Machine Learning Combined with Non-targeted LC-HRMS Analysis for a Risk Warning System of Chemical Hazards in Drinking Water: A Proof of Concept

Saer Samanipour[a,b,*], Sarit Kaserzon[b], Soumini Vijayasarathy[b], Hui Jiang[b], Phil Choi[b], Malcolm J. Reid[a], Jochen F. Mueller[b], Kevin V. Thomas[a,b]

[a]*Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, 0349 Oslo, Norway*
[b]*Queensland Alliance for Environmental Health Science (QAEHS), University of Queensland, 20 Cornwall Street, Woolloongabba, QLD, 4012, Australia.*

Abstract

Guaranteeing clean drinking water to the global population is becoming more challenging, because of the cases of water scarcity across the globe, growing population, and increased chemical footprint of this population. Existing targeted strategies for hazard monitoring in drinking water are not adequate to handle such diverse and multidimensional stressors. In the current study, we have developed, validated, and tested a machine learning algorithm based on the data produced via non-targeted liquid chromatography coupled with high resolution mass spectrometry (LC-HRMS) for the identification of potential chemical hazards in drinking water. The machine learning algorithm consisted of a composite statistical model including an unsupervised

*Saer Samanipour
Email address: saer.samanipour@niva.no (Saer Samanipour)
[1]NIVA, Gaustadalléen 21, 0349 Oslo, Norway
Tel: +47 98222087

component (i.e. principal component analysis PCA) and a supervised one (i.e. partial least square discrimination analysis PLS-DA). This model was trained using a training set of 20 drinking water samples previously tested via conventional suspect screening. The developed model was validated using a validation set of 20 drinking water samples of which 4 were spiked with 15 labeled standards at four different concentration levels. The model successfully detected all of the added analytes in the four spiked samples without producing any cases of false detection. The same validation set was processed via conventional trend analysis in order to cross validate the composite model. The results of cross validation showed that even though the conventional trend analysis approach produced a false positive detection rate of $\leq 5\%$ the composite model outperformed that approach by producing zero cases of false detection. Additionally, the validated model went through an additional test with 42 extra drinking water samples from the same source for an unbiased examination of the model. Finally, the potentials and limitations of this approach were further discussed.

Keywords:

Machine learning; Non-target; LC-HRMS; Drinking water; Statistical modeling

---

1. Introduction

Providing clean drinking water is crucial for sustaining human health and it is therefore defined as one of the UN goals for sustainable development [1].

According to the World Health Organization, providing clean drinking water to the global population may reduce the worldwide disease by $\sim$ 10%. Today, factors such as: urbanization, human chemical footprint (i.e. chemical production, consumption, and release), global water scarcity due to climate change, and population growth are making the production and distribution of clean drinking water to the global population a challenging task [2–4]. The situation is far from static as the challenges grow and change, and this is evident in the ever evolving water quality monitoring programs across the globe. However, during the past two decades, it has become more and more evident that existing water monitoring strategies are not adequate to address these challenges [2–6].

Non-target analysis using liquid chromatography coupled with high resolution mass spectrometry (LC-HRMS) has been the leading analytical strategy to tackle the challenges faced by a diffuse and highly dynamic chemical footprint [7–12]. This approach (i.e. non-targeted LC-HRMS), differently from typically limited and targeted routine monitoring strategies, is not biased towards a small number of target analytes. However, it generates highly complex datasets with thousands of features to be analyzed for each sample. To deal with such large and complex datasets the analysts have to isolate the environmentally relevant features from the generated features lists (i.e. prioritization) [7, 9, 13, 14]. Prioritization may be performed using the intensity of the features and/or based on the statistical significance of those features

when compared to the samples from different origins, for example [7, 13]. Intensity based prioritization is relatively fast, but it ignores the lower intensity features, which may be relevant. Therefore, a statistical approach may be more adequate for analysis of water samples, including drinking water.

Recently, advanced statistical tools such as machine learning algorithms have been utilized for regression, dimension reduction, and sample classification via simple or composite models [15]. This approach is a widely used method for prediction of chemical and physical properties of compounds [16]. However, to our knowledge it has never been used in combination with non-targeted LC-HRMS data for monitoring of water samples.

The aim of this study was to develop a risk warning system of potential chemical hazards in drinking water by combining non-targeted LC-HRMS and machine learning. The drinking water samples (i.e. 82 samples) were divided in three groups: 20 samples for a training set, 20 samples for a validation set, and 42 samples for a test set. The training set was used for the model development whereas the test set and the validation sets were utilized for the model validation. During the model validation and test, we cross validated our model via trend analysis and suspect screening.

2. Methods

2.1. Chemicals

All chemical standards and solvents (ACS grade) were purchased from Novachem Pty Ltd. (Victoria, Australia) whereas the technical grade filters were obtained from Phenomenex. A complete list of the labeled internal standards, their measured retention time, and their measured masses is provided in the Supporting Information, Section S1.

2.2. Environmental Samples and Sample Processing

In total 82 drinking water samples of 1 L each were received from South East Queensland, Australia, during a 6 week time period between March and April 2018. Each sampling day consisted of six water samples, except two instances with five samples, taken during the day with intervals larger than 1 hr. The samples were treated drinking water directly from six treatment plants with the same source water and treatment processes. The samples were delivered to the lab at 4°C and were immediately processed and analyzed (i.e filtered and spiked with internal standards). For the analysis, all 82 drinking water samples were filtered using 2 μm filters and an aliquot of each was transferred into 1.5 mL vials having a final volume of 1 mL, without any further processing. All the samples were spiked with 5 μL of a 1 ppm stock solution of caffeine 13C to obtain an injection standard (i.e. caffeine 13C) concentration of 5 ppb. The sample preparations were kept to minimum in order to avoid any type of cross-contamination of the samples.

