# Contaminant classification using cosine distance based on multiple conventional sensors

Shuming Liu[1*], Han Che[1], Kate Smith[1], Tian Chang[1]

[1]School of Environment, Tsinghua University, Beijing, 100084, China

[*]corresponding author: shumingliu@tsinghua.edu.cn, +86 1062787964

**Abstract**

Emergent contamination events have a significant impact on water systems. After contamination detection, it is important to classify the type of contaminant quickly to provide support for remediation attempts. Conventional methods generally either rely on laboratory-based analysis, which requires long analysis time, or on multivariable-based geometry analysis and sequence analysis, which is prone to being affected by contaminant concentration. This paper proposes a new contaminant classification method, which discriminates contaminants in a real time manner independent of contaminant concentration. The proposed method quantifies the similarities or dissimilarities between sensors' responses to different types of contaminants. The performance of the proposed method was evaluated using data from contaminant injection experiments in a laboratory and compared with a Euclidean distance-based method. The robustness of the proposed method was evaluated using an uncertainty analysis. Results show that the proposed method performed better in identifying the type of contaminant than the Euclidean distance

based method and that it could classify the type of contaminant in minutes without significantly compromising the correct classification rate (*CCR*).

**Keywords**

contaminant classification, conventional sensor, cosine distance, early warning system, water quality

## Introduction

Water systems are vulnerable to contamination accidents[1-2]. For example, in April 2014, crude oil leaked from a petrochemical pipeline in Lanzhou, China, contaminating the water source of a local water plant and introducing hazardous levels of benzene into the city's tap water. Water supply to Lanzhou city was suspended as a result. An intense effort is currently underway to improve analytical monitoring and detection of biological, chemical, and radiological contaminants in water systems. One approach for avoiding or mitigating the impact of contamination is to establish an Early Warning System (EWS). EWS should provide a fast and accurate means of distinguishing between normal variations and contamination events, and should be able to classify the type of contaminant[3].

After an EWS detects the presence of contamination, the next important issue is to classify the type of contaminant. The most commonly used method for contaminant classification is laboratory-based analysis, e.g. ICP-MS. The advantage of this type of analysis is that it can accurately qualify and quantify the contaminant. The disadvantage is that it is time-consuming. In the event of an emergent contamination event, the key to all remediation attempts is time. Therefore, methods of fast classification of contaminants are in great demand. One possible solution is online compound-specific sensors, which need less time than laboratory-based methods[4-8]. However, compound specific sensors can normally only identify one type or a small group of contaminants. In this case, low efficiency or failure in contaminant classification can be expected.

To overcome this drawback, several researchers have attempted to develop real-time contaminant classification methods. Kroll[9] reported the Hach HST approach using multiple types of sensors for event detection and contaminant classification. In the Hach HST approach, signals from 5 separate orthogonal measurements of water quality (pH, conductivity, turbidity, chlorine residual, TOC) were processed from a 5-paramater measure into a single scalar trigger signal. The deviation signal was compared to a preset threshold level. If the signal exceeded the threshold, the trigger was activated[9]. The deviation vector was then used for further classification of the cause of the contamination. The direction of the deviation vector relates to the agent's characteristics. Seeing that this is the case, laboratory agent data can be used to build a threat agent library of deviation vectors. A deviation vector from the monitor can be compared to agent vectors in the threat agent library to see if there is a match within a given tolerance level. This system can be used to classify what caused the trigger event. Yang et al.[10] reported a real-time event adaptive detection, classification and warning (READiw) method for event detection and contamination classification. In this method, four discrimination systems were developed to differentiate the 11 tested contaminants according to the various responses of sensors. The classification process was more based on geometry analysis. The similarity or dissimilarity between examples and classes were not quantitatively evaluated. Oliker and Osfield[11] developed a contamination event detection method for water distribution systems, which comprised a weighted support vector machine for the detection of outliers, and subsequent sequence analysis for the classification of contamination events. It was noticed that either geometry analysis or sequence analysis was prone to being affected by the magnitude of sensor responses, which were normally related to contaminant concentrations. This could then lead to misclassification.

Although effort has been put into developing methods for contaminant classification in recent years, more attention is necessary. Therefore, the objectives of this study are 1) to develop a classification method which is independent of contaminant concentration; 2) to compare the performance of the proposed method with a Euclidean distance-based method.

## Materials and methods

### Data collection

In order to collect contamination data, a pilot-scale contaminant injection experiment (CIE) platform was developed. A process flow schematic of the CIE platform is shown in Figure 1. The water tank is approximately 85 cm high with a diameter of 70 cm, and has a total capacity of 300 L. The tank is linked with online water quality sensors via a peristaltic pump at 0.5 L per minute. Eight types of sensors developed by Hach Homeland Security Technologies were utilized in this study. They can measure the following 8 parameters simultaneously and continuously: temperature, pH, turbidity, conductivity, oxidation reduction potential (ORP), UV-254, nitrate-nitrogen and phosphate. The CIE platform was operated in recirculation mode for baseline establishment. Generally, the process of establishing baseline takes 4-6 hours before any contaminant experiments can be carried out. When operating in single-pass contaminant mode, the contaminant is injected into the pipe connecting the tank and sensors via another peristaltic pump. It is injected at a rate of 2-20 mL per minute depending on concentration requirement. For more information about the CIE platform and the injection experiment, the readers could refer to Liu et al.[12].

