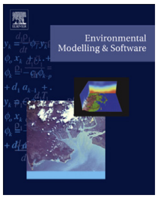
Environmental Modelling & Software 57 (2014) 1–12

Contents lists available at ScienceDirect

Environmental Modelling & Software

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



Minimum volume ellipsoid classification model for contamination event detection in water distribution systems

Nurit Oliker, Avi Ostfeld*

Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 25 August 2013
Received in revised form
14 February 2014
Accepted 26 March 2014
Available online 18 April 2014

Keywords:
Water distribution systems
Water quality
Water security
Event detection
Minimum volume ellipsoid
Sequence analysis

ABSTRACT

The presented study features an event detection model alerting for contamination events in water distribution systems. The developed model comprises a minimum volume ellipsoid (MVE) classifier, detecting outlier measurements, and a following sequence analysis utilizing the MVE binary output, for the classification of events. The model is updated continuously and exploits a constantly growing data base. The MVE enables simultaneous analysis of the water quality parameters. The multivariate analysis explores the relations between water quality parameters and detects changes in their common patterns. The suggested model applied an un-supervised classification method, eliminates the need for simulated events examples in the classifier construction. In the absent of satisfying information regarding the influence of contamination event on the parameter measurements, eliminating the use of any assumption contributes to the model reliability and generality. The model was trained on a real water utility data, and tested on randomly simulated events that were superimposed on the original data base. The model showed high accuracy and detection ability compared to previous studies.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

A water distribution system (WDS) is a vulnerable infrastructure, as it comprises numerous exposed elements which can be simply intruded. Securing drinking water is vital for ensuring society welfare, thus the threat of contaminants intrusion to the network arouses growing concern. Accordingly, water security is one of the most explored topics of the field in recent years. Many resources are invested, both in academy and industry in the development contamination warning systems.

The security of WDS features two major issues: locating sensors around the network, and analyzing their measurements data. The problem of sensor placement is widely explored, featuring over ninety published studies dealing with sensor placing optimization (Hart and Murray, 2010). This research achieved convincing results featuring efficient covering of the network by a minimal number of sensors. Yet, the complementary problem of data analysis is far from being extracted. Most of sensor placement studies assumed the sensors to be "perfect", that is to say if a sensor measures any concentration of a contaminant, it will certainly detect it. Few studies, used a more careful assumption require some threshold concentration to ensure detection, i.e. if a sensor measures more than some set value of concentration, it will detect it. Actually there is no certainty in the detection of contaminants as it is a complex task requires further research. In the overview of warning systems development, there is gap in the data analysis element, and respectively, in the evaluation of the detection ability for the complementary sensor placement problem.

2. Literature review

Some attempts were made to develop systems capable of recognizing specific pollutants intruding the network, according to their unique properties. For example, Adams and Mccarty (2007) utilized light scattering for the detection of spectral signature. The vast verity of pollutants, made it impossible to deal with all yet problematic to focus just on some. Therefore, the specific recognition approach had fallen from grace and a more generic approach was adopted, featuring the use of general water quality parameters (e.g. turbidity, electrical conductivity, pH, etc.) already being measured among utilities, for the aim of event detection. The premise of the last is that abnormal behavior of the general parameters may indicate a contaminant intrusion to the system. Thus, on-line data of water quality parameters is recently used for the development of warning systems. The challenge is then to recognize exceptional behavior of these parameters.

* Corresponding author. Tel.: +972 4 8292782; fax: +972 4 8228898.
E-mail address: ostfeld@tx.technion.ac.il (A. Ostfeld).

http://dx.doi.org/10.1016/j.envsoft.2014.03.011
1364-8152/© 2014 Elsevier Ltd. All rights reserved.

Data mining and event detection methodologies were also developed and applied in atmospheric sciences and for environmental systems. Athanasiadis and Mitkas (2007) presented a hybrid classification-empirical method to aid decision making in air quality management systems. Wang et al. (2010) developed a model for identifying regional atmospheric PM10 transport pathways through an integrated modeling and synoptic pressure pattern analysis. Gross et al. (2010) introduced an open-source software package designed to facilitate the analysis of atmospheric data. Data mining applications were applied to single-particle mass spectrometry data from aerosol particles, and constructed to seamlessly handle large datasets. Gibert and Sanchez-Marre (2011) summarized the Data Mining for Environmental Sciences (DMES) workshop series which initiated in 2006 inside the International Environmental Modelling and Software Society (iEMS), providing guidelines and recommendations for Knowledge Discovery from Data (KDD) techniques and application models. Carslaw and Ropkins (2012) developed an R package for the analysis of air pollution measurement data for enhancing inference possibilities.

Gibert et al. (2008) classified environmental system features for enhancing data analysis and modeling of their complex behavior. Athanasiadis et al. (2010) investigated how data mining methodologies can be incorporated in assuring the quality of decision making processes. Pino-Mejıas et al. (2010) developed and applied data mining models for the prediction of potential habitats for the oak forest type in Mediterranean areas. Hill and Minsker (2010) constructed a real-time anomaly autoregressive data-driven model for identifying deviations from historical environmental data stream patterns. Further to Hill and Minsker (2010), Hill (2013) suggested a real-time method for identifying measurement errors in rain gage data time series based on a dynamic Bayesian network (DBN) model.

Murray et al. (2010), Perelman et al. (2012), and Arad et al. (2013) developed contamination event detection models, based on utilizing general water quality measurements. Perelman et al. (2012) and Arad et al. (2013) applied a parallel analysis for each of the water quality parameters. Utilizing some data-driven method, the models learns the behavior of each parameter time series, and generates a prediction model indicate the expected measured value of the next time step. That way, the models are able to detect deviations from the expected behavior and classify outlier measurements. The outlier estimations of all the parameters are integrated to assess the probability of an event occurrence. Murray et al. (2010) applied several outliers detection algorithm that included both parallel single parameter and multivariate analysis of the time series data. The outlier detection by these algorithms was followed by a statistical tool that calculated event probability according to a sequence of detected outliers.

Horsburgh et al. (2010) developed techniques for characterizing the spatial and temporal variability of low-cost water quality measurements such as turbidity or specific conductance for predicting non-measurable/high-cost water quality parameters. The methodology was applied for estimating phosphorus and total suspended solids in the Little Bear River watershed of northern Utah, USA. Guepie et al. (2012) developed a model based on residual chlorine decay. Their premise was that a contaminant in the WDS will consume a significant fraction out of the measured chlorine and this single parameter provides sufficient valuable data for the detection task.

Hall et al. (2007) conducted an experimental study on sensors response to various contamination intrusions (e.g., Potassium ferricyanide, Malathion formulation, Arsenic trioxide). They showed that at least one water quality inspected parameter (e.g., Chlorine, Turbidity) significantly changed as a result of an intrusion. Szabo et al. (2008) extended Hall et al. (2007) to include sensor responses for Chloraminated water.

Perelman et al. (2012), Arad et al. (2013), Oliker and Ostfeld (2014) utilized supervised classification methods which require the use of event time measurement examples to train the classifier. Unfortunately, for the classification problem there are no published detailed records of real contamination events. Therefore, in supervised classification models, simulated events are been used both for training and testing of the models. For representing the contaminant effect on the general water quality parameters, the models apply some random disturbances to the measured data. Uber et al. (2007) provided guidelines for event simulation based on contaminant reaction kinetics and uncertainty. The random nature of the simulated events maintains generality, but mainly derives from the absence of sufficient knowledge of different contaminants influence.

Klise and McKenna (2006), McKenna et al. (2008) and Murray et al. (2010) applied un-supervised classification methods, requiring the utilization of simulated events only for the assessment of the models performances. In the absent of adequate knowledge of contaminants possible effect on the parameters, eliminating the need in simulated events for the classifier construction contribute to the model reliability and generality.

Nguyen et al. (2013) developed and applied a hybrid methodology of gradient vector filtering, a Hidden Markov Model (HMM), and a Dynamic Time Warping (DTW) algorithm, for pattern recognition of residential water end-use events. Suresh et al. (2013) and Perelman and Ostfeld (2013) investigated how mobile sensor networks can be utilized for optimal event detection and localization in water distribution loop networks. A comprehensive overview on utilizing online water quality data for event detection in water distribution systems was provided by Rusen and Bartrand (2013). They concluded that event detection systems are in their early stages of development, and that investment return is expected if event detection systems will properly address customer concerns on water quality variations.

