Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 11th International Scientific and Practical Conference. Volume I, 77-81

On-line Drinking Water Contamination Event Detection Methods

Sandis Dejus, Alīna Neščerecka, Tālis Juhna
Riga Technical University, Department of Water Engineering and Technology. Address: Kipsalas str. 6B-263, Riga, LV-1048, Latvia.

Abstract. A task of water supply systems is to provide safe drinking water to every customer, which is a basic human need. Aging of water supply networks and increased precaution of terrorism risks led to re-evaluation of drinking water supply system reliability and vulnerability to accidental and intentional contamination. Contamination of drinking water can cause health, social, psychological and economic issues. During the last decade, early warning systems (EWS) were often used to ensure the safety of drinking water. EWS are driven by conventional sets of drinking water quality sensors, and the collected data are analyzed in real time. For detection of contamination events, numbers of algorithms have been developed. Most of the algorithms are based on statistical analysis or machine learning. The aim of this study was to compare existing methods and to identify the method, which is suitable for contamination detection in drinking water from non-compound specific sensors and requires relatively low computational resource.

A detailed review of 11 different algorithms was presented in the current study with the primary focus on detection probability. Cluster analysis in combination with Mahalanobis distances of feature vectors and Canonical correlation analysis (CCA) approach were selected as the most promising methods for application in a new generation of EWS to detect and classify possible contamination events and agents. While canonical correlation analysis method was the most accurate for detection of contamination events, an advantage of Mahalanobis distances was that it not only detects the contamination events but also could identify the type of contaminant. In this study, we conclude that CCA and Mahalanobis distance methods might be applied for detection of contamination events with relatively high and reliable precision.

Keywords: drinking water quality, early warning systems, online monitoring.

I. INTRODUCTION

Drinking water supply systems (DWSS) are vulnerable to deliberate and accidental contamination events. Contamination events might cause health, social, psychological and economic issues to consumers [1]–[4]. There are numerous drinking water deterioration cases reported in different scientific and technological papers during the past decades. For example, more than 1900 drinking water contamination accidents annually have been recorded in China between 1992 and 2006 [5]. A chemical spill in Elk River West Virginia, the United States in 2014 has influenced more than 300 000 residents by the interruption of drinking water service because of deterioration of drinking water quality [6]. To increase the safety and reliability of drinking water supply system the early warning systems (EWS) have been developed [5]–[7]. The tasks of EWS systems include detection of contamination events during drinking water monitoring and following notification to the responsible institutions. EWS consists of a drinking water quality sensors set, data collection and analysis system and alarm triggering algorithm. Measurements, data collection, analysis and possible alarm triggering is aimed to be done on-line [5]. EWS provide not only the real-time detection of possible contamination event but also could classify the type of contaminant occurred in drinking water supply system.

There are two types of sensors used for drinking water quality monitoring. The first type is non-compound specific sensors or conventional type of sensors (pH, temperature, electrical conductivity, etc.) which are used for routine testing in most of the developed countries. These sensors are relatively straightforward and cheap, easy to maintain and install. The other type of sensors is compound specific sensors which are capable of measuring specific drinking water quality parameters with very high precision and amplitude [8]–[10].

Since EWS should be inexpensive, reliable, easy to maintain and integrate into the network [11], in most of the cases exactly non-compound specific sensors are used in such systems. Usually, there are sets of five to eight sensors installed at each monitoring point. Besides the appropriate combination of the sensors, a key factor for properly working EWS is the detection algorithm [5]. Therefore, mathematical algorithms have been developed through the decades to recognize the contamination events between normal periodic fluctuations of drinking water quality. There are

ISSN 1691-5402

http://dx.doi.org/10.17770/etr2017vol1.2627

various emerging algorithms, which differs in precision, reliability and requirement of computing resources. It is important to choose the most advantageous algorithm regarding these parameters for application in a real scale DWSS. However, no comprehensive studies on the comparison of various proposed algorithms and methods were made to evaluate the applicability for EWS.

The aim of this study was to compare and evaluate available and open code algorithms for detection and classification of contamination events with experimental or artificial data acquired by conventional drinking water quality sensors. To do that a cognitive literature study has been done.

