Comparison of two multivariate classification models for contamination event detection in water quality time series

Nurit Oliker and Avi Ostfeld

ABSTRACT

This paper explores two applied classification models alerting for contamination events in water distribution systems. The models perform multivariate analysis of water quality online measurements for event detection. The developed models comprise an outlier detection algorithm and a following sequence analysis for the classification of events. The first model is an unsupervised minimum volume ellipsoid (MVE), which utilizes only normal operation measurements but requires calibration. The second is a supervised weighted support vector machine, which utilizes event examples and performs data-driven optimized calibration. The models were trained and tested on real water utility data with randomly simulated events that were superimposed on the original database. The models showed high accuracy and detection ability compared to previous studies. All in all, the MVE model achieved preferable results.

Key words | event detection, minimum volume ellipsoid, sequence analysis, support vector machine, water distribution systems, water security

Nurit Oliker
Avi Ostfeld (corresponding author)
Faculty of Civil and Environmental Engineering,
Technion – Israel Institute of Technology,
Haifa 32000,
Israel
E-mail: *ostfeld@tx.technion.ac.il*

INTRODUCTION

Securing drinking water is one of the current central issues in the field of water planning and management. Water distribution systems (WDS) are particularly vulnerable as they comprise numerous exposed elements which are prone to contamination events. In recent years, many resources have been invested, both in academia and industry, in the development of monitoring and alert systems. The latest approach features the use of typically monitored water quality parameters, such as turbidity, electrical conductivity, pH, and chlorine concentration, for the detection of exceptional behaviors in networks (Hall *et al.* 2007). The premise of this approach is that abnormal behavior of those parameters may imply an occurrence of a contamination event. Therefore it was established that information from online water quality sensors may provide an early indication of a pollutant presence in the network. The challenge is then to distinguish between normal behavior of the measured parameters, and changes triggered by contaminants' intrusion.

Guepie *et al.* (2012) developed an event detection model based on residual chlorine decay measurements. Their premise was that a contaminant intruding the system will consume a considerable amount of chlorine. Thus a significant decrease in chlorine concentration may indicate the presence of a contaminant. Murray *et al.* (2010), Perelman *et al.* (2012), and Arad *et al.* (2013) developed contamination event detection models based on utilizing multivariate water quality measurements. Perelman *et al.* (2012) and Arad *et al.* (2013) applied a neural network prediction model, detecting deviations from the expected behavior and classifying outlier measurements. The sequence of normal/outlier observations was translated to event probability using Bayes' rule. Murray *et al.* (2010) applied several outlier detection algorithms: a linear filter, a multivariate nearest-neighbor algorithm, and a set-point proximity algorithm, followed by a statistical tool that calculated event probability.

doi: 10.2166/aqua.2014.033

The understanding of the relationships between the parameters and their response to different events is still vague. There are no recorded measurements of a real time contamination event. In order to represent contaminants' effect on the 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 randomly simulated events were superimposed on the original data (Murray *et al.* 2010; Perelman *et al.* 2012; Arad *et al.* 2013).

Perelman *et al.* (2012) and Arad *et al.* (2013) applied supervised classification methods, utilizing the simulated events for both the construction of the classifier and its assessment. Murray *et al.* (2010) applied unsupervised classification methods, requiring the use of simulated events only for the assessment of the model performances.

Support vector machine (SVM) was introduced by Boser *et al.* (1992) for the classification of multivariate data. The classifier utilizes a training data set which includes samples of two known classes, to create a hyperplane which separates the space into two domains. The objectives that moderate the hyperplane location are: maximize the distances between the hyperplane and the vectors, on the one hand; and minimizing the error of misclassified vectors, on the other. Naturally, this is a trade off, since the more errors enabled, the larger the hyperplane margin can be.

The SVM problem can be defined as

$$\begin{aligned} &\underset{W,b,Si\geq 0}{\text{Minimize}}\left[\frac{1}{2}\|W\|^2+C\sum_{i=1}^{n}S_i\right] \\ &\text{Subject to: } y_i(W^{\mathrm{T}}x_i+b)\geq 1-S_i \quad i=1,\ \ldots,n \end{aligned} \tag{1}$$

where W is the normal vector to the hyperplane, C is the classifier parameter, S_i are the slack variables, n is the number of vectors in the training data set, y_i is either 1 or -1 indicating the class to which the vector x_i belongs, and b is a coefficient which determines the axis intercepts.