3. Objectives

The objectives of the presented study are: (1) Apply multivariate analysis of the water quality time series data. The multidimensional analysis provides different description of the system and reveals effects associated with changes in the relations between the parameters. (2) Apply an un-supervised classification method, and (3) develop an accurate and sensitive contamination event detection model.

The model comprises two modular elements: a minimum volume ellipsoid (MVE) classifier, detecting outlier measurements, and a following sequence analysis, utilizing the MVE binary output for the classification of events.

Minimum volume ellipsoid was introduced by Rousseeuw (1985) for the detection of outliers in multidimensional data. This classification method finds the minimal closed quadric surface that contains a given group of vectors. The dimension of the ellipsoid corresponds to the vectors dimension. In most applications, the ellipsoid is required to include some set fraction out of the given vectors. The fraction can be determined according to the measurements degree of reliability (i.e. if the data is more reliable the ellipsoid is required to include a larger fraction of the vectors). After the ellipsoid is situated, any new observation can be classified as normal, if it's lying inside the ellipsoid, or outlier, if it's lying outside of it.

In case of time-depended measurements, the data base is constantly growing, and the ellipsoid contains more vectors with time.

This study is part of an ongoing effort on developing water quality event detection models (Perelman et al., 2012; Arad et al., 2013; Oliker and Ostfeld, 2013, 2014).

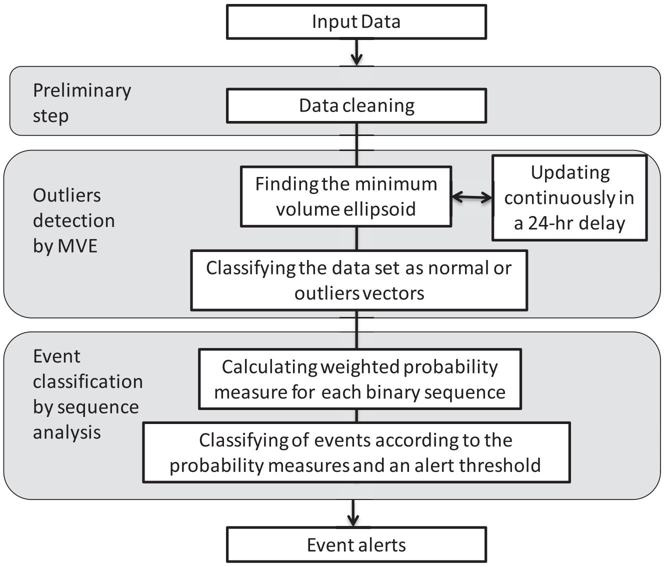

Input Data
Preliminary step
Data cleaning
Outliers detection by MVE
Finding the minimum volume ellipsoid
Updating continuously in a 24-hr delay
Classifying the data set as normal or outliers vectors
Event classification by sequence analysis
Calculating weighted probability measure for each binary sequence
Classifying of events according to the probability measures and an alert threshold
Event alerts

Fig. 1. Model scheme.

4. Model development

The proposed classification model is shown in Fig. 1 and described as follows. The input data of the model consist of water quality time series, as shown in Fig. 2. A preliminary step to the classification process is the data cleansing, removing corrupted values and noise measurements out of the data. The classification process comprises two modular elements: a MVE for the detection of outlier measurements, and a following sequence analysis, utilizing the MVE binary output, for the classification of contamination events.

4.1. Data cleansing

A preliminary step to the analysis of the data is the data cleaning. Every measured data includes some measurement noises which are liable to affect the classifier construction. In this case, the presence of noise measurements may over-expend the ellipsoid and blight the classifier sensitivity. Thus, it seemed essential to filter the data before analyzing it. The model includes a very simple data cleansing, consists of removing non-positive values, and values which exceed 4 standard deviations away from the average. Negative values are physically impossible when referring to water quality parameters, thus surely originate in measurements errors. Values situated more than 4 standard deviations away from the mean are also very unlikely to represent true measurements, and therefore, removed from the data base.

4.2. Ellipsoid construction and outlier detection

The classifier is constructed by finding the minimal ellipsoid, which includes 95% of the vectors in the known data set. The fraction of vectors required to be bounded in the ellipsoid,

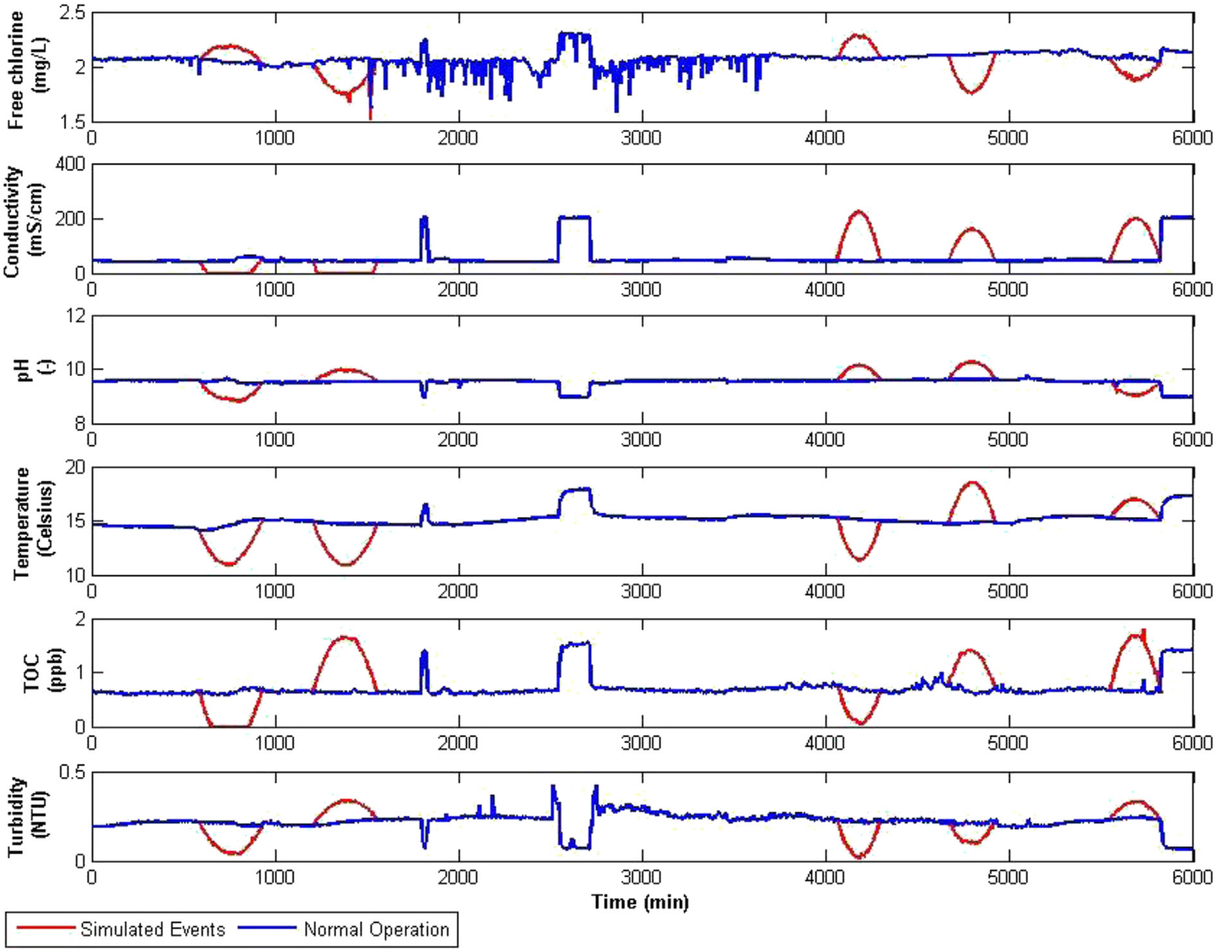


Fig. 2. Water quality time series: normal operation measurements and superimposed simulated events.

determines the relative weight of sensitivity and specificity of the classifier. When the fraction is higher, the obtained ellipsoid is larger, and the accepted classifier is more specific and less sensitive (classifying more of the space as normal representing). Reducing the fraction results in a smaller ellipsoid, detecting more outliers but produces more false alarms. The 95% fraction is a typical value that was set by trial and error.