(Figure 1)

### Contaminants investigated

Specific quantities of various contaminants were injected into the system simulator. The contaminants investigated were determined according to statistical reports on water pollution incidents in urban water supply systems in China over the past 20 years. Three groups of the most common six pollutants were selected: atrazine, glyphosate, cadmium nitrate, nickel nitrate, sodium fluoride and sodium nitrate. They were also selected based on China's national standards regarding source water quality GB3838-2002 and drinking water quality GB5749-2006. The concentration ranges of tested contaminants are provided in the supplementary material (Table T1) and were decided using the concentration limit given in the above national standards.

**Classification method**

Clustering or cluster analysis is the process of grouping a set of objects into classes of similar objects. Objects in any one cluster share similar features. Although definitions of similarity vary from one clustering model to another, in most of these models the concept of similarity is based on distances, e.g., Euclidean distance and cosine distance[13-15].

(Figure 2)

In cluster analysis, similar objects are assumed to have close values. If the distance of an *object* to a particular *class* is shorter than the distances to other classes, the *object* is deemed as belonging to that *class* (Figure 2). In this way, cluster analysis can be used to identify the type of contaminant. An *object* can be an *example* or *instance* of the *class*. In this study, the term *instance* refers to the object in a pre-defined *class*, while *example* refers to the object to be classified. Both *instances* and *examples* are vectors consisting of *features*. The *features* are extracted and derived from the sensor responses for contaminants.

Figure 3 shows the responses to cadmium nitrate and atrazine at time *t1* and *t2* for 8 types of sensors. If the sensor reading is taken as the *feature*, $p^{t1}$, $p^{t2}$, $q^{t1}$ and $q^{t2}$ are 8-dimensional vectors. As shown in Figure 3, the graphs for $p^{t1}$ and $p^{t2}$ are clearly similar to each other, while the graph for $q^{t1}$ is closer to the graph for $q^{t2}$. An essential task of this study is to quantify the similarity or dissimilarity between two vectors, which is then used for contaminant identification.

(Figure 3)

**Similarity measure**

There are several methods of measuring the similarity between two objects (i.e. two *l*-dimensional vectors). In this study, cosine similarity was adopted. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them[16-17]. The cosine of two vectors can be derived by using the Euclidean dot product formula.

$$p \cdot q = \|p\|\|q\|\cos\theta \tag{1}$$

Given two vectors of attributes, p and q, the cosine similarity, cos(θ), is represented using

$$similarity(p,q) = \cos(\theta) = \frac{p \cdot q}{\|p\|\|q\|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}} \tag{2}$$

in which n is the dimension of vector p and q.

This function gives a similarity measure in the sense that the cosine value gets larger as the two vectors become more parallel to each other in the *l*-dimensional space. Or, in other words, as the two data segments become more similar, their cosine similarity approaches 1.0 and their distance approaches 0.0. Therefore, cosine similarity can be used as a distance metric in the following way:

$$D(p,q) = 1 - similartity(p,\mathrm{q}) \tag{3}$$

Since the cosine similarity reflects the magnitude of the angle between two vectors in the *l*-dimensional space, it is a many-to-one function. Compared with the other distance measures, like Euclidean distance, the cosine similarity ignores the magnitude difference between the two vectors, i.e.

$$similarity(Ap,q) = \frac{\sum_{i=1}^{n} Ap_i q_i}{\sqrt{\sum_{i=1}^{n} (Ap_i)^2}\sqrt{\sum_{i=1}^{n} q_i^2}} = similarity(p,q) \tag{4}$$

Therefore, when the cosine distance is used for contaminant identification, the variation range of sensor data need not be predetermined.

**Contaminant Classification**

The distance from a point p to a *class* **c** is given by:

$$D(p,\mathrm{C}) = 1 - similartity(p,\mu_c) \tag{5}$$

in which, $\mathrm{D(p,C)}$ is the distance from a point to a *class* and $\mu_c$ is the mean of all *instances* in *class* C.

The type of contaminant is identified by comparing the distances from examples to *classes*. Assuming there are *n types* of contaminants, $\boldsymbol{C_1}$, $\boldsymbol{C_2}$,…… $\boldsymbol{C_n}$, (or n *classes*), each *class* contains many vectors (i.e. *instance* of *class*). For any example p to be identified, if there exists

$$D(p,\mathrm{C}_i) < D(p,\mathrm{C}_j), j = 1,2,\cdots n, i \neq j \quad (6)$$

then it is deemed that $p \in C_i$.

**Evaluation of classification performance**

The performance of the classification method is evaluated using the correct classification rate (*CCR*). *CCR* can be calculated by

$$CCR = \frac{CC}{CC+IC} \times 100\% \quad (7)$$

where *CC* refers to the correct classification of a contaminant, *IC* is the incorrect classification of a contaminant as another type of contaminant. A greater *CCR* means the method is more capable of contamination identification.

