Anomaly Detection in Water Management Systems

Massimiliano Raciti, Jordi Cucurull, Simin Nadjm-Tehrani

Abstract Quality of drinking water has always been a matter of concern. Traditionally, water supplied by utilities is analysed by independent laboratories to guarantee its quality and suitability for the human consumption. Being part of a critical infrastructure, recently water quality has received attention from the security point of view. Real-time monitoring of water quality requires analysis of sensor data gathered at distributed locations and generation of alarms when changes in quality indicators indicate anomalies. The event detection system should produce accurate alarms, with low latency and few false positives.

This chapter addresses the application of data mining techniques developed for information infrastructure security in a new setting. The hypothesis is that a clustering algorithm ADWICE that has earlier been successfully applied to n-dimensional data spaces in IP networks, can also be deployed for real-time anomaly detection in water management systems. The chapter describes the evaluation of the anomaly detection software when integrated in a SCADA system. The system manages water sensors and provides data for analysis within the Water Security initiative of the U.S. Environmental Protection Agency (EPA). Performance of the algorithm is illustrated and improvements to the collected data to deal with missing and inaccurate data are proposed.

1 Introduction

Water management systems deserve a special attention in critical infrastructure protection due to a number of factors. First, the quality of distributed water affects every single citizen with obvious health hazards. Second, in contrast to some other infrastructures where the physical access to the critical assets may be possible to

Department of Computer and Information Science
Linköping University, Sweden
e-mail: {name.surname}@liu.se

restrict, in water management systems there is a large number of remote access points difficult to control and protect from accidental or intentional contamination events. Third, in the event of contamination, there are few defence mechanisms available. Water treatment facilities are typically the sole barrier to potential large scale contaminations and distributed containment of the event leads to widespread water shortages. Techniques to model the spatial and temporal distribution of the contaminants [30] can be used, but because the scale of the distribution network they are complex to apply.

A recent health hazard was identified in France where 30 cubic metres of fluid containing 12g per litre of low-grade uranium were spilt at the Tricastin facility near Marseilles [1]. In the USA a major initiative has been established by the US Environmental Protection Agency (EPA) in response to Homeland Security Presidential Directive 9, under which the Agency must "develop robust, comprehensive, and fully coordinated surveillance and monitoring systems, including international information, for water quality that provides early detection and awareness of disease, pest, or poisonous agents." [2].

Supervisory control and data acquisition (SCADA) systems provide a natural opportunity to increase vigilance against water contaminations. Specialised event detection mechanisms for water management can be included such that 1) a contamination event is detected as early as possible and with high accuracy with few false positives, and 2) predictive capabilities ease preparedness actions in advance of full scale contamination in a utility.

While research on protection of SCADA systems have seen an increased attention in the past decade, most of the reported works focus on how to deploy detection and mitigation mechanisms in the event of an adversary attack on the power networks or the information and communication (ICT) network on which the SCADA system depends [15]. Published literature in which water management systems is the application area, covers the protection of the ICT related security issues for SCADA in water management systems [29], but little work has been published on detection and anticipation of water contaminations using ICT techniques.

The analogy with ICT threats is however not vacuous. Recent advances in intrusion detection target complex ICT environments where large scale systems are integrated with Internet with no well-determined perimeter. Increase in both accidental and malicious activity creates a changing landscape for emergent information infrastructures; hence the difficulty of modelling the system and the attack patterns statically. In these networks, intrusion detection is either based on modelling and recognising the attacks (misuse detection) or modelling the normal behaviour of the system and detecting potential intrusions as a deviation from normality (anomaly detection) [32, 24]. While misuse detection provides immediate diagnosis when successful, it is unable to detect cases for which no previous data exists (earlier similar cases in history, a known signature, etc.). Anomaly detection, on the other hand, is able to uncover new attacks not seen earlier, but it is dependent on a good model of normality. Misuse detection requires exact characterisation of known constraints on the historical data set and gives accurate matches for those cases that are modelled. Anomaly detection is most often based on learning techniques which creates an approximate model of normality. A typical problem is the high rate of false positives if attacks and normality data are similar in subtle ways.

The available data from water management system sensors are based on biological, chemical and physical features of the environment and the water sources. Sinche these change over seasons, the normality model is rather complex. Also, it is hard to create a static set of rules or constraints that clearly capture all significant attacks since these can affect various features in non-trivial ways and we get a combinatorial problem. Therefore, we propose the application of learning based anomaly detection techniques as a basis for contamination event detection in water management systems.

Since anomaly detection needs a model of normality one could imagine using classification based techniques to extract models of benign and contaminated data samples automatically. However, the clustered data sets would then have to be individually examined by experts to verify the suitability of the normality clusters (representation of benign data). Another approach would be to get the anomaly detector to only learn normality from data that is known to be benign. In water management systems since it may be possible to analyse water quality in test beds and prepare a calibration a normality model based on benign data can be built. In this paper we explore this direction. An interesting question is then whether the detection technique provides fast enough recognition of the contamination events and whether it can be accurate and reliable enough.

The contributions of this chapter are as follows.

- Application of a method for Anomaly Detection With fast Incremental ClustEring (ADWICE) [8] in a water management system based on measured sensor values from the EPA database.
- Analysis of the performance of the approach for two stations using performance metrics such as detection rate, false positives, detection latency, and sensitivity to the contamination level of the attacks.
- Discussion of reliability of the analysis when data sets are not perfect (as seen in real life scenarios), where data values may be missing or less accurate as indicated by sensor alerts.

The chapter is composed of six sections. Section 2 describes the background. Section 3 describes ADWICE, an existing anomaly detection tool. Section 4 describes the application of ADWICE on a water management system, presents the results obtained, and proposes a technique to deal with unreliable data. Section 5 presents related work in this field. The paper is concluded in Section 6, with description of future works.

2 Background

The monitoring of water quality in a distribution system is a highly complex and sensitive process that is affected by many different factors. The different water qualities coming from multiple sources and treatment plants, the multiplicity of paths that water follows in the system and the changing demand over the week from the final users make it difficult to predict the water quality at a given point of the system life time. Water quality is determined by the analysis of its chemical composition: to be safe to drink some water parameters can vary within a certain range of values, and typically the maximum and the minimum values are established by law. Water from different sources have different compositions. Before entering the distribution system, water is treated first in the treatment plants, in order to ensure its safety. Once processed by the treatment plant, water enters the distribution system so it can be directly pumped to the final user, or stored in tanks or reservoirs for further use when the demand on the system is greater than the system capacity. System operations have a consistent impact on water quality. For instance, pumping water coming from two or more different sources can radically modify the quality parameters of the water contained in a reservoir. In general, the water quality (WQ) is measured by the analysis of some parameters, for example:

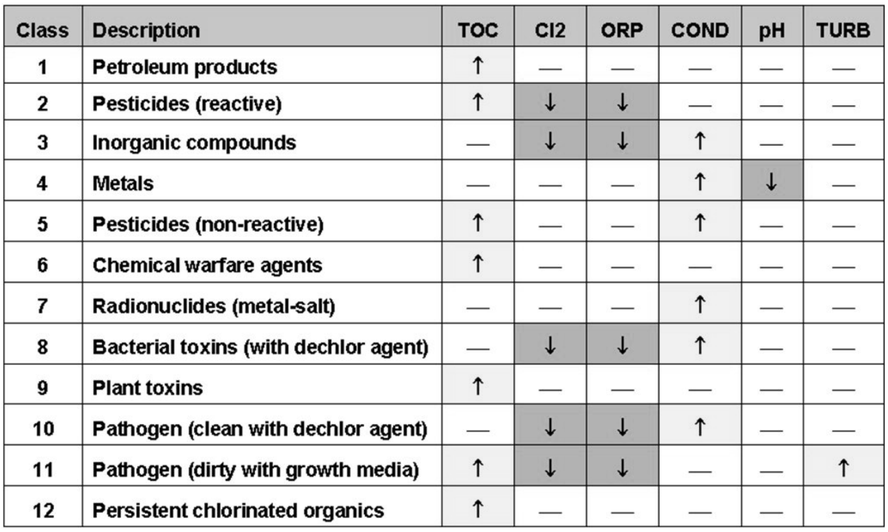
- *Chlorine (CL2) levels*: free chlorine is added for disinfection. Free chlorine levels decrease with time, so for instance levels of CL2 in water that is stagnant in tanks is different from levels in water coming from the treatment plants.
- *Conductivity*: estimates the amount of dissolved salts in the water. It is usually constant in water from the same source, but mixing waters can cause a significant change in the final conductivity.
- *Oxygen Reduction Potential (ORP)*: measures the cleanliness of the water.
- *PH*: measures the concentration of hydrogen ions.
- *Temperature*: is usually constant if measured in short periods of time, but it changes with the seasons. It differs in waters from different sources.
- *Total Organic Carbon* (TOC): measures the concentration of organic matter in the water. It may decrease over the time due to the decomposition of organic matters in the water.
- *Turbidity*: measures how clear the water is.

In normal conditions, it is possible to extract some usage patterns from the system operations relating the changes of WQ parameters with changes of some system configurations: for example the cause of a cyclic increment of conductivity and temperature of the water contained in a reservoir can be related to the fact that water of a well known characteristic coming from a treatment plant is cyclically being pumped into the reservoir. This information must be taken into account to avoid false alarms raised by the Event Detection System (EDS).

The situation changes dramatically when some contaminants are intentionally or accidentally injected in some points of the distribution system. Contaminants cause changes in one or more water parameters at the same time, so event detection systems must be able to detect and classify events caused by normal system operations as well as events caused by contaminants. This makes the monitoring of water quality more complex, and effective tools must be applied for this new situation.

The United States Environmental Protection Agency has launched an Event Detection System challenge to identify the best tools applicable to event detection in the water quality domain. In particular, EPA is interested in the development of Contaminant Warning Systems (CWS) that in real-time proactively detect the presence of contaminants in the distribution system. The goal to take the appropriate countermeasures upon unfolding events to limit or cut the supply of contaminated water to users.

The challenge is conducted by providing water quality data from sensors of six monitoring stations from four US water utilities. Data comes directly from the water utilities without any alteration from the evaluators, in order to keep the data in the same condition as if it would come from real-time sensing of the parameters. Data contains WQ parameter values as well as other additional information like operational indicators (levels of water in tanks, active pumps, valves, etc.) and equipment alarms (which indicate whether sensors are working or not). Each station differs from the others in the number and type of those parameters. A baseline data is then provided for each of the six stations. It consists of 3 to 5 months of observations coming from the real water utilities. Each station data has a different time interval between two observations, ranging in the order of few minutes. The contaminated testing dataset is obtained from the baseline data by simulating the superimposition of the contaminant effects on the WQ parameters. Figure 1 [4] is an example of effects of different types of contaminants on the WQ values.



Class	Description	TOC	Cl2	ORP	COND	pH	TURB
1	Petroleum products	↑	—	—	—	—	—
2	Pesticides (reactive)	↑	↓	↓	—	—	—
3	Inorganic compounds	—	↓	↓	↑	—	—
4	Metals	—	—	—	↑	↓	—
5	Pesticides (non-reactive)	↑	—	—	↑	—	—
6	Chemical warfare agents	↑	—	—	—	—	—
7	Radionuclides (metal-salt)	—	—	—	↑	—	—
8	Bacterial toxins (with dechlor agent)	—	↓	↓	↑	—	—
9	Plant toxins	↑	—	—	—	—	—
10	Pathogen (clean with dechlor agent)	—	↓	↓	↑	—	—
11	Pathogen (dirty with growth media)	↑	↓	↓	—	—	↑
12	Persistent chlorinated organics	↑	—	—	—	—	—

Fig. 1 Effect of contaminants on the WQ parameters

EPA has provided a set of 14 simulated contaminants, denoting them contaminant A to contaminant N. Contaminants are not injected along the whole testing sequence, but the attack can be placed in a certain interval inside the testing data, with a duration limited to a few timesteps. Contaminant concentrations are added following a certain profile, which define the rise, the fall, the length of the peak concentration and the total duration of the attack. Figure 2 shows some examples of profiles.

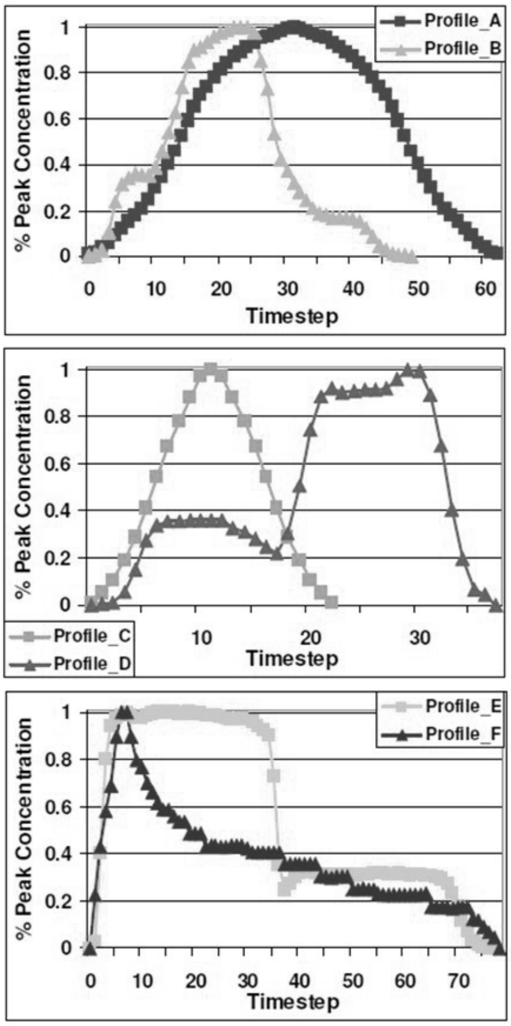


Fig. 2 Example of Event Profiles

To facilitate the deployment and the evaluation of the EDS tools, a software called EDDIES has been developed and distributed by EPA to the participants. EDDIES has four main functionalities:

- Real-time execution of EDS tools in interaction with SCADA systems (collecting data from sensors, analysing them by the EDS and sending the response back to the SCADA tool to be viewed by the utility staff).
- Offline evaluation of EDS tool by using stored data.
- Management of the datasets and simulations.
- Creation of new testing datasets by injection of contaminants.

