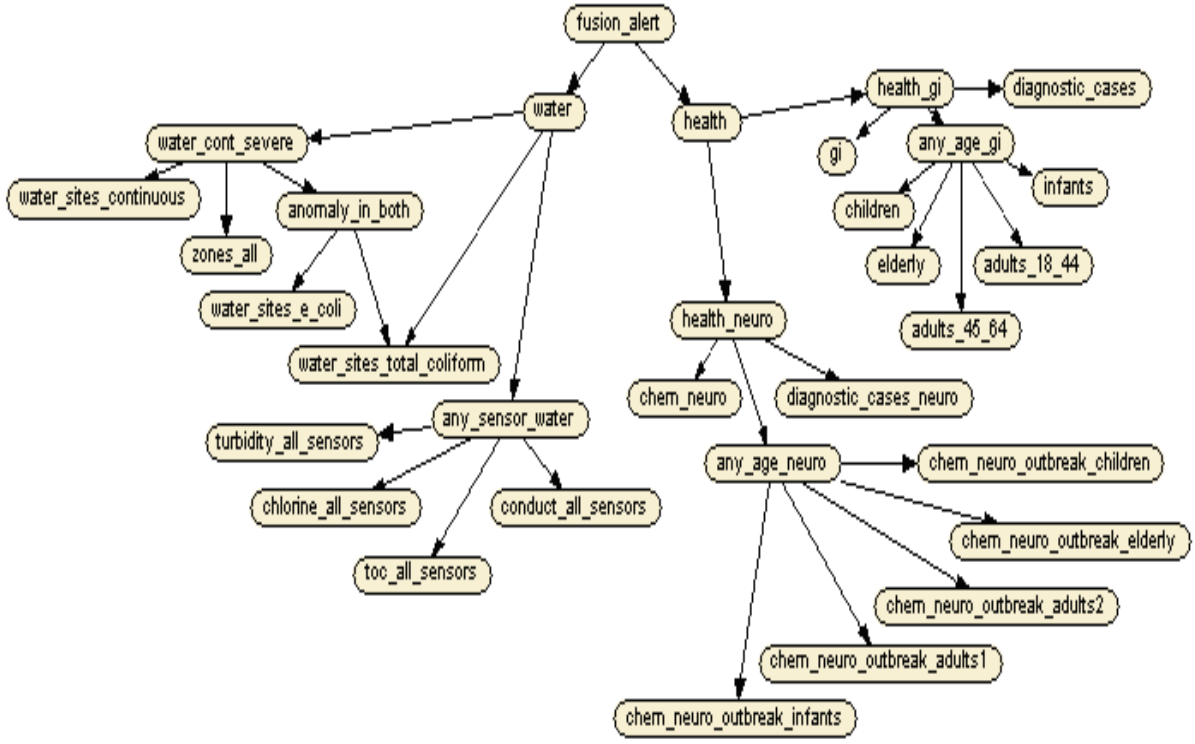


Figure 6 Water Health Fusion Bayesian Network

Patterns in both datasets consistent with a drinking water contamination health event are investigated by grouping anomalous findings in different ways. Descriptions of what types of anomalies are associated with the branches of the Fusion BN are described below.

The water branch of the Fusion BN searches for contamination as follows:

- Severe contamination related to a) anomalous water measurements over multiple sensor types and locations, b) anomalous water measurements in continuous monitor data at multiple sites, or c) positive *E. coli*
- Unacceptable number of positive total coliform results as

determined by the local water utility
- A pattern of anomalous water measurements for a specific sensor type at multiple locations

The health branch of the Fusion BN searches for patterns in the GI-related and Chemical Contamination/Neuro-related queries separately. The types of patterns the Fusion BN looks for in the health data include:

- High probability values at the top-level output nodes of the health BNs
- Anomalies in the health data across one or more age groups that may not show up at the top-level node
- Specific references to things such as reportable diseases, diagnostic cases, or unusual complaints

No specific alert levels are recommended at this time. Instead, the fusion alert values at the highest level output nodes are left as continuous probabilities with values between 0 and 1. Alert levels will probably be system/location dependent and interpretation can be left up to the user. After an appropriate amount of training, users will be able to associate fusion output levels with what they have found to be appropriate alert levels.

3.1.7 Types of Water Quality Data

Drinking water contaminants include microorganisms, inorganic and organic chemicals, toxins, and radionuclides. A variety of physical and chemical parameters (e.g., color, odor, turbidity, dissolved oxygen, temperature, conductivity, pH, alkalinity, coliform bacteria presence, disinfectant levels) can be measured to determine the suitability of water for human consumption. Wells, private water utilities, and public water utilities provide drinking water to the majority of the U.S. population. Many larger cities are supplied by a combination of these sources. According to the EPA publication *FACTOIDS: Drinking Water and Ground Water Statistics for 2007*[15], there are over 4000 community water systems in U.S. urban areas. A community water system is defined as a public water system that supplies water to the same population year round. About three-quarters of the community water systems in the U.S. use ground water (e.g., aquifers) as the water source, while the remainder use surface water (e.g., rivers, lakes). This project involves only community public water utilities that use surface water sources and have agreed to participate.

Water utilities use a variety of approaches in water treatment plants to ensure that the source water is treated prior to distribution for use. To protect drinking water from disease-causing organisms, or pathogens, water suppliers often add a disinfectant, such as chlorine or chloramine, to drinking water. A residual of these chemicals is maintained in the distribution system to prevent microbial re growth. Following treatment, contaminants can enter the distribution in a variety of ways. They may be released from biofilms on the inner surfaces of distribution pipes, or introduced into the distribution system through breaks, leaks, joints, or cracks. Finally, humans can introduce contaminants either accidentally or intentionally. However, drinking water is never sterile; ill health effects are rarely reported.

Water quality measurements including disinfectant levels vary among the different water utilities. The most common measurements taken by utilities are those of temperature, pH, total coliform bacteria presence/absence, and *E. coli* presence/absence, plus additional regulated

parameters. The following list includes descriptions of those common measurements that are addressed in this report:

1. **Chlorine** is one type of disinfectant commonly used and, once added to water, forms hypochlorous acid and hypochlorite ion depending on the pH of the water. Hypochlorous acid is the primary disinfectant form of chlorine. The chlorine residual level (i.e., chlorine residual = chlorine dose – chlorine demand) must be maintained at a specific level (above 0.2 mg/L). The normal maximum average residual EPA typically allows (with some exceptions) is 4.0 mg/L because of unwanted byproducts and health effects. Chlorine residual is typically measured at the water treatment plant discharge point and various points within the distribution system. Chlorine is the principal disinfectant used by two utilities that provided data during different phases of this project, System 1 and System 3.

2. **Monochloramine** (often generally called **chloramines**) is another type of disinfectant that may be used and is the type used in System 2. Chloramines are products of the chemical reaction between chlorine and ammonia. In 2002, about 20% of U.S. drinking water utilities used chloramines[24].

a. Chlorine and monochloramine have different benefits and drawbacks. Monochloramine has the advantage of creating fewer byproducts than chlorine and persisting longer. Using chlorine may result in the formation of disinfection byproducts. On the other hand, monochloramine is a weaker and slower-acting disinfectant than chlorine. Also, certain non-pathogenic bacteria can oxidize the ammonia to form nitrite and nitrate (nitrification) and this may result in chloramine depletion that could allow general bacterial re growth in the water.

b. Water analysis of chloramine systems is described by the terms "combined chlorine residual concentration" to denote chlorine in the form of chloramines and "free chlorine residual concentration" to denote chlorine in the form of hypochlorous acid and hypochlorite ion. System 2 measures only the combined chlorine residual and refer to it as only 'chlorine residual". The water will be re-sampled if the value falls below 0.3 ppm and will notify their water quality manager when it falls below 0.1 ppm. Because there is less of a concern with byproducts when using monochloramine and because of the potential for bacterial nitrification depleting monochloramine, utilities often try to maintain a higher residual (than with chlorine) within the distribution system.

3. **Temperature** may vary with the water source and its depth. Depending on the location and season, temperatures would range from near freezing (near 0 °C) to as high as 35 or 40 °C. Temperature is often measured within the distribution system because of its impact on general water chemistry. Many other parameters, including pH, conductivity, and dissolved oxygen, require temperature compensation to determine their values accurately. Also, higher ambient temperatures tend to cause chlorine to come out of solution (off gas), so a reduction in chlorine may

result from high temperatures rather than contamination.

4. **The pH,** a logarithmic measurement of the amount of hydrogen ion in the water, is commonly measured. It is indicative of its acid/base properties, with a lower pH being acidic and a higher pH being basic (a pH of 7 is neutral, neither acid nor base). The pH can affect chemical reactions in the water, including speciation of metals and other contaminants. Biodegradation of organic compounds may be reduced or accelerated depending on the pH. Chlorine readings obtained by membrane/ amperometric methods are dependent on pH. Chlorine is a more effective disinfectant at lower pHs Depending on the source water, pH values are typically between 6.5 and 9.5. Some utilities maintain higher pH value than the incoming source water as part of their water softening or corrosion control programs. Large changes in pH should not occur naturally because drinking water contains a variety of natural buffers.
5. **Total coliform** measurement is required by federal regulation and indicates the presence or absence of coliform bacteria. Coliform bacteria are naturally occurring bacteria that are usually non-pathogenic, although the category includes some pathogenic types. The presence of coliform bacteria in drinking water is used as an indicator of contamination. According to EPA standards, this indicator is of concern if more than 5% of the water samples per month test positive.
6. **Utilities are required by federal regulation to measure *E. coli*** a particular species of coliform bacteria. While it is typically non-pathogenic, there are strains of *E. coli* that can cause serious disease in humans. Therefore, the presence of *E. coli* in drinking water indicates more serious contamination than the presence of coliform bacteria. According to EPA standards, even one positive test result warrants concern. When a water utility finds a positive result for total coliform or *E. coli*, an investigation by the utility is immediately begun. Then, if necessary, public health authorities are contacted and drinking water alerts (e.g., boil water alerts) are issued to the public.
7. **Conductivity** (or **specific conductance**) indicates ionic compounds which include nutrients, pesticides, cyanide, or fluoride. It is measured by some utilities. The EPA does not regulate conductivity because it can vary widely depending on the source water (e.g., hard, brackish, salt water).
8. **Total Organic Carbon (TOC)** is an indicator of the amount of organic carbon in the water and frequently correlates with the chlorine demand during water treatment. TOC is seldom measured, especially within the distribution system. However ultraviolet (UV) absorbance measurements within the distribution system are becoming more prevalent as a TOC surrogate. The UV absorbance (usually at 254 nm) can be correlated to a corresponding TOC value adjusting for turbidity. The TOC depends upon the source water content and the type of treatment process, so a lot of variation exists among utilities. While this makes it difficult to define an upper limit, it probably should not be higher than 5 mg/L.

9. **Turbidity** is an indicator of suspended solids that might include pathogens and sediment. It is occasionally measured within the distribution system. Increased turbidity might result from a contamination event, but it could also result from system flushing activity or a water-main break. It is difficult to assess what is normal, but values higher than 1.0 should probably be investigated.
10. **Alkalinity** is a measure of water hardness, but it is seldom measured outside of the water treatment plant. Typical values vary depending on the source water. For the Washington, DC, area, typical values are about 35–98 ppm. Alkalinity is an indicator of the buffering capacity of the water and, therefore, the resistance of the water to changes in pH. Water with a higher alkalinity would be less likely to have large changes in pH than water with normal alkalinity.
11. **Heterotrophic Plate Counts (HPC)** is a method used to measure the common bacteria found in water. HPC is not an indicator of health effects or disease. Lower HPC indicates that the water treatment plant is well maintained. Systems using surface water or groundwater under the direct influence of surface water should have an HPC no greater than 500 bacterial colonies/mL.
12. **Dissolved Oxygen (DO)** indicates the volume of oxygen contained in the water and is seldom measured within the distribution system. DO varies depending on the source, water temperature, salinity, and altitude. Water from flowing sources tends to have a higher DO than water from relatively stagnant sources. Water normally contains DO; DO values greater than about 5 mg/L support aquatic life, while DO values dropping below about 2 mg/L for a few days tends to kill aquatic biota. Low values can be caused by algal blooms or by a contaminant causing an increase in oxygen demand.
13. **Oxidation-Reduction Potential (ORP)** indicates sanitizer effectiveness. ORP is seldom measured outside the treatment plant. Because it reflects ionic effects on water chemistry, ORP may have impacts on contaminant chemistry. Positive and negative ORP values indicate oxidation and reduction potential, respectively. ORP is related to pH. High pH (basic) ionized water could have a negative ORP, while most bottled water has a low pH (acidic) and a positive ORP (around +400 mV). It varies widely depending on the source and the natural dissolved mineral content. Typical drinking water from the tap has values between +200 and 600 mV.