2.3. Instrumental Conditions and Analysis

All 82 drinking water samples were analyzed using a Sciex ExionLC chromatography system coupled to a Sciex X500R QTOF mass spectrometer (AB SCIEX, USA). Ten μL of each sample was directly injected into the system and separated with Kinetex Biphenyl column (50 $\times$ 2.1 mm, 2.6 µm, Phenomenex) at 50°C. The separations were carried out using 0.1% formic acid in MilliQ water as mobile phase A and 0.1% formic acid in methanol as mobile phase B at a flow rate of 0.4 mL/min. The gradient started at 0% B for 0.5 min, then ramped up to 100% B in 9.5 min with a non-linear Curve (convex) and maintained at 100% B until 14.5 min before returning to 0% B for equilibration. The mass spectrometer was equipped with a TurboIonSpray ion source and operated employing Electron Spray Ionization (ESI) source in positive mode with data-independent acquisition. During pseudo MS^2 scans, the collision energy (CE) was set at 35 eV (more details are provided elsewhere [12]). These instrumental conditions were previously optimized for these type of analysis [12, 17, 18].

For quality control, all the glassware used during the analysis were baked overnight at 450°C. We did not expect a large level of variability in the samples due to the simplicity of the matrix (i.e. drinking water) as was previously observed for similar matrices [12, 17, 18]. Moreover, each five samples were followed by a blank injection, which consisted of a MilliQ water spiked with the labeled internal standards (Section S1). All the analyzed

blanks were procedural blanks and were treated in the same way as the samples. The samples were injected in a randomized order.

2.4. Experimental Setup

The 82 drinking water samples were divided into three categories: the training set (20 samples), the validation set (20 samples), and the test set (42 samples), Fig. 1. For the training set and validation set, we selected a 50% division of the data in order to avoid any over-training of the model [15, 19]. With regards to the test set, we used a large test set in order to assess if a large enough training set was used for the model generation. In other words, a small training set would result in a large number of false positive detection during the model test. The training samples were employed during the machine learning algorithm development (i.e. the composite model) whereas the test set samples were used for an unbiased performance evaluation of the model. Four out of 20 validation set samples were spiked with a mixture of 15 labeled internal standard at 2.5 ppb, 5 ppb, 10 ppb, and 20 ppb of each internal standard in addition to caffeine 13C. Prior to the model development all these 40 water samples were subject to conventional suspect screening in order to assess their quality (see Section 2.7 for more details). Finally, we employed the developed and validated model to assess the quality of the test set (i.e. 42 water samples). During the model validation and model test steps we included two different cross validation steps, which consisted of processing the same dataset with two conventional methods (i.e. trend

analysis and suspect screening). Further details regarding both the trend analysis and the suspect screening are provided in Sections 2.5 and 2.7). Moreover, the validated model was further examined via synthetic datasets where 5 randomly selected samples from the validation set were added to the test set. This process was repeated 50 times to further test the applicability of the model for different drinking water samples. This implied during each iteration a random combination of the spiked and unspiked samples were added to the test set for further evaluation. Doing so enabled us to truly evaluate the likelihood of false detection of the model. It should be noted that the samples did not go through any sample pre-concentration and the concentration of each spiked standard at the lowest concentration level (i.e. 2.5 ppb) was close to the measured limit of detection for the same standards (i.e. $\sim$ 1.5 ppb or 15 pg on column).

Using this experimental design, we were able to first build our model via the training set, validate the model using the validation set, and test the model through the synthetic test set.

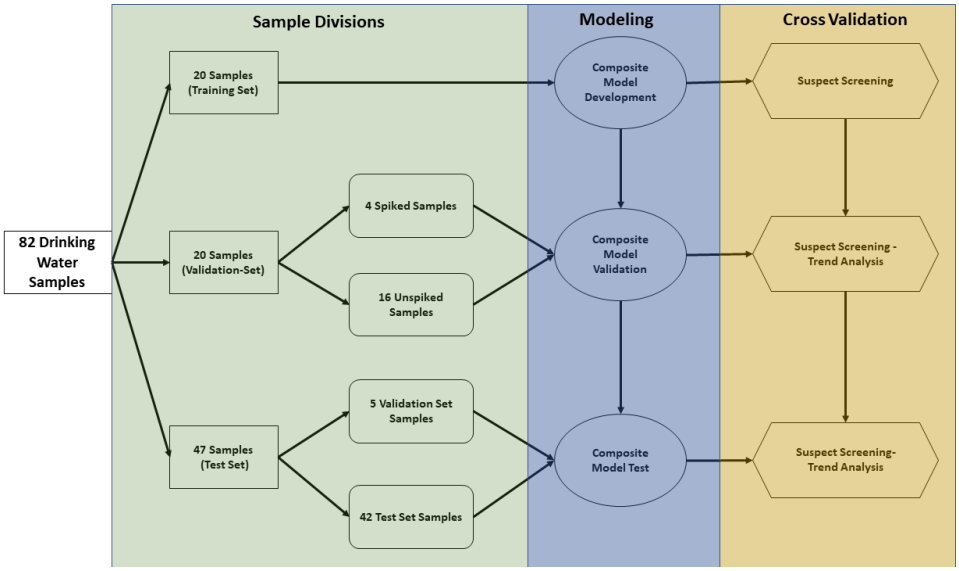


Figure 1: The schematic of the workflow employed in this study including sample division, modeling, and cross validation.

