A Coupled Decision Trees Bayesian Approach for Water Distribution Systems Event Detection

Jonathan Arad¹, Lina Perelman², and Avi Ostfeld³

¹ M.Sc. student, Faculty of Civil and Environmental Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel; PH: +972-4-8295943; FAX: +972-4-8228898; E-mail: aradjon@tx.technion.ac.il

² Ph.D., Research Associate, Faculty of Civil and Environmental Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel; PH: +972-4-8292630; FAX: +972-4-8228898; E-mail: lina@tx.technion.ac.il

³ Associate Professor, Faculty of Civil and Environmental Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel; PH: +972-4-8292782; FAX: +972-4-8228898; E-mail: ostfeld@tx.technion.ac.il

Abstract

Detecting contamination events in water supply systems is a constant concern for utilities. It is reasonable to assume that injection of foreign substances will affect the behaviour of typically measured water parameters. For this reason, identifying contaminants using water quality and hydraulic measurements which are regularly monitored is appealing. A generic framework integrating Decision Trees (DTs) and Bayesian sequential probability updating rule is presented for detecting contamination events in Water Distribution Systems (WDS). The Aquatic Event Detection Algorithm (AEDA) utilizes DTs to depict the correlation between water quality and hydraulic parameters in order to detect possible outliers. The analysis is followed by updating the probability of a contamination event by recursively applying Bayes rule. AEDA is assessed through correlation coefficient ($R^2$), Mean Squared Error (MSE), confusion matrices, Receiver Operating Characteristic (ROC) curves, and True and False Positive Rates (TPR and FPR). AEDA is tested using simulated contamination events, imposed on water parameters, to imitate pollution scenarios in WDS.

Introduction

Event detection is one of the current most challenging topics in water distribution systems analysis. WDS are inherently vulnerable comprising numerous exposed elements which can be exploited for malicious actions (Kroll, 2006). It is apparent that some critical elements of WDS will have to be physically protected; however, securing each apparatuses is not feasible. Thus the development of other methods ensuring the delivery of sufficient (quantity) and adequate (quality) drinking water is needed. Particularly, water quality security research is currently focused on searching for quality threats in WDS through analysing routinely measured hydraulic and water quality parameters, also known as the surrogate approach (Hall et. al., 2007).

This study focuses on interpreting data collected from sensors measuring routine parameters (e.g. Chloramine, pH, Temperature, Electrical Conductivity) to detect outliers indicating possible contamination events. The observations attained from sensors are time series readings of the water parameters. These time series contain part of the inputs from a Supervisory Control and Data Acquisition (SCADA) system. The proposed methodology utilizes DTs to estimate the relationships between water parameters in a WDS.
A number of models have been developed to understand and model multi-components in a WDS, such as chlorine decay, reformation of microbial contaminants, and substrate consumption; however, the application and calibration of these models is a difficult task due to the number of parameters involved and the information required (Chungsying et al., 1994 and Rauch et al., 1999). Several related works are conducted in this field, aiming at finding quality faults to enhance system security. The CANARY event detection software is developed at Sandia National Laboratories in collaboration with the EPA National Homeland Security Research Center providing both off-line and real-time analysis tools for monitored data to detect anomalies and indicate possible contamination (Hart et al., 2010 and Murray et al., 2010). The CANARY incorporates water quality assessment, residual classification, and water quality event determination (Hart et al., 2007). The HACH GuardianBlue™ Early Warning System analyses monitored data, finds significant deviations from baseline, matches the event to previous patterns through events fingerprint, and alarms possible events.

The goal of this work is to provide indications of potential contamination events, not for a specifically monitored pollutant, but for a wide range of possible contaminants. The proposed methodology consists of two main stages: 1) Off-line – Create and assess the DTs model for temporary analysis of multivariate water time series in a WDS, 2) On-line – Combine the developed DTs model with Bayesian sequential analysis for estimating the probability of an event.

Problem description

The detection of a contamination event using the surrogate approach, i.e. identifying contaminants using regular water quality and hydraulic measurements, is comprised of three main processes: estimating future data based on past observations, classifying computed residuals, and making a decision. Figure 1 shows a schematic flowchart of the process. The input contains time series data of water quality and hydraulic parameters. The objective of the approach is to analyse the input attained from sensors spread across the WDS collecting real-time data and concluding whether the data represents normal routine behaviour of the WDS or a contamination event. Significant deviation from normal behaviour or recognisable patterns are an indication of the presence of contaminants.