3.1.8 Measurement of Water Quality Data

Water quality can be measured in the following ways:

1. **Continuous sampling**: There are relatively few water utilities that use automated, nearly continuous measurement sensors within their distribution system. Continuously sampled sites have an obvious advantage in early detection of water quality anomalies because the data can be measured as frequently as every few minutes and sent immediately from a remote location to the water utility. Another advantage is that individual

measurements from a continuously sampled site can be easily compared with many data that are nearby in time from the same site so that anomalies relative to that site's baseline or background can be more easily detected. Using these sensors are also more expensive than using grab samples (described below) because, with the grab sample, the utility only has to maintain one set of centralized laboratory equipment. Because the sensors are remote from the treatment plant and located within a pipe, it can be expensive and time-consuming to calibrate them often. Due to their expense and calibration issues, most water quality measurements are made using grab samples. Of the utilities that have participated in the water module development, only one used continuous measurements within the distribution system, although it mostly relies on grab samples.

2. **Grab samples** are water samples obtained by the water utility from various locations within the water distribution system. Utility employees collect the samples and take them back to their laboratory for analysis. Compared with continuous samples, grab samples have the advantage that their data can be measured in a controlled laboratory setting with a single centralized set of laboratory equipment that can easily be calibrated and maintained. However, there is also the potential for human error when collecting the samples. Grab samples from the utilities that participated in this project were typically obtained on a weekly schedule from an individual site, so several months of data would be required to establish a baseline for the site. Labor hour costs can be high for grab sampling programs because the sample has to be physically transported from the field to the lab.

3. **Practical considerations:** Of the water utilities that participated in this effort, only one used continuous sampling within the distribution system. However, even this utility relies mostly on grab samples because it can afford to operate only a few continuous sampling sensors, compared with dozens of grab sample sites. In addition, despite the higher sample rate, experience with the continuous monitor data quality had been unsatisfactory and has not stimulated efforts to expand this capability. While grab sample sites are widely distributed within a distribution system, the frequency with which they are sampled is limited. The grab sample system was designed for quality control and for testing of adherence to standards, not for frequent water quality surveillance. Grab sample sites for the utilities that participated in this project were typically sampled for quality control data once per week. In populous areas, nearby sites are scheduled for monitoring on different days of the week so that some quality control data are taken on a nearly daily basis. However, missed samples are common at some sampling site.

4. **Operational data:** It is very common for utilities to remotely monitor distribution operating conditions via Supervisory Control and Data Acquisition (SCADA) Systems. Operational parameters include pressure, flow, tank levels, pump and valve status and pump run times. The telemetry for this type data is well established at large

utilities. Water quality parameter monitoring discussed above will make use of this existing communications network as it evolves from the treatment plant to the distribution system. In some cases, this operational data can be used to understand sudden changes in on line water quality data.

3.1.9 Clustering for Baseline Determination of Water Quality Parameters

Because of the limitations on the sampling frequency and on the reliability of grab sample data, there are often few recent measurements available for a given site; this lack of recent data makes the estimation of expected measurement values challenging. Such expectation estimates are essential for a detection algorithm to correctly distinguish anomalous measurements. For this reason, JHU/APL developed an approach to derive estimates by pooling recent measurements of similar sampling sites. The approach is to form clusters of like sampling sites so that measurements within each cluster can be combined to calculate a baseline mean and standard deviation for each quantity of interest, and then current measurements for each site in the cluster can be compared to those baseline values. The following subsection provides details of this approach.

3.1.9.1 Clustering Approach

Examination of time series from different sampling sites showed distinct differences in scale, variability, and cyclic data patterns. Discussions with water utility personnel revealed many reasons for these variations, including differences in engineering operations, source water characteristics, and sensor brands or versions. The discussions made it clear that explicitly adjusting for these differences would be a formidable development task that would have to be repeated for each geographic region monitored. Therefore, the JHU/APL clustering approach was an implicit one based on historical data and on readily available fixed parameters. In other words, sites for which past measurement scales and patterns were dissimilar and which were different for other reasons, such as distant location, were assigned to different baseline clusters, regardless of the reason for the differences.

JHU/APL achieved this sampling site partitioning for baseline calculation using a divisive clustering algorithm that had been developed at JHU/APL for ocean region partitioning.[16] Inputs to the algorithm were the sampling site properties chosen for grouping and the number of clusters desired. The divisive approach began with a single cluster containing all sites and then segregated them into successively smaller clusters using a weighted combination of designated site properties until the desired number of clusters is obtained.

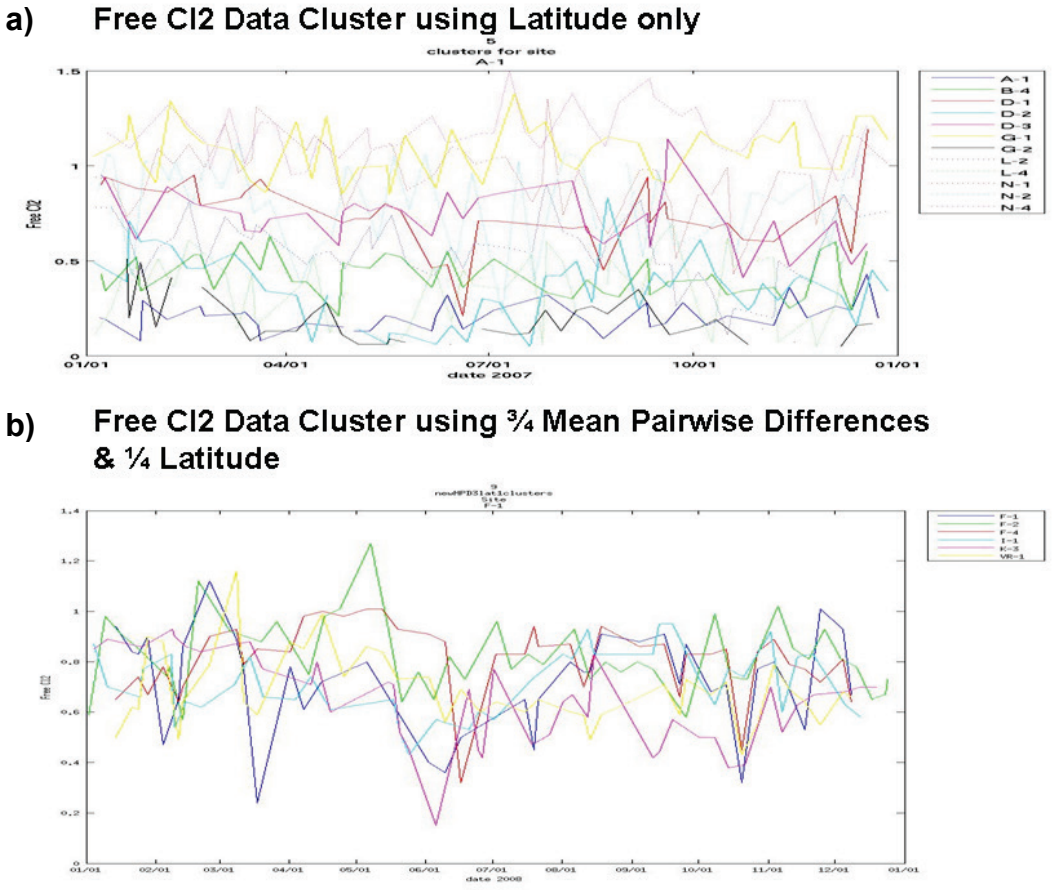
A variety of properties and property combinations were tested, including pressure, distance from the water treatment plant, latitude, longitude, mean and variance of free chlorine concentration, and distance measures between pairs of chlorine time series (e.g., mean squared difference between two time series). Because the algorithm allows fixed linear weighting of the selected properties, JHU/APL compared plausible weighting schemes based on the heuristic relative importance of each classifier.

3.1.9.2 Evaluation of Site Clusters

JHU/APL judged the effectiveness of the various site classifier combinations with a two-step process. The first step of evaluation

of site clusters was a visual inspection of the chlorine concentration time series data from the clustered site groups. Groups of time series that are unacceptable and acceptable (see Figure 7, graphs a and b, respectively) because of the degree of similarity of the series within the cluster. Criteria for choosing the number of clusters and the property combinations were:

a. **Similarity within clusters**: The time series plots should show sufficiently similar behavior for the pooled baseline means and variances to be appropriate for anomaly detection for each site in the cluster at most measurement times.
b. **Parsimony**: To achieve the criteria above, the classification criteria should be as few and as simple as possible and the number of clusters also as few as possible.
c. **Sufficient data representation**: The number of sites in a cluster should be large enough to provide sufficient pooled data to generate a stable baseline. JHU/APL required at least four sites in each cluster.



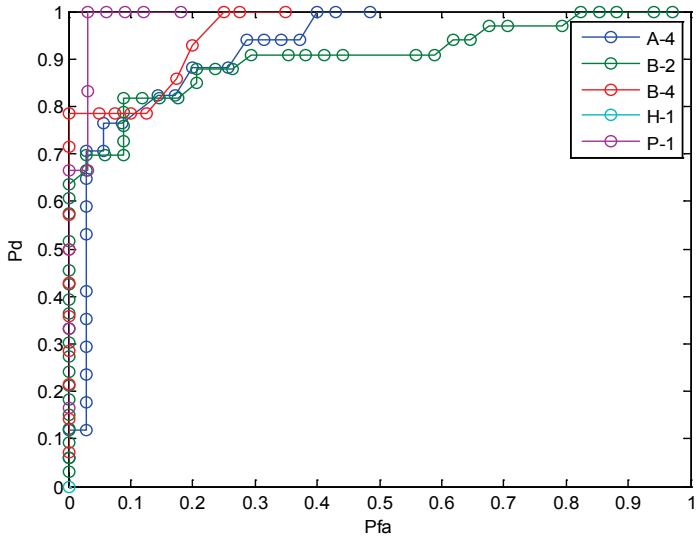
a) Top plot is an example of a cluster in which the free chlorine values are not sufficiently similar to provide a stable baseline; b) Bottom plot shows a cluster that provides a more stable baseline.

Figure 7 Data Cluster Comparison

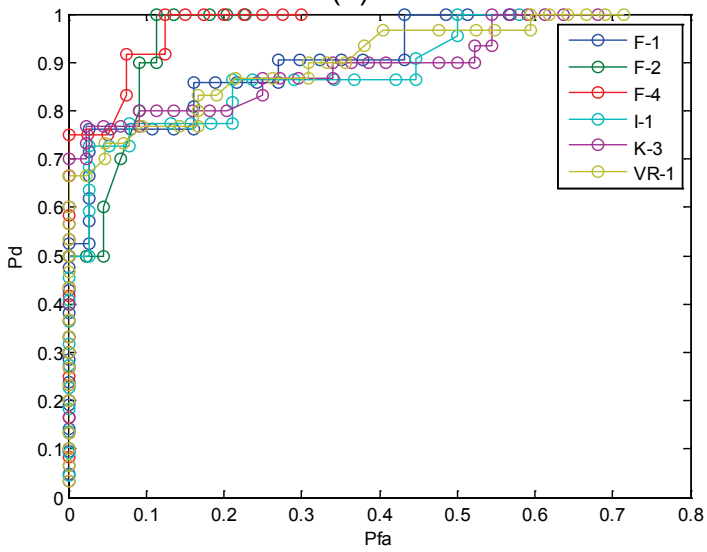
Using visual inspections, a few candidate-weighted property combinations were chosen.

The second step of the process of evaluating site clusters was to apply the event detection software[17] detection algorithms to the clusters derived with the candidate combinations and form separate Receiver Operating Characteristics (ROC) curves using as background data the chlorine concentration time series for each site in each cluster. (ROC curves are plots of *detection rates versus false alarm rates* for different threshold levels.) A detection threshold can be chosen by picking a point on the curve that corresponds to a desired detection rate (or false alarm rate). The ROC evaluation methodology followed that of McKenna et al.[22] and is described in Section 3.4.2.

For the final step of evaluating site clusters, JHU/APL then selected the weighted property combination for site clustering that gave the most consistent detection performance by visual inspection of the ROC curves, and JHU/APL permanently adopted the clusters given by this selected combination for baseline calculation. Figure 8 (a) and (b) show sets of ROC curves for two site clusters chosen for the City 3 system given a set inject signal level. The inject signal was generated by computing the background data's standard deviation and multiplying it by a constant factor (level). Selecting operating points (based on the desired detection or false alarm rate) on the individual curves provides the detection thresholds for each site. In a detection system operating continually over a period of years, the site clustering and resultant baseline determination should be reviewed periodically and when the configuration of sampling sites is changed.



(a)



(b)

Figure 8 Receiver Operating Characteristics (ROC) Curves for City 3 Site Clusters (a) and (b) are examples of ROC curves from two different clusters. The legend labels represent the sites that were assigned to the cluster.

3.1.9.3 Outcome of Sampling Site Clustering

As could be expected, the preceding clustering process did not yield the same cluster criteria for different drinking water distribution systems, and sites in different geographic regions clustered differently. However, the adaptation of the clustering method to a new region would require a relatively small amount of development if at least a year of historical data were available. The geographic transfer of the process also would not require detailed knowledge of the system operation nor would it unduly burden the utility staff.