2.5. Machine Learning Algorithm Workflow

The acquired chromatograms for all the samples, including the training set, validation set, and the test set went through the following steps sequentially: 1) peak picking, 2) peak alignment, 3) correction for the background variability, 4) standardization, and finally 5) modeling. The workflow was divided in two parts pre-processing, which included steps 1 to 4 and the modeling which was the fifth step in the complete workflow.

2.5.1. Data Pre-processing

All the chromatograms were peak picked using Sciex OS 1.4 (AB SCIEX, USA) employing a minimum peak area of 1000 counts and a signal to noise ratio of five. After the peak picking, we used Sciex OS for the alignment of the chromatograms, which employed a maximum peak width of 6 s in the time dimension whereas the mass window was set to 0.003 Da. Both of these parameters were selected based on the reported peak boundaries in the time dimension and the observed mass error in m/z values (i.e. $\pm$0.003 Da). The aligned peak list at this stage went through the correction for the background variability. This step enabled us to correct for the variability observed in the background signal caused by the instrument fluctuations [20]. We employed the C13 labeled caffeine signal for the background variability correction of all the datasets, including the training set, validation set, and the test set. For the background variability correction, the signal of all the features in a sample was divided by the intensity recorded for the injection

standard (i.e. C13 caffeine) in that sample. The last step of the data pre-processing consisted of standardization via Pareto method [20], which divides the intensity of each feature (i.e. the tensor of m/z value, retention time, and intensity) by the square root of the standard deviation of that feature across all the chromatograms. The standardization reduces the variability range of each feature, thus giving the same importance to each feature independently from their intensity. Following the above-mentioned steps enabled us to adequately prepare our data for the modeling steps.

2.5.2. Composite Model Development

The training set, consisting of the pre-processed peak-list (i.e. m/z, retention time, and the relative intensities) of 20 drinking water samples, was used for the composite model development. The purpose of this model was to describe the chemistry of the unspiked water, through modeling the maximum variance in the training set for each feature. Therefore, an observed larger variability for a certain feature during the validation step implied the presence of an abnormality or a potential chemical hazard in that sample. The validation set employed in this study included 20 drinking water samples from which 4 were spiked with 15 labeled internal standards. The validation set was generated in a double blind manner to comprehensively evaluate the capability of the model in distinguishing the clean water samples from the spiked ones. Both the training set and the validation set were also employed for tuning the model parameters. For both the model building, model val-

idation, and model test we employed a non-targeted approach utilizing all the features in the samples. Therefore, our model was based on the complete chemical composition of the LC-HRMS analyzable fraction of the drinking water samples. This implied that theoretically only one statistically meaningful feature was enough for distinguishing the spiked samples from unspiked ones.

Our model consisted of a linear combination of principal component analysis (PCA) [19] and partial least square discrimination analysis (PLS-DA) [21, 22] modeling approaches, that enabled the confident separation of the unspiked drinking water samples from the spiked ones. The PCA modeling approach is an unsupervised method, which enables an unbiased evaluation of the underlying trends in the data. However, given its nature [19], the PCA is less sensitive towards small changes in the data. PLS-DA, on the other hand, is a supervised approach, which takes advantage of the prior knowledge of the data [21, 22]. In other words, this method utilizes the user defined classification in the training set to create the model. This implies that the model is forced to give a higher importance to certain variables, that are causing the separation of the pre-defined groups from each other. However, this method suffers from overfitting issues [21, 22]. We used a linear combination of the two modeling approaches in order to fully harvest the higher sensitivity of PLS-DA and at the same time take full advantage of the robustness of PCA.

For the PCA modeling, we used the singular value decomposition algorithm [23] given the larger number of variables (i.e. features) than the number of the measurements (i.e. the drinking water samples). We used the sum of the absolute values of the scores for the first two PCs as the output of the PCA model (S_{PCA}). The choice of using only the first two PCs was based on the fact that these two PCs combined described $\geq 50\%$ of the observed variability in our dataset, which indicates the existence of an underlying trend [19]. The same pre-processed training set was used for PLS-DA model building. During the training step, the PLS-DA was trained only using the unspiked samples, which enabled the generation of a highly sensitive model. One of the crucial steps in the PLS-DA modeling is the selection of the number of components to generate the model, in order to avoid overfitting issues [21, 22]. This choice was carried out through an optimization process employing the training set. We performed 100 simulations where 15 samples were randomly selected from the training set for each iteration. A new PLS-DA model was generated during each simulation with new score values and components. We also recorded the number of necessary components to describe 95% (i.e. 95% confidence interval) of the variability in the data for each simulation. The results of these simulations indicated that four components were necessary to describe 95% of the variability in all the simulated cases. Therefore, we limited the number of PLS-DA components to three, in order to avoid overfitting issues [21, 22]. When calculating the