**Robustness of the proposed method**

The proposed method relies on the readings of online water quality sensors. Inevitably, fluctuations exist in online readings, which might come from equipment noise or ambient variation. An important issue for a contaminant classification method is how robust it is when dealing with fluctuations in readings. To evaluate the robustness of the proposed method, artificial uncertainties were added to the raw readings. It is assumed that the uncertainty obeys Gaussian distributions. The uncertainty quantification is achieved through a sampling-based method, Latin hypercube sampling (LHS) technique. In LHS[18], values of stochastic tested vectors are generated in a random, yet constrained way. First, the values of variables in the original tested vectors are taken as means (i.e. raw readings of sensors) and the standard deviation is equal to 1% of the mean value (i.e., coefficient of variation $C_v$=0.01, for example). The range of each vector variable can be calculated using a Gaussian distribution equation, which is then divided into *Ns* non-overlapping intervals on the basis of equal probability. After that, a single random value is selected from each interval. This process is repeated for all variables in a *feature* vector. Once that is done, the *Ns* values obtained for the first vector variable are paired in a random manner with *Ns* values obtained for the second vector variable and so on. *Ns feature* vectors are generated from the original *feature* vector. By repeating the same process, *feature* vectors and the associated uncertainty can be obtained for all time steps. The *CCRs* for *feature* vectors with uncertainty can then be obtained. Finally, the robustness is evaluated using equation 8.

$$robustness = \frac{CCR_{confidence\,\square\square\square\square\square}}{CCR_o} \qquad (8)$$

in which, $CCR_{confidence\,\square\square\square\square\square}$ is the 95% confidence limit of the *CCRs* with uncertainty and $CCR_o$ is the original *CCR*. For example, if the 95% confidence limit of the *CCRs* with uncertainty is 0.8 and the original *CCR* is 1, then the robustness value is 0.8/1=0.8. A higher robustness value means that the method is more robust.

## Experiments and Results

### Formation of *classes* of contaminants

In this study, *features* were extracted to facilitate the quantitative evaluation of similarity or dissimilarity between different types of contaminants. For all sensors in this study, the sensor responses obtained at each time step were adopted to form a *feature* vector (8 dimensions). For instance, the vector at the 1$^{st}$ minute for glyphosate was [1.32 7.06 757.67 10.76 276.96 3.16 9.42 0.08] with the vector sequence being turbidity, pH, conductivity, temperature, ORP, nitrate, UV and phosphate. Figure 4 shows the corresponding *feature* vectors at different concentrations. As shown in Figure 4, the extracted *features* share some similarity, but dissimilarity also exists. By extracting such data from all time steps, the *class* for glyphosate was established. The same procedure was repeated for the other contaminants examined in this study and a library containing 6 *classes* was obtained.

(Figure 4)

### Contaminant classification

Glyphosate and cadmium nitrate were chosen to demonstrate the performance of the contaminant classification method. The concentrations for glyphosate were 1.4mg/l, 2.8mg/l, 7.0mg/l and 14.0mg/l. For cadmium nitrate, the concentrations were

0.004mg/l, 0.008mg/l, 0.016mg/l and 0.032mg/l. A new group of contaminant injection experiments were conducted to produce data for contaminant classification. The raw experimental data contained sensor responses for baseline and presence of contaminant at 4 concentrations. As reported in Liu et al.[12], the contamination events were detected 1 minute after introduction of contaminants. The sensor response data after detection were separated from the raw data and used in the classification. They were treated using the procedure above to obtain *example feature* vectors. In total, there were 110 glyphosate and 200 cadmium nitrate *example* data to be tested.

(Figure 5)

For glyphosate, the cosine distances to all *classes* for each *example* (or 1 minute time step) was calculated using equation 4 and are shown in Figure 5. The green dots show the distance between *examples* and the glyphosate *class*. For all time steps (from 1 to 110), it can be noted that, although the concentration varies, the cosine distances from the examples to glyphosate *class* are rather stable and small. They are mostly in the range of [0 0.02]. The distances to the other *classes* are much greater. For example, the distance to chromium nitrate is around 0.16 (the blue circles in Figure 5). This is shown in Table 1, along with mean and standard deviation values of the distances. The proposed method classified the type of contaminant by comparing the cosine distance. The one with the closest distance is deemed to be the correct class. Table 1 and Figure 5 reveal that the examples are closer to the glyphosate *class*. On the basis of equation 6, the *feature* vectors of the example are more similar to the ones for glyphosate. Therefore, it can be concluded that the contaminant is glyphosate. Using equation 7, the *CCR* of the classification was calculated to be 0.918, which suggests that the tested contaminant is correctly classified in 91.8% of situations in this study.

(Figure 6)

For cadmium nitrate, Figure 6 shows the cosine distances to different classes, in which the red dots indicate the distances to the cadmium nitrate *class*. For all time steps, the distances to the cadmium nitrate class were in the range of 0.01 to 0.04 with the mean of 0.0277 (Table 1). It is obvious that the distances to cadmium nitrate are smaller than the ones to other classes in most cases in this study. The *CCR* was calculated to be 0.975.

(Table 1)

In terms of the time needed for classification, once a contamination event is detected by an EWS, the contaminant classification module will be activated. Theoretically, the type of contaminant can be classified within 1 minute (i.e. the sensor reporting step). However, in practice, the time might be a bit longer since the sensor responses to presence of contaminant might sometimes need to stabilize. As shown in Figure 5, the contaminant was classified correctly to be glyphosate 1 minute after the contamination event alarm. This means that the distance to the correct class was the smallest from the 1$^{st}$ minute onwards. For the case of cadmium nitrate, the proposed method can classify correctly 6 minutes after activation. In the first 5 minutes, the tested examples were incorrectly classified. The key strength of the proposed method is that it classifies the type of contaminant in a real time manner. Compared to laboratory-based methods, classification in 6 minutes with no significant compromise of *CCR* is an advantage.