In application of this process to water qulaity data two ultilitiesin injunction with JHU/APL determined site clusters using weighted combinations of geographic

location, pressure zone location, and distance from the water treatment plant. However, for the other distribution system data, this combination of factors did not yield distinct groups of similar chlorine time series measurement as required above. Possible reasons for different classifiers were the topography and geographic orientation of city, however, the JHU/APL approach does not require detailed regional modeling. JHU/APL tested several dozen combinations of parameters and settled on two: the site latitude and the mean pair wise difference between site-specific chlorine time series data. Analysis of the chlorine time series data led the JHU/APL developers to formulate the latter classifier, which requires a matrix of mean pair wise differences based on a sizable data sample. To calculate the mean pair wise difference between time series data from two sites, JHU/APL used data from each site over a 1-year period for time intervals when both sites had data (ignoring times when one site was missing a measurement). The sites that yielded measurements for less than half the weeks were excluded; nearly all of these sites were temporary ones. JHU/APL then calculated the mean absolute difference between the measurements for those weeks.

Following the test procedure described above, the chosen metric used to cluster the single site was the weighted sum: $0.75*D+0.25*L$. where D is the mean pair wise difference and L is the site latitude. In addition, JHU/APL found that nine clusters gave distinct groups of similar chlorine behavior with at least four sites in each cluster. In the resulting surveillance system, the baseline mean and standard deviation used for each site measurement corresponded to the values from the derived cluster containing that site. ROC curves, as shown in Figure 8, corroborated this clustering for robust distributed detection.

3.1.10 Protocol for "Real-Time" Data Acquisition

The users' responses to the beta test evaluation indicated no issues with data transfer or entry protocols. Currently, health and water quality data are transmitted to JHU/APL via File Transfer Protocol (ftp) or Secured File Transfer Protocol (sftp). For example, one utility sent data to the health department behind the firewall via ftp, and then the health department sends the data to JHU/APL via sftp. Another utility sent both grab sample and continuous data. One type of water quality data was sent by sftp and the other type are sent via Secure Hypertext Transfer Protocol (https). Data are sent after a request is submitted to the utility's site. The health data from arrives via ftp according to each jurisdiction's data sharing agreement with JHU/APL. The sources of these data are independent entities and no enforced standards exist. JHU/APL has no authority to enforce any protocols that subsequently may be determined and is dependent upon the data source to use the protocol that is deemed appropriate by the utility. Fortunately, JHU/APL has been able to use the ftp protocols as they currently exist.

Real-time data acquisition can be broken into two areas for the EPA Water Security Module as illustrated in the following outline.

1. The first is the additional health processing during data acquisition.
    a. A processor is run on the public health data to categorize the data specifically for conditions related to the EPA system. This processor produces water-related health conditions that can be queried.

b. The health data are run through the BN system, which queries the specific data, runs analysis detection algorithms on those data, and sets values on the BN for further processing.

2. The second is processing the water data as a data source.

a. The water data is collected, processed, and divided into two categories:
   i. Grab Sample
   ii. Continuous

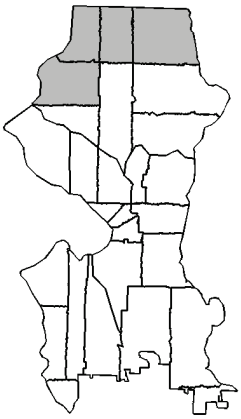
b. Water data are clustered by collection site to get the most appropriate baseline data for the current environment. Each installation at a utility will require baseline clustering analysis since each site has different water properties.

c. Detection algorithms are run on the water data and the information stored for later use.

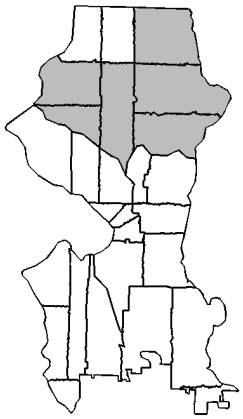
b. The water detection results are then utilized in the BN processing

3.1.11 Water Area Selection in a select city

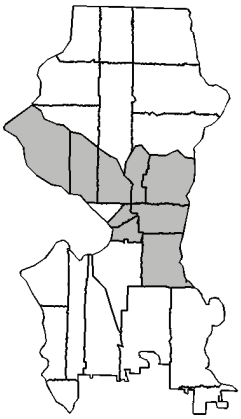
JHU/APL achieved spatial discrimination in the waterborne disease module by running the BN separately in five spatial regions, which were chosen based on advice from both the Public Health Department and Water Utility. The selection process was used to guarantee that each region was represented in the health data and that each region received a common drinking water supply whose quality was tested in the available input data. The BN complexity requires a rich data supply, and no attempt was made to apply the BNs at finer spatial granularity, which would have been unrealistic for detection of rare disease events. The five areas chosen are shown in Figure 9. Each figure shows the region divided into zip codes outlined by black lines. The zip codes assigned to the different health areas are shaded in gray. Some of the zip codes were assigned to two areas to take into account the uncertainty associated with water sites that were close to the boundary between the areas.



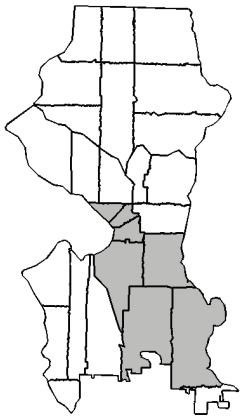
Area 1



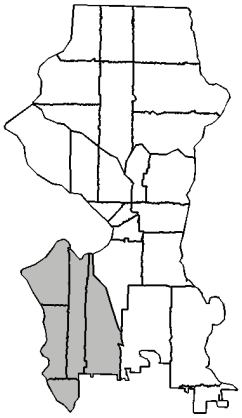
Area 2



Area 3



Area 4



Area 5

Figure 9 Health Areas

3.1.12 General Description of ESSENCE Health Indicator Data

ESSENCE is an automated syndromic surveillance system that uses electronic health data from a wide variety of sources. For the purposes of the Water Security Module, these data are currently derived from hospital emergency department (ED) visits. The basic ED data include the following:

1. Hospital ED location
2. Date of visit
3. Time of visit
4. Patient residential zip code
5. Patient gender (male/female)
6. Patient age (usually a range such as 0-4, 5-17, 18-44, 45-65, 65+)
7. Disposition i.e., whether the patient status was admitted, deceased, or discharged/sent home (Disposition is sometimes included but usually this information is lacking.)
8. Free text containing a chief complaint for each patient

A chief complaint is a short phrase that describes the main concern expressed by the patient as the reason for coming to the ED. Sometimes these complaints are recorded just as the patient states them, but other times they are recorded as the nurse's or physician's best interpretation of the patient's health concern. In cases in which the patient cannot communicate, the chief complaint data instead may contain the apparent primary presenting sign of the patient. For example, chief complaints can include such phrases as "abdominal pain," "gunshot wound," "diabetes," "coma," and "threw up." The chief complaint text data can be parsed into syndrome categories using specialized free text processing algorithms that can take into account spelling errors, abbreviations, and acronyms. The data can also be queried directly using features in ESSENCE. This querying feature allows for logical data joining such as "and" and "or" and also allows for wildcard characters. For example, one can query all ED data that contain the phrase "^fever^,and,^vomit^" where the ",and," is a logical operator and the "^" is used as a wildcard.

Note that individual patient identifiers are typically excluded from ESSENCE data to protect patient privacy. Because a syndromic surveillance system is population based, it relies on daily counts of health records that have certain similarities. Those similarities are usually chosen to be related to the same or similar causes. It is also important to realize that ED data are pre-diagnostic, which is both an advantage and a disadvantage. It is an advantage because it might allow earlier detection rather than waiting for a confirmed diagnosis. In this case, a health data anomaly could be detected before more of the population is affected. It is a disadvantage because the number and nature of the cause(s) of the health anomaly is. Therefore, such syndromic surveillance systems work best as a supplement to traditional public health surveillance rather than a substitute for it.

To utilize these syndromic surveillance systems, similar chief complaints may be grouped together by user-designed queries or by pre-packaged queries. These groupings can be called syndromes, sub syndromes, or case definitions in order of increasing specificity. For example, a gastrointestinal syndrome can be defined as including any chief complaint containing the phrases (or their synonymical variants): abdominal or stomach pain, nausea, vomiting, or diarrhea. While such a grouping could be very sensitive for the purposes of anomaly detection, abdominal or stomach pain is one of the most common chief complaint seen in

the ED and might have been caused non-gastrointestinal reasons. Even nausea and vomiting could be triggered by non-gastrointestinal causes. In addition, since different patients with the same disease might present their symptoms differently, some patients with the same disease could be excluded if the predefined grouping is too specific. The challenge in defining a chief complaint grouping is to find the proper balance between sensitivity and specificity.

3.1.13 Description/Rationale for Selection of Chemical Neurological Syndrome

The project used BN fusion methods to detect health effects related to human exposure to chemically contaminated water (See Section 3.1.6). To do this, processed outputs from two data sources were used as input to sets of sub syndrome-based queries. The probabilistic outputs from those queries were in turn used as inputs to the BNs. This section provides 1) a brief description of the chief complaints to sub syndrome and syndrome binning process, 2) the chief complaints processing steps to create specific sub syndromes for capturing water-contamination-associated human symptoms, and 3) the process for building queries whose outputs served as the raw material for the BNs.

To conduct temporal detection, the ESSENCE bins chief complaints into two levels of progressively more sensitive groups. The first level groupings are called sub syndromes and the second, syndromes. For instance, a chief complaint record containing "bad food," "food," or "food poison" are assigned to the "food poisoning" sub syndrome and the "food poisoning" sub syndrome along with several others sub syndromes, such as "diarrhea," "vomiting," and "gastroenteritis," make up the Gastrointestinal (GI) syndrome. The hierarchical process of binning chief complaints to syndromes maximizes sensitivity to health conditions that can present in different ways while retaining the ability to narrow down by sub syndrome.

The fundamental process of binning chief complaints to sub syndromes must compensate for the differences in various terms and manners by which patients describe chief complaints, and for the negation of certain terms by other terms in the string of chief complaints. Weights are assigned to particular chief complaint terms based on how relevant the terms are (or are not) to the sub syndrome. Higher weights are assigned to chief complaints terms presumed to have a higher likelihood of being related to the sub syndrome; likewise, lower weights or negative weights are assigned to terms with lower likelihoods or terms which alone have no effect but when combined with other terms have an additive or subtractive effect. The combined weighted values for each of the terms are compared against a preset absolute value to determine whether or not the terms will contribute to a particular sub syndrome. For example, "swallowing" alone will carry a weighted value that might not be binned to a sub syndrome. However, the presence of this term along with the term "difficulty" (which would have been assigned its own weight) would increase the combined score so that the chief complaint string would at least be binned to the "dysphasia" sub syndrome. For this project, the complex process of assigning chief complaints to sub syndromes was used to create weighting tables for twenty-three new sub syndromes including: chloracne, confusion, vertigo, hyperpigmentation, metallic taste, and dysphasia (See Appendix E). These and other sub syndromes were chosen based on common human symptoms associated with exposure to the list of contaminant classes of interest to the EPA.[13]

The final processing step was to design custom queries based on these sub syndromes. Six queries were selected to provide probabilistic outputs to BNs. To determine these optimal queries, JHU/APL epidemiologists developed sample queries likely to capture water contamination-related health effects for the contaminant classes of interest. Some queries were designed to capture rare gastrointestinal and neurological heath conditions that would be unusual in the absence of a deliberate contamination. Others were designed to capture common gastrointestinal and neurological health conditions that can be caused by natural or intentional contamination. The queries were further specified for adults and children. The custom query and on-the-fly detection capabilities of ESSENCE were used to carefully assess background counts and temporal alerting algorithm outputs for each of the sample queries. Via this process, the six queries likely to provide optimal probabilistic inputs to the BNs for identifying an intentional drinking water contamination event were chosen.

3.2 User Interface

The ESSENCE Water Security System interface was designed to provide the users with the ability to trace fusion alerts back to the source data. As seen in Sections 3.1.3-3.1.6, tracing BN output values back to the original source data (input nodes) can become complicated when there are multiple connections among the nodes. This interface aims to allow the user to drill down from the output nodes through the relevant BN paths. Nodes without high anomaly probabilities can be ignored.