Having the baseline data and the possibility to create simulated contaminations, EDS tools can be tuned and tested in order to see if they suite this kind of application. In the next sections we will explain how we adapted an existing anomaly detection tool and we will present the results obtained by applying ADWICE to data from two monitoring stations.

3 Anomaly detection with ADWICE

ADWICE is an anomaly detector that has been developed in an earlier project targeting infrastructure protection [7]. The basic idea is that a normality model is constructed as a set of clusters that summarise all the observed normal behaviour in the learning process. Each cluster comprises a set of points and it is represented through a summary denoted cluster feature (CF). The points are multidimensional numeric vectors where each dimension represents a feature in data. CF is a data structure that has three fields: the number of points in the cluster, the sum of the points in the cluster, and the sum of the squares of the points. The first and second element can be efficiently used to compute the average for the points in the cluster used to represent the centroid of the cluster. The third element, the sum of points can be used to check how large is a circle that would cover all the points in the cluster, and using this radius, how far is a new point from the centre of the cluster. This is used for both building up the normality model (is the new point close enough to any existing clusters or should it form a new cluster?), and during detection (is the new point close enough to any normality clusters or is it an outlier?).

In both cases, and more specifically during detection, the search through the existing clusters needs to be efficient (and fast enough for the application). In order to find the closest cluster we need an index that helps to find the closest cluster to a given point efficiently. The cluster summaries, that constitute the normality observations, are therefore organised in a tree structure. Each level in the tree summarises the CFs at the level below by creating a new CF which is the sum of them.

ADWICE uses an adaptation of the original BIRCH data mining algorithm which has been shown to be fast for incremental updates to the model during learning, and efficient when searching through clusters during detection. The difference is the indexing mechanism used in one of its adaptations (namely ADWICE-Grid), which has been demonstrated to give lower false positive rates due to fewer indexing errors [8].

The implementation of ADWICE consists of a Java library that can be embedded in a new setting by feeding the preprocessing unit (e.g. when input are alphanumeric and have to be encoded into numeric vectors) from a new source of data. The algorithm has three parameters that have to be tuned during the pre-study of data (with some detection test cases) in order to "optimise" the search efficiency: the maximum number of clusters (M), and the threshold for comparing the distance to the centroid (E). The threshold implicitly reflects the maximum size of each cluster. The larger a cluster (with few points in it) the larger the likelihood that points *not* belonging to the cluster are classified as part of the cluster – thus decreasing the detection rate. Too small clusters, on the other hand, lead to overfitting and increase the likelihood that new points are considered as outliers, thus adding to the false positive rate.

In the experiments for this application we have used ADWICE with a setting in which M has been set to 150 in one case and 200 in another one, and E has been varied between 1 and 2.5 as it will be shown in the RoC curves in the results section.

While deploying machine learning based anomaly detectors for detection of attacks in networks is known to face considerable challenges [35], we show in this chapter that it is worth exploring the technique in data collected from sensors in critical infrastructures such as water management systems.

4 Training and detection

4.1 Training

The training phase is the first step of the anomaly detection. It is necessary to build a model of normality of the system to be able to detect deviations from normality. ADWICE uses the approach of pure anomaly detection, meaning that training data is supposed to be unaffected by attacks. Training data should also be long enough to capture as much as possible the normality of the system. In our scenario, the data that EPA has provided us is clean from contaminants, the baseline data contains the measurements of water quality parameters and other operational indicators over a period of some months. Pure anomaly detection is thereby applicable.

For our purpose, we divided the baseline data into two parts: the first is used to train the anomaly detector, while the second one is first processed to add the contaminations and then used as testing data. To see how the anomaly detector reacts separately to the effect of each contaminant, 14 different testing datasets, each one with a different contaminant in the same timesteps and with the same profile, are created.

4.1.1 Feature selection

A feature selection is made to decide which parameters have to be considered for the anomaly detection. In the water domain, one possibility is to consider the water quality parameters as they are. Some parameters are usually common to all the stations (general WQ parameters), but some other station-specific parameters can be helpful to train the anomaly detector on the system normality. The available parameters are:

- *Common WQ Parameters*: Chlorine, PH, Temperature, ORP, TOC, Conductivity, Turbidity
- *Station-Specific Features:* active pumps or pumps flows, alarms, CL2 and PH measured at different time points, valve status, pressure.

Sensor alarms are boolean values which indicate whenever sensors are working properly or not. The normal value is 1, while 0 means that the sensor is not working or, for some reason, the value is not accurate and should not be taken into account. The information regarding the pump status could be useful to correlate the changes of some WQ parameter with the particular kind of water being pumped to the station. There are other parameters that give information about the status of the system at different points, for example the measurement of PH and CL2 of water coming from other pumps.

Additional features could be considered in order to improve the detection or reduce the false positive rates. Those features can be derived from some parameters of the same observation, or they can consider some characteristic of the parameters along different observations. For instance, to emphasise the intensity and the direction of the parameter changes over the time, one possible feature to be added would be the difference of the value for a WQ parameter with the value in previous observations. This models the derivative function of the targeted parameter. Another feature, called sliding average, is obtained by adding for each observation a feature whose value is the average of the last *n* values of a WQ parameter. Feature selection and customisation must be made separately for each individual station, since they have some common parameters but they differ in many other aspects.

ADWICE assumes the data to be in numerical format to create an n-dimensional space state vector. So the timestep series of numerical data from water utilities suit the input requirements of ADWICE. This means that our testing data does not require any particular preprocessing phase before feeding it to the anomaly detector.

4.1.2 Challenges

The earlier application of ADWICE has been in IP networks. In its native domain, the main problem is finding a pure dataset, not affected by attacks, but the quantity and quality of data is always high. Network traffic generates a lot of data, which is good for having a reasonable knowledge of normality as long as resources for labelling the data are available. Feature selection from IP headers, for example, is easy and does not lead to many problems, while the difficult issues would arise if payload data would need to be analysed, where we would face privacy concerns and anonimisation. In a SCADA system, sensors could give inaccurate values and faults can cause missing observations. This makes the environment more complicated, thus feature selection and handling is complex. Dealing with inaccurate or missing data requires more efforts to distinguish whenever an event is caused due to those conditions or due to contamination. Furthermore, the result of the anomaly detection is variable depending on where the attack is placed. It is not easy, for example, to detect a contamination when at the same time some evaluations about some WQ parameters are inaccurate and some others are missing. Training the system with a limited dataset can result in a sub-optimal normality model, and this causes raising of a lot of false alarms when testing with data that resembles uncovered normality conditions of the system. In the next section we show some results that we obtained testing our anomaly detector with data from two different monitoring stations, proposing some possible solutions for the kinds of problems described.