## Discussion

### Comparison to Euclidean distance based method

In previous studies, Liu et al.[12] reported that the magnitudes of sensor responses vary with the concentration of contaminant (or see Figure F1, F2, F3, F4 and F5 in supplement documents). This is typically obvious for pH, nitrate, phosphate and ORP. For example, the pH and ORP values for the glyphosate concentration of 1.4, 2.8, 7.0, 14.0mg/l are 6.89, 6.71, 6.41, 6.10 and 277.66, 283.33, 291.67, 299.29 mV respectively. The aim of this study is to establish a method to classify the type of contaminant by evaluating the similarity between examples and classes. The classification method should be independent of or less related to the concentration of the contaminants since this is not known in advance in a real event. In other words, the distance evaluation method should not be too dependent of contaminant concentration. If the distance evaluation is closely related to magnitude of sensor response, the classification method might fail to differentiate events caused by the same type of contaminant with different concentrations.

There are several types of evaluation methods for the distance of vectors. The most commonly used one is the Euclidean distance, which is the "ordinary" distance between two points[19-20]. The Euclidean distance between points p and q is the length of the line segment connecting them, which can be calculated using

$$E(q,p) = \|q-p\| = \sqrt{(q-p)\cdot(q-p)} \quad (9)$$

in which $E(q,p)$ is the Euclidean distance between points p and q.

Figure 7 schematically shows the Euclidean distances and cosine distance between points p1, p2, q1 and q2. Points p1 and p2 are the sensor response vectors of contaminant 1 at concentrations 1 and 2. Points q1 and q2 are the vectors for contaminant 2 at concentrations 1 and 2. As shown in Figure 7, the Euclidean distance between p1 and p2 is $||p1-p2||$ and the cosine distance is 0. For p2 and q2, the Euclidean distance is $||p2-q2||$ and the cosine distance is 1- $\cos(\theta)$. Therefore, by using the cosine distance method, p1 and p2 (also q1 and p2) can be classified to the correct class. However, if the Euclidean distance were used, it might group p2 and q2 into the same class because $||p2-q2|| < ||p1-p2||$. To further explain this, the vectors associated with glyphosate and cadmium nitrate at different concentrations were taken as examples to calculate the Euclidean and cosine distances.

(Figure 7)

Table 2 shows the cosine and Euclidean distances between points *a, b, c* and *d*, in which *a* is the vector of sensor responses to glyphosate at concentration of 1.4mg/l, *b* is the vector for glyphosate at 14.0 mg/l, *c* is the vector for cadmium nitrate at concentration of 0.008 mg/l and *d* is the vector for cadmium nitrate at concentration of 0.032 mg/l. In Table 2, the numbers above the diagonal are cosine distances and the ones below are Euclidean distances. As shown in Table 2, the cosine distances for points from the same class are smaller than those for points from different classes. For example, D(a,b)=0.0027, while D(a,c)=0.1091. This explains the correctness of the assumption in Figure 2, i.e. similar objects have shorter distance.

It is also observed that the cosine distance is not 'sensitive' to the magnitude the vector (in other words, the concentration of the contaminants). As shown in Figure 4, the magnitude of sensor response vectors at 1.4 mg/l and 14mg/l is obviously different. However, their cosine distances to other points are close. For example, D(a,c)=0.1091, D(b,c)=0.1081. Euclidean distance, on the other hand, is related to the magnitude of the vector. For example, E(a,c)=92.3888, while E(b,c)= 158.4424. Furthermore, the case may arise where the Euclidean distance between points from the same class might be greater than that between points from two different classes. This is shown in Table 2. For example, E(a,b) = 95.5981 and E(a,c) = 92.3888. In this case, incorrect classification would occur if Euclidean distance were used for contaminant classification. Point c would be wrongly classified as being in the same class as point *a* if Euclidean distance was adopted. Therefore, it was concluded that cosine distance is more suitable than Euclidean distance for classifying the type of contaminant since the evaluation for similarity is more related to the contaminant's characteristics rather the magnitude of sensor responses.

(Table 2)

**Robustness**

The level of uncertainty is given by the value of *Cv*. In this study, four values of *Cv* (0.005, 0.01, 0.02 and 0.03) were used. The value of *Ns* is determined according to the literature. For a given *Cv*, by setting *Ns*=2000, 220000 *feature* vectors with uncertainty were finally generated for glyphosate. These *feature* vectors were divided into 2000 groups. Each contains 110 *feature* vectors. By feeding the 2000 groups of *feature* vectors into the proposed contaminant classification method, the *CCR*s for every group were obtained. The histograms of these *CCRs* are displayed in Figure 8, which shows that the proposed method has robustness of over 0.82 for uncertainty *Cv*=0.005, *Cv*=0.01 and *Cv*=0.02. This suggests that the performance of the proposed contaminant classification method is steady and reliable and can cope well with the uncertainty from the online sensors. For the case of *Cv*=0.03, the performance of the method is less satisfactory. The *CCR* for this case is 0.75, which is much lower than the original *CCR* (0.92). It should be noted that the uncertainty examined in this study is assumed to be from equipment noise or ambient variation. A change of sensor reading due to sudden sensor failure or presence of contaminant is not treated as noise, but instead as an event, which normally means a 1-20% change of sensor reading. Therefore, it is deemed that the uncertainty levels adopted in this study are significant enough.