Minimum volume ellipsoid (MVE) was introduced by Rousseeuw (1985) for the detection of outliers in multidimensional data. This is a classification method based on finding the minimal closed quadric surface which contains some group of vectors. Usually, the ellipsoid is required to include some set fraction out of the samples, where the fraction corresponds to the certainty level of the measurements (i.e., if the data are more reliable the ellipsoid is required to include a larger fraction of the samples).

The minimal ellipsoid problem can be formulated by

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

where Pi is a measured vector (required to be bounded by the ellipsoid), c is the ellipsoid center coordinates vector, and A is the matrix of coefficients in the ellipse equations.

After the ellipsoid is found, any new observation is classified as normal if located inside the ellipsoid, or outlier if located outside of it.

This paper aims to explore and compare two developed models for event detection in WDS. The first is a supervised weighted SVM model (Oliker & Ostfeld 2014a) and the second is an unsupervised MVE model (Oliker & Ostfeld 2014b). This paper is an extension of Oliker & Ostfeld (2014c) that reviewed and compared the SVM model and a previous version of the MVE model. This paper aims to elaborate the comparison of the two models conducted in Oliker & Ostfeld (2014c) to previous studies and further analyze the results.

METHODOLOGY

The general scheme of the two models is presented in Figure 1. The models comprise an identical preliminary procedure of data cleansing, an outlier detection algorithm (SVM or MVE), and a following sequence analysis, which exploits the outlier detection binary output for the classification of events. Both models are updated continuously and utilize a constantly growing database. The working assumption of the study was that after 24 hours, the true nature of the measurements is clarified. Thus, in a 24-hour delay any measured data are added to the training data set. The algorithms are briefly described below.

Data cleansing

Almost all measured data include some measurement noise. This noise creates a bias which is likely to affect the classifier

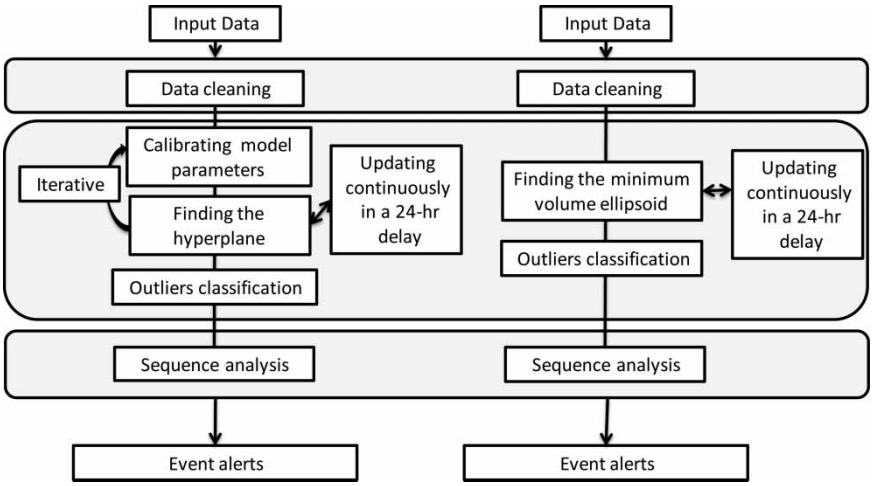

Figure 1 | Scheme of the two models.

performance. Therefore, the data are filtered before being analyzed. The two models include a very simple data cleaning, which consists of removing non-positive values, and values which exceed some standard deviation away from the mean. Negative values, or zeros, are physically impossible when referring to water quality parameters, and surely originate from measurement error. Thus, all non-positive measurements are removed from the database. Values that exceed 4 standard deviations away from the mean are also removed as they feature very exceptional samples unlikely to represent true measurements. Of course, the data cleaning is conducted for each of the given water quality parameters independently. Note that the exclusion threshold (i.e., 4 standard deviations) should be re-evaluated for each applied database in order to make sure only a slight fraction of the suspiciously exceptional data is removed.

The SVM model

This two-step classification model includes a weighted SVM and a following sequence analysis. The SVM is a supervised model, using known event examples for the construction of the classifier. Furthermore, the model utilizes the known data set for the calibration of its parameters. Namely, the model is iteratively constructed with different parameters and evaluates their suitability.

The classic SVM gives equal weight to all samples. The applied weighted SVM aimed at blurring differences between the two class data set sizes, and dealing with the time factor attribute. The data size differences are blurred by a weight inversely correlated to the class size. That way the minor class vectors get higher importance in the classifier construction. A time decay factor gives higher weight to the more recent observations. That way, the classifier gives higher importance to recent observations, yet exploits the existing database fully.