4.2 Detection Results

Over six available stations, we have chosen to test our anomaly detection with the easiest one and the hardest one. As mentioned before, we have generated testing datasets by using the second half of the baseline data and adding one contamination per dataset. The contamination has been introduced in the middle of the dataset according to the profile A, depicted in Figure 2, which is a normal distribution of the concentration during 64 timesteps. Details about the single stations will be presented separately. Each testing dataset then contains just one attack along several timesteps. The detection classifies each timestep as normal or anomalous. The detection rate (DR) is calculated from the number of timesteps during the attack that are detected as anomalous, according to the following formula: $DR = TP/(TP+FN)$, where TP refers to the number of true positives and FN refers to the number of false negatives. The false positive rate (FPR) are the normal timesteps that are erroneously classified as anomalous according to the formula $FPR = FP/(FP+TN)$, where FP is the number of false positives and TN refers to the number of true negatives.

4.2.1 Station A

Station A is located at the entry point of a distribution system. It is the best station in terms of reliability of values. It only has the common features and three pump status indicators. Values are not affected by inaccuracies and there are no missing values both in the training and testing datasets. The baseline data consists of one observation every 5 minutes during the period of five months. The first attempt in generating a dataset is done by injecting contaminants according to the normal distribution during 64 timesteps, in which the peak contaminant concentration is 1 mg/L. Table 1 shows the results that we obtained doing a common training phase and then running a test for each of the contaminants. Training and testing have been carried out using a threshold value E set to 2 and the maximum number of clusters M is set to 150. Considering the fact that the amount of contaminant is the lowest possible the results from Table 1 are not discouraging. Some contaminants affect more parame-

Contaminant ID	False Positive Rate	Detection Rate
A	0.057	0.609
B	0.057	0.484
C	0.057	0
D	0.057	0
E	0.057	0
F	0.057	0
G	0.057	0
H	0.057	0.422
I	0.057	0
J	0.057	0.547
K	0.057	0
L	0.057	0.406
M	0.057	0.156
N	0.057	0.516

Table 1 Station A detection results of 1mg/L of concentration

ters at the same time and their effect is more evident; some others only affect few parameters with slight changes. Contaminant F for instance only affects the ORP, which is not measured in this station, so this contaminant is undetectable in this case. The anomaly detector must be tuned in order to fit the clusters over the normality points and let the furthest points to be recognised as attacks. To determine the best threshold values the ROC curves can be calculated by plotting the detection rate as a function of the false positive rate while changing the threshold value. Evaluation of the ROC curves of all the contaminants can give hints to determine the best trade-off that gives good detection rates and false positives, but all of those curves refer

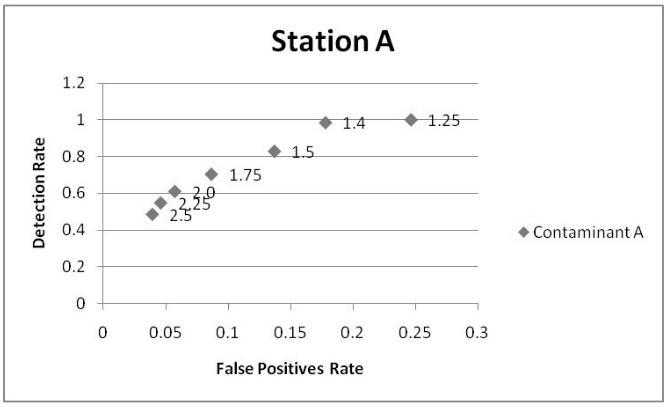


Fig. 3 Station A contaminant A ROC curve

to a contaminant concentration peak of 1 mg/L. As non-experts it was not clear to us whether this could be a significant level of contamination. For this reason we have tested the sensitivity of the anomaly detection by incrementally increasing the contaminant concentration. In our tests, we increased the concentration in steps of 4 mg/L a time, up to 24 mg/L. Figure 4 shows the variation of the detection rates of three significant contaminants with respect to the increase of the concentration. Contaminant A is the easiest to detect, Contaminant L is medium and Contaminant E is difficult to detect since it only sligthly affects the TOC. In this figure the false positive rate is not considered since with the higher concentration of contaminants it is easier to detect the deviation from normality without any increase in the false alarms. These results confirm that even if ADWICE is not designed for this kind of application, by finding the optimal tradeoff between detection and false positive rates for 1mg/L, this anomaly detector would give good results for any other greater concentration. We conclude therefore that ADWICE is a good candidate tool to be applied as EDS for this station.

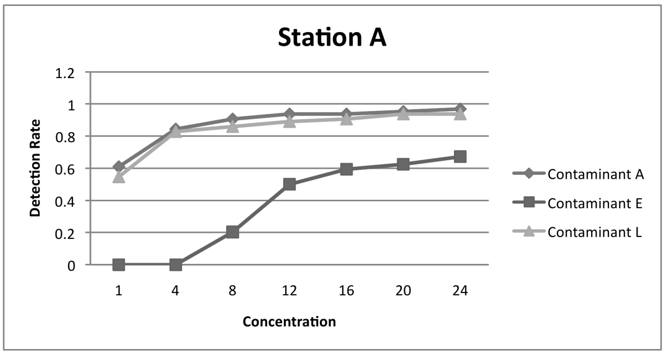


Fig. 4 Concentration Sensitivity of the Station A

4.2.2 Station D

The situation becomes more complicated when a source of uncertainty is added to the system. Station D is located in a reservoir that holds 81 million gallons of water. The water quality in this station is affected by many operational parameters of co-located pump stations and reservoirs. Station D contains more parameters than station A and some sensors are affected by inaccuracy. In detail, Station D has the following parameters:

- *Common features*: PH, Conductivity, Temperature, Chlorine levels, TOC, Turbidity.
- *Alerts*: CL2, TOC and Conductivity; 1 means normal functioning, while 0 means inaccuracy.

- *System indicators*: three pump flows, two of them supply the station while the third is the pipe which the station is connected to.
- *Valves*: indicates the position of the key valve; 0 if open, 1 if closed.
- *Supplemental parameters*: Chlorine levels and PH measured in water coming from pump1 and pump2.