The ellipsoid can be formulated by:

$$(x-c)^T \times A \times (x-c) = 1 \tag{1}$$

where x is the variables vector, c is the ellipsoid center coordinates vector and A is the matrix of coefficients in the ellipse equations.

The minimal ellipsoid problem can be expressed by:

$$\begin{aligned}&\text{Minimum}\ \ \log|A|\\ &\text{Subject to}: (Pi-c)^T \times A \times (Pi-c) \le 1 \quad \forall i\end{aligned} \tag{2}$$

where Pi is a measured vector (required to be bounded by the ellipsoid). This formulation features a simplified convex optimization problem (Moshtagh, 2005).

The ellipsoid is found by the Khachiyan Algorithm (Khachiyan, 1996), an efficient method for the MVE problem (Todd and Yıldırım, 2007). The Khachiyan algorithm finds the minimum volume ellipsoid (characterized by the matrix A and its center coordinates as shown in (1)) by an iterative procedure constructing a sequence of decreasing ellipsoids. The algorithm starts from a feasible solution, i.e. a large volume ellipsoid containing all given vectors. Iteratively, the fraction of samples included in the ellipsoid is repeatedly checked for guaranteeing its pre-defined value (this case, 95% of the vectors). The algorithm is converging in a polynomial bounded number of iterations (Todd and Yıldırım, 2007). The most extreme 5% vectors are excluded from the calculations. That is to say, the minimal ellipsoid that includes any combination of 95% of the samples is found.

As mentioned, the ellipsoid construction exploits only the normal vectors (i.e. measured in normal operation time) in the known data set and doesn't require an abnormal (i.e. events representing vectors) examples. An example of a 6-dimensional ellipsoid projection on 3-D space is shown in Fig. 3.

With time the "known" data base is continuously growing. The working assumption of the study was that after 24 h, the true native of the measurements is clarified, and those taken in normal operation conditions can be used for the ellipsoid construction. Therefore, the ellipsoid is re-constructed constantly, utilizing a growing data base, adding the "on-line" measured data in a 24-h delay, without eliminating all previous measurements. The updating of the ellipsoid is presented in Fig. 4.

After the ellipsoid parameters are found, the incoming unknown measured data can be classified as normal (if situated inside the ellipsoid) or outliers (situated outside of it). The output of the MVE classifier is a binary sequence, consist the normal (0) and outliers (1) classification of each time step measurements vector.

4.3. Sequence analysis

The binary output of the MVE classifier includes the normal or outlier classification of the measurements. There is a difference between the classification of outliers and events, as a succession of outliers is surely stronger evidence to an event occurrence than a single one. Therefore, the classification of every time step is based on sequence analysis of a segment out of the MVE binary output.

The sequence analysis parameters, as all the other parameters of the model were set by trial and error. Since this is an un-supervised classification method the calibration of the model could not be made by the model itself.

The length of the analyzed sequence for the classification of every time step is 6, corresponding to time duration of 30 min for measurement taken every 5 min. When classifying a certain time step measurements the sequence segment ending in the same time step is utilized for its classification as event time or normal time measurement.

The sequence segment is evaluated by a probability measure composed of two elements, *proportion* and *continuity*. The measure is calculated by the expression:

$$\text{Probability measure} = 0.75 \times \text{proportion} + 0.25 \times \text{continuity} \tag{3}$$

where the *proportion* is the fraction of outliers out of the segment, and the *continuity* is their succession. The last is expressed by the longest sequence of outliers within the analyzed sequence, as a fraction relative to the segment length. The *proportion* element represents the exceptionality of the sequence among other measurements. The *continuity* element represents the reliability of the outliers, where a sequential sequence of outliers is a stronger event indicator than a fragmented one. The two elements summed-up to unity in order to reflect the probability that the sequence indicates an event occurrence. The coefficients of the measure (i.e., the 0.25 and 0.75 values) were determined by trial and error. A sensitive analysis is shown on Table 1.

For example, if the analyzed sequence segment is [0 1 1 1 0 1] then the *proportion* of outliers in the sequence equals 4/6. The

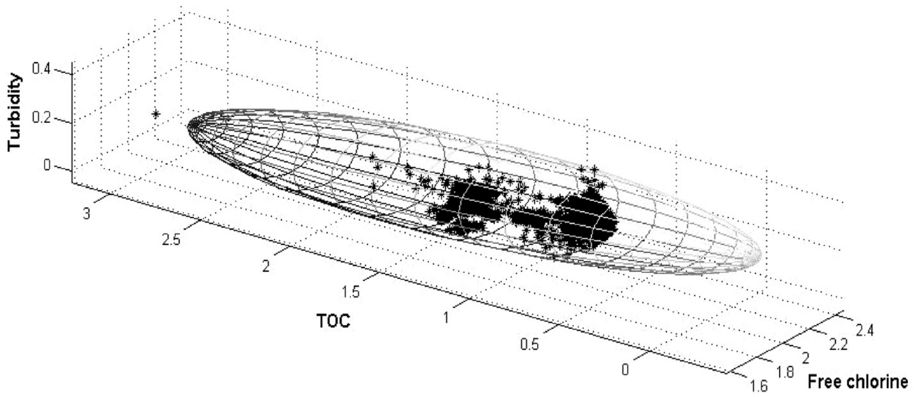


Fig. 3. Example of 3-D ellipsoid projection on the Turbidity-TOC-Free chlorine space.