II. CONTAMINATION DETECTION ALGORITHMS

The main part of the EWS is the contamination detection algorithm. Numerous studies have been done to develop and select the most precise event detection algorithm. The accuracy of the event detection method is defined by its ability to place the current measurement of water quality parameters into one of two classes: background – clear and safe water, event – contaminated water [12]. The detection methods during last decades have evolved and expanded from single factor correlation analysis to generic algorithms and artificial neural network analysis. A summary of these methods that were developed and tested in last decades and described in scientific papers is shown in Table 1. However, it should be noted that the methods, based only on the theoretical probability of detection of a potential contamination event, e.g. without any specific sensor installed in DWSS and actual measurements, were not studied in the present review.

Usually, the evaluation of event detection methods was done by assessing the trade-offs between false positive (FP) and false negative (FN) decisions as a function of the detection methods. The adopted received operating characteristic (ROC) curve has been chosen as an evaluation tool [13]. This curve has been used in all evaluated studies. The ROC curve defines the probability of detection (PD) that can be obtained as a function of the corresponding false alarm rate (FAR). FAR is equal to the number of FPs divided by the total number of values that are actually below the detection threshold as in equation (1). The PD is defined as the number of true positives divided by all events that exceed the detection threshold equation (2) [12]. TP represents a true positive detection, TN – true negative. A greater PD means the method is more capable of detecting a real event, while a smaller FPR implies the method is less likely to classify a routine operation as an event. FAR and PD values varies between 0 and 1 [5].

$$PD = \frac{TP}{TP + FN} \quad (1)$$

$$FAR = \frac{FP}{FP + TN} \quad (2)$$

This approach was applied in all observed studies in this paper. Comparison of PD and FAR results, reported in the reviewed papers, is shown in Table 1. The higher is PD, the higher is the probability that the event would be detected in a real contamination situation. Thus responsible organizations could take preventive actions. In contrast, high FAR represent a likelihood of the cases, when the alarm would be triggered wrong. A false alarm could lead to a financial loss and decrease of society confidence in the organization. Thus, an ideal algorithm should have PD value close to 1 and low FAR values. Overall these parameters should be considered by water utilities for integration of EWS into the online monitoring system of drinking water quality.

The type and number of drinking water quality parameters and contamination agents could also affect the results, as shown in the reviewed studies (Table 1). A good example of the impact of the sensor set of detection results is reported for PE method where PD values of 0,76 and 1,00 for nickel and atrazine respectively as contamination agents was reported [5]. Although experimental data sets and real scale data were reported in several studies, some works were based on the artificial data sets with simulated contamination events. It is related to the fact that it is not always feasible to simulate a contamination event experimentally since it requires special facilities and could be unsafe. Therefore, the overall knowledge about the contamination event influence on drinking water quality parameters is limited and actual disturbances to the measurements are unknown [14]. For example, the results of experimental and real scale studies might be affected by sensor calibration, signal noises, signal processing and intensity of data collection [2].

To gain more reliable comparison, PD and FAR values for each method were acquired from multiple studies and data sets and summarized in Table 1. It demonstrates a high variety on contamination detection probability, obtained by different algorithms (Table 1). First generation methods were developed earlier, and have mostly simple algorithms. The highest detection probability (PD) were 0,89, 0,92, 0,587 and FAR of 0,88, 0,82, 0,093 for MED, LPF and ANN (Multivariate time series) methods respectively. Thus high PD values correspond with high FAR, and vice versa, which indicates that either normal signal fluctuations would be assumed for contamination events, or missed. It is explicable with relatively simple algorithms used in MED and LPF methods. Although ANN (Multivariate time series) method is based on artificial neuron network and showed very low FAR, it was not effective for detection contamination events. At this stage, it is unsuitable for drinking water monitoring. However,

there is a great potential for improvement of the approach.

Second generation methods that contain more complicated algorithms shows higher PD, in some cases (CCA, MVE, Canary, SVM, DSM, PE) reaching even 1,00 that means 100% of contamination events will be detected. Though the FAR values of 0,34 and 0,1 for Canary and SVM methods raises doubts on reliability and detection capabilities, PE method has been applied for different types of contamination, and the overall results were ambiguous PD and FAR varied a lot, and were between 0,69 – 1,00 and 0 – 0,87 respectively. PE method can be suggested as an applicable tool in certain conditions. However, the overall usability should be considered. Moreover, the detection of the real scale event was not accurate, resulting with PD = 0,83 and FAR = 0,33. CCA, MVE and DSM methods demonstrated very low FAR values of 0, 0,008 and 0,032 that shows a high potential to be implemented in EWS. It should be emphasized that for methods PE, CCA, DSM experimental data sets have been used leading that those methods have shown a high potential for real scale events. DSM is the only method with relatively high results that has been analyzed for real contaminants.