By checking the data that EPA has provided, we noted that the only sensor inaccuracy alert that is sometimes raised is the TOC alert, but in general we will assume that the other alerts could be raised as well. There are some missing values in different points scattered within the baseline file. The baseline data consists of one observation every 2 minutes during the period of three months. The same procedure for station A has then been applied to this station. Figure 5 shows the ROC curve obtained with the peak concentration of 1 mg/L and the same profile (profile A, Figure 2). An accurate feature selection has been carried out to get reasonable results, since trying with all the station parameters the false positive rate is very high. This makes it not worthwhile to explore threshold variations with such bad results. To mitigate the effects of the missing data and the accuracy, the derivatives and sliding averages of the common parameters have been added as new features. The outcome was that the derivatives emphasise the intensity of the changes, thus improving the detection of the effects of the contaminations, while the sliding window averages mitigated the effect of the abrupt changes in data caused by the inaccuracies or missing data. Some parameters have been ignored, like the pumps flows and the key valve, since they caused lots of false positives if included as features. The same training and testing procedure for station A has then been applied to this station. Figure 6 shows the sensitivity of the detection with the increase of the concentrations. As in the case for station A, ADWICE was run with the parameter E=2. Since the dataset is more complex and there are more possible combinations of data to represent, the maximum number of clusters M was set to 200.

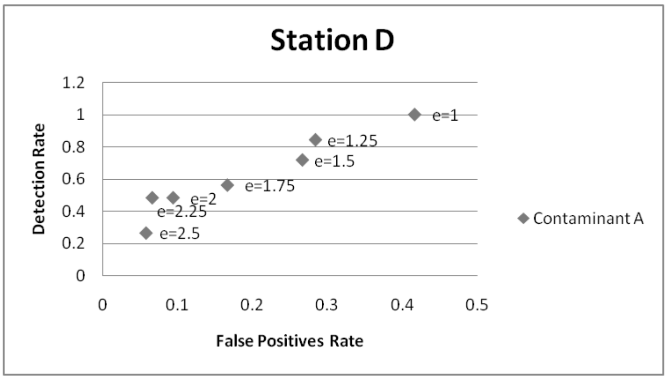


Fig. 5 ROC curve station D contaminant A

Data from from the above two stations have resulted in clustered models consisting of 115 clusters for station A and 186 clusters for station D.

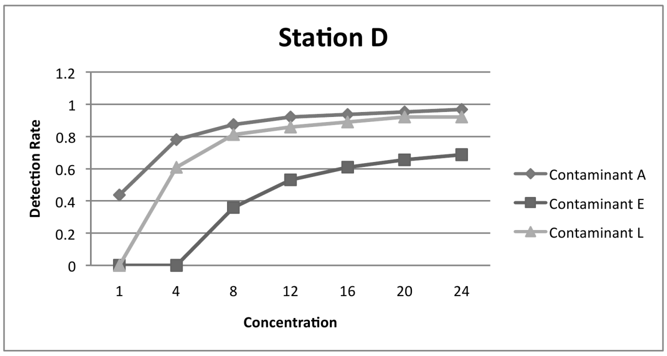


Fig. 6 Concentration sensitivity station D

4.3 Detection Latency

This section focuses on adequacy of the contaminants detection latency. As mentioned in section 2, the final goal of the EPA challenge is to apply the best EDS tools in real water utilities to proactively detect anomalous variations of the WQ parameters. Real-time monitoring allows to take opportune countermeasures upon unfolding contaminations. This makes the response time to be as crucial as the correctness of the detection in general, since even having a good detection rate (on average) a late response may allow contaminated water to leave the system and be delivered to users causing severe risks for their health.

The first issue that comes when measuring the detection latency is from when to start counting the elapsed amount of time before the first alarm is raised. This problem is caused by the fact that different event profiles make it necessary to consider the latency in different ways. In case of the normal distribution depicted as profile A (Figure 2), a possible approach could be counting the latency of the detection event from the initiation of the contamination event, since the concentration rapidly reaches its peak. If the peak concentration was reached very slowly, the evaluation of latency based on the first raised alarm from the beginning would result in an unnecessary pessimism (e. g. see profile D in Figure 2). In this case it would be more appropriate to start counting the reaction time from the time when the peak of the event has taken place. An earlier detection would then give rise to a negative delay and this would signal a predictive warning. For the purpose of our experiments, the normal distribution of profile A suits the computation of the latency based on counting the number of samples from the beginning of the event.

Since in the baseline data time is represented as discrete timesteps, we measure the latency by counting the number of timesteps elapsed before the first alarm is raised. Figure 7 and 8 show the measured latencies for Station A and Station D respectively, with respect to the detection of the three contaminants presented in the previous section. The curves indicate that in the case of the lowest concentration the latencies are high. For instance the detection of contaminant A and L in station

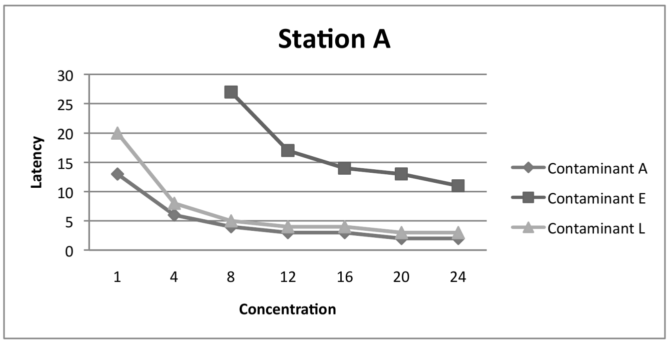


Fig. 7 Detection latency in station A

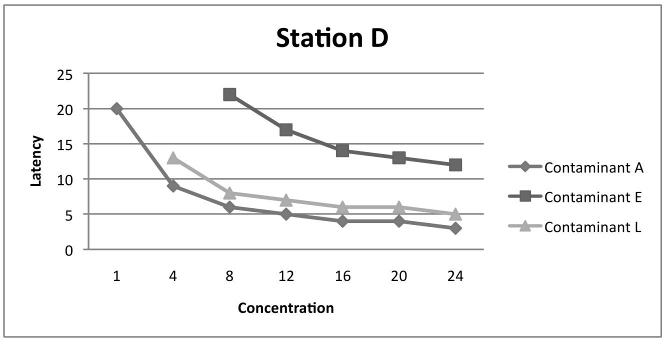


Fig. 8 Detection latency in station D

A has a latency of 13 and 20 timesteps respectively. They are around one fourth and one third of the total duration of the contamination event (64 timesteps). Contaminant E is not detected at all, so its latency is not represented in the graph. The situation changes positively when the concentrations are gradually increased. The latencies of Contaminant A and L drop sharply until a concentration of 8 mg/L is reached, decreasing 60% and 75% respectively. At the same time, there are some detections of Contaminant E, characterised by a high latency. From this point the latencies of Contaminant A and L steadily drop, while the latency of Contaminant E decreases more rapidly. Latencies for the station D follow the same pattern, although the values are slightly higher.