The weight is given by

$$Mi(j, k, \Delta t) = \mathrm{e}^{-0.0001\Delta t} + \frac{\mathrm{size}(k)}{\mathrm{size}(j)} \tag{3}$$

where M_i is the weight given to vector i, j is the class to which vector i belongs, k is the opposing class, and Δt is the number of time steps between the measurement vector and the current time. The first part of the expression is the time decay factor, and the second is the class size blurring.

Including the weights vector M_i in the model expends the formulation of the objective function given in Equation (1) to be

$$\underset{W,b,Si\geq 0}{\mathrm{Minimize}} \left[\frac{1}{2}\|W\|^2 + \mathrm{C}\sum_{i=1}^{n} M_i S_i\right] \tag{4}$$

After the hyperplane is found, the space is divided to normal and outlier regions. Any incoming vector is classified as normal or outlier, depending on its location.

A sequence of outliers is clearly much stronger evidence of an event occurrence than a single one, thus the classification of events is performed by a sequence analysis which exploits the binary output of the SVM.

The sequence analysis is performed by calculating a probability measure composed of three elements: outliers proportion, outlier continuity, and history based probability. For the classification of each time step a sequence ending in that time step is used to calculate the probability measure. The length of the analyzed sequence is determined by data-driven optimization. The measure is given by:

$$\text{Probability measure} = \alpha_1 \times \text{proportion} + \alpha_2 \times \text{continuity} + \alpha_3 \times \text{history} \quad (5)$$

where the 'proportion' element is the fraction of outliers within the sequence. The 'continuity' is the longest series of outliers within the analyzed sequence, divided by the sequence length. The 'history' element is based on the appearances of this exact sequence in the training data set, which equals the number of appearances while an event occurrence is divided by the total number of appearances.

The elements coefficients are sums into unity and determined by data-driven optimization. The 'proportion' element expresses how much it is exceptional relative to normal operation time, the 'continuity' element represents the reliability of the outlier's indications, and the 'history' is the known experience of the sequence appearances.

If the probability measure exceeds a threshold value, the time step is classified as an event, otherwise it is classified as normal operation. The threshold value is also determined by an optimization. The model which does not require operator interference and is automatically calibrated is described in detail in Oliker & Ostfeld (2014a).

The MVE model

The general structure of the model (shown in Figure 1) comprises a MVE classifier for the detection of outliers and a following sequence analysis for the classification of events.

The MVE construction includes finding the minimal ellipsoid that contains 95% of the training data set vectors. The 95% value was set by trial and error. The dimension of the ellipsoid corresponds to the number of measured water quality parameters. After the ellipsoid is found, vectors are classified as normal if lying inside the ellipsoid or outliers if lying outside of it.

The classification of each time step as normal or an event is determined according to sequence analysis of the 6-bit length binary sequence ending at the classified one. Similarly to the SVM model, the sequence analysis is based on the calculation of a probability measure. This measure differs from Equation (5) in the lack of the history element. The MVE does not utilize a known database so this element cannot be calculated. The measure is given by:

$$\text{Probability measure} = 0.75 \times \text{proportion} + 0.25 \times \text{continuity} \quad (6)$$

where the 'proportion' and 'continuity' elements are identical to those described in Equation (5).

The model parameters: sequence length, measure weights (i.e., the 0.75 and 0.25 values), and alert threshold values were all determined by trial and error.

Application

The models were applied on a real database that was attained by a utility in the United States and available from CANARY (2013). The data (shown in Figure 2) include online water quality measurements taken every 5 minutes during 4 weeks (approximately 8,000 time steps). All measurements were taken in normal operating conditions. The data include the following six water quality parameters: free chlorine, electrical conductivity, pH, temperature, total organic carbon (TOC), and turbidity.