The results for MD shows not only accurate detection of contamination events but also the correct classification of certain contamination agents. For example, PD of 0,73 - 0,79 means that in the case of DWSS contamination with four different contaminants, three would be identified correctly.

SVM, PE, CCA, MD, and MVE methods approaches allow not only detection of the contamination event, but also the classification of contamination types, detected in a certain event. Still, the studies of classification are only in preliminary phase and numerous experiments with different contaminants, concentrations, flows should be accomplished to develop a working algorithm for this issue.

The methods, which were applied for experimental studies, shows modest results in comparison to methods with artificial data. This could be explained by the diversity of data generated in artificial data sets and additional data distortion in experiments sensors and its properties [6], [15].

Based on PD and FAR values, Canonical correlation analysis (CCA) method provides the most accurate contamination detection. Thus it has a potential to be applied for real DWSS monitoring. Although FAR data are not available for Mahalanobis distance (MD) method, relatively high PD values and its ability to categorize the contamination agents also make this method very promising for EWS.

Table I
Evaluation of Contamination Detection Algorithms

Method	PD	FAR	Data source	Contamination agent	Parameters	Ref.
Multivariate Euclidean distance (MED)	0,52 - 0,89	0,22 - 0,88	exp	Cadmium nitrate	T, pH, NTU, EC, ORP, UV-254, nitrate, phosphate	[16]
Linear prediction filters (LPF)	0,38 - 0,92	0,24 - 0,82	exp	Cadmium nitrate		[16]
Pearson correlation Euclidean distance (PE)	0,97	0,025	exp	Cadmium nitrate		[16]
	0,83	0,33	r	Phenol		[12]
	0,76 - 1,00	0 – 0,1	exp	Herbicides, pesticides, lead nitrate, nickel nitrate, trivalent chromium		[5]
	0,69 - 0,74	0,78 – 0,87	art			[6]
Canonical correlation analysis (CCA)	0,90 - 1	0	exp	Acrylamide		[17]
Minimum ellipsoid classification (MVE)	0,66 - 1	0,05 - 0,08	art	-	T, pH, NTU, EC, TOC, chlorine	[14]
Artificial Neural Networks (ANN) Multivariate time series	0,085 - 0,587	0,001 - 0,093	art	-		[18]
Artificial Neural Networks (ANN) Dynamic thresholds scheme	0,38-0,99	0,04 - 0,15	art	-		[15]
Canary default algorithm	0,63 - 1	0,17 - 0,34	art	-		[2], [14], [19]
Support vector machine (SVM)	0,75 - 1	0,02 - 0,1	art	-		[2]
Mahalanobis distances (MD)	0,73 - 0,79	-	exp	Herbicides, heavy metals, inorganic salts	T, pH, NTU, EC, ORP, UV-254, nitrate, phosphate	[7]
Extended Dempster-Shafer method (DSM)	0,27 - 1	0,006 - 0,032	exp	Potassium ferricyanide, ferric ammonium sulfate	EC, pH, free chlorine, total chlorine, nitrate, sulphate, TOC, COD	[20]

Legend: PD – probability of detection, FAR – false alarm rate, T – temperature, NTU – turbidity, EC – electrical conductivity, ORP – oxidation-reduction potential, UV-245 – ultraviolet light sensor, TOC – Total Organic Carbon, COD – Chemical Oxygen Demand, exp – data acquired in experiments, r – data acquired in real contamination event, art – artificial data sets used

The present review demonstrates the overall comparison between different approaches and algorithms for contamination detection. However, it should be noted that the observed studies were performed within various conditions, used different data sets, types of contaminants and detection sensors, which should be taken into account for selection of contamination detection approach. However, each of studies analyzed in this paper has been aimed to find the best mathematical approach and compare it to previously used algorithms that mean a reliable data and methods comparison in data analysis done by previous authors [7], [12]

III. DISCUSSION

11 algorithms for detection of contamination event of drinking water were compared in this paper. Although the probability of contamination detection varied between different studies, generally all algorithms could reach 0,5 probability coefficient under certain conditions. While PD lower limits below this value were mostly obtained with first generation approaches, PD could even reach 1 with the second generation algorithms. The recent methods showed more precise results than older algorithms, which shows a positive tendency in methods development. Although the methods were capable of detecting the contamination events, some shortages and drawbacks were found.