Checking the results against reality, a latency of 13 discrete timesteps for Contaminant A in Station A would correspond to a latency of 65 minutes, which is quite a long time. One should note that time interval between two observations has a high impact on the real latency, since for example 20 timesteps of detection latency for Contaminant A in Station D with a concentration of 1 mg/L corresponds to 40 minutes of latency, 25 minutes less than the latency in Station A. Even in this case the results are definitely improved by the increases in the contaminant concentrations, but domain knowledge is required to evaluate whether the selected increments to a certain concentration are meaningful.

4.4 Missing and inaccurate data problem

In Section 4.2 we have seen that data inaccuracies and missing data were a major problem in station D. Our approach for the tests carried out so far has been to use workarounds but not provide a solution to the original problem.

More specifically, our workaround for missing data was as follows. We have replaced the missing data values with a zero in the preprocessing stage. When learning takes place the use of a derivative as a derived feature helps to detect the missing data and classify the data points in its own cluster. Now, if training period is long enough and includes the missing data (e.g. inactivity of some sensors or other operational faults) as normality, then these clusters will be used to recognise similar cases in the detection phase as long as no other sensor values are significantly affected. During our tests we avoided injecting contaminants during the periods of missing data.

Sensor inaccuracies are indicated with a special alert in the provided data set (a 0 when the data is considered as inaccurate, i.e. the internal monitoring system warns for the quality of the data). According to our experiments it is not good to train the system during periods with data inaccuracies, even when workaround are applied. First, learning inaccurate values as normality may result in excessive false positives when accurate values are dominant later. Second, the detection rate can be affected if the impact of the contaminant is similar to some of the inaccurate values. Thus a more principled way for treating this problem is needed.

Our suggestion for reducing the impact of both problems is the classical approach in dependability, i.e. introducing redundancy. Consider two sensor values (identical or diversified technologies) that are supposed to provide measurements for the same data. Then the likelihood of both sensors being inaccurate or both having missing values would be lower than the likelihood of each sensor “failing” individually. Thus, for important contaminants that essentially need a given sensor value’s reliability we could learn the normal value based on both data sets. When a missing data is observed (0 in the alert cell) the preprocessing would replace both sensor values with the “healthy” one. When one sensor value is inaccurate the presence of the other sensor has an impact on the normality cluster, and vice versa. So, on the whole we expect to have a better detection rate and lower false positive rate with sensor replicas (of course at the cost of more hardware).

In the experiments so far we have not yet been able to create the duplicate data sets since the generation of the base line requires domain knowledge of the water management experts. However, we are working towards incorporating a new base line with the replicated sensor and showing its impact on accuracy.

5 Related work

In this section we first describe work that is closely related to ours (section 5.1), and then we continue with an overview of other works which are related to the big picture of water quality and monitoring (sections 5.2 to 5.6).

5.1 Water quality anomalies

The security issues in water distribution systems are typically categorised in two ways: hydraulic faults and quality faults [12]. Hydraulic faults (broken pipes, pump faults, etc.) are intrinsic to mechanical systems, and similar to other infrastructures, fault tolerance must be considered at design time to make the system reliable. Hydraulic faults can cause economic loss and, in certain circumstances, water quality deterioration. Online monitoring techniques are developed to detect hydraulic faults, and alarms are raised when sensors detect anomalous conditions (like a sudden change of the pressure in a pipe). Hydraulic fault detection is often performed by using specific direct sensors and it is not the area of our interest. The second group of security threats, water quality faults, has been subject to increased attention in the past decade. Intentional or accidental injection of contaminant elements can cause severe risks to the population, and Contamination Warning Systems (CWS) are needed in order to prevent, detect, and proactively react in situations in which a contaminant injection occurs in parts of the distribution system [5]. An EDS is the part of the CWS that monitors in real-time the water quality parameters in order to detect anomalous quality changes. Detecting an event consists of gathering and analysing data from multiple sensors and detecting a change in the overall quality. Although specific sensors for certain contaminants are currently available, EDS are more general solution not limited to a set of contaminants.

Byers and Carlsson are among the pioneers in this area. They tested a simple online early warning system by performing real-world experiments [9]. Using a multi-instrument panel that measures five water quality parameters at the same time, they collected 16.000 data points by sampling one measurement of tap water every minute. The values of these data, normalised to have zero as mean and 1 as standard deviation, were used as a baseline data. They then emulated a contamination in laboratories by adding four different contaminants (in specific concentrations) to the same water in beakers or using bench scale distribution systems. The detection was based on a simple rule: an anomaly is raised if the difference between the measured values and the mean from the baseline data exceeds three times the standard deviation. They evaluated the approach comparing normality based on large data samples and small data samples. Among others, they evaluated the sensitivity of the detection, and successfully demonstrated detection of contaminants at concentrations that are not lethal for human health. To our knowledge this approach has not been applied in a large scale to multiple contaminants at multiple concentrations.

Klise and McKenna [22] designed an online detection mechanism called multivariate algorithm: the distance of the current measurement is compared with an expected value. The difference is then checked against a fixed threshold that determines whether the current measurement is a normal value or an anomaly. The expected value is assigned using three different approaches: last observation, closest past observation in a multivariate space within a sliding time window, or by taking the closest-cluster centroid in clusters of past observations using k-mean clustering [17]. The detection mechanism was tested on data collected by monitoring four water quality parameters at four different locations taking one measurement every hour during 125 days. Their contamination has been simulated by superimposing values according to certain profiles to the water quality parameters of the last 2000 samples of the collected data. Results of simulations have shown that the algorithm performs the required level of detection at the cost of a high number of false positives and a change of background quality can severely deteriorate the overall performance.

A comprehensive work on this topic has been initiated by U.S. EPA resulting in the CANARY tool [18]. CANARY is a software for online water quality event detection that reads data from sensors and considers historical data to detect events. Event detection is performed in two online parallel phases: the first phase, called state estimation, predicts the future quality value. In the state estimation, history is combined with new data to generate the estimated sensor values that will be compared with actually measured data. In the second phase, residual computation and classification, the differences between the estimated values and the new measured values are computed and the highest difference among them is checked against a threshold. If that value exceeds the threshold, it is declared as an outlier. The number of outliers in the recent past are then combined by a binomial distribution to compute the probability of an event in the current time step.

While in our case the model of the system is based on observations from the training phase, CANARY integrates old information with new data to estimate the state of the system. Thus, their EDS is context-aware. A change in the background quality due to normal operation would be captured by the state estimator, and that would not generate too many false alarms. Singular outliers due to signal noise or background change would not generate immediately an alarm, since the probability of raising alarms depends on the number of outliers in the past, that must be high enough to generate an alarm. Sensor faults and missing data are treated in such way that their value does not affect the residual classification: their values (or lack thereof) are ignored as long as the sensor resumes its correct operational state.

CANARY allows the integration and test of different algorithms for state estimation. Several implementations are based on the field of signal processing or time series analysis, like time series increment models or linear filtering. However, it is suggested that artificial intelligence techniques such as multivariate nearest neighbour search, neural networks, and support vector machines can also be applied. A systematic evaluation of different approaches on the same data is needed to clearly summarise the benefits of each approach. This is the target of the current EPA challenge of which our work is a part.