contribution of the score values of the components on the S_{PLS-DA} value, the first component contributed more than 50% of the S_{PLS-DA}. Consequently, for simplicity we only included the X-score (i.e. the score value associated to the predictor block) of the first component in the PLS-DA score calculations (i.e. S_{PLS-DA}). The selection of the number of components in the PLS-DA model is case dependent and must be evaluated during the model creation for each dataset. Finally, we generated a score value for the final model, hereafter referred to as final score (S_{final}), for each drinking water sample in the training set. The S_{final} was a weighted linear combination of the S_{PCA} and S_{PLS-DA} (Eq. 1). In Eq. 1 the S_{PCA}, S_{PLS-DA}, and S_{final} were the score values from PCA model, PLS-DA model, and the final model, respectively while the w_{PCA} and w_{PLS-DA} were the weight value associated with PCA and PLS-DA score values (Eq. 1). The training set was employed to optimize the weight values as such to produce S_{final} values ranging between -1 and 1. While performing the weight value optimization, we utilized the likelihood of false positive detection as the selection criteria for the tested weight values. The details of this process is described below, Section 3.1.

$$S_{final} = w_{PCA} \cdot S_{PCA} + w_{PLS-DA} \cdot S_{PLS-DA} \tag{1}$$

2.6. Trend Analysis

We performed trend analysis [5, 12, 24–26] on the validation set in order to compare the performance of the composite statistical model with a more

conventional approach. During these analysis, we produced the signal intensity of each feature for all the samples including the pre-processed training set and validation set. In this case we singled out the features that were enriched at a statistically significant levels through the comparison of the median of a feature across all the samples (i.e. background) to the intensity of that feature in each sample (signal). For a feature to be considered statistically significant, it had to produce a signal to background ratio of five in the validation set. Consequently, a feature that met all these criteria was considered a statistically significant feature and was selected for post-processing (e.g. identification). The signal to background ratio of five was selected based on the observed variability of the features in the training set, which enabled us to minimize the likelihood of false positive detection.

2.7. Suspect Screening

The samples for the case study were suspect screened using a local library of pesticides, pharmaceuticals, personal care products, illicit drugs, and industrial chemicals (3000 chemicals), provided with the vendor software package. We employed LibraryView package provided by Sciex OS for these analyses. We utilized a mass accuracy of $\pm$ 0.003 Da and at least 3 matched fragments, in order to confidently identify a suspect analyte. These criteria were previously shown to be effective in processing such datasets [10, 12, 17, 18, 27, 28].

2.8. Rate of False Detection

We also evaluated the rate of false detection (i.e. false positive and/or false negative) [29, 30] of the features that were isolated via the composite model and/or trend analysis. A selected feature was considered a false positive when its accurate mass, retention time, or the sample order, during the analysis, did not match the same parameters of the added internal standards. On the other hand, a feature was assumed a false negative if it was added into a sample as an added internal standard and it was not selected by either the composite model or trend analysis as a statistically significant feature. This appeared to be reasonable given that the thresholds for positive detection in both the composite model and the trend analysis were set as such to produce zero cases of false positive detection for the training set.

Using the rates of false detection, we were able to comprehensively compare the performance of the composite statistical model to the more conventional approach of trend analysis.

2.8.1. Computations

All the data manipulations and modeling were performed on a personal computer with an i7 processor and 16 GB of memory using Matlab 2015b [31].

3. Results and Discussion

A machine learning algorithm was developed and validated for a risk warning system for chemical hazards in drinking water, using the data produced via non-targeted LC-HRMS. The machine learning algorithm took advantage of a composite statistical model, which used a linear combination of a supervised method (i.e. PLS-DA) and an unsupervised approach (i.e. PCA). The composite statistical model utilized all the features present in the sample, thus a non-targeted approach. This composite model utilizes the training set to learn about the variability range of each feature in drinking water samples. Consequently, if one or more of the features in the validation/test samples has a larger intensity compared to its observed variability in the training set, the model will generate a large S_{final} value, which is translated into a trigger for the risk warning system. We validated the developed model employing a validation set of 20 drinking water samples from which 4 were spiked with 15 labeled internal standards at different concentration levels, ranging from 2.5 ppb to 20 ppb. The spiked samples were used for evaluation of false negative and false positive detection rates while the unspiked samples were used for the assessment of false positive detection. We also compared the performance of the model with the conventional trend analysis, typically used for processing this type of data. Finally, the validated model was further tested in processing of 42 water samples alongside with conventional suspect screening. This is, to our knowledge, the first study using the combination of machine learning and non-target analysis for

a risk warning system of chemical hazards in water samples. Also it should be noted that this is a proof of concept study and further implementation of this approach on more complex samples are necessary and will be subject of our future studies.

3.1. Optimization of the Machine Learning Algorithm

The training set was used to select the weight values as well as the thresholds of false positive detection for the composite model. In order to select these parameters, we ran 50,000 (400×125) simulations where for each iteration 18 randomly selected samples out of 20 samples in the training set were used to generate the final composite model. In order to perform this optimization, a squared matrix of weight values varying between 0 and 2 with steps of 0.1 was generated (i.e. a matrix of 20×20, thus 400 members in the matrix). At each point in this matrix 125 simulations took place for false detection calculations. Employing this approach, we generated a distribution of S_{final} values for unspiked drinking water samples enabling us to calculate the rate of false positive detection for different weight values. This optimization process indicated that the best weight values were 0.1 and 1 for w_{PCA} and w_{PLS-DA}, respectively, producing the smallest cases of false positive detections.