The methods, based on artificial neural networks, require simultaneous data collection from all sensors that could lead to technical issues in real DWSS conditions [18].

Most of the reviewed studies in the present work concluded that the methods proposed by authors are capable of detecting certain contamination events[2], [5]–[7], [12], [14]–[18], [20]. However, further research is needed to test these methods for conditions, which could influence the accuracy of the methods. For example, it would be important to know how these methods would respond in the real or laboratory scale conditions, with the presence of different types of contaminants, different or changing spreading velocities and contaminant concentrations. Moreover, detection limits for each method should be found in experimental sessions and setups. Detection limits are essential for detection methods because the even low concentration of contamination agents could significantly affect consumers in long term perspective if continuously or periodically appear in the DWSS [17]. Also, the thresholds used in each method should be verified experimentally.

From the reviewed methods, only MD and CCA could classify contamination agents detected in DWSS. Identification of contamination agent is of particular importance for the development of actions and scenarios that should be applied by water utilities during and after the contamination event to ensure the quality of drinking water at consumption point.

None of the methods proposed in the previous studies have addressed the potential contamination of DWSS with biological agents (*Escherichia coli, Pseudomonas aeruginosin*, *Clostridium perfringens*). It is surprising since biological contamination could affect the health of drinking water consumers even more than chemical contamination. Moreover, biological water quality monitoring is obligatory for drinking water and is regulated by the European Union legislation [21]. Furthermore, no investigation on possible correlations between microbiological and physical-chemical parameters of drinking water quality have been done. It clearly shows a need of methods' evaluation for microbiological parameters detection.

Although the results of the second generation detection methods are rather high and precise, the computing resource of running them must be taken into account. Detection and classification of contamination events by using proposed MD and PE methods can reach up to 4 and 9 minutes delay respectively [5], [7] for a single monitoring point with a set of 9 surrogate sensors. For a real scale network, the time and resource needed for compilation of algorithms might increase rapidly. The relation of detection precision, costs of sensors sets and computational resources should be taken into account while developing each of proposed methods for integration in EWS for real DWSS. The methods used for detection of contamination events should be robust, simple and relatively computing resource friendly to ensure the functionality and possibility to implement them in an EWS for real scale system and hydraulic conditions by using fewer sensors. Linking detection tools with a hydraulic modeling would provide a unique next generation monitoring tool for drinking water quality, which could predict possible contaminant distribution in DWSS and identify the contamination point. Within the given situation only Canary default algorithm was tested and implemented in a real scale DWSS by linking it with a real scale hydraulic DWSS models [19].

IV. CONCLUSIONS

During the last decade, many studies on contamination event detection methods for the drinking water supply system were carried out. Numerous methods, based on different approaches including statistical analysis, clustering, and artificial neural networks have been proposed. High detection probability and low false alarm rate are the main parameters to select the algorithms. The ability of classification of different contamination types should also be taken into account. Therefore, CCA and MD methods have been chosen as the most promising methods.

Although the methods have shown good results of detection probability in the reported studies, more tests and experiments in the pilot and real scale

should be done to ensure the stability and functionality in real scale conditions. Additionally, detection of biological contamination should be evaluated.

As the most promising methods CCA and MD were selected.

There is a lack of measurement data and information about real contamination events reported all over the world. More accurate and detailed reports on each event should be done to improve the capabilities of proposed methods.

V.ACKNOWLEDGMENTS

This work was supported by the Latvian National research program SOPHIS under grant agreement Nr.10-4/VPP-4/11

REFERENCES

[1] A. Rasekh and K. Brumbelow, "Drinking water distribution systems contamination management to reduce public health impacts and system service interruptions," Environ. Model. Softw., vol. 51, pp. 12–25, 2014.