To train the SVM model and assess both models, events were simulated and superimposed on the data set. The events were characterized by their magnitude, direction, and effect duration. Those properties were determined by a random with a uniform distribution selection from a set range of values. The shape of the superimposed events is the Beta distribution function shape (Keepings 2010), where its two parameters (the multiplier and exponent) were set with the value 2. The average frequency of events was from once to twice a day, and the occurrence timing had no restriction (enabling long normal operation times

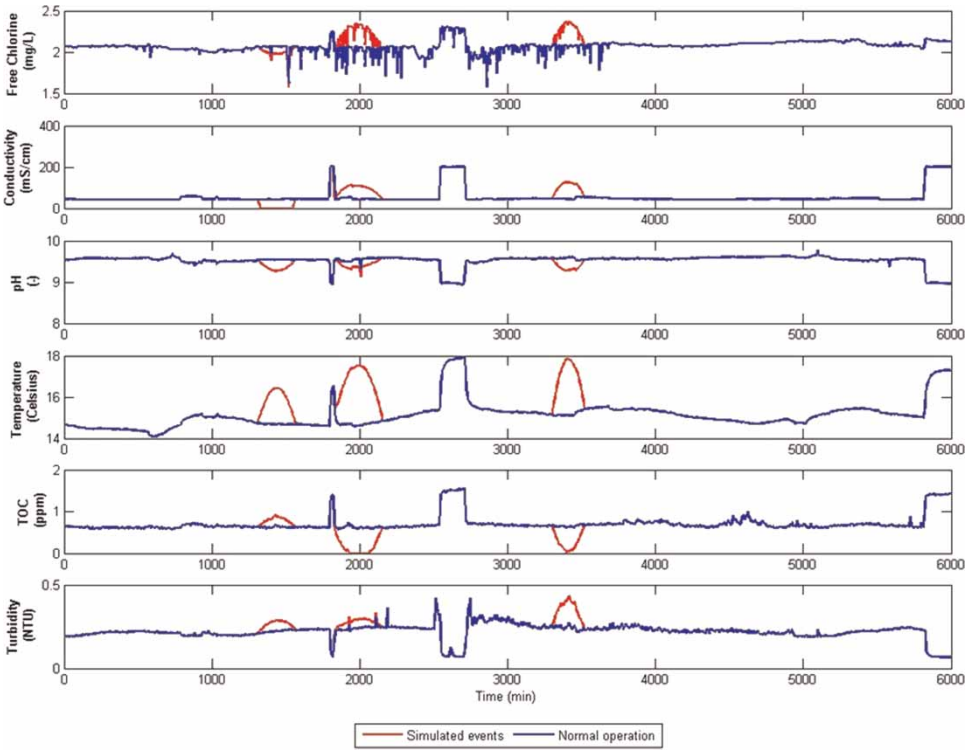

Figure 2 | A segment of the normal operation database and superimposed simulated events.

together with overlapping events). The direction of the deviations was also set randomly, positively or negatively superimposed on the data. Both magnitude and direction were set independently for each parameter.

Three types of scenarios were simulated, producing events with different intensities. The types differ in the events' duration and magnitude. 'High' included events with a duration of 4–6 hours, and disturbance magnitudes of 1–2.5 standard deviations. 'Medium' included events of 3–6 hours, with 0.5–2 magnitudes of standard deviations. 'Low' included events of 2–4 hours, with magnitudes of 0–1 standard deviations.

A segment of the used database, with an example of superimposed 'Medium' type simulated events is shown in Figure 2. The data segment includes 6,000 minutes of measurements including six parameters: free chlorine (mg/l), electrical conductivity (mS/cm), pH(-), temperature (Celsius), TOC (ppm), and turbidity (NTU).

The database was conventionally divided into training and testing data sets. The training data set included 70% of the data and was used to construct the classifier. The remaining 30% were left untouched in order to simulate real time operation and enable model testing.

In the SVM model the training data set was sub-divided into two sub-sets, 70% for training and 30% for validation. That is to say, the whole database was divided into 50% training, 20% validation, and 30% testing. This division enabled the calibration of the model, as the training data set was used to train the model with different parameters, and the validation data set was used to evaluate their suitability.

RESULTS

The models were evaluated according to their performance on the testing data set. Their assessment was done using two measures described by

$$\text{Accuracy} = \frac{\text{Well-classified vectors}}{\text{Total vectors number}}$$
$$\text{Detection ratio} = \frac{\text{Detected events}}{\text{Total events number}} \quad (7)$$

where the 'accuracy' measure indicates the model reliability and the 'detection ratio' measure presents its sensitivity.

The averaged results of 45 comparison runs are shown in Figure 3. The runs include three types of event intensity, with 15 scenarios of each type (i.e., Low, Medium, and High). The MVE model showed clear superiority, reflected in higher 'detection ratio' and 'accuracy' values for all types of events.