The S_{final} values of 1, 1.2, and 1.5 resulted in false positive detection likelihoods of 5.0%, 1.0%, and 0.1%, respectively, employing the optimized

weight values (Fig. 2). In order to further evaluate the likelihood of false positive detection, 5,000 simulations were performed using the pre-processed training set and the optimized weight values, which resulted in a distribution of the S_{final} values. These values in the generated distribution then were converted into the likelihood of false positive detection [29, 30]. For this study, a S_{final} value of 1.2 was selected as the threshold for a statistically significant warning for a potential chemical hazard risk. The selected likelihood of false positive detection enabled us to associate a high level of confidence to the samples that produced an S_{final} value of ≥ 1.2.

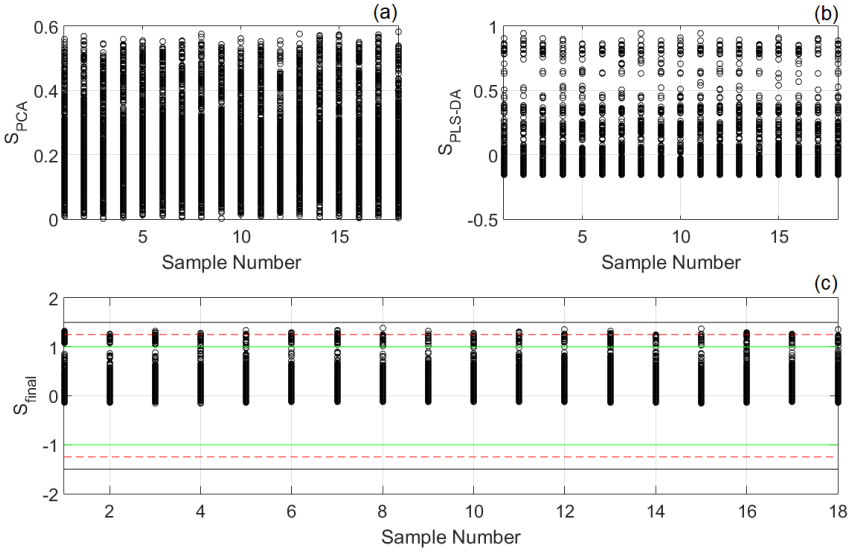


Figure 2: The (a) weighted S_{PCA}, (b) weighted S_{PLS-DA}, and the S_{final} values calculated via Eq. 1 for 5,000 simulations with weight values of 0.1 and 1 for PCA and PLS-DA models. The green line, red dotted line, and black line in panel (c) define the S_{final} values of 1, 1.2, and 1.5, respectively.

3.2. Model Validation via Validation Set (i.e. Spiked Samples)

The previously developed model was validated using the pre-processed validation set. For the model validation, we replaced one of the samples in the training set (randomly selected between sample 2 and sample 19 of the training set) with one of the samples in the validation set. At this stage the PLS-DA model was forced to consider the added sample as a spiked sample whether it was spiked or not. This implied that for a spiked validation sample both models (i.e. PCA and PLS-DA) produced larger score values, and consequently a large final score. On the other hand, for a non-spiked validation sample, considered as spiked sample, the PLS-DA model produced a large score value whereas the PCA model generated a small score, which resulted in a small final score. The random selection of the location of the added validation sample into the training set was due to the fact that we wanted to be sure that the location of the sample addition did not affect the outcome of the algorithm. We evaluated each sample in the validation set using the above mentioned procedure in an iterative way.

The proposed machine learning algorithm (i.e. the composite statistical model) detected all the 4 spiked samples without producing any cases of false positive and/or false negative detections, Fig. 3. The S_{final} values ranged from 1.27 (Fig. S1) for the sample in the validation set with the lowest spike level (i.e. 2.5 ppb) to 2.5 (Fig. S2) for the sample spiked with 20 ppb of the standard mixture. For all the samples that were not spiked with internal

standards the S_{final} was $\leq$1, Fig. S3. By looking at the ratio of the loading values of PCA and PLS-DA model, we were able to identify the features that were the cause of the abnormality (Fig. S4). Based on the absolute intensity of the loading ratios, the top 95.0% of the features were selected for isolating those that were describing the large S_{final} values. This resulted in selection of 15 features, which belonged to the labeled standards. For example, a feature with loading value ratio of 8325 and 9330 for PCA and PLS-DA, respectively, was associated with the signal of carbamazepine D10, which was one of the added internal standards. Additionally, we evaluated the model limit of detection (LOD_{model}) for the tested 15 standards using the response factor calculated based on the slope of the standard addition calibration curve of the spiked samples. The composite model resulted in an averaged LOD_{model} of $\sim$ 1.8 $\pm$ 0.3 ppb for evaluated internal standards, which was comparable to the measured instrument LOD for these standards of $\sim$ 1.5 ppb. This was performed by artificially reducing the signal of each internal standard in the validation set employing 0.01 ppb steps until the model was not able to distinguish the spiked samples from the unspiked training set. The last detectable signal for an internal standard was considered the LOD_{model} for that standard. Furthermore, we compared the LOD_{model} of the composite model (i.e. combined PCA and PLS-DA) to each of the models individually. The limit of detection of the PCA model (LOD_{PCA}) alone appeared to be $\sim$ 12.0 $\pm$ 1 ppb across all 15 labeled analyts, which was around 6 times larger than the LOD_{model}. When using the PCA model alone for analysis of the

validation set, this model produced four cases of false negative detections and no cases of false positive detection. On the other hand, for the PLS-DA, this model resulted in 6 cases of false positives and zero cases of false negative detection for the processing of the validation set. This was due to the lower LOD of PLS-DA model (LOD_{PLS-DA}) of $\sim$ 1.0 $\pm$ 0.2 ppb. These results indicated the higher performance of the composite model compared to each of the individual models suggesting high sensitivity and robustness of the final composite model.