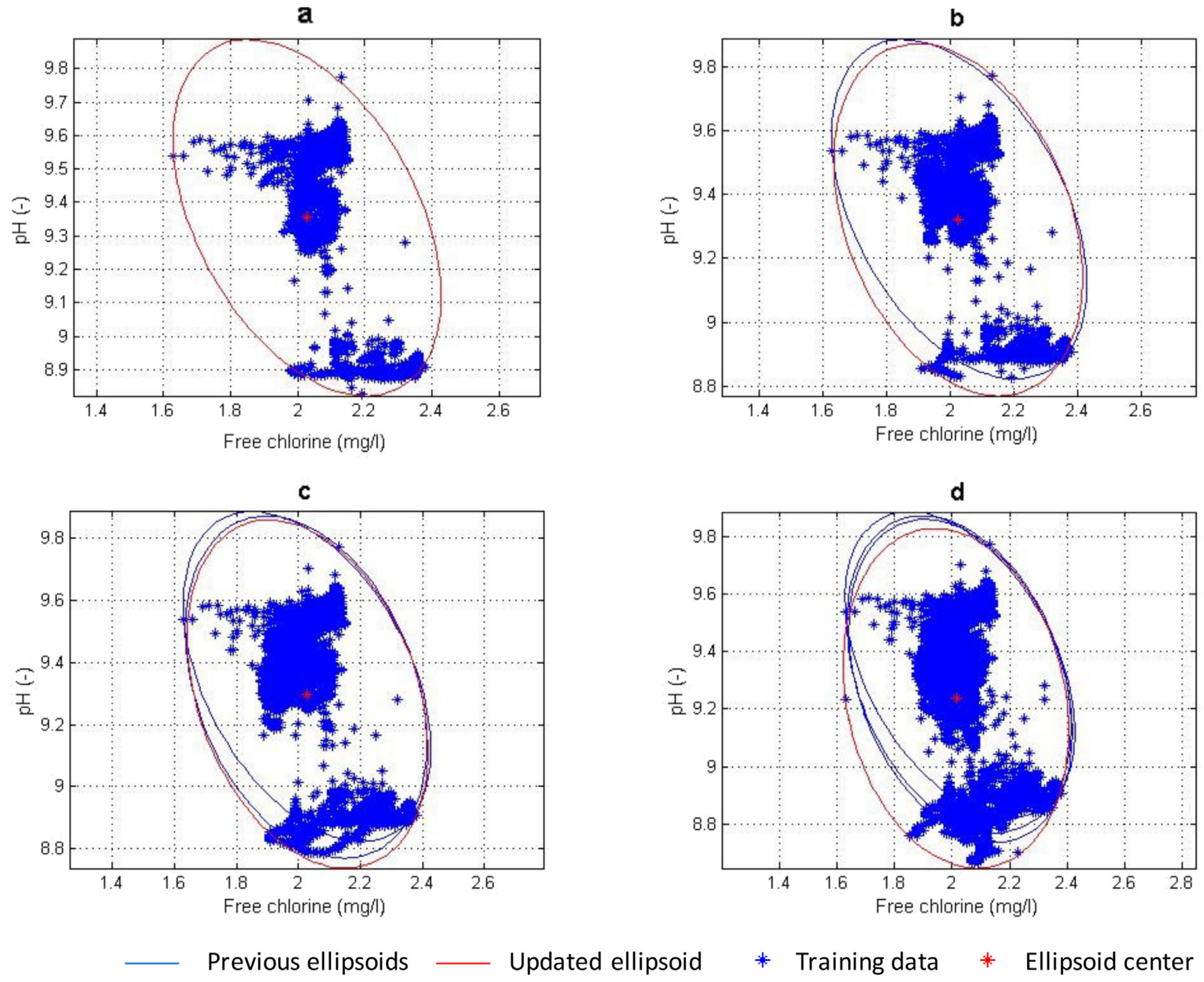


Fig. 4. The ellipsoid updating: (a) is the initial ellipsoid according to the training data set, (b) (c) and (d) are the updates after 7, 20 and 35 days, respectively.

continuity equals 3/6 (according to the longest sequence of outliers divided by the segment length). That case, the probability measure equals 0.625 (0.75 × 0.667 + 0.25 × 0.5). This analysis is used for the classification of the time step represented by the last element in the segment.

After being calculated the probability measure is compared to a threshold value. If the probability measure exceeds the threshold value, the time step is classified as event, otherwise it is classified as normal operation. The classification decision rule can be represented by:

$$\text{If Probability measure} \geq \text{Threshold then event alert is raised else normal operation is reported} \quad (4)$$

The transition from outlier detection to event classification is shown in Fig. 5, presenting the binary state of the data base together with its binary classification. Each time step representing either normal operation or event time, and superimposed is the corresponding classification of it as (a) normal or outlier, and (b) normal or event. The transition from outlier's classification to event classification achieves the clearance of false alarms from the results. The model avoids from turning short or inconsistent outlier's detection into an event alert. Fig. 5 shows an example of false alarm removal, where the outlier detection includes 3 false alarms and the following event classification filters 2 out of them. There may be cases in which an event is barely detected by the MVE and consequently classified as non-event time. From the model development trials those cases were very few. It seemed that the

Table 1
Sensitive analysis of the probability measure given in (3). w1 and w2 are the *proportion* and *continuity* coefficients, respectively. The detection ratio and accuracy are calculated according to (6). The gray cells show the selected coefficients.

Event type	w1	w2	w1	w2	w1	w2	w1	w2
	1	0	0.75	0.25	0.25	0.75	0	1
	Detection ratio	Accuracy	Detection ratio	Accuracy	Detection ratio	Accuracy	Detection ratio	Accuracy
High	0.58	0.917	0.663	0.918	0.585	0.918	0.577	0.917
Medium	1	0.944	1	0.954	1	0.947	1	1
Low	1	0.94	1	0.944	1	0.94	1	1

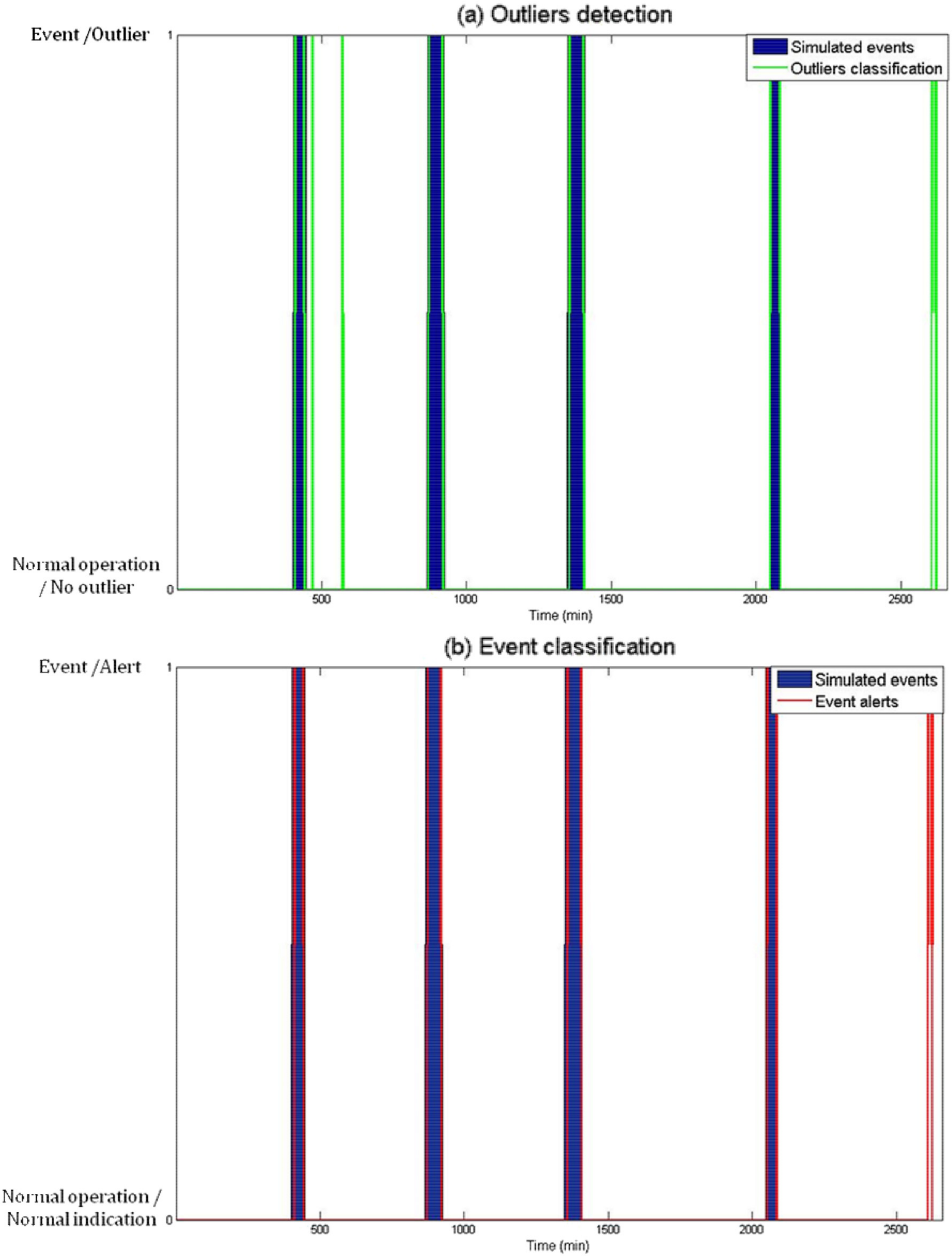


Fig. 5. Example of an outliers detection (a) together with the following event classification (b) for a testing data set segment.

clearance of false alarms provided a clear advantage to the described transition.

The complete computer code of the proposed methodology and metadata on the program structure are provided as Supplementary files. Contact the authors for any additional details.

5. Application

The described methodology was applied on a real data base that was attained by a utility in the United States and available from CANARY (2013). The data (shown in Fig. 2) was collected by a

SCADA system, measuring water quality data from a WDS and includes online multivariate water quality measurements taken every 5 min during four weeks (approximately 8000 time steps). All measurements were taken in normal operating conditions. The data includes the following 6 water quality parameters: Total chlorine, Electrical conductivity, pH, Temperature, Total organic carbon (TOC) and Turbidity.

The data base was divided into two sub-sets: 70% were allocated for training the classifier and the rest 30% for testing. The training data set was used as "off-line" data for the construction of the initial ellipsoid. The testing data set was left un-touched in order to simulate the "on-line" real-time operation of the model and enable its assessment.