The following description steps through examples of the different screens and navigation panes. The user interface provides a walkthrough for investigating a BN fusion alert that may indicate a water contamination-related health event. The introductory screen displays the area-based fusion alert page (Figure 10). The matrix of red and green bars represents probabilities of alert based on the top-level Fusion BN node. The color convention for the bars is green equals ‘no alert’ and red equals ‘alert,’ where the probability of ‘alert’ (or ‘no alert’) is proportional to the percentage of the bar that is colored red (or green). The percentage is drawn directly from the BN node probabilities and there is not a threshold so that individual users can determine the level of concern. The probabilities on the introductory screen in Figure 10 represent the fusion of both water and health data anomalies that were calculated based on a set of child nodes in the BN structure. This page enumerates the top-level fusion node alert by areas and dates, where the areas are represented in rows and the dates in columns. In this way, the user is able to select a date and area of interest for investigation. The page also displays maps corresponding to the areas that are listed in the matrix and provides the capability to select a different set of data ranges for visualizing alerts and correlating data. The same matrix screen is also available as a configuration option for all BN nodes.

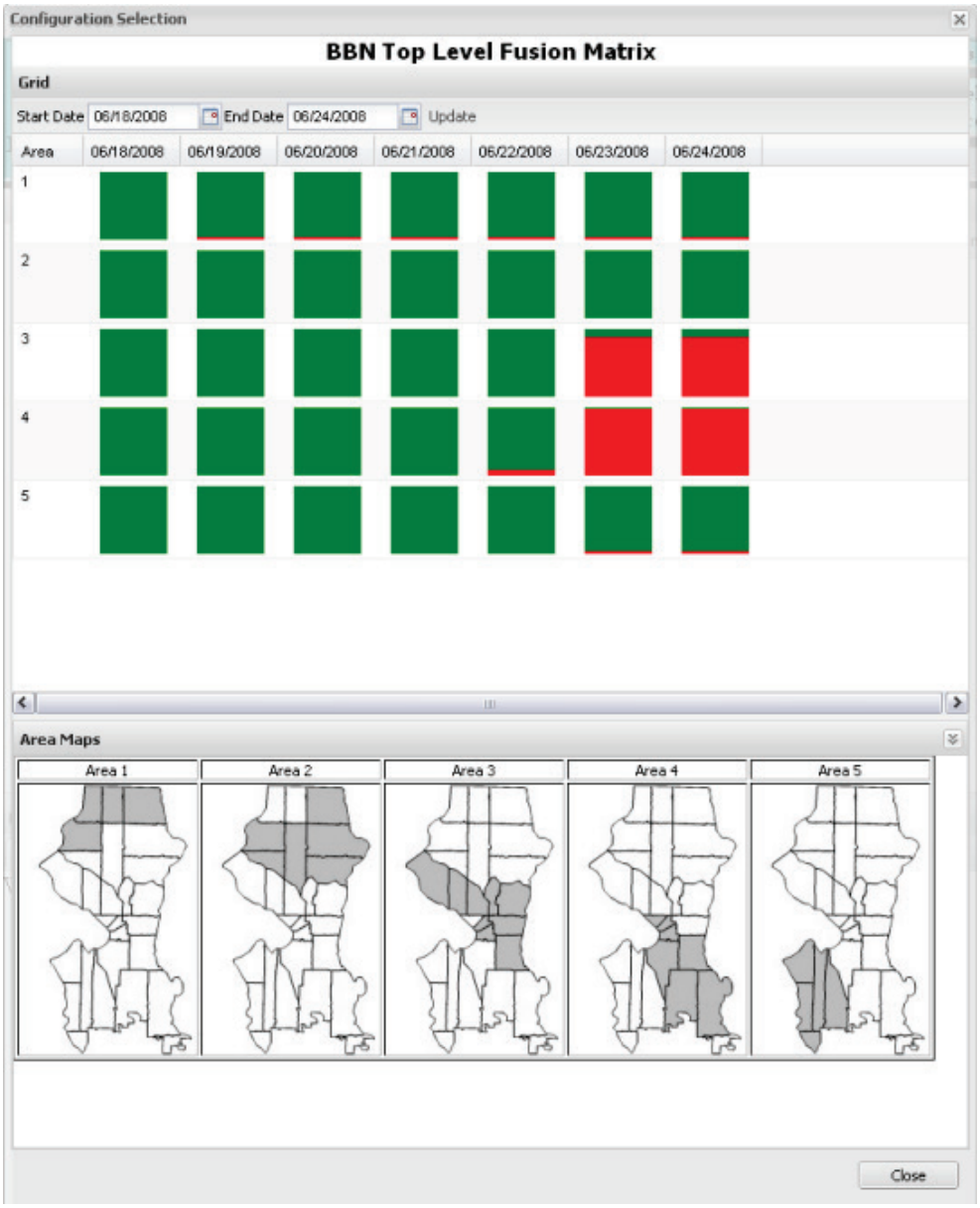


Figure 10 Introductory Screen

The investigation process includes two main concepts: drilling down through the BN structure and analyzing data that is presented. Drilling down allows the user to explore the probabilities for different nodes in the BN structure. In the EPA system, access to drill-down data can be limited according to user roles. So, water utility users could be prevented from seeing health data and/or health department users could be prevented from accessing water data. The system can also provide any other restriction that may be necessary.

After selecting an area and date of interest on the introductory screen, screen shown in Figure 11 would be displayed. There are three main panes on this screen: the navigation pane on the left side, the BN graph on the top right side, and the details pane on the bottom of the screen. The BN graph pane displays of two levels of nodes in the BN structure. The top node (parent

node) in the graph, 'Health Water Fusion', represents the node that was selected based on the matrix on the previous page. The nodes on the bottom level are the child nodes of the top row node. The red/green values for these nodes have the same meaning as those for the bars on the introductory matrix screen. If available, selecting the 'Show Details' button for a node will switch the details data at the bottom of the screen to that of the selected node. Selecting the 'Drill Down' button for a node will update the BN graph pane to display that node as the top-level node with any corresponding children nodes on the lower row. The navigation pane provides a more familiar method for navigating through the BN. The details pane displays the actual time series data or probabilities in graph or table forms.

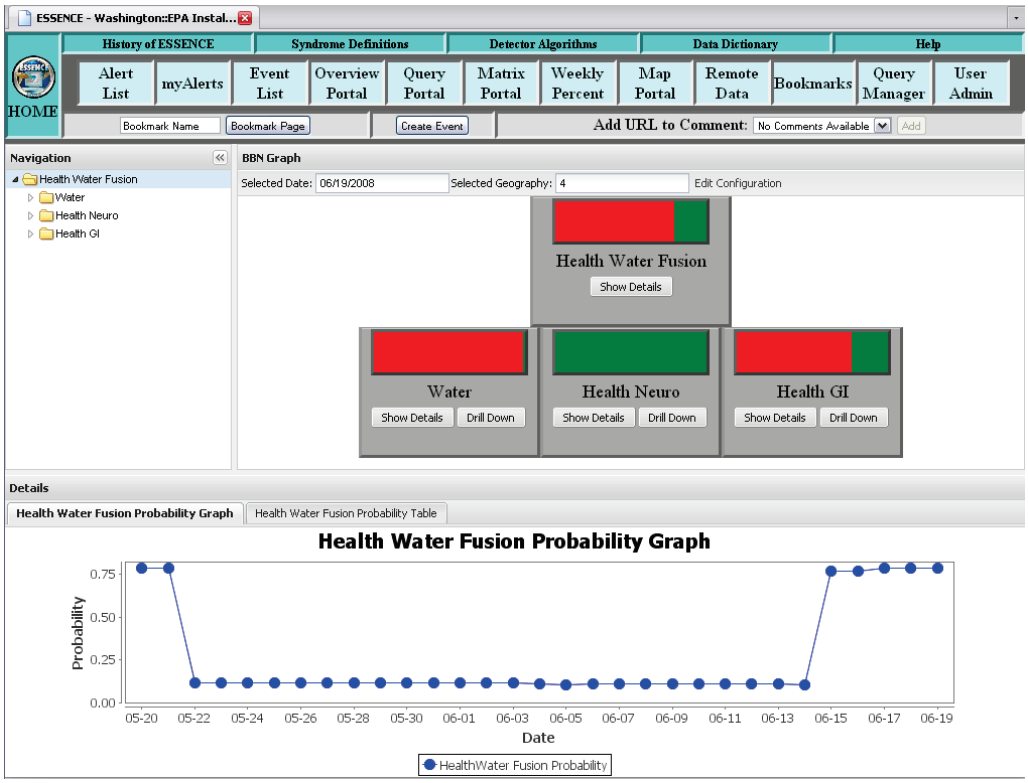


Figure 11 Secondary Screen

Figure 12 shows another example of a navigation screen. The drill-down screen shows the investigation process after drilling into the Health Chem-Like area of the BN. In the screen, the user can see that the 'Adults 1 Chem-Like' node shows a significant probability increase and the child node 'Adults 1 GI' is the BN's highest contributor. Note the parallels between the parent and child nodes in the BN graph pane and the folder structure in the navigation pane. The highlighted node in the navigation pane is the top-level node in the BN graph.

Figure 12 Sample Drill Down

Navigation is also provided to the user in the form of tree and graph visualization (Figure 13). The visualization forms are linked to each other and provide similar user interaction for consistency. However, the two visualization types perform two distinct functions. The navigation tree allows the user to immediately look at a particular node in the entire BN structure. It also maintains a trail of investigation levels that have been drilled through to trace the steps of a user. With this the user can backtrack to any previous level simply by clicking on the parent folder. The folder concept is utilized as it is familiar navigation on common computer file systems.

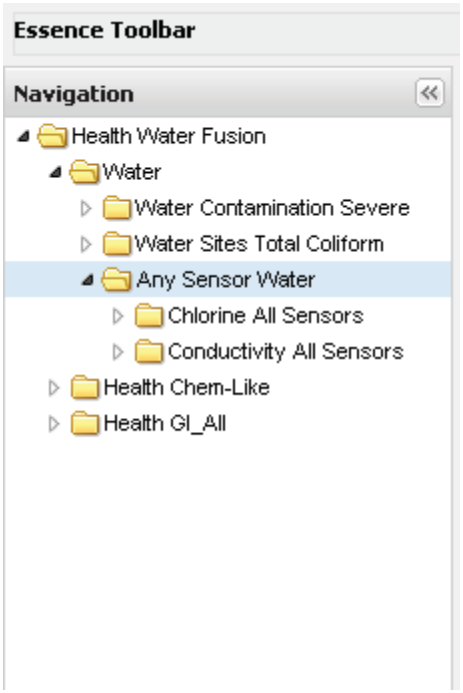


Figure 13 Folder Navigation Pane

The BN graph navigation pane (Figure 14) displays the data for the currently selected geography and date for one point in the water branch of the BN structure. Note that the 'Chlorine All Sensors' node is yellow (not just true or false). For some BN nodes, there are multiple levels. In this example, the yellow bar corresponds to the number of sites within the area that are alerting. The corresponding time series plot in the details pane (not shown here) will contain a legend for the colors in the BN graphs.

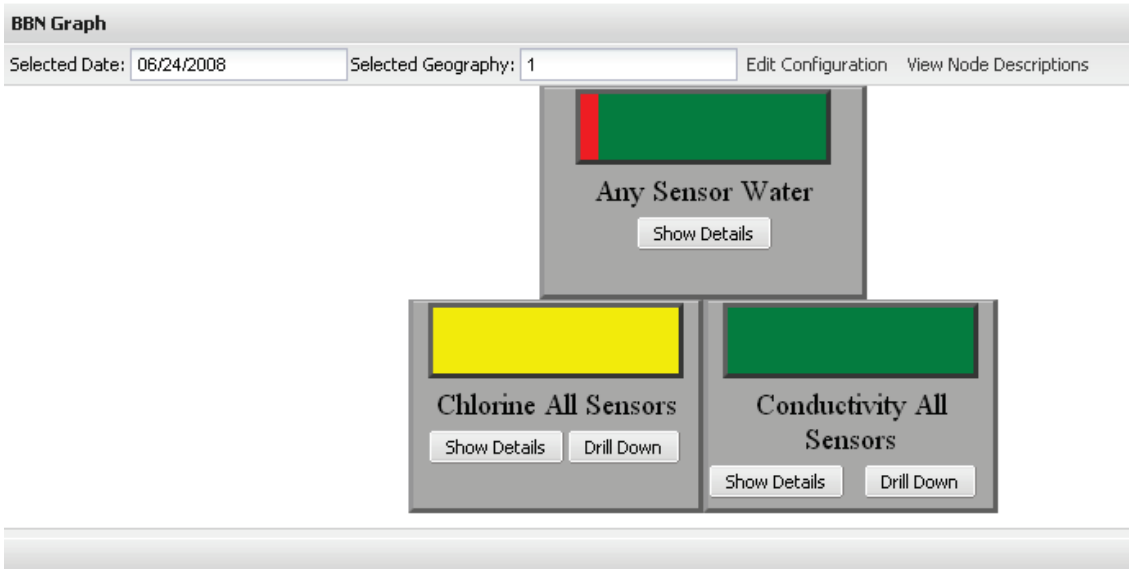


Figure 14 Bayesian Network Graph Navigation Pane

At certain points during the navigation, the user will be presented with a selection matrix for the water data, which is specific to the water data and fusion process. Water data are analyzed per site and then aggregated up to an area of interest. In drilling down through the BN, the reverse process needs to occur. In this case, users will be presented a selection matrix that will let them choose a specific water site to investigate.