![Figure 1. Process overview](image)

Methodology

Decision Trees

This study utilizes regression trees for detecting outliers in water quality measurements. The method used to construct the DTs is shortly described hereafter and fully detailed in (Arad et al., 2011). A pruning process was conducted in order to minimize the number of splitting rules taking into account the cost of the tree. The cost of a DT is the sum over all terminal leafs of the estimated probability of a node times the cost of a node. The cost of a node is the average squared error over the observations in that node. Pruning is performed in order to optimize tree size to achieve a balance between performance and utilization (Breiman, et al., 1984.).
Model assessment

Traditionally, classification model performance assessment is accomplished using metrics derived from a confusion matrix. The confusion matrix represents the model's classification of all observations to appropriate classes giving four possible labels to each observation – True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). A number of model performance metrics can be derived from the confusion matrix, for example sensitivity and specificity of the model. Sensitivity is defined as the True Positive Rate (TPR) and specificity is the True Negative Rate (TNR), where 1-Specificity is a more commonly used metric stating the False Positive Rate (FPR) as shown in Eq. 1.

\[
\text{TPR} = \frac{TP}{TP + FN} = \text{Sensitivity} \quad \text{FPR} = \frac{FP}{FP + TN} = 1 - \text{Specificity} \quad (1)
\]

TPR and FPR are also used to construct ROC curves, which visually depict the same information as the confusion matrix, demonstrating the fundamental performance trade-off between TPR and FPR in an intuitive fashion.

Sequential Bayesian updating

In sequential analysis the number of observations is not known in advance, instead, observations come in sequence, and a decision needs to be made on the current state. Each decision can take one of three forms: 'Non-event', 'Event', or 'Take an additional observation'. Initially, the probability of an event is assumed rare and with each new observation the posterior probability of an event is sequentially updated using Bayes rule, as shown in Eq. 2.

Given \( P(\theta = \theta_1) = \pi_j \) and \( \pi_j(t) = P(\theta = \theta_1 | y_{j,t}) \) the posterior probability is:

\[
\pi_j(t+1) = \frac{P(y_{t+1} | \theta_1)\pi_j(t)}{P(y_{t+1} | \theta_1)\pi_j(t) + P(y_{t+1} | \theta_0)(1-\pi_j(t))}
\]

where \( \theta_1 = \text{Event} \), \( \theta_0 = \text{Non-event} \), and \( y \in \{\text{outlier, normal}\} \).

The outcome of the probability projects on the decision considering a possible event. Initial prior probability is set to \( \pi_j(0) = \pi_0 \). \( P(y_{t+1} | \theta_1) \) and \( P(y_{t+1} | \theta_0) \) are taken from the confusion matrix, namely the TPR and FPR, respectively.

Aquatic Event Detection Algorithm (AEDA)

The proposed scheme relies on multivariate time series data collected by SCADA system sensing hydraulic and water quality data gathered from a WDS. Historical data is used for the data driven model training, error threshold setting, and model assessment, termed the Off-line procedure. New incoming observations are used to discover possible faults in real-time, termed the On-line procedure. The main steps of AEDA are described below and depicted in Figure 2.

**Off-line procedure**

1. **Data driven model**

DTs are used for processing characteristics and model the relationships between multivariate water parameters in a WDS. The governing concept is that during normal operation conditions the parameters exhibit some associative behavior and correlation, whereas during events these relations are violated. DTs are suitable for this purpose since their development...
does not require a priori knowledge of the physical and chemical laws governing the parameters. A DT is constructed for each target quality parameter with the input vector containing measured time series of all predictive parameters and lagged target parameter, as formulated in Eq. 3.

\[ \hat{x}_i(t) = f(x_i(t), x_{i-1}(t), x_i(t-1), x_{i+1}(t), \ldots, x_n(t)) \]  

(3)

Where \( x_i(t) \) and \( \hat{x}_i(t) \) are the measured and estimated water parameters at time \( t \), respectively, and \( f(\cdot) \) is defined by the DT. The fit of the model is evaluated through mean, Standard Deviation (STD), and correlation (\( R^2 \)) between the measured and estimated parameters.

2. Residual estimation and classification

Residuals are estimated as the difference between measured and estimated parameters' values, as shown in Eq. 4, represented as time series.

\[ \text{ER}_i(t) = x_i(t) - f_i(\cdot) = x_i(t) - \hat{x}_i(t) \]  

(4)

where \( \text{ER}_i(t) \) is the estimated residual for parameter \( i \) at time step \( t \). The residuals of the model under normal conditions (i.e. without events) are evaluated through MSE.

For each DT, the estimated residuals are bounded, such that the majority of the errors rest within the upper and lower limits. Observations exceeding the threshold are considered to be outliers. Identified outliers are not yet labeled as events and undergo further analysis.

