Journal of Environmental Management 161 (2015) 385–391

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

# Performance evaluation for three pollution detection methods using data from a real contamination accident

Shuming Liu*, Han Che, Kate Smith, Musuizi Lei, Ruonan Li

*School of Environment, Tsinghua University, Beijing, 100084, China*

ARTICLE INFO

*Article history:*
Received 21 May 2015
Received in revised form
9 July 2015
Accepted 10 July 2015
Available online xxx

*Keywords:*
Contamination detection
Pearson correlation
Euclidean distance
Linear prediction filter

ABSTRACT

Early warning systems have been widely deployed to safeguard water security. Many contamination detection methods have been developed and evaluated in the past decades. Although encouraging detection performance has been obtained and reported, these evaluations mainly used artificial or laboratory data. The evaluation of detection performance with data from real contamination accidents has rarely been conducted. Implementation of contamination event methods without full assessment using field data might lead to failure of an early warning system. In this paper, the detection performance of three contamination detection methods, a Pearson correlation Euclidean distance (PE) based detection method, a multivariate Euclidean distance (MED) method and a linear prediction filter (LPF) method, was evaluated using data from a real contamination accident. Results improve understanding of the implementation of detection methods to field situations and show that all methods are prone to yielding worse detection performance when applied to data from a real contamination accident. They also revealed that the Pearson correlation Euclidean distance based method is more capable of differentiating between equipment noise and presence of contamination and has greater potential to be used in real field situations than the MED and LPF methods.

## 1. Introduction

Protection of drinking water systems from accidental and intentional contamination events has increased in importance in recent years due to security concerns (Liu et al., 2014; Yang et al., 2009). Between 1992 and 2006, an average of 1906 contamination accidents occurred per year in China (Yang et al., 2010). For example, the Songhua River was contaminated by nitrobenzene from a chemical plant explosion in 2005, which resulted in a 4 day suspension of water supply to Harbin, China (Wang et al., 2012). One approach for avoiding or mitigating the impact of contamination is to establish an early warning system (EWS).

A key part of an EWS is the detection algorithm, which utilizes data from online sensors to evaluate water quality and detect the presence of contamination. Many studies have been conducted to develop detection algorithms using signals from conventional water quality sensors. As summarized by McKenna et al. (2008), there are two approaches to developing and testing event detection methods using water quality sensor signals. First, laboratory and test-loop evaluation of sensors and associated event detection algorithms provides direct measurement of chemical changes in background water quality caused by specific contaminants (Hall et al., 2007; Yang et al., 2009). Results from these physical experiments can be used to quantify which deviations from background water quality signals are indicative of contamination events. These responses can then be integrated into event detection methods. For example, Yang et al. (2009) proposed a real-time event adaptive detection, identification and warning (READiw) methodology in a drinking water pipe. The suggested adaptive transformation of sensory measurements reduced background noise and enhanced contaminant signals.

The second approach to event detection is based on signal processing and data-driven techniques (McKenna et al., 2008). For example, Kroll (2006) reported the Hach HST approach using multiple sensors for event detection and contaminant identification. Hart et al. (2007) reported a linear prediction filter (LPF). The LPF method predicts the water quality at a future time step and evaluates the residual between predicted and observed water quality values. Klise and McKenna (2006) developed an algorithm to classify the current measurement as normal or anomalous by calculating the multivariate Euclidean distance (MED). The MED

* Corresponding author.
*E-mail address:* shumingliu@tsinghua.edu.cn (S. Liu).

http://dx.doi.org/10.1016/j.jenvman.2015.07.026
0301-4797/© 2015 Elsevier Ltd. All rights reserved.

approach provides a measure of the distance between the sampled water quality and the previously measured samples contained in the history window. Liu et al. (2015c) presented a new detection method that identifies the existence of contamination by comparing Euclidean distance of correlation indicators, which are derived from the correlation coefficients of multiple water quality sensors. Allgeier et al. (2005) and Raciti et al. (2012) utilized artificial neural networks (ANN) and support vector machines (SVM) to classify water quality data into normal and anomalous classes after supervised learning. Perelman et al. (2012) and Arad et al. (2013) reported a general framework that integrates a data-driven estimation model with sequential probability updating to detect quality faults in water distribution systems using multivariate water quality time series. In general, these algorithms process the water quality data at each time step and compare this data with a preset threshold. If the deviation is greater than the preset threshold value, an alarm is triggered.