So far, detection has been carried out on single monitoring stations. In a water distribution network, several monitoring stations could cooperate on the detection of contaminant event by combining their alarms. This can help to reduce false alarms and facilitate localisation of the contamination source. Koch and McKenna have recently proposed a method that considers events from monitoring stations as values in a random time-space point process, and by using the Kulldorffs scan test they identify the clusters of alarms [23].

5.2 Hydrodynamical aspects and distribution network topology

Modelling hydraulic water flow in distribution systems has always been an aspect of interest when designing and evaluating water distribution systems [36]. A water distribution system is an infrastructure designed to transport and deliver water from several sources, like reservoirs or tanks, to consumers. This infrastructure is characterised by the interconnection of pipes using connection elements such as valves, pumps and junctions. Water flows through pipes with a certain pressure, and valves and pumps are elements used to adjust this to desired values. Junctions are connection elements through which water can be served to customers. The flow of water through the distribution system can be described by mathematical formulation of fluid dynamics [14].

Water distribution networks are modelled using graphs where nodes are connection elements and edges represent pipes between nodes. Computer-based simulation has become popular to study the hydraulic dynamics as well as the water quality through the network. Notwithstanding the problem of intentional or accidental contaminations, water has always been monitored for quality, and the distribution system must be studied to compute the quality decay over the network [21]. System modelling has been performed for finding the appropriate location to place treatment facilities.The most popular tool to model and evaluate water quality in distribution systems is EPANET [3].

5.3 Contamination diffusion

Modelling water quality in distribution networks allows the prediction of how a contaminant is transported and spread through the system. Using the equations of advection/reaction Kurotani et al. initiated the work on computation of the concentration of a contaminant in nodes and pipes [25]. They considered the topographical layout of the network, the changing demand from the users, and information regarding the point and time of injection. Although the model is quite accurate, this work does not take into account realistic assumptions like water leakage, pipes aging, etc. A more realistic scenario has been considered by Doglioni et al. [11]. They evaluate the contaminant diffusion on a real case study of an urban water distribution network that in addition to the previous hypothesis considers also water leakage and contamination decay.

5.4 Sensor location problem

The security problem in water distribution systems was first addressed by Kessler et al. [20]. Initially, the focus was on the accidental introduction of pollutant elements. The defence consisted of identifying how to place sensors in the network in such way that the detection of a contaminant can be done in all parts of the distribution network. Since the cost of installation and maintenance of water quality sensors is high, the problem consists of finding the optimal placement of the minimum number of sensors such that the cost is minimised while performing the best detection. Research in this field has been accelerated after 2001, encompassing the threat of intentional injection of contaminants as a terrorist action. A large number of techniques to solve this optimisation problem have been proposed in recent years [28, 31, 6, 34, 12].

Latest work in this area [13] proposes a mathematical framework to describe a wider number of water security faults (both hydraulic and quality faults). Furthermore, it builds on top of this a methodology for solving the sensor placement optimisation problem subject to fault-risk constraints.

5.5 Contamination source identification

Another direction of work has been contamination source identification. This addresses the need to react when a contamination is detected, and to take appropriate countermeasures to isolate the compromised part of the system. The focus is on identifying the time and the unknown location in which the contamination started spreading.

Laird et al. propose the solution of the inverse water quality problem, i.e. backtracking from the contaminant diffusion to identify the initial point. The problem is described again as an optimisation problem, and solved using a direct nonlinear programming strategy [27, 26]. Preis and Ostfeld used coupled model trees and a linear programming algorithm to represent the system, and computed the inverse quality problem using linear programming on the tree structure [33].

Guan et al. propose a simulation-optimisation approach applied to complex water distribution systems using EPANET [16]. To detect the contaminated nodes, the system initially assumes arbitrarily selected nodes as the source. The simulated data is fed into a predictor that is based on the optimisation of a cost function taking the difference between the simulated data and the measured data at the monitoring stations. The output of the predictor is a new configuration of contaminant concentrations at (potentially new) simulated nodes, fed again to the simulator. This process in iterated in a closed-loop until the cost function reaches a chosen lower bound and the actual sources are found. Extensions of this work have appeared using evolutionary algoritms [37].

Huang et al. use data mining techniques instead of inverse water quality or simulation-optimisation approaches [19]. This approach makes possible to deal with systems and sensor data uncertainties. Data gathered from sensors is first processed with an algorithm to remove redundancies and narrow the search of possible initial contaminant sources. Then using a method called Maximum Likelihood Method, the nodes are associated with the probability of being the sources of injection.

5.6 Attacks on SCADA system

A further security risk that must be addressed is the security of the event detection system itself. As any other critical infrastructure, an outage or corruption of the communication network of the SCADA infrastructure can constitute a severe risk, as dangerous as the water contamination. Therefore, protection mechanisms have to be deployed in response to that threat, too. Since control systems are often sharing components and making an extensive use of information exchange to coordinate and perform operations, several new vulnerabilities and potential threats emerge [10]. This is a wide area of study and the reader is referred to several sources including other chapters in this book for further studies.

6 Conclusion

In this chapter an existing learning based anomaly detection technique has been applied to the detection of contamination events in water distribution systems. These systems are monitored by water quality sensors that provide chemical properties of the water which are processed and used to feed the detector.

The introduction of this system is challenging since the chemical properties of the water can change along the time depending on its source and can be confused as a contamination event. Nevertheless, the use of a learning based anomaly detection technique, which allows the characterisation of all the variations of the system normality, has proved to be effective. Besides, additional features based on sliding windows and derivatives of the data analysed have been introduced to improve the efficiency of the solution under certain circumstances.

The performance of the approach has been analysed using real data of two water stations together with synthetic contaminants superimposed with the EDDIES application provided by EPA. The first results, in terms of detection rate and false positive rate, have shown some contaminants are easier to detected than others. The sensitivity of the anomaly detector has also been been studied by creating new testing data sets with different contaminant concentrations. The results have shown that with more contaminant concentration the detector obtains higher detection rates with low false positive rates. The latency of the detection has also been analysed, showing reasonable results that are qualitatively improved as the contaminant concentration is increased. The inaccuracy of the data provided in one of the stations has negatively affected the performance, but the potential to improve the outcomes have been discussed.

Further research must be done in the analysis of the performance with different event profiles, since the current analysis has considered just one of them. Besides, in some cases a lack or inaccuracy of the monitored data from the chemical sensors has been observed. A solution based on redundancy of sensor values is proposed and it will be applied and evaluated in the future. Finally, the detector algorithm used has retraining and forgetting capabilities, which can be enabled to adapt the normality model to changes in the topology of the water distribution system. Further research must be done to evaluate the effects of the adaptability in this environment.

Acknowledgements We would like to thank Katie Umberg, from EPA, for the provision of data, knowledge on water management systems and the review of our chapter. Support by Dan Joy, from Critigen, in connection with the EDDIES software, is gratefully acknowledged. The financial support of the CUGS graduate school during the preparation of this study is also acknowledged.