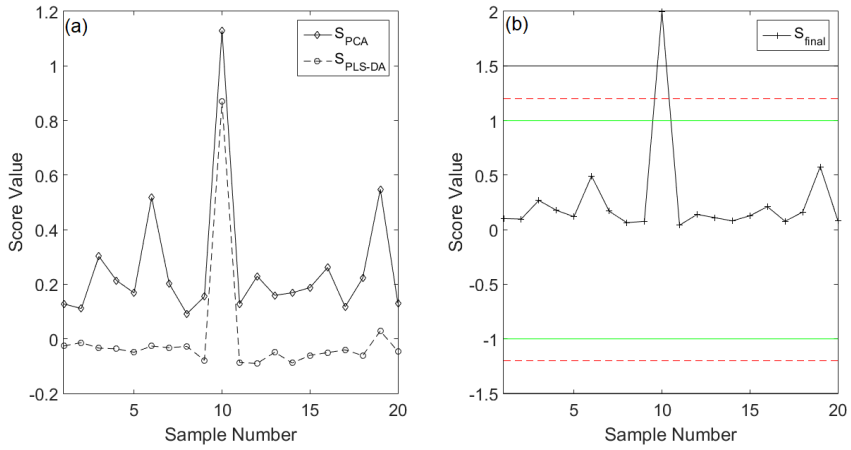


Figure 3: The score values for (a) PCA model (S_{PCA}), PLS-DA model (S_{PLS-DA}) and (b) the composite model score value calculated using Eq. 1. In this instance the sample number 10 was the spiked sample.

3.3. Comparison Between the Composite Model and Conventional Trend Analysis

We compared the performance of the composite model with the conventional trend analysis, which is commonly used for detection of pulsed point source into the water samples [24]. During the trend analysis, we selected a signal to background ratio of five, for a feature to be considered statistically significant. We used the pre-processed training set and the validation set for the comparison between the two methods (Section 2.6).

The trend analysis approach produced 30 cases of false positive detections (i.e. a false positive rate of $\leq 5\%$ [29, 30]) without producing any cases of false negatives. On the other hand the composite model was able to detect all 15 spiked analytes in all 4 samples without producing any cases of false positive and/or false negative. For 27 out of 30 (i.e. 90%) of the false positive cases manual inspection of the features caused their elimination from the list. These features appeared to have low intensity and high level of variability across all the samples, including the training set, Fig S5. The remaining three features identified as false positives appeared to be the isotopes of a real features. For example, a feature identified as a false positive with m/z value of 181.083 and a retention time of 4.50 min was the M+2 isotope of atrazine desisopropyl D5 with an accurate mass of 179.085 and the retention time of 4.48 min. We further investigated the 3 meaningful features isolated via trend analysis in the composite model. For those three features, their

loading values were smaller than 95%. Therefore, the composite model did not consider these features as statistically significant, which further indicates the robustness of this approach compared to the conventional method (i.e. the trend analysis).

The composite model (i.e. the machine learning algorithm) performed better than the conventional trend analysis when applied to the validation set. This method was able to capture all the added analytes in the spiked samples without producing any cases of false detections. This method showed to be less sensitive to the high variability in the data compared to the conventional trend analysis method. However, further tests are necessary to comprehensively evaluate the effect of noise on such a model. Overall, this showed to be a sensitive, accurate, and reliable tool for capturing contamination in the drinking water.

3.4. Further Testing via Test Set

We further tested the capability of the composite model in distinguishing a spiked water sample from an unspiked one. Additionally, this final test enabled us to evaluate the applicability of the same training set for a different batch of water samples taken from the same source (Section 2.4).

The composite model produced 3 cases of false negative and zero cases of false positive detection out of the total 2350 (i.e. 47 samples $\times$ 50 sim-

ulations) evaluations during the test. All 3 cases belonged to the spiked water samples at lowest concentration level (i.e. 2.5 ppb). Both the composite model and the conventional suspect screening did not produce any abnormality cases for the test set, which was expected considering that these samples were treated drinking water.

The outcome of the composite model was in agreement with the conventional suspect screening, which is indicative of its robustness. However, more complex matrices should be tested in order to further evaluate the applicability of this method. Analysis of more complex matrices such as ground water and surface water will be subject of our future study. Finally, it should be noted that this study is a proof of concept for applicability of such an approach for water related matrics.

