WATER RESEARCH 80 (2015) 109–118

Available online at www.sciencedirect.com

ScienceDirect

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

A multivariate based event detection method and performance comparison with two baseline methods

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

School of Environment, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 27 January 2015
Received in revised form
30 April 2015
Accepted 5 May 2015
Available online 12 May 2015

Keywords:
Contaminant classification
Conventional sensor
Early warning system
Euclidean distance
Pearson correlation
Water quality

ABSTRACT

Early warning systems have been widely deployed to protect water systems from accidental and intentional contamination events. Conventional detection algorithms are often criticized for having high false positive rates and low true positive rates. This mainly stems from the inability of these methods to determine whether variation in sensor measurements is caused by equipment noise or the presence of contamination. This paper presents a new detection method that identifies the existence of contamination by comparing Euclidean distances of correlation indicators, which are derived from the correlation coefficients of multiple water quality sensors. The performance of the proposed method was evaluated using data from a contaminant injection experiment and compared with two baseline detection methods. The results show that the proposed method can differentiate between fluctuations caused by equipment noise and those due to the presence of contamination. It yielded higher possibility of detection and a lower false alarm rate than the two baseline methods. With optimized parameter values, the proposed method can correctly detect 95% of all contamination events with a 2% false alarm rate.

1. Introduction

China has suffered thousands of water contamination events over the past few decades. Between 1992 and 2006, an average of 1906 contamination accidents occurred per year (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). More recently, in February 2012, the drinking water source of a city in the lower Yangtze River area of Jiangsu province was contaminated by a phenol spill from a South Korean cargo ship. One approach for avoiding or mitigating the impact of contamination is to establish an Early Warning System (EWS), which normally includes online sensors, a connected supervisory control and data acquisition (SCADA) system, a detection algorithm and a

Abbreviations: ANN, artificial neural network; ARMA, autoregressive moving average; CIE, contaminant injection experiment; 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; READiw, real-time event adaptive detection, identification and warning; ROC, receiver operating characteristic; SCADA, supervisory control and data acquisition; SVM, support vector machine; 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.2015.05.013
0043-1354/

decision support system (Hasan et al., 2004). EWS should provide a fast and accurate means to distinguish between normal variations in water parameters and actual contamination events.

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. Conventional water quality sensors have been playing a growing role in EWS because they are easy to maintain, reliable and cost-effective. 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; Kroll and King, 2006a, b; Liu et al., 2015a, b). For example, Hall et al. (2007) carried out a sensor response experiment for 9 types of contaminants and realized that more than one sensor responded to each tested contaminant. After noticing this phenomenon, researchers have attempted to develop contaminant detection methods using responses from multiple sensors. Yang et al. (2009) developed a real-time event adaptive detection, identification and warning (READiw) methodology in a drinking water pipe. The suggested adaptive transformation of sensor measurements reduced background noise and enhanced contaminant signals. In the method employed by Yang et al. (2009), the relative value of concentrations of free and total chlorine, pH and oxidation reduction potential (ORP) are used for contaminant classification. This allows for contaminant detection and further classification based on chlorine kinetics. Kroll (2006) developed the Hach HST approach using multiple sensors for event detection and contaminant identification. In this approach, signals from 5 separate orthogonal measurements of water quality (pH, conductivity, turbidity, chlorine residual, total organic carbon (TOC)) are processed from a 5-parameter measure into a single scalar trigger signal. The deviation signal is then compared to a preset threshold level. If the signal exceeds the threshold, the trigger is activated (Kroll, 2006). In Kroll's method, although responses from multiple sensors are utilized, their internal relationship is not explored. McKenna et al. (2008) 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 actual water systems. Another drawback of these types of methods is that the threshold level is site dependent. When applied to a situation different from the one for which the method is developed, field calibration is necessary.

The second approach to event detection is based on signal processing and data-driven techniques (McKenna et al., 2008). For example, Hart et al. (2007) reported a linear prediction filters (LPF) method. 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 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) compared the performance of LPF, MED and a time series increments method. These algorithms process water quality data at each time step to identify periods of anomalous water quality and the probability of a water contamination event having occurred at that time step. The averaged deviation between the observed and predicted responses from time series data for each sensor is compared with a preset threshold. If the averaged deviation is greater than the preset threshold value, an alarm is triggered. Allgeier et al. (2005) and Raciti et al. (2012) used 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. A common feature of signal processing and data-driven methods is that they rely mainly on pure mathematical data analysis. The characteristics of sensor responses to contaminants and the connections between these are not considered by these methods. For online water quality sensors, fluctuations can either be caused by equipment noise, variability in hydraulics and water demand, or the presence of contaminant. Signal processing and data-driven methods have very limited ability to differentiate between these two types of fluctuations, which can lead to false positive alarms.

To overcome this drawback, this paper describes a new method for real-time contamination detection using multiple conventional water quality sensors for source water. The proposed method aims to achieve contamination detection by using Pearson correlation coefficients to explore the correlative relationship between signals from multiple sensors and their Euclidean distances. A Pearson correlation coefficient is a measure of the strength and direction of the linear relationship between two variables (Mudelsee, 2003). In recent years, it has been used for classification purpose. For example, Monedero et al. (2011) applied Pearson correlation coefficients to the problem of detecting fraud and other non-technical losses in a power utility. Benesty et al. (2008) used the coefficients to reduce noise in speech estimation, a topic which has attracted a considerable amount of attention over the past few decades. However, the application of this correlation coefficient in the field of contamination detection has not been explored. The method proposed in this study is tested using data from a laboratory contaminant injection experiment (CIE). Its detection performance is evaluated and compared with two baseline methods.

2. Methods and materials

2.1. *The proposed event detection method*

The proposed event detection method, called Pearson correlation Euclidean distance-based method (PE), includes three steps: calculation of Pearson correlation coefficients, calculation of correlation indicators and calculation of Euclidean distances.