[2] N. Oliker and A. Ostfeld, "A coupled classification - Evolutionary optimization model for contamination event detection in water distribution systems," Water Res., vol. 51, pp. 234–245, 2014.

[3] the United States Environmental Protection Agency, "Response Protocol Toolbox : Planning for and Responding to Drinking Water Contamination Threats and Incidents," Epa, no. December, 2003.

[4] R. M. Grayman, W.M., Deninger, R.A., Clark, "Vulnerability of water supply to terrorist activities," CE News, vol. 14, pp. 34–38, 2002.

[5] S. Liu, H. Che, K. Smith, and L. Chen, "Contamination event detection using multiple types of conventional water quality sensors in source water," Environ. Sci. Process. Impacts, vol. 16, no. 8, pp. 2028–2038, 2014.

[6] S. Liu, R. Li, K. Smith, and H. Che, "Why conventional detection methods fail in identifying the existence of contamination events," Water Res., vol. 93, no. February, pp. 222–229, 2016.

[7] S. Liu, H. Che, K. Smith, and T. Chang, "A real time method of contaminant classification using conventional water quality sensors," J. Environ. Manage., vol. 154, pp. 13–21, 2015.

[8] C. P. Marshall, S. Leuko, C. M. Coyle, M. R. Walter, B. P. Burns, and B. a Neilan, "Carotenoid analysis of halophilic archaea by resonance Raman spectroscopy," Astrobiology, vol. 7, no. 4, pp. 631–643, 2007.

[9] P. R. Hawkins et al., "A review of analytical methods for assessing the public health risk from microcystin in the aquatic environment," J. Water Supply Res. Technol. - Aqua, vol. 54, no. 8, p. 509 LP-518, Dec. 2005.

[10] J. Jeon, J. H. Kim, B. C. Lee, and S. D. Kim, "Development of a new biomonitoring method to detect the abnormal activity of Daphnia magna using automated Grid Counter device," Sci. Total Environ., vol. 389, no. 2–3, pp. 545–556, 2008.

[11] M. Brussen, "On-line Water Quality Monitoring. Review of Sydney's Current Status and Future Needs. Sydney Water Report," Sydney, 2007.

[12] S. Liu, H. Che, K. Smith, M. Lei, and R. Li, "Performance evaluation for three pollution detection methods using data from a real contamination accident," J. Environ. Manage., vol. 161, pp. 385–391, 2015.

[13] S. A. McKenna, M. Wilson, and K. A. Klise, "Detecting changes in water quality data," J. / Am. Water Work. Assoc., vol. 100, no. 1, pp. 74–85, 2008.

[14] N. Oliker and A. Ostfeld, "Minimum volume ellipsoid classification model for contamination event detection in water distribution systems," in Procedia Engineering, 2014, vol. 70, pp. 1280–1288.

[15] J. Arad, M. Housh, L. Perelman, and A. Ostfeld, "A dynamic thresholds scheme for contaminant event detection in water distribution systems," Water Res., vol. 47, no. 5, pp. 1899–1908, 2013.

[16] S. Liu, K. Smith, and H. Che, "A multivariate based event detection method and performance comparison with two baseline methods," Water Res., vol. 80, no. May, pp. 109–118, 2015.

[17] R. Li, S. Liu, K. Smith, and H. Che, "A canonical correlation analysis based method for contamination event detection in water sources," 12th Int. Conf. Hydroinformatics, no. July, 2016.

[18] L. Perelman, J. Arad, M. Housh, and A. Ostfeld, "Event detection in water distribution systems from multivariate water quality time series.," Environ. Sci. Technol., vol. 46, no. 15, pp. 8212–9, 2012.

[19] D. B. Hart, S. A. Mckenna, U. S. Epa, and N. Homeland, User ' s Manual for CANARY, no. September. Cincinnati: U.S Environmental Protection Agency, 2012.

[20] D. Hou, H. He, P. Huang, G. Zhang, and H. Loaiciga, "Detection of water-quality contamination events based on multi-sensor fusion using an extented Dempster–Shafer method," Meas. Sci. Technol., vol. 24, p. 55801, 2013.

[21] The Council of the European Union, "Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption," Off. J. Eur. Communities, vol. L330, pp. 32–54, 1998.