In order to evaluate the developed models it was necessary to compare their performance to other models when tested on identical event scenarios. The models were compared to Arad *et al.* (2013) and CANARY (2013). Arad *et al.* (2013) include independent outliers' detection element for each water quality parameter, based on an artificial neural network algorithm. This model has two optional decision rules for triggering an event alert: decision rule 1 warrants at least three outliers out of the six water quality parameters, and decision rule 2 requires at least five. CANARY (2013) is the best known model in the field featuring outlier detection by a combination of a few algorithms: a linear filter, a multivariate nearest-neighbor algorithm, and a set-point proximity algorithm.

Tables 1 and 2, respectively, present the averaged 'detection ratio' and 'accuracy' results of the presented SVM and MVE models, Arad *et al.* (2013) two decision rules and CANARY (2013), for 15 identical event scenarios with low, medium, and high intensities (i.e., five events of each type). The CANARY (2013) software requires the operator selection of model parameters, and therefore the comparison included five sets of parameters: the default parameters, two changes

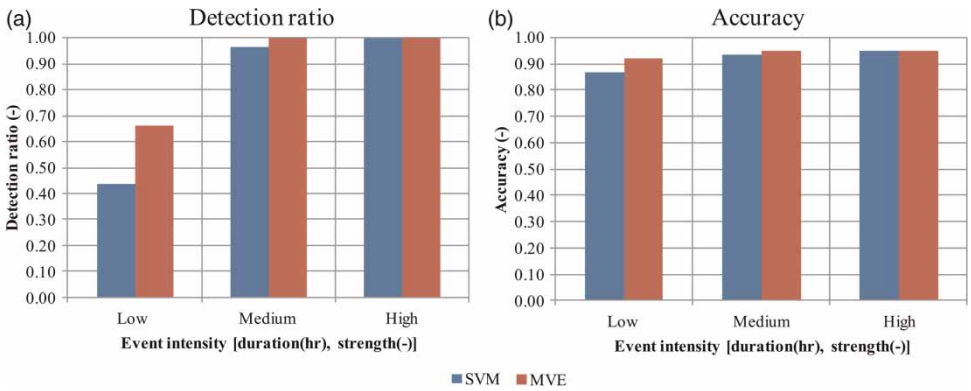

Figure 3 | Comparison of the MVE and SVM 'detection ratio' (a) and 'accuracy' (b). The values are an average of 15 runs for each event type.

Table 1 | Comparison of all models averaged detection ratio for events with low, medium, and high intensity (five runs for each type)

					Event scenario		
Model					**Low**	**Medium**	**High**
MVE					0.66	1	1
SVM					0.54	0.95	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							
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 2 | Comparison of all models averaged accuracy for events with low, medium, and high intensity (five runs for each type)

Model					Event scenario		
					Low	Medium	High
MVE					0.92	0.95	0.94
SVM					0.87	0.93	0.95
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							
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

of the event threshold, and two changes of the precision values, which determine the outlier detection sensitivity. The bolded values in the CANARY (2013) parameter set are the changes conducted in the default parameters for each set of runs. The CANARY (2013) default parameters produced an over-sensitive classifier. Thus, the precision values of the free chlorine, conductivity, and pH were enlarged. However, the results have not shown any dramatic improvement.

All in all, the MVE showed the best results with preferable event detection ability and classification accuracy. The 'detection ratio' of the MVE was the best for all types of events, whereas for the high intensity events, the same perfect detection ratio of 1 (i.e., detecting all events) was achieved by the SVM model as well. The SVM model showed excellent detection ability for the medium and high events, but the worst results for the low event type. The MVE achieved the highest 'accuracy' with an average of 0.94, compared to 0.92, 0.89, and 0.92 of the SVM and Arad *et al*. (2013) first and second decision rules, respectively. The CANARY (2013) showed a significantly lower accuracy with an overall average of 0.76.

DISCUSSION AND CONCLUSIONS

This paper presents the comparison of two developed classification models for event detection in WDS. The two models perform multivariate analysis of the data, explore the relations between parameters, and detect abnormal behavior in their mutual patterns. The models showed improved performance relative to previous works.

The simulated events applied in this field are completely generic, assuming an event causes some unknown disturbances to the measurement. The use of an unsupervised classification method seems to provide a fundamental advantage as it reduces the need of any assumptions regarding the event's influence. The unsupervised method utilizes only the real normal operations measurements, when trained to recognize any abnormal behavior compared to the normal data set.