It is worth noting that previous studies about online sensors in water supply systems generally assumed perfect sensors, which means that sensors worked in good condition[21-23]. Although this assumption has allowed researchers to make significant progress in early warning system design, sensor failures can significantly impact the reliability of an early warning system design. It is commonly known that sensor readings do contain uncertainties as they are easily affected by ambient variation and equipment condition. From an implementation perspective, it is essential that an early warning system using online sensor data is robust enough and can cope with uncertainty from sensors. The analysis here shows that the proposed method has good capacity of tolerate uncertainty in sensor readings.

(Figure 8)

(Table 3)

**Future studies**

This study proposed a concentration-independent contaminant classification method based on conventional water quality sensors. The basis of this method is that points in one class stay close and are separate from other classes. In spite of great improvement in recent years, readings from online sensors are still affected by noise and ambient variation. For a method based on online sensor readings, it is important to understand the impact of uncertainty from sensor readings on the model output. Although this study demonstrated the robustness of the proposed method in the event of sensor uncertainty or ambient variation through an initial uncertainty analysis, a global sensitivity analysis would be more helpful to understand the extent of uncertainty from each sensor. This should be conducted in future study.

Meanwhile, since the proposed method classifies by comparing the distances to predefined classes, incorrect classification error would occur if two (or more) classes overlap each other. This study involved a limited number of contaminants and no overlaps were noticed, but the possibility does exist. In a future study, this has to be addressed. A possible solution to this is that the classification decision could be made based on distances from more than one type of features. For example, if the features using original sensor responses from two types of contaminants overlap, another type of feature (e.g. the deviation between real readings and baseline) can be employed to differentiate these two classes.

## Conclusion

By using data from online water quality sensors, this study proposed a real time and concentration-independent contaminant classification method. From the analysis, the following conclusions were drawn.

1) The proposed method classifies the type of the contaminant by comparing their cosine distances to predefined classes. Results from the analysis show that the

proposed method can identify glyphosate and cadmium nitrate 1 and 6 minutes after detection with the *CCR* of 91.8% and 97.5%. Compared to laboratory-based methods, classification in minutes without significant compromising the *CCR* is an advantage.

2) Results show that the performance of the proposed method was not related to the contaminant concentration. This implies that the proposed method is more suitable than the Euclidean distance method for contaminant classification since the concentration of contaminant is not known a priori.

**Acknowledgements**

This work is jointly supported by Tsinghua Independent Research Program (2011Z01002) and Water Major Program (2012ZX07408-002).

**Reference**

1. USEPA, *Baseline threat information for vulnerability assessments of community water systems*, Washington, DC, 2002.

2. USEPA, *Planning for and responding to drinking water contamination threats and incidents*, Washington, DC, 2003.

3. M. V. Storey, B. van der Gaag and B. P. Burns, *Water Res*, 2011, 45, 741-747.

4. C. J. de Hoogh, A. J. Wagenvoort, F. Jonker, J. A. Van Leerdam and A. C. Hogenboom, *Environ. Sci. Technol.*, 2006, 40, 2678-2685.

5. J. Jeon, Kim, J. H. Lee, B.C., S. D. Kim, *Sci. Total Environ.*, 2008, 389, 545-556.

6. R. K. Henderson, A. Baker, K. R. Murphy, A. KHambly, R. M. Stuetz and S. J. Khan, *Water Res.*, 2009, 43 (4), 863-881.

7. P. R. Hawkins, S. Novic, P. Cox, Neilan, B. A. Burns, B. P. Shaw, G., W. Wickramasinghe, Y. Peerapornpisal, W. Ruangyuttikarn, T. Itayama, T. Saitou, M. Mizuochi and Y. Inamori, *J. Water Supply: Res. Technol.-AQUA*, 2005, 54, 509-518.

8. C. P. Marshall, S. Leuko, C. M. Coyle, M. R. Walter, B. P. Burns and B. A. Neilan, *Astrobiology*, 2007, 7, 631-643.

9. D. Kroll, *Securing our water supply: Protecting a vulnerable resource*, Pennwell, 2006.

10. J. Y. Yang, R. C. Haught and J. A. Goodrich, *J. Environ. Manage.*, 2009, 90, 2494-2506.

11. N. Aliker and A. Ostfeld, *Water Res.*, 2014, 51, 234-245.

12. S. M. Liu, H. Che, K. Smith and L. Chen, Environmental Science Processes & Impacts, 2014, 16, 2028-2038.

13. M. Tabacchi, C. Asensio, I. Pavon, M. Recuero, J. Mir, and M.C. Artal, Applied Acoustics,2013, 74, 1022-1032.

14. Z.P. Zhao, P. Li, and X.Z. Xu, Applied Mathematics & Information Sciences, 2013, 7, 1243-1250.

15. K.A. Nguyen, R.A. Stewart, and H. Zhang, Environmental Modeling & Software 47, 108-127.

16. A.A. Pascasio, Linear Algebra Appl, 2001, 325, 147-159.

17. M.W. Sohn, *Soc Networks*, 2001, 23, 141-165.

18. G. Manache, C.S. Melching, *J Water Res Plan Man*, 2004, 130, 232-242.

19. L. Liberti, C. Lavor, N. Maculan and A. Mucherino, *Siam Rev*, 2014, 56, 63-69.

20. C. Stoecker, S. Welter, J.H. Moltz, B. Lassen, J.M. Kuhnigk, S. Krass and H.O. Peitgen, *Med Phys*, 2013, 40, 091912.