Figure 15 shows the list of water sites for a specific area under investigation. The visualization shows a high probability for two sites, J-2 and J-3.

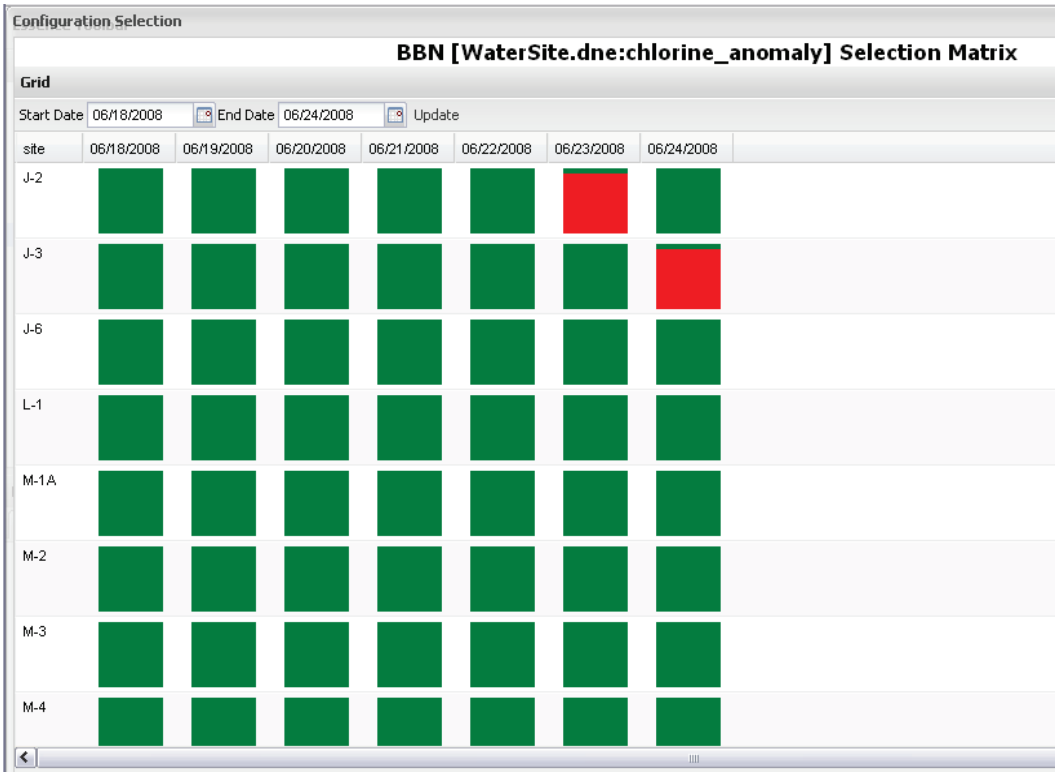


Figure 15 Water Site Selection Matrix

The selection matrix is also available as an "Edit Configuration" option (Figure 16), which allows the user to change the currently selected geography and date at any point during the investigation. The example shown here shows alerts for a node in the Chem-Like/Neurological BN across several areas. The ability to go back to this selection matrix is useful if the user encounters a previous date of concern and would like to investigate that area. Users are also able to quickly visualize other areas or sites for the node that they are currently investigating, giving them the capability of comparing various geographies.

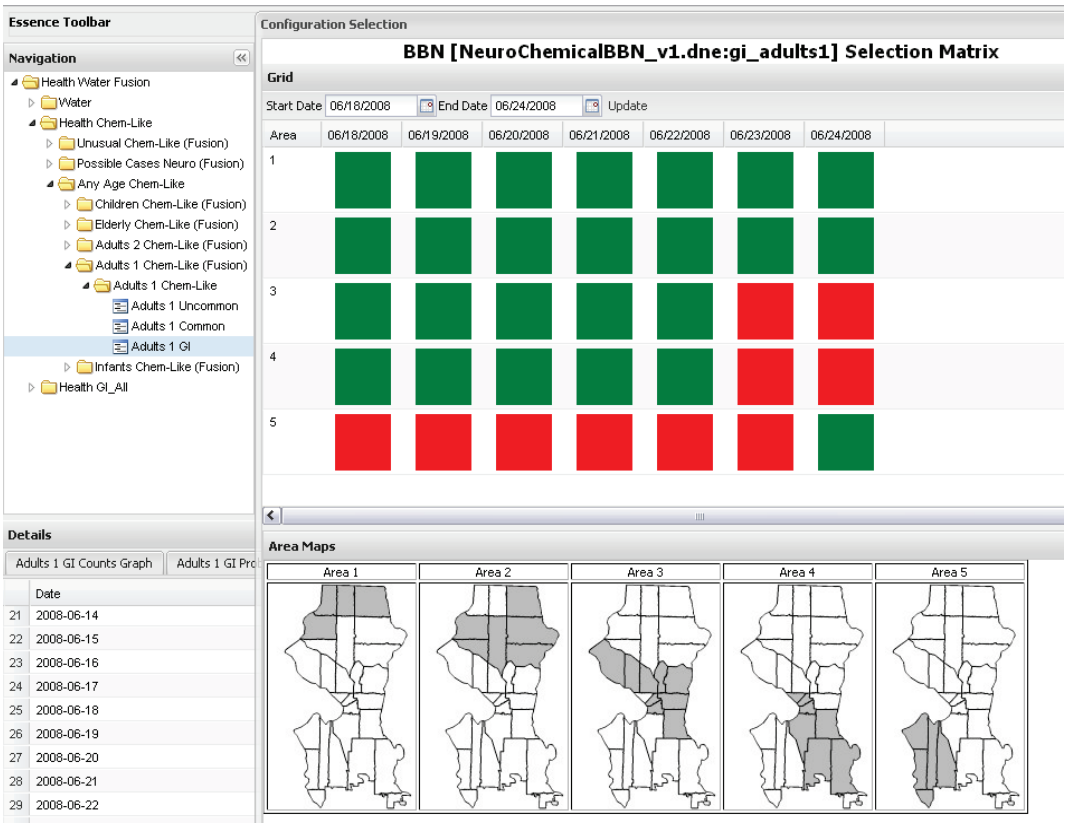


Figure 16 Example of Alerts Across Areas

Data can be presented to the user in several ways in the Detail section of the page as shown in Figure 17. Currently, each node will display a time series graph and a data table of the BN probabilities for the currently selected geography. This provides a history of the BN probabilities to the user. The histories are useful in determining if a level of BN concern is out of the ordinary for a particular node.

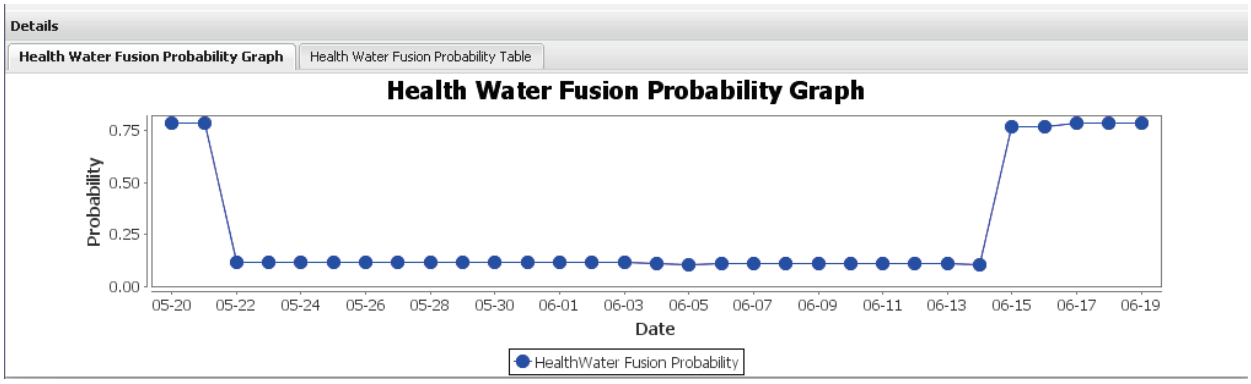


Figure 17 Example of Detail Section in User Interface

The Details section for each node in the BN can be externally configured to display other data as well. This will primarily occur with low-level nodes that take in raw data to perform analysis; however, high-level nodes could also display aggregated data and this

configurability can be utilized to do so. The displays will consist of raw tabular data, as well as time series graphs that display specific raw data relevant to the given node. In addition, the display will take into account user access privileges.

The screen in Figure 18 shows the lowest level node of the water branch of the BN. The red bar indicates that a drop was found in the free chlorine raw data. These data are shown on the time series page (Figure 18) as well as in tabular form (Figure 19).

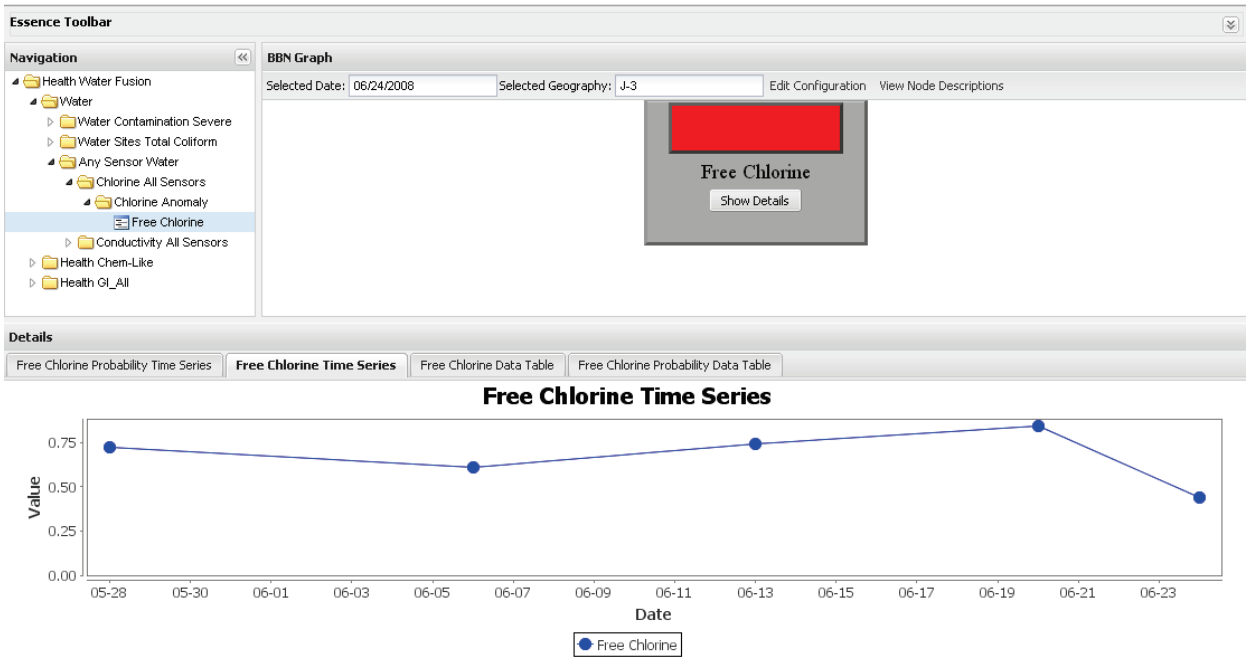
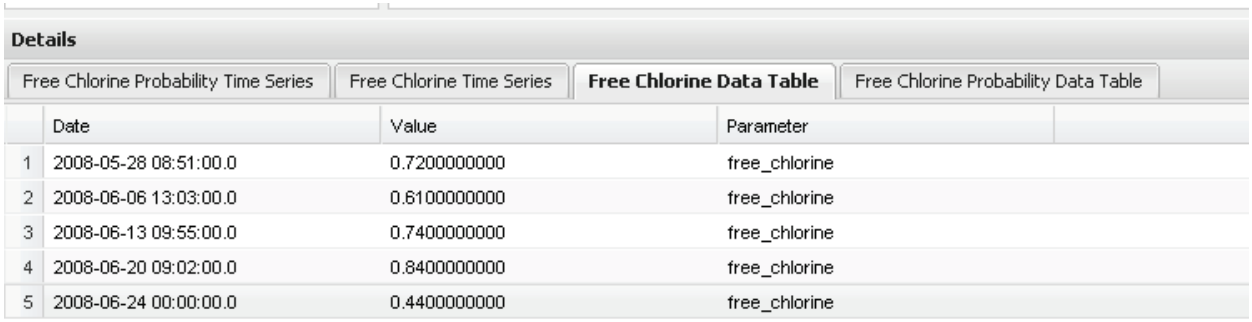


Figure 18 Example of Free Chlorine Drop Shown in Time Series Form



	Date	Value	Parameter
1	2008-05-28 08:51:00.0	0.7200000000	free_chlorine
2	2008-06-06 13:03:00.0	0.6100000000	free_chlorine
3	2008-06-13 09:55:00.0	0.7400000000	free_chlorine
4	2008-06-20 09:02:00.0	0.8400000000	free_chlorine
5	2008-06-24 00:00:00.0	0.4400000000	free_chlorine

Figure 19 Example of Free Chlorine Drop Shown in Tabular Form

After drilling down to the lowest level nodes, the user can also navigate to various ESSENCE pages to investigate detailed data for the specific node. The user can also navigate to other nodes in the network and edit the configuration to view other areas and other dates for the currently investigated node.