3.5. Potential and Limitations

The developed and validated composite model was shown to be a reliable, robust and accurate method for detection of anomalies (i.e. potential contaminants) in drinking water samples. The thresholds for the risk warning could be set by the acute and adverse toxicity of the drinking water samples, which will expand the applicability of this method to monitoring of both the produced drinking water as well as the source water used for producing drinking water. At the current state, the samples were injected as is into the instrument for analysis without any pre-concentration. However, addi-

tion of a pre-concentration step would drastically increase the sensitivity of this method, based on the achieved model LODs that were similar to the analytical LODs. In other words, the pre-concentration step may potentially increase the sensitivity of the model by increasing the instrument sensitivity. Moreover, this method is designed to screen the samples rapidly for anomalies. In addition to the triggered warning, the model will produce a list of features that are causing anomalies, which should be evaluated by the analyst. In practical terms, the analyst can focus only on the samples that triggered a warning and the selected features rather than all the features and samples, therefore simultaneous sample and feature prioritization. Finally, this method could be employed for continuous monitoring of more complex aqueous matrices as long as the observed variability in the training set is representative of the normal state of that matrix.

It should be noted that this method was applied to the peak list in the current study due to the cleanness of the drinking water matrix. However, for more complex matrices, this method should be applied to the raw data in order to be able to model the variability observed in the data. This implies a drastic increase in its computational cost. The warning thresholds are highly dependent on the observed within feature variability of the training set. Consequently, the analyst must assure that the variability in the training set is representative for the variability present in the test set in a normal state, which is also necessary for the conventional trend analysis. In other words, if

the variability in the training set is too large, the model would lose sensitivity (i.e. producing false negatives) whereas if the variability in the training set is too small, then the model will become too sensitive (i.e. producing false positives). Similarly to the trend analysis, given the dependency of the explored chemical space on the analysis conditions[14, 32], the training sets are specific to a sample set and analysis conditions. Therefore, a good understanding of both the matrix and the analytical instrument is crucial to the success of this approach.

4. Acknowledgement

The authors are thankful to the the Research Council of Norway for the financial support of this project (RESOLVE, 243720). We are grateful to Dr. Sharon Grant and Dr. Jake O'Brien for their comments during the project development.

5. Supporting Information

Supporting Information containing the list of standards and complementary figures is available as stated in the text.

[1] UN, UN Goals for Sustainable Development, https://www.un.org/sustainabledevelopment/water-and-sanitation/ (May 2010).

[2] S. R. Newton, R. L. McMahen, J. R. Sobus, K. Mansouri, A. J. Williams, A. D. McEachran, M. J. Strynar, Suspect screening and non-targeted analysis of drinking water using point-of-use filters, Environmen. Pollu. 234 (2018) 297–306.

[3] A. Pal, Y. He, M. Jekel, M. Reinhard, K. Y.-H. Gin, Emerging contaminants of public health significance as water quality indicator compounds in the urban water cycle, Environ. Int. 71 (2014) 46–62.

[4] S. D. Richardson, T. A. Ternes, Water analysis: emerging contaminants and current issues, Anal Chem. 90 (1) (2017) 398–428.

[5] S. D. Richardson, Water analysis: emerging contaminants and current issues, Anal. Chem. 81 (12) (2009) 4645–4677.

[6] Y. Peng, S. Hall, L. Gautam, Drugs of abuse in drinking water–a review of current detection methods, occurrence, elimination and health risks, TrAC Trends Anal. Chem. 85 (2016) 232–240.

[7] T. Bader, W. Schulz, T. Lucke, W. Seitz, R. Winzenbacher, Application of non-target analysis with lc-hrms for the monitoring of raw and potable water: Strategy and results, in: Assessing Transformation Prod-

ucts of Chemicals by Non-Target and Suspect Screening- Strategies and Workflows Volume 2, ACS Publications, 2016, pp. 49–70.

[8] N. A. Alygizakis, S. Samanipour, J. Hollender, M. Ibáñez, S. Kaserzon, V. Kokkali, J. A. van Leerdam, J. F. Mueller, M. Pijnappels, M. J. Reid, et al., Exploring the potential of a global emerging contaminant early warning network through the use of retrospective suspect screening with high-resolution mass spectrometry, Environ. Sci. Technol. 52 (9) (2018) 5135–5144.

[9] E. L. Schymanski, H. P. Singer, J. Slobodnik, I. M. Ipolyi, P. Oswald, M. Krauss, T. Schulze, P. Haglund, T. Letzel, S. Grosse, et al, Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis, Anal. Bioanal. Chem. 407 (21) (2015) 6237–6255.

[10] S. Samanipour, M. J. Reid, K. Bæk, K. V. Thomas, Combining a deconvolution and a universal library search algorithm for the nontarget analysis of data-independent acquisition mode liquid chromatography-high-resolution mass spectrometry results, Environ. Sci Technol. 52 (8) (2018) 4694–4701.

[11] E. L. Schymanski, H. P. Singer, P. Longrée, M. Loos, M. Ruff, M. A. Stravs, C. Ripollés Vidal, J. Hollender, Strategies to characterize polar organic contamination in wastewater: exploring the capability of high

resolution mass spectrometry, Environ. Sci. Technol. 48 (3) (2014) 1811–1818.

[12] S. L. Kaserzon, A. L. Heffernan, K. Thompson, J. F. Mueller, M. J. G. Ramos, Rapid screening and identification of chemical hazards in surface and drinking water using high resolution mass spectrometry and a case-control filter, Chemosphere 182 (2017) 656–664.

[13] S. Samanipour, M. J. Reid, K. V. Thomas, Statistical variable selection: An alternative prioritization strategy during the non-target analysis of LC-HR-MS data, Anal. Chem. 89 (10) (2017) 5585–5591.