The unsupervised method comes at a price as it precludes model automatic calibration. The absence of known event examples prevents the model from verifying testing and perform self-tuning of its parameters. Thus, all parameters of the MVE model were set by trial and error. In contrast, the SVM parameters were set autonomously by data-driven optimization, reducing the need for any interference of the operator.

All in all, the MVE model showed superiority in all aspects featuring the highest detection ability and accuracy. Despite the complexity of the SVM model, its performance was inferior. A possible cause is over-fitting of the model and the validation data set, as the SVM parameters were selected according to their performance for the validation data set.

The SVM is a linear classifier that features a separating surface of a hyperplane. In cases where the database is not linearly separable a kernel function is required for the transformation of the data. Conversely, the MVE is a non-linear classifier, featuring a quadratic separating surface. As the database seemed to have a roughly Gaussian distribution in most parameters, it seems the MVE quadratic surface encompass it better. Furthermore, the ellipsoid in the MVE model is required to include 95% of the given data set and thus is highly associated with the data set location in space. Conversely, the hyperplane of the SVM model has no restriction regarding its location relative to the data set, as the number of support vectors is unlimited but only minimized by the objective function shown in Equation (1). Consequently, in some cases, the hyperplane may be located away from most of the data set and perform badly. The relation of the ellipsoid to the data set might contribute to its relative stability.

All models were tested on a database of a single water utility. However, the simulated events include various scenarios with randomly determined properties. The models apply generic classification methods and the conclusions are expected to fit other WDS applications.

Similar to previous studies in the field (Murray *et al.* 2010; Perelman *et al.* 2012; Arad *et al.* 2013), neither the network topology nor the hydraulic model were taken into account in the presented models. In fact, all models to date feature the application of generic classification methods with no consideration of the hydraulic understanding of the system in the decision process of event detection. The models were calibrated on the water quality data but could actually be adjusted to any kind of time series data.

All mentioned event detection models were developed to analyze data collected in a single location in the network, applied on measurements taken in a single node. Further research aims to applying spatial analysis, extending the model for the analysis of several sensors spread in the network, while including network topology and hydraulic model in the event detection model.

ACKNOWLEDGEMENTS

This study was supported by the joint Israeli Office of the Chief Scientist (OCS) Ministry of Industry, Trade and Labor (MOITAL), and by the Germany Federal Ministry of Education and Research (BMBF), under project number GR_2443. Funds were also received from the Technion Grand Water Research Institute, and from the Technion Funds for Security Research.

REFERENCES

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.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. 1992 A training algorithm for optimal margin classifiers. In: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (D. Haussler, ed.). July 27–29, Pittsburgh, PA, pp. 144–152.

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

Guepie, B. K., Fillatre, L. & Nikiforov, I. 2012 Sequential Monitoring of Water Distribution Network. *Paper presented at the IFAC Proceedings (IFAC-Papers Online)*, Vol. 16, Part 1, pp. 392–397.

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

Keepings, E. S. 2010 *Introduction to Statistical Inference*. Dover Publications, New York.

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*. EPA/600/R-10/036, US Environmental Protections Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, OH, USA. http://cfpub.epa.gov/si/si_public_record_report.cfm?address=nhsrc/&dirEntryId=221394.

Oliker, N. & Ostfeld, A. 2014a A coupled classification – evolutionary optimization model for contamination event detection in water distribution systems. *Water Res.* **51**, 234–245.

Oliker, N. & Ostfeld, A. 2014b Minimum volume ellipsoid classification model for contamination event detection in water distribution systems. *J. Environ. Modell. Softw.* **57**, 1–12.

Oliker, N. & Ostfeld, A. 2014c Comparison of multivariate classification methods for contamination event detection in water distribution systems. In: *Proceedings of the 12th International Conference on Computing and Control for the Water Industry – CCWI2013*. Procedia Engineering, Elsevier, Perugia, 70, pp. 1271–1279.

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.

Rousseeuw, P. J. 1985 Multivariate estimation with high breakdown point. In: *Mathematical Statistics and Applications* (W. Grossmann, G. Pflug, I. Vincze & W. Wertz, eds). Reidel Publishing Company, Dordrecht, The Netherlands, pp. 283–297.

Uber, J. G., Murray, R., Magnuson, M. & Umberg, K. 2007 Evaluating real-time event detection algorithms using synthetic data. In: *Paper presented at the Restoring our Natural Habitat – Proceedings of the 2007 World Environmental and Water Resources Congress*, May 15–19, Tampa, FL, USA.

First received 20 March 2014; accepted in revised form 5 June 2014. Available online 24 July 2014