3.3 TRAINING AND EXERCISE

Training was provided a few weeks before the actual exercise in the form of a webinar

in which the participants were walked through the user interface. However, the participants did not have any hands-on experience before the exercise and the JHU/APL test team quickly learned that users should have been allowed more time to become familiar with the system. Therefore, the test team needed to provide additional training before the exercise started on the second day. Appendices A and B (Exercise Brochure and Facilitators Guide, respectively) contain detailed information regarding the exercise and scenarios.

3.3.1 Webinar Description

The purpose of the Water Security Simulation Module was to help public health and water utility users narrow down and investigate possible waterborne illnesses caused by contamination events.

Webinars scheduled over three consecutive days included training, an exercise that covered four days of real and simulated data, and a debriefing at the conclusion of the final day's exercise.

3.3.2 Operational Utility Assessment

The Operational Utility Assessment (OUA) brought together end-users from the Utilities and the Public Health Department in a no-fault setting to use the Water Security Simulation Module in a realistic fashion. By using the outcome of module training in conjunction with existing policies, plans, and standard operating procedures, users evaluated and provided feedback about the system's effectiveness.

Objectives of the OUA were to:

- Demonstrate the system's ability to detect conditions that strongly suggested the presence of a hazardous substance in the water system.
- Assess the ability of the system to deliver appropriate information that would enable the user to make an informed decision or take action.
- Determine user proficiency as a result of system training

3.4 EVALUATION

The evaluation focused on the system in an attempt to validate the system's strengths and identify areas that needed improvement. To evaluate the demonstration:

- Evaluators will observe the demonstration and collect supporting data.
- User feedback will be solicited.
- Data will be objectively analyzed against expected outcomes.
- An after action report will document the strengths as well as changes that need to be made to the Water Security Simulation Module. Similarly, it will evaluate associated plans, policies, procedures, staffing, training, and communications and coordination to ensure expected outcomes are delivered. Because few meaningful quantitative standards exist for many of the critical functions and tasks that the Water Security Simulation Module system was designed to support, the assessment consists largely of qualitative metrics.

3.4.1 Graphical User Interface

The following subsections discuss the utility assessment, the chosen criteria for evaluating the Graphical User Interface (GUI), and evaluation results.

3.4.1.1 Description of the Operational Utility Assessment

During the assessment, the participants were asked to navigate through the interface to become more familiar with the system and investigate any alerts and associated data. Users were encouraged to ask questions and notify the JHU/APL test team of any issues at any time during the exercise. Detailed descriptions of the assessment and evaluation results can be found in the appendices.

3.4.1.2 Graphical User Interface Evaluation Criteria

The graphical user interface (GUI) was evaluated based on the following criteria:

- Was the interface intuitive and easy to use?
- Were the navigation features adequate for reviewing the data?
- Did the fusion outputs make sense and could they be interpreted in context?

3.4.1.3 Results

After the first day of the exercise, participants either disagreed or were neutral regarding all of the criteria. By the end of the exercise, answers to these questions moved toward agreement or remained neutral. A full summary of the GUI evaluation, including user comments, can be found in Appendix D. The Lickert Scale[20] used in the evaluation forms (discussed in Appendix C) was as follows:

1 - Strongly Agree
2 - Agree
3 - Neutral
4 - Disagree
5 – Strongly Disagree
N/A – Not Applicable.

3.4.2 Water Quality Algorithms

The detection algorithms used for processing the water data were based on the algorithms from the CANARY event detection software.[17] The CANARY event detection software was designed to provide early alerts based on automated online water sensor data that were assumed to be continuously available, periodic, synchronous sensor measurements. However, the real data required preprocessing steps and other modifications to the detection methods to deal with the non-ideal data.

For the continuously sampled water data, hourly averages were computed in order to smooth the noisy data. The averaged continuous measurements were normalized as:

$$y(n) = (x(n)-\mu)/\sigma$$

where $y(n)$ is the normalized value at time index n (hourly increments), $x(n)$ is the averaged value, and μ and σ are the mean ands standard deviation, respectively, for all samples at that site within a previous 3-day window. The autoregressive estimate, $y'(n)$, was computed using the Yule-Walker autocorrelation method[21]. Then the difference, $y(n) - y'(n)$, was compared to a threshold for the desired false-alarm rate.

The grab sample water data was processed using a slightly different method than the

continuously sampled water data. For a given measurement from a single site and day, a mean and standard deviation were computed using water sample data from all sites within the same predetermined cluster. The method used for determining the clusters is described later in this report. For each grab sample measurement at sites numbered $m = 1, 2, \ldots, M$, the normalized value used for anomaly detection at time index n for site is denoted as $y_m(n)$ and calculated as:

$$y_m(n) = (x_m(n)-\mu)/\sigma$$

where μ is the mean and σ is the standard deviation for all cluster samples within a previous 7-day window.

A threshold, T, is set using the background data to meet a desired false alarm rate. An anomaly is detected for that site measurement when $y_m(n) > T$.

3.4.2.1 Methodology for Measuring Performance of Water Quality Data Anomaly Detection Algorithms

Performance of the anomaly detection algorithms described in the preceding subsection was measured by injecting simulated single sample values into the historical data streams. Two sets of grab sample data were available for this analysis. One city was able to a set of continuous monitor data.

The methodology for measuring the performance of both the grab sample and continuous monitor anomaly detection algorithms is as follows:

- Inject single spikes into data (negative for chlorine measurements) that are some multiple of the standard deviation of the background data for the current measurement type.
- Generate the ROC curves for each site/cluster combination by sweeping the detection threshold through both the background and injected time series to obtain the probability of detection (pDs) and probability of false alarm (pFA) for each threshold.
- Set pFA to a desired level and find the corresponding pD.

3.4.2.2 Sensitivity at Practical Alert Rates

The cells in Table 3 contain average pDs for the same measurement type over all sites for injects that were four times the background signal's standard deviation (4-sigma). The pDs are computed as the number of injects that are detected divided by the total number of injects. The table columns correspond to a given pFAs of 0.01, 0.05, and 0.10 for injects. A pFA equal to 0.01 corresponds to 1 false alarm for every 100 background (i.e. non-inject) samples. Performance will degrade for smaller amplitude signals and improve for higher amplitude signals assuming the same pFA. Performance of both grab sample data and continuous data are shown. Only chlorine- and pH-related measurements are shown here, since the other water quality measurement types did not have adequate data for analysis purposes. Data provided from two different types of continuous chlorine monitors (labeled 1 & 2 in the table) were kept separate in the analysis for comparison.

Table 3 Average Anomaly Detection Performance for 4-sigma Injects

Measurement Type	pFA = 0.01*	pFA = 0.05	pFA = 0.10
Continuous Free Chlorine (1)	-	0.71	-
Continuous pH	-	0.67	-
Continuous Free Chlorine (2)	-	0.68	-
City 1Grab Free Chlorine	0.66	0.74	0.80
City 1 Grab Total Chlorine	0.66	0.75	0.80
City 1 Grab pH	0.40	0.58	0.67
City 2Grab Free Chlorine	0.54	0.70	0.78
City 2 Grab pH	0.67	0.81	0.89

*pFA = False Alarm

Results for the chlorine-related measurements appear consistent, but results for the pH values are very different for the two cities. Note that the results shown here only provide a snapshot of how the water quality detection algorithms perform with the probability of false alarm held constant. In practice, the users can determine if pD or pFA should be controlled while keeping in mind the tradeoff between high probability of detection and low probability of false alarm. In addition, water utility experts should also determine what amount of change in specific water quality parameters would be anomalous for their system.

3.4.3 Detection Performance of Bayesian Networks

Performance of the BNs is dependent on whether or not the detection algorithms on both the water and health side detect anomalies in the time series data. This section discusses the performance of the BNs separately from the detection algorithms by presenting a set of selected scenarios that are variations on the original scenario used during the exercise. The number of possible input-to-output combinations needed to fully characterize the BN is too large to cover here, so the BN outputs presented here provide an overview of how the BN will react to different anomalies in the data. Because the structures of the BNs are the same for all regions, the results here are for only one of the five areas defined in Section 2.1.11.

The scenarios used to analyze BN performance are as follows:

1. Scenario 1 is described in Appendix B. The event included drops in free chlorine at two sites within one area (and an additional site in another area) and injects associated with common GI complaints associated with EPA's contaminant classes, and neurological specific complaints.

Scenario 1 plus anomalies in conductivity data at two water sampling sites
E. coli alert only plus Scenario 1 health injects
Scenario 1 water injects plus GI only
Scenario 1 water injects plus conductivity injects and GI injects only
E. coli only and GI only
Scenario 1 water injects and health injects for adults only
Scenario 1 water plus conductivity injects and health injects for adults only
E. coli only and injects for adults only
Baseline water (no detections) and Scenario 1 health injects
Scenario 1 water injects and baseline health (no detections)

Results for the Health, Water, and Fusion BNs are discussed in the following sections.

3.4.3.1 Health Bayesian Networks

Health BN output values are only dependent on changes in their inputs. Therefore, changes in the water data will not affect the outputs at this level. Values for the Chemical Contamination/Neurological BN are shown in Table 4. Scenarios 1, 2, 3, and 10 result in the same probabilities because they all have the same health input values. As expected, scenarios with GI only injects result in low probability values since this BN looks for patterns associated with Chem/Neuro symptoms. Finally, the scenarios with detections in the adult age groups have higher probabilities than the GI only scenarios but lower probabilities than when anomalies are found across three age groups.

Values for the GI BN are shown in Table 5. Scenarios 1 through 6 and 10 result in the same probabilities since they all have the same GI-related input values. As expected, the scenarios with detections in the adult age groups only have lower probabilities than the original scenarios with additional age groups.

Table 4 Sample Chemical/Neurological Bayesian Network Outputs

Scenario	Output Values						
	Infants	Children	Adults1	Adults2	Elderly	Diagnostic Cases	Chem/ Neuro
1. Original Water and Health Injects	0.20	0.99	0.94	0.79	0.12	0.33	0.74
2. Original Water + Conductivity and Original Health Injects	0.20	0.99	0.94	0.79	0.12	0.33	0.74
3. E. coli Only and Original Health Injects	0.20	0.99	0.94	0.79	0.12	0.33	0.74
4. Original Water and GI Only	0.00	0.01	0.02	0.02	0.00	0.00	0.01
5. Original Water + Conductivity and GI Only	0.00	0.01	0.02	0.02	0.00	0.00	0.01
6. E. coli Only and GI Only	0.00	0.01	0.02	0.02	0.00	0.00	0.01
7. Original Water and Detections in Adults Only	0.03	0.01	0.33	0.16	0.02	0.06	0.14
8. Original Water + Conductivity and Adults Only	0.03	0.01	0.33	0.16	0.02	0.06	0.14
9. E. coli Only and Detections in Adults Only	0.03	0.01	0.33	0.16	0.02	0.06	0.14
10. Baseline Water and Original Health	0.20	0.99	0.94	0.79	0.12	0.33	0.74
11. Original Water and Baseline Health	0.00	0.00	0.00	0.00	0.00	0.00	0.01

Table 5 Sample Gastrointestinal Bayesian Network Output Values

Scenario	Output Values						
	Infants	Children	Adults1	Adults2	Elderly	Diagnostic Cases	GI
1. Original Water and Health Injects	0.44	0.94	0.84	0.63	0.43	0.04	0.88
2. Original Water + Conductivity and Original Health Injects	0.44	0.94	0.84	0.63	0.43	0.04	0.88
3. E. coli Only and Original Health Injects	0.44	0.94	0.84	0.63	0.43	0.04	0.88
4. Original Water and GI Only	0.44	0.94	0.84	0.63	0.43	0.04	0.88
5. Original Water + Conductivity and GI Only	0.44	0.94	0.84	0.63	0.43	0.04	0.88
6. E. coli Only and GI Only	0.44	0.94	0.84	0.63	0.43	0.04	0.88
7. Original Water and Detections in Adults Only	0.07	0.07	0.20	0.07	0.07	0.02	0.25
8. Original Water + Conductivity and Detections in Adults Only	0.07	0.07	0.20	0.07	0.07	0.02	0.25
9. E. coli Only and Detections in Adults Only	0.07	0.07	0.20	0.07	0.07	0.02	0.25
10. Baseline Water and Original Health	0.44	0.94	0.84	0.63	0.43	0.04	0.88
11. Original Water and Baseline Health	0.02	0.02	0.02	0.02	0.02	0.02	0.16

3.4.3.2 Water Quality Bayesian Networks

Table 6 contains Water Quality BN results. Scenarios that involve injects for the free chlorine data and/or conductivity data have probabilities near or equal to 1.0 for the specific sensor measurement types (e.g., "chlorine anomaly," "conductivity anomaly"). When an *E. coli* inject occurs, the output probability at the top-level node ("water quality contaminant" column) is similar to the output values when there are injects in both the chlorine and conductivity data. Note that the output probability at the 'Water Quality Contaminant' node is only 0.42 when anomalies are only found in the chlorine data. This node is looking for patterns across multiple sensor measurement types. However, the information from chlorine anomalies at multiple sites is still captured in the 'Chlorine Anomaly' node.