References

1. Http://www.independent.co.uk/news/world/europe/contamination-fears-after-leak-from-french-nuclear-waste -plant-863928.html, Accessed 26 April 2010
2. Http://cfpub.epa.gov/safewater/watersecurity/initiative.cfm, Accessed 26 April 2010
3. Http://www.epa.gov/nrmrl/wswrd/dw/epanet.html Accessed 19 November 2010
4. Allgeier, S.C., Umberg, K.: Systematic evaluation of contaminant detection through water quality monitoring. In: Water Security Congress Proceedings. American Water Works Association (2008)
5. ASCE: Interim voluntary guidelines for designing an online contaminant monitoring system. American Society of Civil Engineers, Reston,VA (2004)
6. Berry, J.W., Fleischer, L., Hart, W.E., Phillips, C.A., Watson, J.P.: Sensor placement in municipal water networks. Journal of Water Resources Planning and Management **131**(3), 237–243 (2005)
7. Burbeck, K., Nadjm-Tehrani, S.: ADWICE: Anomaly detection with real-time incremental clustering. In: Proceedings of 7th International Conference on Information Security and Cryptology (ICISC 04), *LNCS*, vol. 3506, pp. 407–424. Springer (2004)
8. Burbeck, K., Nadjm-Tehrani, S.: Adaptive real-time anomaly detection with incremental clustering. Information Security Technical Report - Elsevier **12**(1), 56–67 (2007)
9. Byer, D., Carlson, K.: Real-time detection of intentional chemical contamination in the distribution system. Journal American Water Works Association **97**(7) (2005)
10. Cárdenas, A.A., Amin, S., Sastry, S.: Research challenges for the security of control systems. In: Proceedings of the 3rd conference on Hot topics in security, pp. 6:1–6:6. USENIX Association, Berkeley, CA, USA (2008)
11. Doglioni, A., Primativo, F., Giustolisi, O., Carbonara, A.: Scenarios of contaminant diffusion on a medium size urban water distribution network. pp. 84–84. ASCE (2008)
12. Eliades, D., Polycarpou, M.: Security of water infrastructure systems. In: R. Setola, S. Geretshuber (eds.) Critical Information Infrastructure Security, *Lecture Notes in Computer Science*, vol. 5508, pp. 360–367. Springer Berlin / Heidelberg (2009)

13. Eliades, D., Polycarpou, M.: A fault diagnosis and security framework for water systems. Control Systems Technology, IEEE Transactions on **18**(6), 1254 –1265 (2010)
14. Friedlander, S., Serre, D. (eds.): Handbook of mathematical fluid dynamics, Vol. 1. Elsevier B.V (2002)
15. Goetz, E., Shenoi, S. (eds.): Critical Infrastructure Protection. Springer (2008)
16. Guan, J., Aral, M.M., Maslia, M.L., Grayman, W.M.: Identification of contaminant sources in water distribution systems using simulation–optimization method: Case study. Journal of Water Resources Planning and Management **132**(4), 252–262 (2006)
17. Han, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
18. Hart, D., McKenna, S.A., Klise, K., Cruz, V., Wilson, M.: Canary: A water quality event detection algorithm development tool. pp. 517–517. ASCE (2007)
19. Huang, J.J., McBean, E.A.: Data mining to identify contaminant event locations in water distribution systems. Journal of Water Resources Planning and Management **135**(6), 466–474 (2009)
20. Kessler, A., Ostfeld, A., Sinai, G.: Detecting accidental contaminations in municipal water networks. Journal of Water Resources Planning and Management **124**(4), 192–198 (1998)
21. Khanal, N., Speight, V.: Increasing application of water quality models. pp. 514–514. ASCE (2008)
22. Klise, K.A., McKenna, S.A.: Multivariate applications for detecting anomalous water quality. pp. 130–130. ASCE (2006)
23. Koch, M.W., McKenna, S.: Distributed sensor fusion in water quality event detection. to appear in Journal of Water Resource Planning and Management **137**(1) (2011)
24. Kruegel, C., Valeur, F., Vigna, G.: Intrusion Detection and Correlation Challenges and Solutions. Springer (2005)
25. Kurotani, K., Kubota, M., Akiyama, H., Morimoto, M.: Simulator for contamination diffusion in a water distribution network. In: Industrial Electronics, Control, and Instrumentation, 1995., Proceedings of the 1995 IEEE IECON 21st International Conference on, vol. 2, pp. 792 –797 vol.2 (1995)
26. Laird, C.D., Biegler, L.T., van Bloemen Waanders, B.G.: Mixed-integer approach for obtaining unique solutions in source inversion of water networks. Journal of Water Resources Planning and Management **132**(4), 242–251 (2006)
27. Laird, C.D., Biegler, L.T., van Bloemen Waanders, B.G., Bartlett, R.A.: Contamination source determination for water networks. Journal of Water Resources Planning and Management **131**(2), 125–134 (2005)
28. Lee, B.H., Deininger, R.A.: Optimal locations of monitoring stations in water distribution system. Journal of Environmental Engineering **118**(1), 4–16 (1992)
29. Luiijf, E., Ali, M., Zielstra, A.: Assessing and improving SCADA security in the dutch drinking water sector. In: Critical Information Infrastructure Security Revised papers from Third International Workshop on Critical Information Infrastructure Security (CRITIS 08), *LNCS*, vol. 5508, pp. 190–199 (2009)
30. Murray, R., Uber, J., Janke, R.: Model for estimating acute health impacts from consumption of contaminated drinking water. J. Water Resource Planning and Management **132**(4), 293–299 (2006)
31. Ostfeld, A., Salomons, E.: Optimal layout of early warning detection stations for water distribution systems security. Journal of Water Resources Planning and Management **130**(5), 377–385 (2004)
32. Pietro, R.D., Mancini, L.V.: Intrusion Detection Systems. Springer (2008)
33. Preis, A., Ostfeld, A.: Contamination source identification in water systems: A hybrid model trees–linear programming scheme. Journal of Water Resources Planning and Management **132**(4), 263–273 (2006)
34. Propato, M.: Contamination warning in water networks: General mixed-integer linear models for sensor location design. Journal of Water Resources Planning and Management **132**(4), 225–233 (2006)

35. Sommer, R., Paxson, V.: Outside the closed world: On using machine learning for network intrusion detection. In: Security and Privacy (SP), 2010 IEEE Symposium on, pp. 305 –316 (2010)
36. Walski, T.M., Chase, D.V., Savic, D.A., Grayman, W., Beckwith, S., Koelle, E. (eds.): Advanced water distribution modeling and management. Haestead Press (2004)
37. Zechman, E.M., Ranjithan, S.R.: Evolutionary computation-based methods for characterizing contaminant sources in a water distribution system. Journal of Water Resources Planning and Management **135**(5), 334–343 (2009)