5.1. Simulated events

Unluckily for the classification task, there are no published detailed records of real event time measurements in WDS. Therefore, contamination events were simulated in order to test the model. Following Klise and McKenna (2006), McKenna et al. (2008), Perelman et al. (2012), Arad et al. (2013), contamination events were simulated and superimposed on the measured routine patterns (as shown in Fig. 2).

The presented model had utilized the simulated events only for the assessment of the model performance. Therefore, the simulated events were super-imposed only on the testing data set.

The simulated events, presented in Table 2, included various scenarios in order to test the model performance in different situations. The events were characterized by their magnitude, direction, and effect durations, determined randomly by a uniform distribution selection from a set range of values (as shown in Table 2).

The superimposed events were in the shape of Beta distribution function (Keeping, 2010) that may be expressed by:

$$f(y|a,b) = \frac{1}{B_{(a,b)}} y^{(a-1)} y^{(b-1)}$$
$$B_{(a,b)} = \int_0^1 t^{a-1}(1-t)^{y-1}\mathrm{d}t \qquad (5)$$

where y is the random value for which the distribution is calculated for, a and b are the distribution shape parameters, and t is the integration variable. For the event simulation the two shape parameters (a and b), were set as 2, features a Gaussian shape with no infinite marginal. That way, the events influence was limited to the selected event duration and had no residual effect on the data base.

The events duration was assumed to be between two and six hours. The magnitude strength ranged from zero to 2.5 standard deviations. The direction of the deviations was set randomly either to the positive or negative direction. Both magnitude and direction were separately set for each of the parameters.

Table 2
The properties of the simulated events. The frequency of all events ranged between once and twice a day. The beta distribution parameters where both set as 2.

Event type	Duration		Strength		No. of influenced parameters	
	min	max	min	max	min	max
High	48 (4 h)	72 (6 h)	1	2.5	6	6
Medium	36 (3 h)	72 (6 h)	0.5	2	6	6
Medium 2	36 (3 h)	72 (5 h)	0.5	1.5	6	6
Medium 3	36 (3 h)	72 (5 h)	0.5	1	6	6
Low	24 (2 h)	48 (4 h)	0	1	6	6
Low 2	24 (2 h)	48 (3 h)	0	0.8	6	6
Low 3	24 (2 h)	48 (3 h)	0	0.5	6	6
Partial	36 (3 h)	72 (6 h)	0.5	2	1	6

Table 3
The initial ellipsoid parameters according to the 'center form' presented in (1).

Matrix A					
6.39	0	0.63	0	−0.22	2.21
0	0	0.03	0	−0.01	0
0.63	0.03	10.62	0.39	−0.34	−1.96
0	0	0.39	0.13	−0.06	0.61
−0.22	−0.01	−0.34	−0.06	0.65	1.25
2.21	0	−1.96	0.61	1.25	33.38
Vector C					
Free chlorine	Conductivity	pH	Temperature	TOC	Turbidity
2.04	105.97	9.24	16.96	1.17	0.18

Table 4
The model performace for different event types, including events that influenced only 1, 2 and 3 of the parameters and events in different intensities. The values are averaged results of 10 scenarios of each event type. All event properties are given in Table 2.

Event type	Detection ratio	Accuracy	False alarms
Partial – 1 influenced parameter	0.44	0.90	0.9
Partial – 2 influenced parameters	0.64	0.90	1
Partial – 3 influenced parameters	0.91	0.93	1
Low 3	0.13	0.89	1
Low 2	0.30	0.91	1
Low	0.66	0.92	0.7
Medium 3	0.76	0.90	1
Medium 2	0.97	0.93	1
Medium	1	0.95	1
High	1	0.94	1

The average frequency of events was between once to twice a day, and the occurrence timing had no restrictions. The times of events within the data base were set randomly enabling long normal operation times along with overlapping events.

5.2. Results and discussion

The model was trained and tested on the data base described above, comprises six water quality parameters measurements. The initial ellipsoid, constructed according to the initial training data set, is presented in Table 3. Matrix A includes the ellipsoid coefficients, and the vector C includes the center coordinates, according to the ellipsoid formulation in (1). The ellipsoid projection in the space of Free chlorine-TOC-Turbidity is shown if Fig. 3.

The model was evaluated by two measures: Accuracy and detection ratio, calculated by:

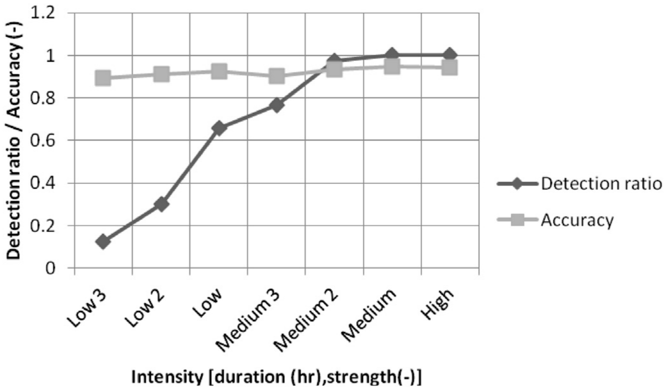


Fig. 6. Averaged Performance of the model for events with different intensities. The-shown values appear in Table 4 and the events properties are given in Table 2.

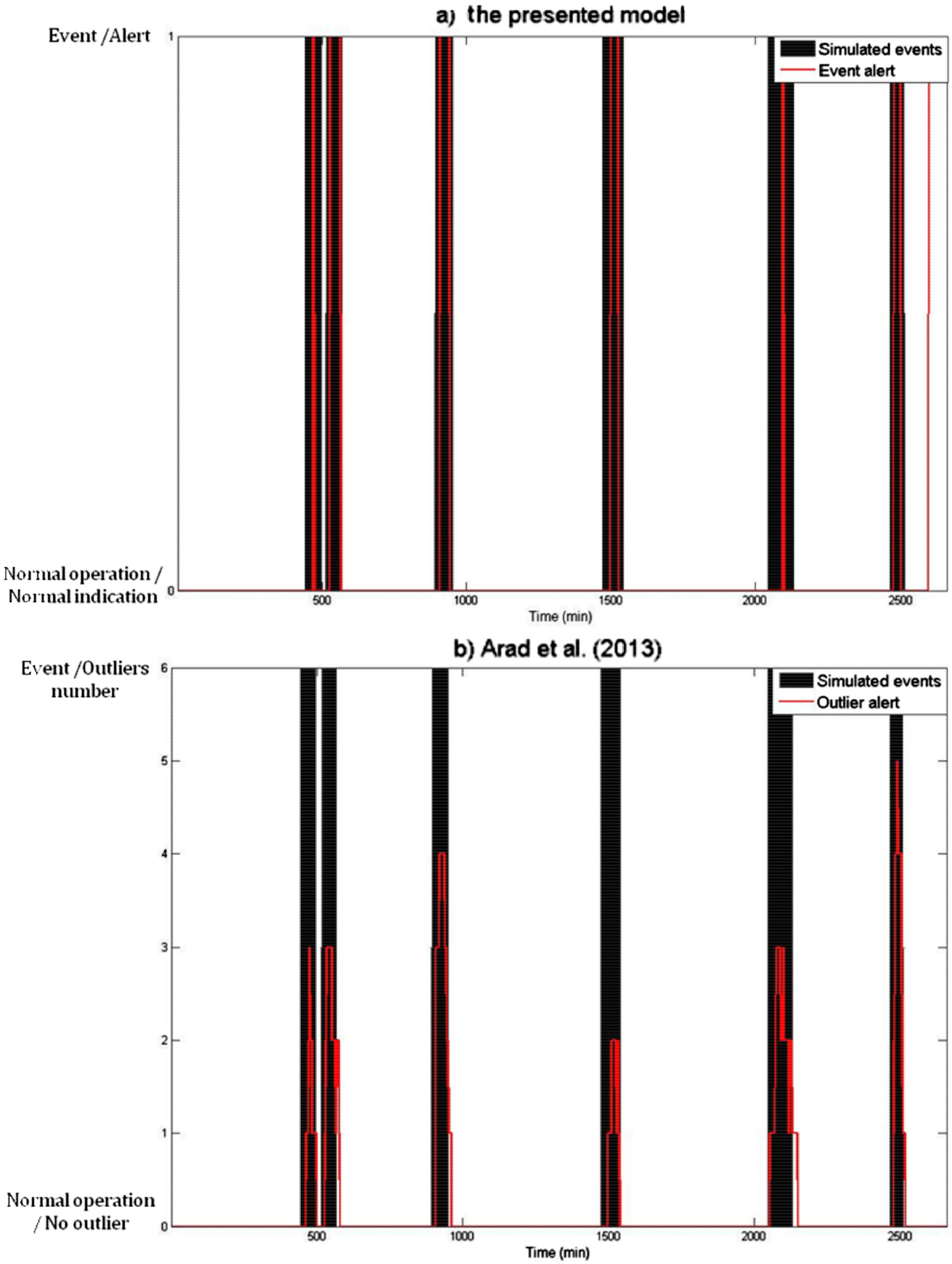


Fig. 7. A comparison example between (a) the current model and (b) Arad et al. (2013) performance for the same scenario of the type 'high' (described in Table 2).

