Water Research 93 (2016) 222–229

Contents lists available at ScienceDirect

Water Research

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

Why conventional detection methods fail in identifying the existence of contamination events

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

School of Environment, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 31 August 2015
Received in revised form
11 February 2016
Accepted 12 February 2016
Available online 16 February 2016

Keywords:
Early warning system
Contamination detection
Water security
Pearson correlation

ABSTRACT

Early warning systems are widely used to safeguard water security, but their effectiveness has raised many questions. To understand why conventional detection methods fail to identify contamination events, this study evaluates the performance of three contamination detection methods using data from a real contamination accident and two artificial datasets constructed using a widely applied contamination data construction approach. Results show that the Pearson correlation Euclidean distance (PE) based detection method performs better for real contamination incidents, while the Euclidean distance method (MED) and linear prediction filter (LPF) method are more suitable for detecting sudden spike-like variation. This analysis revealed why the conventional MED and LPF methods failed to identify existence of contamination events. The analysis also revealed that the widely used contamination data construction approach is misleading.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Water systems are vulnerable to accidental and intentional contamination events. One example is the 2014 chemical spill involving crude 4-methylcyclohexanemethanol, which contaminated the Elk River in the state of West Virginia, United States of America (Whelton et al., 2015). The chemical spill occurred 1 mile upstream from West Virginia American Water's drinking water intake and resulted in the suspension of water service for up to 300,000 residents. One approach for mitigating the impact of contamination is to establish an early warning system, which would provide a fast and accurate means of detecting the existence of contamination events (Hasan et al., 2004; US EPA, 2005). A key part of an early warning system 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. These methods can generally be summarized into two groups.

Methods in the first group are mainly based on signal-to-noise principles. Laboratory and test-loop evaluation of sensors and associated event detection algorithms provide direct measurement of chemical changes in background water quality caused by specific contaminants (Hall et al., 2007; Kroll, 2006; Kroll and King, 2006; Yang et al., 2009). McKenna et al. argued that 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 an actual water system (McKenna et al., 2008). Meanwhile, although the absolute magnitude may be part of a detailed quantitative analysis, these methods mainly rely on qualitative observations. When performing a quantitative analysis, it is not only the absolute magnitude of the change that is important, but also the slope of the change and the magnitude relative to the size and fluctuations in the baseline. Thus, quantitative evaluation makes the signal-to-noise principles difficult to use since it is location-specific.

The second group of methods is based on signal processing and data driven based techniques (Allgeier et al., 2005; Kroll, 2006; McKenna et al., 2008; Raciti et al., 2012). Most early warning systems currently being used employ this type of detection method. 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

Abbreviations: COD, chemical oxygen demand; DO, dissolved oxygen; EWS, early warning system; FAR, false alarm rate; FN, false negative; FP, false positive; LPF, linear prediction filters; MED, multivariate Euclidean distance; ORP, oxidation reduction potential; PE, Pearson correlation Euclidean distance based method; PD, probability of detection; ROC, received operating characteristic; TN, true negative; TOC, total organic carbon; TP, true positive.

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

http://dx.doi.org/10.1016/j.watres.2016.02.027
0043-1354/© 2016 Elsevier Ltd. All rights reserved.

distribution systems using multivariate water quality time series. These algorithms generally process water quality data at each time step and compare the data with a preset threshold. If the deviation is greater than the preset threshold value, an alarm is triggered. Hart et al. (2007) developed a linear prediction filter (LPF) method, which 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). Water Research Foundation (2014) reported that early detection systems using sensor data from multiple sites may offer significant benefits over the ones using data from single site. The MED approach provides a measure of the distance between the sampled water quality and the previously measured samples contained in the history window. McKenna et al. (2008) evaluated the detection performance of the MED and LPF methods against simulated anomalous water quality data containing 10 levels of spike strength. Area under the receiver operating characteristic (ROC) curve was adopted for performance assessment. In McKenna et al.'s study, resulting areas under the ROC curve ranged from 0.46 for spike strengths of 1.0 (background) to 0.98 for strength of 3.5 standard deviations from the mean, where an ROC curve area of 1.0 indicates perfect detection. Liu et al. (2014, 2015a) presented a detection method that identifies the existence of contamination based on correlative coefficients. Later, Liu et al. (2015b,c) improved this method by employing Euclidean distance to evaluate the differences between background and contamination and developed a Pearson correlation Euclidean distance based (PE) method. By comparing the detection performance of the PE, MED and LPF methods against data from laboratory experiments and a real contamination event, Liu et al. (2015b,c) declared that the PE 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.