21. M. Weickgenannt, Z. Kapelan, M. Blokker and D.A. Savic, *J Water Res Plan Man*, 2010, 136, 629-636.

22. A. Kessler, A. Ostfeld, G. Sinai, *J Water Res Plan Man*, 1998, 124, 192-198.

23. A. Ostfeld, E. Salomons, *J. Water. Res. Pl.-Asce*, 2004, 130(5), 377-385.

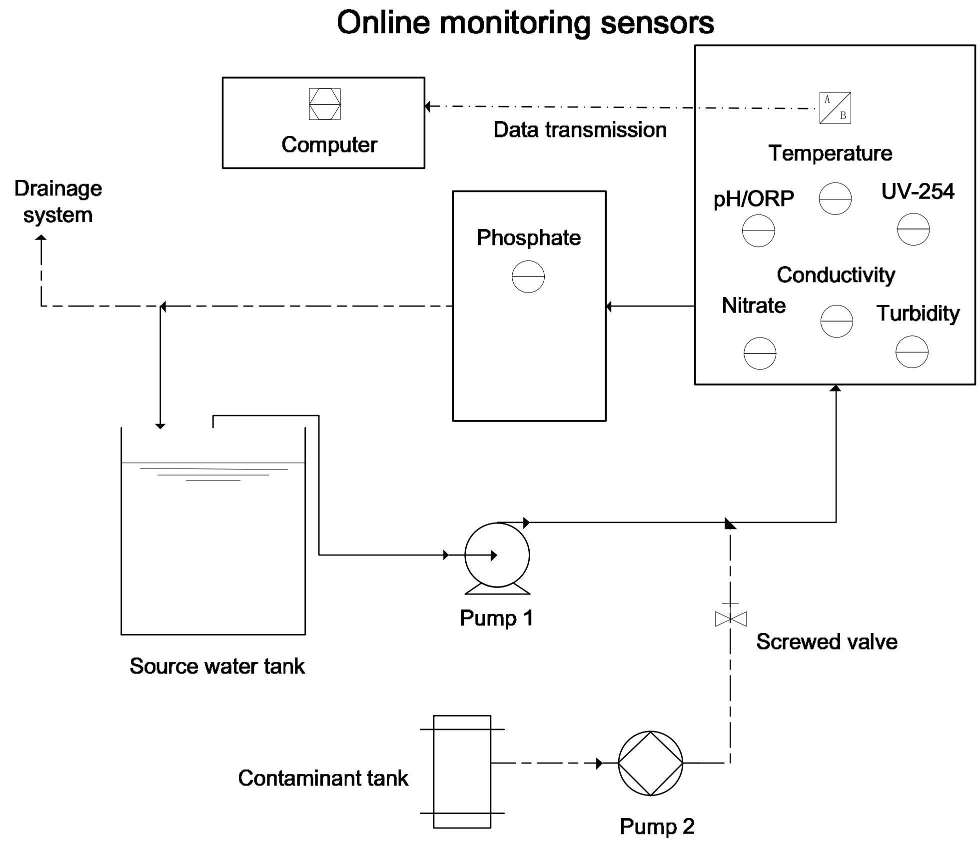


Figure 1 A process flow schematic of the pilot-scale system

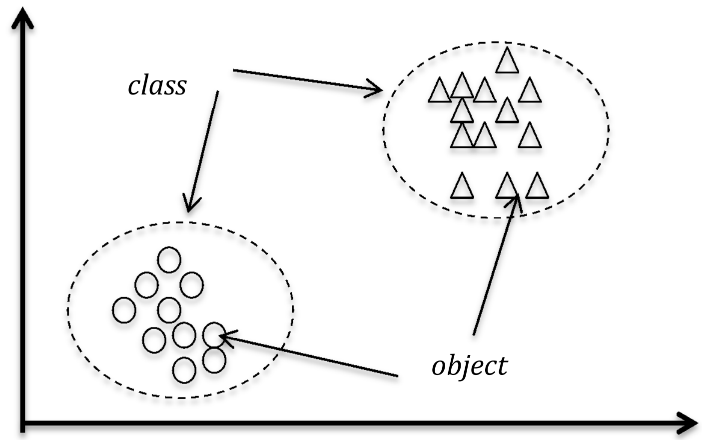


Figure 2 Schematic graphs of *class* and *instance*

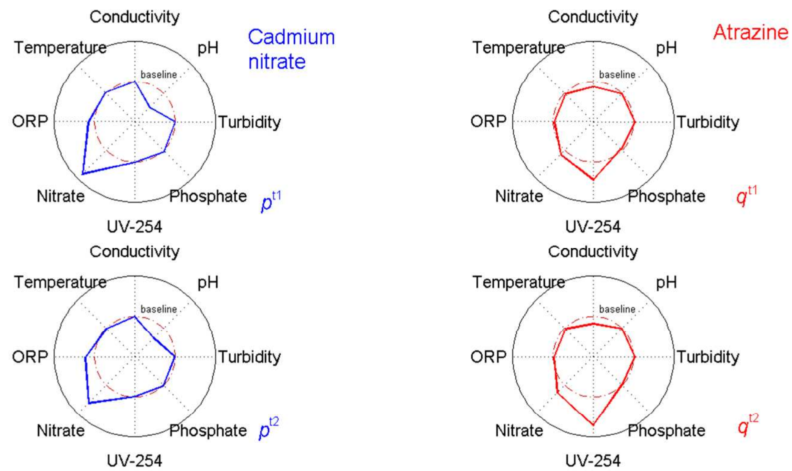


Figure 3 Four instances of *features* of cadmium nitrate and atrazine

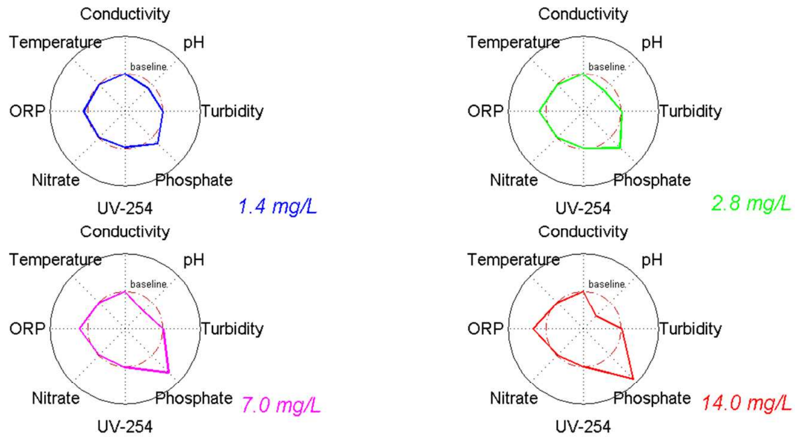


Figure 4 The demonstration of *feature* vectors at glyphosate 1.4, 2.8, 7, 14mg/l