$$\text{Accuracy} = \frac{\text{Well-classified vectors}}{\text{Total vectors number}}$$
$$\text{Detection ratio} = \frac{\text{Detected events}}{\text{Total events number}} \qquad (6)$$

where the accuracy indicates the model sensitivity and the detection ratio presents the model detection ability. Another measure that was analyzed and presented is the index of false alarms (FA), features the number of event alerts raised during normal operation conditions.

The presented model was tested by the different scenarios described in Table 2. The full results, average of 10 runs for each event type, are presented in Table 4. Each run includes a different scenario of simulated events, applied on the identical testing data set. The measures of detection ratio, accuracy and false alarms were calculated for each run separately, and the arithmetical means of 10 runs for each event type are presented. The model showed stability, reflected by the corresponding of performance to the intensity of the event (i.e. as the event intensity was higher the model showed higher accuracy and detection ratio). This trend is shown in Fig. 6. The event intensity was set by its length and magnitude (meaning higher intensity events were longer with stronger disturbances to the routine patterns). The average accuracy and detection ratio of six different event types are presented. The detection ratio showed monotonic ascent with event intensity, as the accuracy ascent was not monotonic but showed improvement. For partial influence events, i.e. events that effects only part of the parameters, the detection ratio showed high variability, equals 0.44 for events that influence a single parameter, 0.64 for events influence two of the parameters and 0.91 for events influencing three parameters. On the whole, the accuracy showed low diversity with all values ranging between 0.89 and 0.95. The model gave about one false alarm for all event scenarios.

With time the ellipsoid is necessarily extended when more normal defined data is included in it. If a significant change occurs, the data will be classified as an event. However if a moderate increase or decrease trend is measured it may expand the ellipsoid and thereby make the classifier less sensitive. Using a “time window” for the construction of the classifier (i.e., utilizing only a certain segment of the latest measured data) will not address this issue as the ellipsoid will only shift faster with the measured trend. A simple way for the model to partially cope with shifts of the data is to use hard threshold values. If the measured data exceed a predetermined threshold, whether for a specific parameter or a combination of parameters, an alert is raised. This threshold will be determined by the operator according to the system properties.

5.3. Model comparison

In order to evaluate the model it was necessary to compare its performance to other models when tested on identical events scenarios. The model was compared to Arad et al. (2013) and CANARY (2013). Arad et al. (2013) suggested an event detection model that includes two optional decision rules for event alerts. The model comprises independent outlier’s detection element for each of the water quality parameters, based on an artificial neural network algorithm. The decision rules differ in the minimal number of parallel detected outliers required in order to trigger an event alert. Decision rule 1 required at least three outliers out of the six water quality parameters, and decision rule 2 required at least five outliers. Inherently, decision rule 1 detects at least, and mostly more events than decision rule 2. On the other hand decision rule 1 produce more false alarms. Representing the statistic measures of sensitivity and specificity (Altman and Bland, 1994), decision rule 1 is more sensitive and decision rule 2 is more specific. CANARY (2013) is the most known model in the field feature outlier detection by a combination of a few algorithms: a linear filter, a multivariate nearest-neighbor algorithm, and a set-point proximity algorithm. The CANARY (2013) comprises parameters that are user-defined and thus requires the operator calibration.

An example of the comparison between the presented model and Arad et al. (2013) for the same event scenario is shown in Fig. 7. In the featured scenario the presented model detected all six events and gave one false alarms. Arad et al. (2013) decision rule 1 detected five of the six events (all except the forth) and decision rule 2 detected just one of the six (the last one). Neither of the decision rules gave any false alarm.

Fig. 8 shows the averaged performance of the presented model and Arad et al. (2013) for events which influenced only one, two or three out of the six parameters. For both models the performances were expectedly better when more parameters were influenced. The accuracy of the models for all scenarios showed minor differences, revolving around 0.9, where the detection ratio had showed more variety. The presented model showed the highest detection ratio for all types of partial influencing events, where decision rule 2 showed the lowest for all.

Tables 5 and 6 show the full comparison results of the presented model, Arad et al. (2013) and CANARY (2013) by the aspects of accuracy and detection ratio respectively. The comparison include the suggested model, Arad et al. (2013) two decision rules and five variations of CANARY parameters selection. The presented values are the averaged results for 10 runs of each event type. The presented model showed the highest detection ratio with an overall average of 0.87 for all event types, compared with 0.79 by Arad et al. (2013) and 0.82 by CANARY (2013). The model accuracy was also perefereble with an overall average of 0.94 compared to 0.91 by Arad et al. (2013) and 0.77 by CANARY (2013). The drawback of the presented model is an average false alarm of 0.93, where Arad et al. (2013) decision rule 1 gave almost null (0.02), decision rule 2 gave null and CANARY (2013) gave (0.14). Fig. 9 presents this trade-off between high detection ratio and low false alarms for the presented model and Arad et al. (2013). The presented model showed higher sensitivity expressed in higher detection ability but also in roughly one false alarm more than Arad et al. (2013) two decision rules.

Unlike the suggested model and Arad et al. (2013) which doesn’t require any calibration, and produces a sole injective classification, the CANARY (2013) is a user-interactive software requires parameters calibration. Thus, the comparisons included five variations of parameter selection. However, note that the results are not

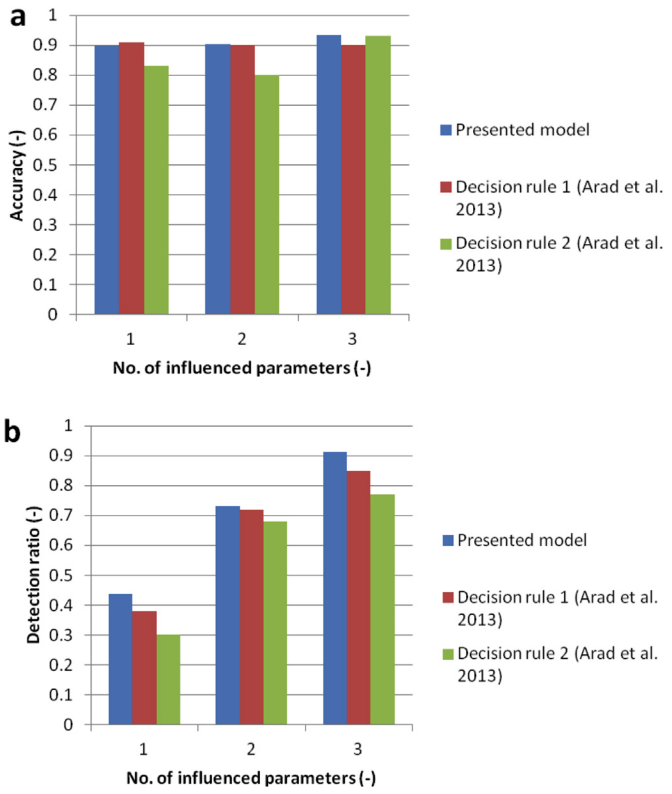


Fig. 8. A comparison between the presented model and Arad et al. (2013). The averaged accuracy (a) and detection ratio (b) for events that influence only 1, 2 and 3 of the parameters. Events properties are given in Table 2.