Table 6 Sample Water Quality Bayesian Network Outputs

Scenario	Output Values (Per Site)				
	Chlorine Anomaly	Conductivity Anomaly	pH Anomaly	Coliform Anomaly	Water Quality Contaminant
1. Original Water and Health Injects	0.91	0.01	0.01	0.00	0.42
2. Original Water + Conductivity and Original Health Injects	0.99	0.99	0.36	0.00	0.76
3. E. coli Only and Original Health Injects	0.05	0.01	0.01	1.00	0.75
4. Original Water and GI Only	0.91	0.01	0.01	0.00	0.42
5. Original Water + Conductivity and GI Only	0.99	0.99	0.36	0.00	0.76
6. E. coli Only and GI Only	0.05	0.01	0.01	1.00	0.75
7. Original Water and Detections in Adults Only	0.91	0.01	0.01	0.00	0.42
8. Original Water + Conductivity and Detections in Adults Only	0.99	0.99	0.36	0.00	0.76
9. E. coli Only and Detections in Adults Only	0.05	0.01	0.01	1.00	0.75
10. Baseline Water and Original Health	0.00	0.00	0.00	0.00	0.03
11. Original Water and Baseline Health	0.91	0.01	0.01	0.00	0.42

3.4.3.3 Fusion Bayesian Network

Table 7 shows output results for the Fusion BN. The main point to take away from these results is that the highest level fusion alert will only have a high probability when there is a water alert and at least one of the health categories (GI or Chem/Neuro) alerts exist. The level of the fusion alert is also scaled according to the probability levels of the water and health alerts.

Table 7 Sample Fusion Bayesian Network Outputs

Scenario	Output Values (Per Area)			
	Water Alert	Health - GI Alert	Health - Chem/Neuro Alert	Fusion Alert
1. Original Water and Health Injects	0.99	1.00	1.00	0.99
2. Original Water + Conductivity and Original Health Injects	1.00	1.00	1.00	1.00
3. E. coli Only and Original Health Injects	1.00	1.00	1.00	1.00
4. Original Water and GI Only	0.95	1.00	0.00	0.95
5. Original Water + Conductivity and GI Only	0.99	1.00	0.00	0.99
6. E. coli Only and GI Only	1.00	1.00	0.00	0.99
7. Original Water and Detections in Adults Only	0.21	0.03	0.03	0.11
8. Original Water + Conductivity and Detections in Adults Only	0.69	0.10	0.10	0.35
9. E. coli Only and Detections in Adults Only	0.98	0.16	0.04	0.26
10. Baseline Water and Original Health	0.01	1.00	1.00	0.11
11. Original Water and Baseline Health	0.13	0.00	0.00	0.02

3.4.4 User Assessment

Exercise participants provided valuable feedback in the form of comments during the exercise and on the evaluation forms provided to them. A complete write up on the OUA Evaluation can be found in Appendix D: Assessment Evaluation.

3.4.4.1 Ease of Use

Additional training and a hands-on step through demonstration were required on the second day of the exercise. The participants provided several recommendations related to improving ease of navigation and interpretation of displays.

3.4.4.2 Effectiveness of the Module

Although the assessment experienced some delays at the beginning, it was generally well received. Survey responses from the second day were mostly in the '2 - Agree' category (See Appendix D). The training, in the form of a Webinar conducted the second day of the assessment, proved to be very valuable. Participants could see the value in

having a tool like the ESSENCE Water Security Module and the potential to be able to quickly glance at the GUI to see if a problem might exist. Participants believe with some additional development work this will be a very powerful tool.

3.4.4.3 Wish List

The following items would improve the effectiveness of the assessment method:

- Additional training, delivered at the appropriate time, as well as additional hand-on practice would increase end-users' level of comfort with the Module.
- A simplified dashboard with a quick, drill-down capability would greatly reduce issues with navigating the Module. Adding back button functionality would help as well.
- An ability to view what has been reviewed and what remains to be reviewed would increase the Module's usefulness. Enabling the capability to hover over nodes and view supplemental information would help as well.

3.5 CONCLUSIONS

During the exercise, participants provided useful comments and suggestions regarding both the user interface and algorithm outputs. It became apparent early on that, if this work is to be continued in the future, involving users early in the design process would be the best approach to developing a useful system. Input from a diverse group of users from different utilities would help guide the next iteration in the right direction. Allowing the users to test the module for ease of use and for comprehension of algorithm outputs (even in demonstration mode) over an extended period of time would provide valuable information.

On the algorithm side, a lack of realistic injects (or real event data) limited the ability to truly measure the performance of the detection algorithms and BNs. A set of simulated event data (including both water and health data) developed by an independent source would be a necessary component for validation.

Although the grab sample water quality data were adequate for developing the prototype module and running the exercise as a proof-of-concept, continuous monitor data are really what is needed for detecting contamination in the drinking water supply. In the grab sample data observed to date, measurements were taken weekly at individual sites. Until continuous data are collected widely at water utilities, the likelihood of detecting anomalies related to intentional contamination in the water data is small.

In conclusion, a water security module was developed and exercised by participants from water utility and public health departments. Feedback from the users indicated that they did see value in the tool, but also suggested areas for improvements. Significant effort is still required to turn this into a fully functional system, but the base functionality is available to be built upon.

4 System architecture and expansion to other locations

A definition and description of the general system architecture for the ESSENCE Drinking Water Surveillance System are provided in this section. The feasibility of expansion to other ESSENCE cities, estimated costs to deploy the system in these cities, possible benefits of the system, and recommendations for future work are also discussed.

4.1 GENERAL SYSTEM ARCHITECTURE FOR ESSENCE WATER SECURITY INITIATIVE – CONTAMINATION WARNING SYSTEM

The high-level design of the software developed for the Drinking Water Surveillance System is described in this section. The overall system includes the core ESSENCE system, the Intelligent Decision Support Network (IDSN) Manager, and the BN visualization extensions integrated into the ESSENCE system for visualizing, drilling down into, and analyzing the results of the Bayesian detection algorithms.

4.1.1 Scope

In this section, all of the major components and their interfaces are described. The core ESSENCE system is only described in detail as it relates to the BN tools and the related visualization and analysis components.

4.1.2 Background

This project takes the existing ESSENCE Disease Surveillance System and integrates the new visualizations to support the IDSN Bayesian analysis results. New visualization tools provide coordinated drill-down through the water analysis and health analysis branches of the Bayesian analysis network.

4.1.3 System Overview

Three major components of the Drinking Water Surveillance System can be seen in Figure 20: the Enterprise ESSENCE system, the Intelligent Decision Support System (IDSS) system, and the ESSENCE Water Safety Web Module. Enterprise ESSENCE and its various components form the foundation on which the Drinking Water Surveillance System is built. The water data-source(s) are integrated into this foundation. The IDSS system manages the definition, processing, evaluation, and result generation for the Bayesian Fusion process (in the blue box). When the BN network runs, the results from the Fusion BN are accumulated and then passed to the ESSENCE Water Safety Web Module. The Water Safety Web Module also uses some of the IDSS configuration to aid in the visualizations. Within the ESSENCE Web interface, the new web module provides the user with a visualization of the analysis results, drill-down capabilities, direct navigation, and analysis and details of the underlying data.

4.1.4 System Architectural Design

Figure 20 is a block diagram of the system architecture. Not all data sources (e.g. Retail Sales) shown in this diagram are used for the Water Module. However, the diagram does show how the water data is integrated in relation to the other components of the

Enterprise ESSENCE system. The tan colored components in the diagram are existing ESSENCE components. The blue shaded boxes represent additional components for the Water Module.

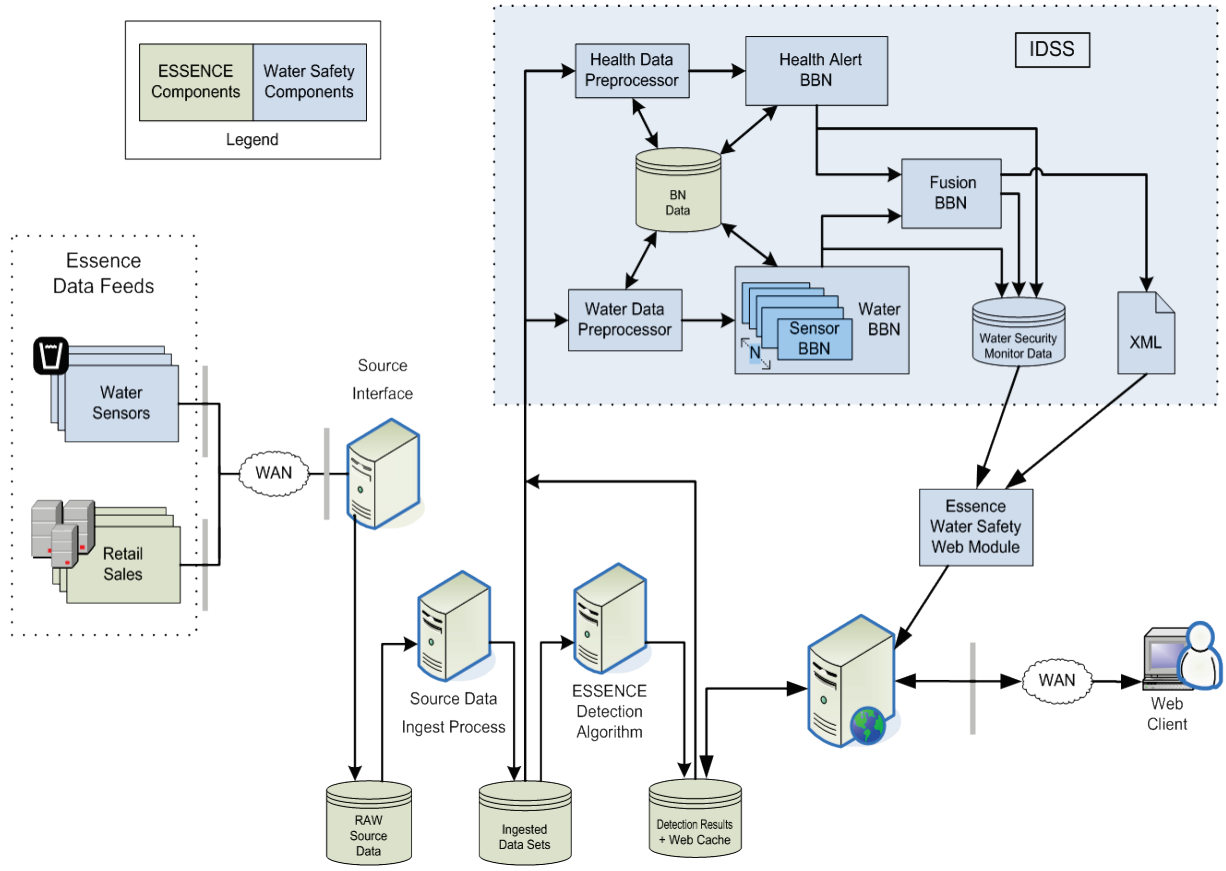


Figure 20 System Architecture

WAN = Wide Area Network, XML = Extensible Markup Language, BNN = Bayesian Belief Network

4.1.5 System Components

The following subsections describe the three main components of the Drinking Water Surveillance System in more detail.

4.1.5.1 Intelligent Decision Support System

The Intelligent Decision Support System (IDSS) implements a generic framework resulting in maximum flexibility for managing and executing distributed networks of BNs (Figure 21). IDSS is composed of a property file, a configuration file, one or more Databases, one or more BNs, and a Java™ (Sun Microsystems, Inc.) application for processing the network of BNs.