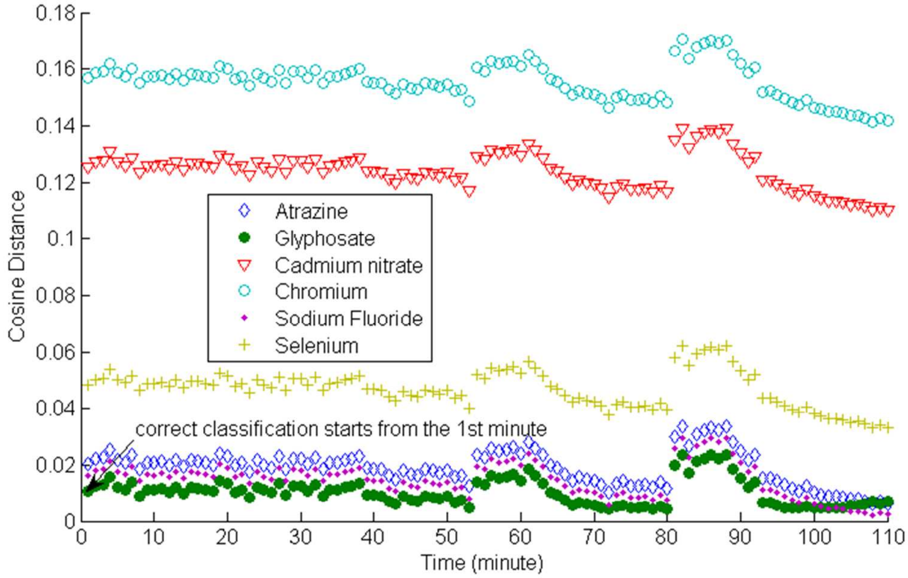


Figure 5 The cosine distance of glyphosate to 6 *classes*

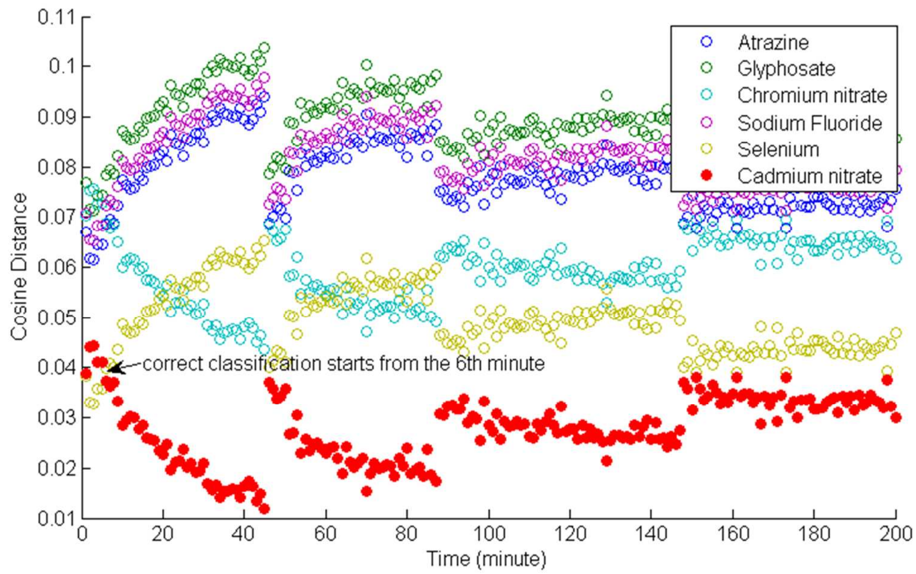


Figure 6 The cosine distance of cadmium nitrate to 6 *classes*

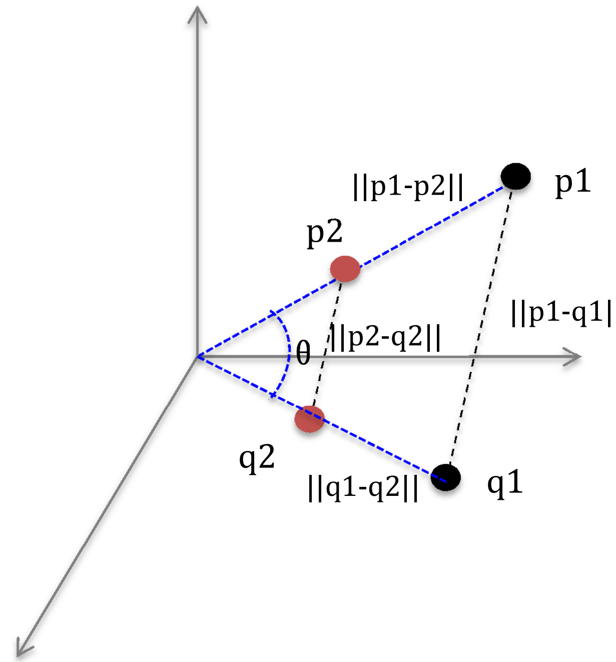


Figure 7 Schematic drawing of cosine and Euclidean distances

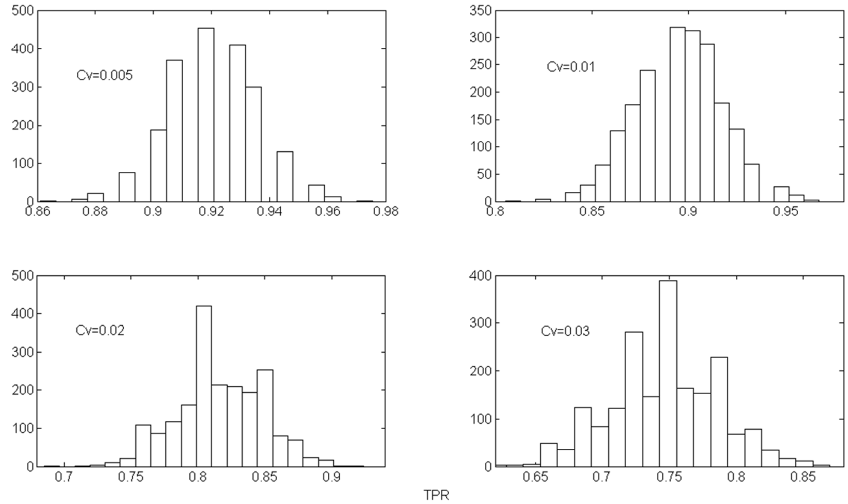


Figure 8 The histogram of *CCRs* with uncertainty and robustness

Table 1 Averaged cosine distances from examples to *classes*

| Tested example | | Cosine distances from examples to classes | | | | | |
|---|---|---|---|---|---|---|---|
| | | Atrazine | Glyphosate | Cadmium nitrate | Nickel nitrate | Sodium fluoride | Sodium nitrate |
| Glyphosate | *mean* | 0.0190 | **0.0105** | 0.1240 | 0.1555 | 0.0148 | 0.0470 |
| | *Standard deviation* | 0.0063 | 0.0048 | 0.0065 | 0.0065 | 0.0063 | 0.0065 |
| Cadmium nitrate | *mean* | 0.0781 | 0.0880 | **0.0277** | 0.0593 | 0.0819 | 0.0494 |