In the process of developing an event detection algorithm, it is essential that the detection performance be evaluated against contamination data. Generally, rigorous assessment of event detection algorithms requires a set of known events against which these algorithms can be evaluated. However, water quality datasets that originate from operating utilities, contain contamination events and include information on the location, timing and contaminant associated with each change rarely exist. Therefore, nearly all performance assessments of event detection algorithms developed over the past few decades have been based on artificial contamination datasets. One recent exception was that of Liu et al. (2015b) which evaluated the performance of detection methods using data from a real contamination incident. The most common way of creating an artificial contamination dataset is to include two parts: background water quality data and event data. The former is commonly from field observation, while the latter is artificially generated (McKenna et al., 2008).

Although detection algorithms in the literature have been shown to perform well when applied to artificial data, their detection performance in practice is less satisfactory. A typical complaint is the low true positive rate and high false positive rate. Many water contamination incidents have been discovered due to the death of aquatic organisms or the existence of strange odor, rather than through results from an early warning system. For example, the spill in the Elk River was reported by several local residents who began to notice a 'sweet smell' in the air (USA Today, 2014). Similar situations have occurred in China. Although over 1000 water contamination incidents occur each year, it is rare that a contamination incident is detected by an early warning system (Song, 2014).

This has raised questions about the effectiveness of early warning systems in the water industry and why early warning systems fail to trigger the alarm in the case of a real contamination event. There are two possible reasons for the failure. One is the reliability and accuracy of online sensors, which has been a major concern for early warning systems. In recent years, the performance of online sensors has significantly improved. Therefore, failure of early warning systems caused by online sensors is not discussed in this study. The other reason for failure is the detection algorithm. As discussed, many detection algorithms have been developed over the past few decades, but the performance of these algorithms has only been evaluated against artificial contamination data before implementation. Their performance in real contamination incidents has not been examined. It is unclear whether artificial contamination data can truly represent a real contamination incident or whether the method of designing these data leads to detection performance bias.

Therefore, the objectives of this study are: 1) to examine whether artificial contamination data can represent a contamination event in practice; 2) to investigate why conventional detection algorithms fail in detecting the presence of contamination. These objectives are achieved in this study through investigation of the detection performance of three algorithms against real and artificial contamination datasets.

2. Materials and methods

The PE, MED and LPF methods have been reported with promising detection performance in the literature. The PE achieved an ROC curve area of 0.97 (Liu et al., 2015c). The MED and LPF can reach an ROC curve area of 0.98 (McKenna et al., 2008). Therefore, these three methods were selected for discussion in this study. The following sections present a brief introduction of these three methods. For more information, the readers can refer to Liu et al. (2015c), Klise and McKenna, (2006) and McKenna et al. (2008).

2.1. The PE method

Liu et al. (2015b,c) proposed the PE method, which includes three steps: calculation of Pearson correlation coefficients, calculation of correlation indicators and calculation of Euclidean distances. In the PE method, the Pearson correlation coefficient, r, was adopted to quantify the extent of correlation. The number of data involved to calculate the Pearson correlation coefficient is denoted by a parameter n, i.e. *window size*. The PE method employs a *correlation indicator* C_{XY} to denote whether two vectors are closely related. The value of C_{XY} is either 0 or 1, which is obtained by comparing r with a pre-set indicator threshold C^*. For the case of s sensors, the correlation indicator forms an s x s matrix. The correlation indicators above the diagonal are taken to construct a correlation indicator vector. A contamination alarm will be triggered if the Euclidean distance of the correlation indicator vector from the origin point ($_{-PE}$) is greater than a *detection threshold* ($^{*}_{-PE}$).

2.2. The MED method

The MED method considers changes in water quality by comparing two successive distances in a multivariate space defined by the water quality signal (McKenna et al., 2008; Klise et al., 2006). The distance measure, $_{-MED}$, is the difference between the Euclidean distances of successive points to the mean of previous time steps (P_{MED}). A constant detection threshold, $^{*}_{-MED}$, is applied to determine if an event has occurred. The MED method identifies a contamination event when $_{-MED}$ is above this threshold. Otherwise, measurements are considered background.

Download English Version:

https://daneshyari.com/en/article/4480953

Download Persian Version:

https://daneshyari.com/article/4480953

Daneshyari.com