Table 5
Comparison between the average detection ratio of the presented model, Arad et al. (2013), and CANARY (2013) (10 scenarios of each event type, the bold values in the CANARY (2013) parameters comprise the changes conducted in the default parameters for each set of runs).

Model					Event type		
					Low	Medium	High
Current model					0.66	1	1
Arad et al. (2013) Decision rule 1					0.63	0.9	0.98
Arad et al. (2013) Decision rule 2					0.6	0.94	0.85
CANARY (2013)							
CANARY Parameters	Event threshold	Precision – free Cl	Precision – Conductivity	Precision – pH			
Default parameters	0.9	0.0035	1	0.01	0.63	0.88	0.92
Canary – parameters sensitivity analysis	**0.8**	0.0035	1	0.01	0.63	0.92	0.92
	0.7	0.0035	1	0.01	0.63	0.92	0.92
	0.9	**0.01**	**1.5**	**0.05**	0.69	0.86	0.92
	0.9	**0.1**	**2**	**0.1**	0.66	0.86	0.9

Table 6
Comparison between the average accuracy of the presented model, Arad et al. (2013), and CANARY (2013) (10 scenarios of each event type, the bold values in the CANARY (2013) parameters comprise the changes conducted in the default parameters for each set of runs).

Model					Event scenario		
					Low	Medium	High
MVE					0.92	0.95	0.94
Arad et al. (2013) Decision rule 1					0.85	0.87	0.96
Arad et al. (2013) Decision rule 2					0.93	0.93	0.91
CANARY (2013)							
CANARY Parameters	Event threshold	Precision – free Cl	Precision – Conductivity	Precision – pH			
Default parameters	0.9	0.0035	1	0.01	0.78	0.79	0.75
Canary – parameters sensitivity analysis	**0.8**	0.0035	1	0.01	0.77	0.77	0.74
	0.7	0.0035	1	0.01	0.77	0.77	0.74
	0.9	**0.01**	**1.5**	**0.05**	0.81	0.8	0.76
	0.9	**0.1**	**2**	**0.1**	0.66	0.81	0.77

deterministic and CANARY performance might be improved by additional tuning of its parameters.

6. Conclusions

This paper presented the development of a two-step classification model based on minimal volume ellipsoid classifier for the outlier detection and a following sequence analysis, for the event classification. The model supplies a complete decision support system for contamination event detection.

A preliminary step to the classification procedure is the data cleansing which includes the removing of both negative and very exceptional (i.e., exceeds 4 standard deviations away from the mean) measurements. It should be noted that if the data base distribution is far from normal, this step may remove a significant fraction of the data. For the described data base this was not the case, as only a slight fraction of very exceptional data was removed. However, that point should be verified in the calibration step.

The model applied multivariate analysis of the water quality parameters data differs from the parallel analysis of the parameters conducted in most previous studies. This is an application of an unsupervised classification method, utilizing only the real normal operation measured data and reducing the need of simulated events example for the classifier construction.

The multi-dimensional data analysis produce different description of the system and reveals phenomena associated with the mutual relations of the quality parameters. This analysis enables the examination of the measurements relatively to the multivariate system, i.e. evaluating a parameter value not only in relation to the same parameter previous values, but also relatively to the other parameters current and previous values. In addition,

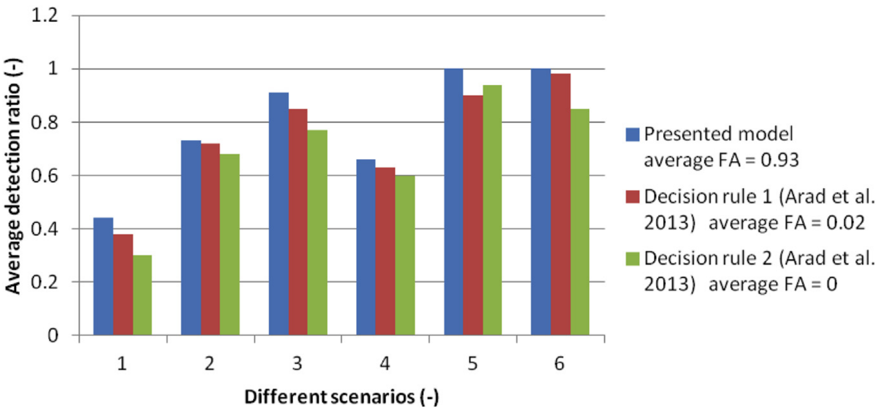


Fig. 9. A comparison of the trade-off between detection ratio and false alarms (FA) in the presented model and Arad et al. (2013).

the simultaneous analysis reduce the later integration required in the parallel analysis approach.

The existing knowledge regarding the events expression in water quality parameters is far from being satisfactory. Therefore, the simulated events applied in this field are completely generic, assuming the event cause some unknown disturbances to the measurements. This random nature of the simulated events maintains generality, as no specific reaction of the quality parameters is assumed, yet it has no physical connection to the water distribution system behavior, and thus quite unfounded. The use of an unsupervised classification method seems to provide a fundamental advantage as it reduces the need in any assumptions regarding the events influence. The unsupervised method utilizes only the known normal operations measurements, trained to recognize any abnormal behavior.

The unsupervised method has its price as it prevents the model autonomic calibration. That is due to the absent of known event examples, prevents the model from verifying testing and perform tuning of its parameters. Therefore, all of the model parameters were set by trial and error.

The presented model achieved good results reflected in high accuracy and detection ratio compared to Arad et al. (2013) and CANARY (2013). The model showed superiority in those aspects for different events scenarios, especially in its detection ability. The relative disadvantage of the model was an average of one extra false alarm in the performance of the "on-line" data set. These results represent the known trade-off between sensitivity and specificity (Altman and Bland, 1994), where the presented model is clearly more sensitive but less specific.

It should be noted, that the model is sensitive to any kind of abrupt shifts. The model is trained on a given data set and if the following classified data is significantly different it is liable to be classified as an event. Inter alia if a second source supplies the water the model should be calibrated accordingly. This implies that the classifier should be trained on the new source data in order to be calibrated as normal operation data. It should be emphasized that without proper calibration any changes in the data caused by source replacement, hydraulic events, etc., may cause an alert and will subsequently require the operator interfering.

Furthermore if the system includes more than one source of water, the data base may be segmented in the measurements space. Using more than a single ellipsoid for the classifier construction is a considerable possibility as it could encircle the data set more compactly. The drawback of multiple ellipsoids is the enlargement of the number of parameters in the model. For the presented data base, each ellipsoid is defined by 42 parameters [according to the formulation in (1)]. The risk of over-fitting of the model increases. Additionally, if there are several separate ellipsoids, the space between them is defined as abnormal. On the physical aspect of water systems, defining intermediate values as outliers seems problematic. Water quality parameters mostly have a continuous range of normal values.

Further research is suggested to examine other classification methods for their suitability to the event detection problem. A parallel study of the authors includes the development of a coupled support vector machine-evolutionary optimization classification model (Oliker and Ostfeld, 2013, 2014). It is a supervised classification method, requiring the use of simulated events for the classifier construction, but enables the autonomic calibration of the model parameters.

Acknowledgments

This work was supported by the Technion Funds for Security research and by the Technion Grand Water Research Institute.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2014.03.011.

References

Adams, J.A., Mccarty, D., 2007. Real-time on-line monitoring of drinking water for waterborne pathogen contamination warning. Int. J. High Speed Electron. Syst. 17 (4), 643–659. http://dx.doi.org/10.1142/S0129156407004850.

Altman, D.G., Bland, J.M., 1994. Diagnostic tests 1: sensitivity and specificity. BMJ 308 (6943), 1552. http://dx.doi.org/10.1136/bmj.308.6943.1552.

Arad, J., Housh, M., Perelman, L., Ostfeld, A., 2013. A dynamic thresholds scheme for contaminant event detection in water distribution systems. Water Res. 47 (5), 1899–1908. http://dx.doi.org/10.1016/j.watres.2013.01.017.

Athanasiadis, I.N., Mitkas, P.A., 2007. Knowledge discovery for operational decision support in air quality management. J. Environ. Inform. 9, 100–107. Available online at: http://www.idsia.ch/~ioannis/publications/pdf/jei2007.pdf (accessed 09.02.14.).