3. Model assessment

Next, the ability of constructed DTs to detect outliers is evaluated based on their performance during events. As mentioned, it is expected that during contamination events the interplay between some or all of the parameters will change compared to normal conditions, and the model will result in larger errors. The performance of DT models is measured through confusion matrices and ROC curves. The confusion matrix represents the model's classification of all observations to one of four classes: TP – the residual was classified as an outlier during an actual event; FP – the residual was classified as an outlier during routine operation; TN – the residual was classified as plausible model error during routine operation; and FN – the residual was classified as plausible model error during an actual event. ROC curves can be constructed based on these confusion matrices to graphically represent the fundamental tradeoff between sensitivity and specificity, i.e. TPR and FPR, respectively. The results of the analysis are used in the On-line stage of AEDA.

On-line procedure

1. Residual evaluation and classification

For each new observation and for each target parameter: A) The DT created in Eq. 3 is used to estimate the value of the target parameter \( \hat{x}_i(t) \). B) The residual \( \text{ER}_i(t) \) is evaluated according to Eq. 4 and classified as a possible outlier or as an expected error bounded in the thresholds.

2. Event probability updating

The probability of an event is updated using sequential Bayesian rule according to Eq. 2, depending on the new observation being classified as an outlier or not, and on the TPR and FPR calculated in step 3 of the Off-line procedure. The posterior probability is the univariate
event probability, i.e. it is updated independently for each parameter and labels the likelihood of a contamination event based solely on the target parameter. An event is declared when the probability exceeds a threshold value which can be defined in number of orders from the initial probability $O(n10^{-d})$, where $d$ is the power of the initial probability and $n$ is order of the posterior probability.

3. Fuse decision

At each time step, univariate event probabilities are fused to give a unified multivariate event probability reflecting the likelihood of an event based on all parameters. Each parameter is assigned a weight reflecting its influence on the synchronized decision, for example, uniform weights given no prior information, proportional to the area of its ROC curve, or based on experts' opinion. Again, a critical probability $P_{\text{Thres}}$ is defined to declare a contamination event.

![Aquatic Event Detection Algorithm (AEDA) scheme](image)

**Figure 2.** Aquatic Event Detection Algorithm (AEDA) scheme

Application and results

Data preparation

1. Data acquisition and partitioning

The methodology is tested on real data collected by a utility in the United States and is available from CANARY. The attained data was collected by the SCADA system sensing hydraulic and water quality data gathered from a WDS. The data contains roughly 3,000 hours (4 months) of water parameters under normal operating conditions and includes the following parameters: Chloramine, Electrical Conductivity (EC), pH, Temperature, Total Organic Carbon (TOC), and Turbidity. The data was divided into two subsets, 67% for training and 33% for testing. The training subset is used for the Off-line process, to create and
train the data driven model. The testing subset is used for the On-line process, to assess the frameworks' power through its ability to identify contamination events.

2. Event simulation

Contamination events in water distribution systems heavily depend on environmental factors making the events hard to be pre-specified. To cope with this difficulty, contamination events were simulated and superimposed on routine patterns, characterized by their magnitude, direction, and duration of deviation (Klise and McKenna, 2006). In this work, the deviations from routine patterns were randomly sampled receiving values in the range of \{0.5, 2.5\}. The shape of the events was assumed to be Gaussian with duration of 25 time steps. This work presents the outcome of the proposed algorithm based on contamination events with randomly sampled deviations, resulting in different magnitudes of event for all parameters.

Off-line procedure

1. Data driven model

Six DTs were created and trained, one for each parameter \(x_i(t)\) \(i = 1,\ldots,6\) with water parameters \{Chloramine, EC, pH, Temperature, TOC, Turbidity\} respectively, to estimate the target water quality parameters and the relationships between them. One tree was constructed for each target quality parameter with the input vector according to Eq. 3. For example, the model for estimating Chloramine takes the following inputs:

\[
\hat{x}_{\text{Chloramine}}(t) = f \left[ x_{\text{EC}}(t), x_{\text{pH}}(t), x_{\text{Temperature}}(t), x_{\text{TOC}}(t), x_{\text{Turbidity}}(t), x_{\text{Chloramine}}(t-1) \right] \tag{5}
\]

The trees then undergo a pruning process. Figure 3 depicts the DT pruning process of Electrical Conductivity DT. The first tree consists of 452 nodes whilst the tree chosen for the estimation process contains only 34 nodes. The dotted line denotes one standard deviation from the cost of the full tree. The selected pruned DT has minimum cost.