Researchers have attempted to evaluate the performance of these methods. The first group of methods has generally been evaluated using data from laboratory contamination injection experiments (Hall et al., 2007; Kroll, 2006; Liu et al., 2015a,b; Yang et al., 2009). As argued by McKenna et al. (2008), a drawback of the laboratory and test-loop results and the resulting algorithms is that variation of the background water quality in these systems may be considerably less than the variation observed in actual water systems. Evaluation of the performance of the second group of methods has mainly used artificial data or data from injection experiments in laboratory. The artificial data normally contains actual background data and artificial event data. For example, in work by McKenna et al. (2008), water quality data collected in a water utility in the United States were used to represent background water quality conditions. Simulated anomalous water quality events (or spikes) were then added to these data. Using observed hydraulics data from CANARY and simulated contamination event data, Perelman et al. (2012) reported that an ANN based detection method yielded a true positive rate of 90% with three false alarms. The READiw method developed by Yang et al. (2009) was capable of correctly detecting all contamination events for the experimental data under discussion. In a study by Liu et al. (2015c), the Pearson correlation Euclidean distance based method was applied to data from an injection experiment and it detected 95% of contamination events correctly with a 2% of false alarm rate. In general, the performance of these approaches is encouraging. However, these evaluations were conducted using only artificial water quality data or laboratory data. It is unclear how these approaches would perform in real contamination situations, in which water quality data contains much more background noise and fluctuations.

To understand the applicability of contamination detection methods, evaluation of these methods using data from actual contamination accidents is necessary. The objective of this paper was to evaluate and compare the performance of three detection methods using data from an actual contamination accident in a water source.

## 2. Methods and materials

The three methods evaluated in this study were Pearson correlation Euclidean distance (PE) based detection method, multivariate Euclidean distance (MED) method and linear prediction filter (LPF) method. These three methods are briefly introduced here.

### 2.1. The PE method

In a parallel study, Liu et al. (2015c) proposed the PE method, which includes three steps: calculation of Pearson correlation coefficients, calculation of correlation indicators and calculation of Euclidean distances.

**Step 1**: Pearson correlation coefficients for multiple sensor signals are calculated. In a previous study, Liu et al. (2014) reported that multiple water quality sensors could respond to a contamination event simultaneously. This is defined as a *correlative response* and is utilized in this study for event detection. Step 1 involves quantifying the extent of correlation using Pearson correlation coefficients, *r*, which are calculated as follows

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2} * \sqrt{\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2}} \tag{1}$$

in which $X$ and $Y$ refer to signal series from two separate water quality sensors (e.g. pH and ORP). $x_i$ and $y_i$ are the $i$th numbers in the signal series. $\overline{X}$ and $\overline{Y}$ stand for mathematical expectation. The number of data or *window size* is given by $n$. The *window size* is the number of past observations used to calculate the Pearson correlation coefficient.

**Step 2**: The value of $r_{XY}$ is between −1 and 1. If the value of $r_{XY}$ is close to 0, the correlation between $X$ and $Y$ is deemed to be weak. In this study, a *correlation indicator* $C_{XY}$ is used to denote whether two vectors are closely related. The value of $C_{XY}$ is either 0 or 1, which is obtained, as shown in Equation (2), by comparing $r_{XY}$ with a pre-set indicator threshold $C^*$.

$$\begin{cases} C_{XY} = 0 & if\ |r_{XY}| < C^*\ or\ X = Y \\ C_{XY} = 1 & if\ C^* \le |r_{XY}| \le 1 \end{cases} \tag{2}$$

**Step 3**: For the case of $s$ sensors, the correlation coefficient forms an $s$ x $s$ matrix, as does the correlation indicator. The correlation indicators above the diagonal are taken to construct a 1 × $m$ dimension vector $V$, which is called the correlation indicator vector (Equations (3) and (4)).

$$\begin{bmatrix} 1 & C_{12} & C_{13} & \cdots & C_{1s-1} & C_{1s} \\ & 1 & C_{23} & \cdots & C_{2s-1} & C_{2s} \\ & & 1 & \cdots & C_{3s-1} & C_{3s} \\ & & & \cdots & \cdots & \cdots \\ & & & & 1 & C_{s-1s} \\ & & & & & 1 \end{bmatrix} \tag{3}$$

$$\underbrace{[C_{12}C_{13}\cdots C_{1s-1}C_{1s}C_{23}\cdots C_{2s-1}C_{2s}\cdots C_{3s-1}C_{3s}\cdots C_{s-1s}]}_{\text{m}} \tag{4}$$

$m$ is determined by

$$m = \sum_{i=1}^{s-1} i \tag{5}$$

The Euclidean distance of the correlation indicator vector from the origin point, $\delta_{PE}$, is calculated using