[14] S. Samanipour, J. A. Baz-Lomba, M. J. Reid, E. Ciceri, S. Rowland, P. Nilsson, K. V. Thomas, Assessing sample extraction efficiencies for the analysis of complex unresolved mixtures of organic pollutants: A comprehensive non-target approach, Anal. Chim. Acta 1025 (2018) 92–98.

[15] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, Vol. 1, MIT press Cambridge, 2016.

[16] T. H. Miller, J. A. Baz-Lomba, C. Harman, M. J. Reid, S. F. Owen, N. R. Bury, K. V. Thomas, L. P. Barron, The first attempt at non-linear in silico prediction of sampling rates for polar organic chemical integrative samplers (pocis), Environmen. Sci. Technol. 50 (15) (2016) 7973–7981.

[17] S. Kaserzon, E. O'Malley, K. Thompson, C. Paxman, G. Elisei, G. Eaglesham, M. Gallen, J. Mueller, Catchment and drinking water quality micro pollutant monitoring program–passive sampling. report 6–summer 2017 and summary report.

[18] S. Kaserzon, C. Gallen, K. Thompson, C. Paxman, J. O'Brien, G. Eaglesham, M. J. G. Ramos, M. Gallen, D. Drage, X. Wang, et al., Catchment and drinking water quality micro pollutant monitoring program-passive sampling. report 1 2014.

[19] R. G. Brereton, Applied chemometrics for scientists, John Wiley & Sons, 2007.

[20] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC genomics 7 (1) (2006) 142.

[21] R. Kramer, Chemometric techniques for quantitative analysis, CRC Press, 1998.

[22] R. G. Brereton, G. R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemometrics 28 (4) (2014) 213–225.

[23] G. H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numerische mathematik 14 (5) (1970) 403–420.

[24] M. Ruff, M. S. Mueller, M. Loos, H. P. Singer, Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river rhine using high-resolution mass-spectrometry–identification of unknown sources and compounds, Water Res. 87 (2015) 145–154.

[25] Z. Li, S. L. Kaserzon, M. M. Plassmann, A. Sobek, M. J. G. Ramos, M. Radke, A strategic screening approach to identify transformation products of organic micropollutants formed in natural waters, Environ. Sci. Proc. & Imp. 19 (4) (2017) 488–498.

[26] T. Bader, W. Schulz, K. Kuummerer, R. Winzenbacher, Lc-hrms data processing strategy for reliable sample comparison exemplified by the assessment of water treatment processes, Anal. Chem. 89 (24) (2017) 13219–13226.

[27] S. Samanipour, K. Langford, M. J. Reid, K. V. Thomas, A two stage algorithm for target and suspect analysis of produced water via gas chromatography coupled with high resolution time of flight mass spectrometry, J. Chromatogra. A 1463 (2016) 153–161.

[28] S. Samanipour, J. A. Baz-Lomba, N. A. Alygizakis, M. J. Reid, N. S. Thomaidis, K. V. Thomas, Two stage algorithm vs commonly used approaches for the suspect screening of complex environmental samples analyzed via liquid chromatography high resolution time of flight mass spectroscopy: A test study, J. Chromatogr. A 1501 (2017) (2017) 68–78.

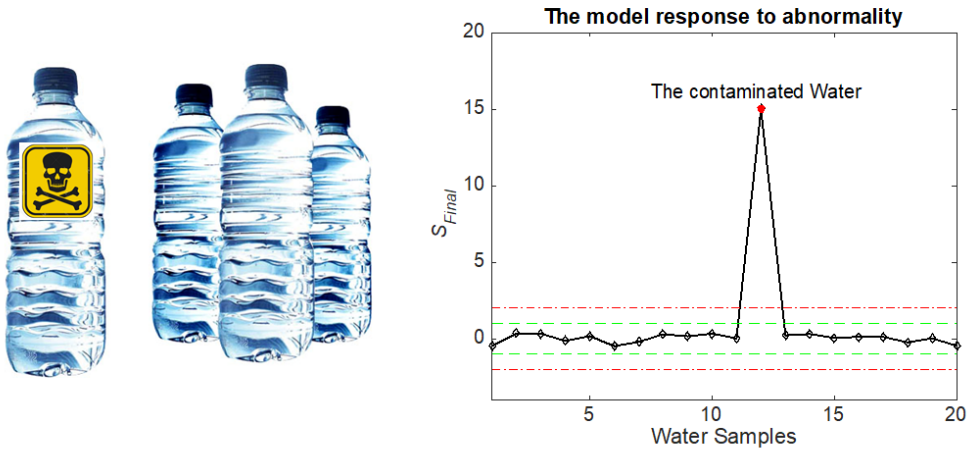
[29] D. S. Burke, J. F. Brundage, R. R. Redfield, J. J. Damato, C. A. Schable, P. Putman, R. Visintine, H. I. Kim, Measurement of the false positive rate in a screening program for human immunodeficiency virus infections, N. Engl. J. Med. 319 (15) (1988) 961–964.

[30] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc.: Ser. B (Methodol.) (1995) 289–300.

[31] MATLAB version 9.1 Natick, Massachusetts: The MathWorks Inc., **2018**.

[32] S. Samanipour, M. Hooshyari, J. A. Baz-Lomba, M. J. Reid, M. Casale, K. V. Thomas, The effect of extraction methodology on the recovery and distribution of naphthenic acids of oilfield produced water, Sci. Tot. Environ. 652 (2019) 1416–1423.

6. TOC



TOC for review only.