Table 1 lists the mean, STD, and \(R^2\) statistics for measured and estimated training data of all six parameters. It can be seen that all DTs have similar means and STD to measured data, and high \(R^2\) (above 0.886), with an exception of Turbidity with \(R^2=0.322\)

![Figure 3. Tree cost versus tree size](image-url)
2. Residual estimation and classification

Residuals are calculated for each new observation of each quality parameter as the total error (difference between measured and estimated values) according to Eq. 4. Table 1 lists the MSE of the DT models (training and testing) of all six parameters. The results demonstrate small MSE for all parameters considering their measuring scale, with the exception of EC.

3. Model assessment

The final stage of the Off-line procedure is to evaluate the sensitivity and specificity of the model. For each parameter TPR and FPR are calculated using Eq. 1 based on residuals classification and outlier identification from the previous step. Table 1 lists the TPR, FPR, and the area under ROC curves for all parameters. The results show that TPR is in the range of [0.194-0.758] with Turbidity having the highest TPR and Chloramine – the lowest; FPR is in the range of [0.000-0.100] with TOC having the lowest rate and Chloramine – the highest; and ROC area in the range of [0.507-0.777] with Chloramine and Turbidity having the smaller and the larger areas, respectively.

On-line procedure

1. Residual evaluation and classification

The remaining unseen data was used for testing. For each new observation and for each parameter, the value of the target parameter and the error were estimated using the DT models created in step 1 of the Off-line procedure. Table 1 lists models’ and errors' statistics of the testing data. Results show that the mean and STD of the DTs model for all parameters during testing remain relatively close to the measured data, however, $R^2$ and MSE differ significantly. This can be addressed to the fact that $R^2$ and MSE are highly susceptible to outliers and noisy data, hence should be replaced with more robust estimators.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.946 (1.946)</td>
<td>78.141 (78.141)</td>
<td>9.067 (9.037)</td>
<td>17.243 (17.243)</td>
<td>0.952 (0.952)</td>
<td>0.222 (0.222)</td>
</tr>
<tr>
<td>STD</td>
<td>0.027 (0.289)</td>
<td>48.676 (48.795)</td>
<td>1.109 (1.110)</td>
<td>2.340 (2.345)</td>
<td>0.268 (0.270)</td>
<td>0.059 (0.104)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.886</td>
<td>0.995</td>
<td>0.999</td>
<td>0.994</td>
<td>0.991</td>
<td>0.322</td>
</tr>
<tr>
<td>MSE</td>
<td>0.009 (0.009)</td>
<td>11.537 (11.537)</td>
<td>0.000 (0.000)</td>
<td>0.025 (0.025)</td>
<td>0.000</td>
<td>0.007</td>
</tr>
<tr>
<td>TPR</td>
<td>0.194</td>
<td>0.33</td>
<td>0.384</td>
<td>0.146</td>
<td>0.244</td>
<td>0.758</td>
</tr>
<tr>
<td>FPR</td>
<td>0.100</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.000</td>
<td>0.075</td>
</tr>
<tr>
<td>ROC Area</td>
<td>0.507</td>
<td>0.693</td>
<td>0.608</td>
<td>0.655</td>
<td>0.765</td>
<td>0.777</td>
</tr>
<tr>
<td><strong>Testing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.997 (2.011)</td>
<td>89.499 (88.670)</td>
<td>9.167 (9.150)</td>
<td>18.031 (18.086)</td>
<td>0.966 (0.990)</td>
<td>0.199 (0.257)</td>
</tr>
<tr>
<td>STD</td>
<td>0.061 (0.062)</td>
<td>53.169 (52.563)</td>
<td>0.211 (0.215)</td>
<td>1.150 (1.190)</td>
<td>0.278 (1.045)</td>
<td>0.058 (2.078)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.499</td>
<td>0.978</td>
<td>0.826</td>
<td>0.856</td>
<td>0.048</td>
<td>0.003</td>
</tr>
<tr>
<td>MSE</td>
<td>0.002</td>
<td>62.684</td>
<td>0.008</td>
<td>0.208</td>
<td>0.104</td>
<td>4.305</td>
</tr>
</tbody>
</table>

2. Event probability updating

For each new observation and for each parameter, the probability of the event is updated using sequential Bayesian rule based on Eq. 3, depending on the new observation being classified as an outlier or not, and on the TPR and FPR calculated in step 3 of the Off-line procedure. In this application, the initial probability of a contamination event was set to $\pi_j(0) = \pi_0 = 10^{-5}$ and the threshold probability for declaring an event was set to $\pi_j(T) = \pi_T = 7 \times 10^{-4}$. At this stage, the probability of an event is updated for each parameter individually and an event is declared based on a single parameter.