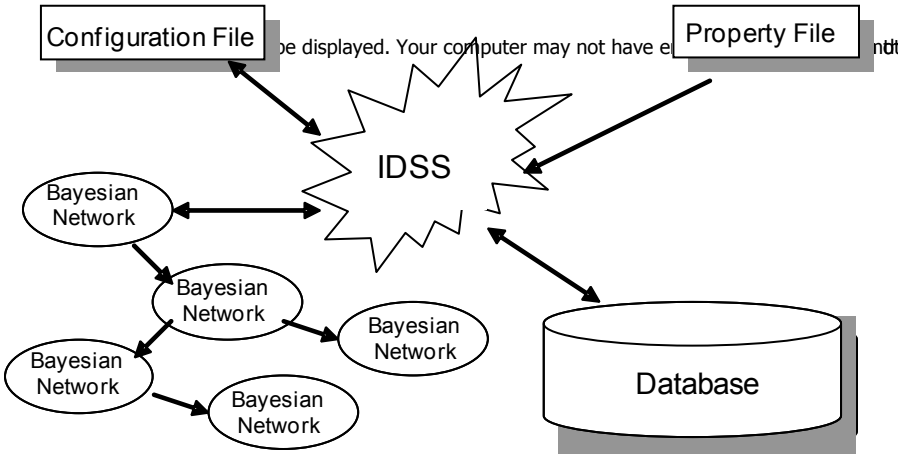


Figure 21 Intelligent Decision Support System (IDSS) Framework

4.1.5.1.1 Network Processor

The Network Processor is a Java-based application that utilizes a configuration file to process one or more BNs. For each BN in the IDSS, it reads or calculates the input nodes' values, processes the BN using the Netica™[18] (Norsys Software Corporation). Interface, and stores the value of the output nodes for the next course of action. The next course of action can be displaying the node's value in an application or submitting it as an input to the next BN in the system.

4.1.5.1.2 Property File

The property file defines the configuration file and its location, the Netica files, and other system-related information. The network processor needs all this information to access the data and call Netica's Application Programming Interface (API).

4.1.5.1.3 Configuration File

The configuration file is an Extensible Markup Language (XML) document that describes the Intelligent Decision Support Network (IDSN) and all of its components. A section on visualization of the data is also included. At the highest level, the file defines all the BNs and the order for processing them and instructs the IDSS to include or exclude BNs for each run of the system. For each network, the configuration file describes every input node and output node and the methods for acquiring the value for the input nodes and storing the result of the output nodes. Some output nodes may have additional instruction on displaying them in an application.

The configuration file lists constraints that apply to the incoming data. For example, it can instruct the IDSS to analyze the data for a defined period of time, or limit the data that is collected to a particular location.

4.1.5.1.4 Database

IDSS can communicate with multiple databases at once. The data for each BN can come from one or more databases. In fact, each node of the network in the system can query different databases. The values of the output nodes are also stored in one or more databases.

4.1.5.1.5 Bayesian Networks

BNs are graphical probabilistic models. Each BN is directed acyclic graph with the nodes representing variables and the edges showing conditional dependencies between variables. A Java API by Netica is used for accessing and processing these networks.

4.1.5.2 Water Module

4.1.5.2.1 Water Data Feeds

Water sensor data is continuously collected from provided data feeds and are stored as raw source data. The raw data is then processed to produce usable data for the Water Safety System (Ingested Data).

4.1.5.2.2 Water Data Preprocessor

A server is utilized to provide water processing. The server runs an executable Java application that is configured using an external configuration file that details the data to be processed as well as the processing steps. Using this configuration file, certain steps in the analysis process can be controlled when running. The processing can include several analysis pieces, performed in batches as declared in the configuration file. The server will run water data clustering as the first task based on specific algorithmic analysis of several data points. The detection algorithms then process the clustered data and store the resultant data in a database for processing by the IDSS.

4.1.5.2.3 Health Data Preprocessor

Health data is processed on ESSENCE data collection servers using ESSENCE techniques to store data in appropriate formats for analysis.

4.1.5.2.4 Water Bayesian Network

The IDSS operates on a server to retrieve water data from both processed results and raw data stores to perform final analysis using several IDSS Bayesian belief techniques. The Water BN operates over a set of water sites and stores the results in BN Data.

4.1.5.2.5 Health Alert Bayesian Network

Health data are utilized by the IDSS to perform final analysis using several IDSS Bayesian belief techniques. The Health Alert BN operates over a set of grouped health data and stores results in BN Data.

4.1.5.2.6 Fusion Bayesian Network

The Fusion BN incorporates results from both Water and Health Alert BNs to provide a correlated Bayesian belief result set for specific locations.

4.1.5.2.7 Bayesian Network Data

A BN Data server houses the results of all IDSS processes with resultant data. These data can then be displayed by the ESSENCE Web Security Web Module to allow specialists to analyze results.

4.1.5.2.8 ESSENCE Web Security Web Module

The Web module runs on an ESSENCE server and hosts the results of the Water Security Processing pieces. These results are displayed through a custom Web visualization site.

4.1.5.3 ESSENCE

The system architecture of ESSENCE is out of the scope of this document. For more detail on the ESSENCE system architecture, please refer to ESSENCE documentation.

4.2 FEASIBILITY OF EXPANSION TO OTHER CITIES

The current system is a prototype developed for a proof-of-concept demonstration. The system has not undergone extensive testing on the software implementation side and the detection and fusion algorithms have only been tested on a limited set of simulated injects. To thoroughly validate the performance of the water security module, a set of realistic data that simulates both injects into the drinking water system and the resulting human health effects is needed. Although the software/information technology (IT) portion of this project was developed to be flexible, an extended break in development time would increase the difficulty of testing this system at other locations.

Expansion to other cities is dependent on the willingness of users to a) learn a new system, b) accept and trust (with time) an unfamiliar method for detecting anomalies in their data, and c) invest up-front time to develop appropriate region definitions for their particular city. In addition, maintaining and updating the system would involve ongoing costs.

4.2.1 Cost Estimate per City

As mentioned in the preceding section, the Water Security Module is still a prototype system that 1) has not been thoroughly tested and validated with data that realistically simulates waterborne outbreaks and 2) has not been run through rigorous software testing to identify all bugs in the user interface. Additional effort is required to automate the process of getting the water quality data into the water security module for processing. The estimated cost, given in Staff Months (SM), for testing and installing the current prototype module in other cities currently using ESSENCE can be broken down into the following elements:

0.25 – 0.50 SM	Coordinating and setting up data transfer protocols with the water utilities. (Effort is dependent on types [continuous and/or grab sample] and amount of data.)
0.50 – 1.00 SM	Analyzing historical data from both the public health and water utility sides for determining appropriate regions. (This includes initial clustering of grab sample data (clustering is not needed for the continuous data) to obtain baselines and testing current algorithms on data.)
0.25 SM	Consulting local public health and water utility officials to understand area specific factors and to receive their input on how to divide the region spatially.
0.50 – 1.00 SM	Building and populating water quality database with historical data. (Effort is dependent on amount and types of data. As continuous monitors eventually come on-line at different water utilities, the amount of effort to build and populate the database may increase.)
0.50 – 1.00 SM	Additional software/IT effort associated with any bugs that may come up during installation to initial support on how to use the tool

The preceding estimates do not include costs for any additional hardware. As with most software, there are ongoing costs associated with maintenance, support, and possible upgrades. These costs depend on the time period (weeks, months, or years) that the system would be used at the install location. Additional cost can be associated with fixing module bugs found during the exercise, automating data ingest to the database, and

implementing improvements (additional functionality, algorithm improvements, and GUI improvements) suggested by exercise participants.

4.2.2 Perceived Benefits

The BN approach provides a quantitative framework to use non-statistical evidence such as media and intelligence reports to improve prior probabilities for decision-making. Such evidence can be incorporated by means of BN nodes that can influence the output probabilities only when information is received. The approach also has the following important advantages:

- Structural capacity to take advantage of engineering and epidemiological expertise along with available historical data
- Management of algorithmic false positives resulting from the sheer number of statistical tests performed, with weighted probabilistic corroboration of consensus among these tests
- Probabilistic accommodation of differences in relevance, data quality, data rate, reliability, and other factors complicating the cognitive fusion of information
- Transparency that can help overcome reluctance to use automated decision-support tools, since the BN diagram can give prompt visual explanation of the logic behind the outbreak warnings.

5 References

1. Liang J., Dziuban E.J., Craun G.F., Hill V., Moore M.R., Gelting, R.J., Calderon R.L., Beach M.J., Roy S.L., "Surveillance for Waterborne Disease and Outbreaks Associated with Drinking Water and Water Not Intended for Drinking–United States, 2003–2004," *MMWR Surveill Summ* 55(SS12), 31–58, 2006.

2. Lombardo J.S., and Ross D., "Disease Surveillance, A Public Health Priority," Chap. 1, in: *Disease Surveillance: A Public Health Informatics Approach,* J. S. Lombardo and D. L. Buckeridge (eds.), John Wiley and Sons, Hoboken, NJ, 20–21, 2007.

3. White House Office of the Press Secretary, Homeland Security Presidential Directive (HSPD), HSPD-7: *Critical Infrastructure Identification, Prioritization, and Protection,* 17 December 2003.

4. White House Office of the Press Secretary, Homeland Security Presidential Directive (HSPD), HSPD-9: *Defense of United States Agriculture and Food,* 3 February 2004.

5. *U.S. General Accounting Office, "Drinking Water: Experts' Views on How Future Federal Funding Can Best Be Spent to Improve Security," Report to the Committee on Environment and Public Works, U.S. Senate, GAO-04-029, U.S. Government Printing Office, October 2003.*

6. *Jackson M.L., Baer A., Painter I., Duchin J., "A Simulation Study Comparing Aberration Detection Algorithms for Syndromic Surveillance," BMC Med Inform Decis 7, 6, 2007.*

7. Reis B.Y., Kohane I.S., Mandl K.., D. "An Epidemiological Network Model for Disease Outbreak Detection," *PLoS Med* 4(6): e210, 2007.

8. Buckeridge D.L., Okhmatovskaia A., Tu S., O'Connor M., Nyulas C., Musen M.A., "Understanding *Detection Performanc*e in Public Heal*t*h Surveillance: Modeling Aberrancy-Detection Algorithms," *JAMIA* 15(6), 760-769, 2008 Nov-Dec.

9. Burkom H., Ramac-Thomas L., Babin S., Holtry R., Mnatsakanyan Z., Yund C., "An Integrated Approach for Fusion of Environmental and Human Health Data for Disease Surveillance," accepted in *Statistics in Medicine*.

10. Lin J.S., Burkom H.S., Murphy S.P., Elbert Y., Hakre S*., Babin S, Feldman A, ,* "Bayesian Fusion of Syndromic Surveillance with Sensor Data for Disease Outbreak Classification," Chap. 6, in: *Life Science Data Mining*, S. Wong and C.-S. Li (eds.), Science, Engineering, and Biology Informatics – Vol. 2, World Scientific Publishing, Inc., Hackensack, NJ, 2007.

11. Pearl J., "Fusion, Propagation, and Structuring in Belief Networks," *Artif Intell* 29(3), 241–288, 1986.

12. Babin S.M., Burkom H.S., Mnatsakanyan Z.R. Ramac-Thomas L.C., Thompson M.W., Wojcik R.A., Lewis S.H., Yund C., "Drinking Water Security and Public Health Disease Outbreak Surveillance," *Johns Hopkins APL Tech Digest* 27(4), 403-411, 2008.

13. *Water System Architecture U.S. Environmental Protection Agency. WaterSentinel System Architecture, Draft, Version 1.0 December 12, 2005. EPA 817-D-05-003. p. vii http://epa.gov/watersecurity/pubs/watersentinel_system_architecture.pdf*

14. Hales C., Sniegoski C., Coberly, J., Jerome Tokars J., "Defining Clinical Condition Categories for Biosurveillance," *Advances in Disease Surveillance* 4, 95, 2007.

15. *FACTOIDS: Drinking Water and Ground Water Statistics for 2007*, Office of Water (4606M), EPA 816-K-07-004, March 2008, www.epa.gov/safewater/data

16. Mandelberg M.D. and Frizzell-Makowski L.J., "Acoustic Provincing of Ocean Basins," *OCEANS 2000 MTS/IEEE Conference and Exhibition* 1, 105-108, 2000.

17. Hart, D. B., McKenna, S.A., Klise,K., Wilson, M.P., and Murry, R. CANARY User's Manual and software upgrades. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-08/040A, 2009.

18. The Netica APIs are a family of Bayesian Network toolkits. http://www.norsys.com/netica_api.html

19. "Public Health Syndromic Surveillance Project ESSENCE Demonstration for NHSRC Cincinnati" (contract title). John Hopkins University, Prime Contractor. EPA EP-C-06-074 Task Deliverable 1(b), Dec 22, 2006.

20. Lickert, R., "A Technique for the Measurement of Attitudes," *Archives of Psychology*, 140, 1-55, 1932.

21. Hayes, M. *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons, Inc., New York, 1996.

22. McKenna S.A., Wilson M., and Klise K.A., "Detecting Changes in Water Quality Data," *J Am Water Works Assoc*, 100(1), 74–85, 2008.

23. CDC case definitions as a web reference: http://www.cdc.gov/ncphi/disss/nndss/casedef/case_definitions.htm

24. Lennteck Water Treatment Solutions. "Disinfectant Chloramines" as a web reference: http://www.lenntech.com/processes/disinfection/chemical/disinfectants-chloramines.htm.

United States
Environmental Protection
Agency

PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT NO. G-35

Office of Research and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
$300