| | *Standard deviation* | 0.0065 | 0.0066 | 0.0065 | 0.0065 | 0.0065 | 0.0066 |

Table 2 The cosine and Euclidean distances

| Euclidean \ Cosine | | Glyphosate | | Cadmium nitrate | |
|---|---|---|---|---|---|
| | | *a* - 1.4mg/l | *b* - 14mg/l | *c* - 0.008mg/l | *d* - 0.032mg/l |
| Glyphosate | *a* - 1.4mg/l | 0 | 0.0027 | 0.1091 | 0.1391 |
| | *b* - 14mg/l | 95.5981 | 0 | 0.1081 | 0.1381 |
| Cadmium nitrate | *c* - 0.008mg/l | 92.3888 | 158.4424 | 0 | 0.0302 |
| | *d* - 0.032mg/l | 113.1857 | 166.8406 | 25.9895 | 0 |

Note: The numbers above the diagonal are cosine distances, while the ones below are Euclidean distances.

Table 3 Statistics of *CCR* under uncertainty (original *CCR*: 0.92)

| *CCR* | $Cv$ =0.005 | $Cv$ = 0.01 | $Cv$ = 0.02 | $Cv$ = 0.03 |
|---|---|---|---|---|
| Mean | 0.92 | 0.90 | 0.82 | 0.75 |
| Standard deviation | 0.01 | 0.02 | 0.03 | 0.04 |
| Robustness | 0.97 | 0.93 | 0.82 | 0.73 |