Figure 4 graphically illustrates six plots which demonstrate the likelihoods of a contamination event based on Bayesian sequential updating for each quality parameter separately, and true events. For example, TOC detects nine of ten events with few false alarms, Turbidity – seven out of ten events with zero false alarms, and Chloramine does not detect any of the events. Figure 4A demonstrates an example true event versus estimated event probability. The lags in the beginning and the end of the event result from the number of observations required to update the probability of an event.

![Figure 4](image_url)

**Figure 4.** Univariate event probability

3. Fuse decision

At each time step, the univariate event probabilities are fused to give the multivariate event probability, reflecting the likelihood of an event based on all measured parameters. Figure 5 shows six event probability plots. Each subplot illustrates the mutual probability of an event taking into account one or more equally weighted quality parameters. For example, in Figure 5 Alarm (1), alarm is raised even if only a single parameter signaled an event. In this case, all events are detected (i.e., high TPR), however, there are many instances of false alarms (i.e., high FPR). Figure 5 Alarm (3) shows that an alarm is raised when three parameters or more signaled an event. In this case, nine of ten events are detected with only one false alarm. Table 2 summarizes averaged results of multiple contamination events. It can be seen that the FPR does not increase with an increase in TPR.
Discussion and future work

A generic framework (AEDA) aimed at detecting anomalous behavior of water quality parameters was presented. The example application has shown DTs to be a useful tool for water quality assessment and classification. The DT model combined with the Bayesian updating rule, constitute a powerful procedure attaining a comprehensive tool for the decision maker. AEDA can be adopted to fit any WDS given a set of tailored parameters (e.g. event initial probability, critical probability). The presented framework uses estimation of measured parameters using DTs; However, AEDA is built as a generic platform which can utilize any estimation model (e.g. artificial neural networks).

Further work needs be conducted to test, generalize, and improve AEDA's performance. Residual classification should be further investigated by, for example, applying a dynamic threshold optimized to the number of historical observations (time period) and their statistics.

Figure 5. Multivariate event probability

Table 2. Detection table

<table>
<thead>
<tr>
<th>Event strength</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of events</td>
<td>10</td>
</tr>
<tr>
<td>Single indicator</td>
<td></td>
</tr>
<tr>
<td>Predicted events</td>
<td>TP</td>
</tr>
<tr>
<td>Chloramine [mg/L]</td>
<td>0</td>
</tr>
<tr>
<td>EC [mS/cm]</td>
<td>7</td>
</tr>
<tr>
<td>pH</td>
<td>4</td>
</tr>
<tr>
<td>Temperature [°C]</td>
<td>9</td>
</tr>
<tr>
<td>TOC [ppb]</td>
<td>9</td>
</tr>
<tr>
<td>Turbidity [NTU]</td>
<td>7</td>
</tr>
<tr>
<td>Multiple indicators</td>
<td></td>
</tr>
<tr>
<td>Predicted Events</td>
<td>TP</td>
</tr>
<tr>
<td>Alarm (1)</td>
<td>10</td>
</tr>
<tr>
<td>Alarm (2)</td>
<td>10</td>
</tr>
<tr>
<td>Alarm (3)</td>
<td>9</td>
</tr>
<tr>
<td>Alarm (4)</td>
<td>5</td>
</tr>
<tr>
<td>Alarm (5)</td>
<td>0</td>
</tr>
<tr>
<td>Alarm (6)</td>
<td>0</td>
</tr>
</tbody>
</table>

Alarm (1) * – at least one quality indicator raised an alarm

Discussion and future work

A generic framework (AEDA) aimed at detecting anomalous behavior of water quality parameters was presented. The example application has shown DTs to be a useful tool for water quality assessment and classification. The DT model combined with the Bayesian updating rule, constitute a powerful procedure attaining a comprehensive tool for the decision maker. AEDA can be adopted to fit any WDS given a set of tailored parameters (e.g. event initial probability, critical probability). The presented framework uses estimation of measured parameters using DTs; However, AEDA is built as a generic platform which can utilize any estimation model (e.g. artificial neural networks).

Further work needs be conducted to test, generalize, and improve AEDA's performance. Residual classification should be further investigated by, for example, applying a dynamic threshold optimized to the number of historical observations (time period) and their statistics.
(mean and STD). Additionally, when the decision about a contamination event is based on multiple quality indicators, weights should be assigned to each indicator relative to its prediction power to provide better results, for example, using the area under ROC curves as possible indication of importance.

References


CANARY, Event Detection Software, EPA. Website: https://software.sandia.gov/trac/canary


HACH GuardianBlue™ Homeland Security Technologies. Website: http://hachhst.com