Athanasiadis, I.N., Rizzoli, A.-E., Beard, D.W., 2010. Data mining methods for quality assurance in an environmental monitoring network. In: Diamantaras, K., Duch, W., Iliadis, L.S. (Eds.), Artificial Neural Networks: ICANN 2010, Part III, LNCS 6354. Springer Berlin Heidelberg, pp. 451–456.

CANARY, 2013. Event Detection Software. EPA, Sandia Corporation. https://software.sandia.gov/trac/canary (accessed 05.02.14.).

Carslaw, D.C., Ropkins, K., 2012. Openair – an R package for air quality data analysis. Environ. Model. Softw. 27–28, 52–61. http://dx.doi.org/10.1016/j.envsoft.2011.09.008.

Gibert, K., Spateb, J., Sanchez-Marre, M., Athanasiadis, I.N., Comas, J., 2008. Chapter twelve data mining for environmental systems. Dev. Integr. Environ. Assess. 3, 205–228. http://dx.doi.org/10.1016/S1574-101X(08)00612-1.

Gibert, K., Sanchez-Marre, M., 2011. Outcomes from the iEMSs data mining in the environmental sciences workshop series. Environ. Model. Softw. 26, 983–985. http://dx.doi.org/10.1016/j.envsoft.2011.01.009.

Gross, D.S., Atlas, R., Rzeszotarski, J., Turetsky, E., Christensen, J., Benzaid, S., Olson, J., Smith, T., Steinberg, L., Sulman, J., Ritz, A., Anderson, B., Nelson, C., Musicant, D.R., Chen, L., Snyder, D.C., Schauer, J.J., 2010. Environmental chemistry through intelligent atmospheric data analysis. Environ. Model. Softw. 25, 760–769. http://dx.doi.org/10.1016/j.envsoft.2009.12.001.

Guepie, B.K., Fillatre, L., Nikiforov, I., 2012. Sequential Monitoring of Water Distribution Network, vol. 16, Part 1, pp. 392–397. http://dx.doi.org/10.3182/20120711-3-BE-2027.00114. Paper Presented at the IFAC Proceedings Volumes (IFAC-Papers Online).

Hall, J.S., Zaffiro, A.D., Marx, R.B., Kefauver, P.C., Krishnan, E.R., Haught, R.C., Herrmann, J.G., 2007. On-line water quality parameters as indicators of distribution system contamination. J. Am. Water Works Assoc. 99 (1), 66–77.

Hart, W.E., Murray, R., 2010. Review of sensor placement strategies for contamination warning systems in drinking water distribution systems. J. Water Resour. Plan. Manag. 136 (6), 611–619. http://dx.doi.org/10.1061/(ASCE)WR.1943-5452.0000081.

Hill, D.J., Minsker, B.S., 2010. Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. Environ. Model. Softw. 25, 1014–1022. http://dx.doi.org/10.1016/j.envsoft.2009.08.010.

Hill, D.J., 2013. Automated Bayesian quality control of streaming rain gauge data. Environ. Model. Softw. 40, 289–301. http://dx.doi.org/10.1016/j.envsoft.2012.10.006.

Horsburgh, J.S., Jones, A.S., Stevens, D.K., Tarboton, D.G., Mesner, N.O., 2010. A sensor network for high frequency estimation of water quality constituent fluxes using surrogates. Environ. Model. Softw. 25, 1031–1044. http://dx.doi.org/10.1016/j.envsoft.2009.10.012.

Khachiyan, L.G., 1996. Rounding of polytopes in the real number model of computation. Math. Operations Res. 21 (2), 307–320. http://www.jstor.org/stable/3690235 (accessed 05.02.14.).

Keeping, E.S., 2010. Introduction to Statistical Inference. Dover Publications, 78-0486685021.

Klise, K.A., McKenna, S.A., 2006. Multivariate application for detecting anomalous water quality. In: Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium, WDSA, Cincinnati, Ohio, USA, pp. 1–11. http://ascelibrary.org/doi/abs/10.1061/40941%28247%29130.

McKenna, S.A., Wilson, M., Klise, K.A., 2008. Detecting changes in water quality data. J. Am. Water Works Assoc. 100 (1), 74–85. http://www.awwa.org/publications/journal-awwa/abstract/articleid/15801.aspx.

Moshtagh, N., 2005. Minimum Volume Enclosing Ellipsoids (accessed 05.02.14.). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.7691&rep=rep1&type=pdf.

Murray, R., Haxton, T., McKenna, S.A., Hart, D.B., Klise, K.A., Koch, M., Vugrin, E.D., Martin, S., Wilson, M., Cruze, V.A., Cutler, L., 2010. Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems: Development Testing and Application of CANARY. U.S. Environmental Protections Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, Ohio, USA. EPA/600/R-10/036. http://oaspub.epa.gov/eims/eimscomm.getfile?p_download_id=496189 (accessed 05.02.14.).

Nguyen, K.A., Stewart, R.A., Zhang, H., 2013. An intelligent pattern recognition model to automate the categorisation of residential water end-use events.

Environ. Model. Softw. 47, 108–127. http://dx.doi.org/10.1016/j.envsoft.2013.05.002.

Oliker, N., Ostfeld, A., 2013. Classification – optimization model for contamination event detection in water distribution systems. World Environ. Water Resour. Congr. 2013, 626–636. http://dx.doi.org/10.1061/9780784412947.061.

Oliker, N., Ostfeld, A., 2014. A coupled classification-evolutionary optimization model for contamination event detection in water distribution systems. Water Res. 51 (15), 234–245. http://dx.doi.org/10.1016/j.watres.2013.10.060.

Perelman, L., Arad, J., Housh, M., Ostfeld, A., 2012. Event detection in water distribution systems from multivariate water quality time series. Environ. Sci. Technol. 46 (15), 8212–8219. http://dx.doi.org/10.1021/es3014024.

Perelman, L., Ostfeld, A., 2013. Operation of remote mobile sensors for security of drinking water distribution systems. Water Res. 47 (13), 4217–4226. http://dx.doi.org/10.1016/j.watres.2013.04.048.

Pino-Mejıas, R., Cubiles-de-la-Vega, M.D., Anaya-Romero, M., Pascual-Acosta, A., Jordan-Lopez, A., Bellinfante-Crocci, N., 2010. Predicting the potential habitat of oaks with data mining models and the R system. Environ. Model. Softw. 25, 826–836. http://dx.doi.org/10.1016/j.envsoft.2010.01.004.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. Math. Stat. Appl. B, 283–297.

Rusen, J.S., Bartrand, T., 2013. Using online water quality data to detect events in a distribution system. J. Am. Water Works Assoc. 105 (7), 22–26.

Szabo, J.G., Hall, J.S., Meiners, G., 2008. Sensor response to contamination in Chloraminated drinking water. J. Am. Water Works Assoc. 100 (4), 33–40.

Suresh, M.A., Stoleru, R., Zechman, E.M., Shihada, B., 2013. On event detection and localization in acyclic flow networks. IEEE Trans. Syst. Man. Cybern. Syst. 43 (3), 708–723. http://dx.doi.org/10.1109/TSMCA.2012.2210411.

Todd, M.J., Yıldırım, E.A., 2007. On Khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids. Discrete Appl. Math. ISSN: 0166-218X 155 (13), 1731–1744. http://dx.doi.org/10.1016/j.dam.2007.02.013.

Uber, J.G., Murray, R., Magnuson, M., Umberg, K., 2007. Evaluating Real-time Event Detection Algorithms using Synthetic Data. Paper presented at the Restoring our Natural Habitat – Proceedings of the 2007 World Environmental and Water Resources Congress. http://ascelibrary.org/doi/abs/10.1061/40927%28243%29499.

Wang, F., Chen, D.S., Cheng, S.Y., Li, J.B., Li, M.J., Ren, Z.H., 2010. Identification of regional atmospheric PM10 transport pathways using HYSPLIT, MM5 CMAQ and synoptic pressure pattern analysis. Environ. Model. Softw. 25, 927–934. http://dx.doi.org/10.1016/j.envsoft.2010.02.004.