Water Air Soil Pollut (2022) 233: 241
https://doi.org/10.1007/s11270-022-05713-7

Evaluating Anthropogenic Origin of Unidentified Volatile Chemicals in the River Rhine

Yangwei Ying · Geert van Kollenburg · André van den Doel · Sanne Brekelmans · Hong Zhou · Gerard Stroomberg · Jeroen Jansen

Received: 25 March 2022 / Accepted: 11 June 2022 / Published online: 22 June 2022
© The Author(s) 2022

Abstract Surface water of rivers like the Rhine is a highly relevant environmental and an important source of the Dutch drinking water. To improve protection of the environment and drinking water supply, it is important to have a continuous overview of the chemical composition of the river. Such an overview may be obtained with contemporary, untargeted analytical platforms like gas chromatography-mass spectrometry. Interpretation of such untargeted data is however challenged by the presence of many chemicals of natural origin. We developed a novel approach to screen for anthropogenic chemicals using non-parametric tests on the time trends of yet unidentified chemicals. The approach uses PARAFAC2 to extract unknown components present in GC–MS data and provides an assessment of whether such components may be anthropogenic. This significantly reduces screening efforts required by human laboratory staff. In total, out of twelve suspect unknown components, eleven were classified as anthropogenic, providing compelling evidence that studying unknown components can be highly valuable for regulatory bodies. This approach filters out many naturally occurring compounds, leaving more resources available for wet-lab identification of suspected anthropogenic chemicals.

Keywords River Rhine · GC–MS · Statistical tests · Pollution detection · Untargeted identification

Yangwei Ying and Geert van Kollenburg contributed to this work equally.

Y. Ying · H. Zhou
Key Laboratory for Biomedical Engineering of Ministry of Education, Zhejiang University, Hangzhou 310027, China

G. van Kollenburg · A. van den Doel · J. Jansen
Institute for Molecules and Materials, Radboud University, Heyendaalseweg, 6525 AJ Nijmegen, The Netherlands

G. van Kollenburg (✉)
Interconnected Resource-Aware Intelligent Systems, Eindhoven University of Technology, Den Dolech 2, 5612AZ Eindhoven, The Netherlands
e-mail: chemometrics@science.ru.nl

S. Brekelmans · G. Stroomberg
RIWA-Rijn, Groenendael 6, 3439 LV Nieuwegein, The Netherlands

S. Brekelmans
University of Applied Sciences Leiden, Zernikedreef, 2333 CK Leiden, The Netherlands

1 Introduction

The river Rhine is one of the largest rivers in Europe with a catchment area of 185.000 km^2 and an average discharge of 2300 m^3/s (Diehl et al., 2005; Ruff et al., 2015). The Rhine is used in many ways: as a source for leisure and recreation, as a waterway, for the discharge of wastewater, and as a source of drinking water. It comes as no surprise that the surface

water of the Rhine has a strong influence on the quality of drinking water (Loos et al., 2009) if not treated adequately.

Pollution, caused by industry, agriculture, or wastewater, can have grave consequences for river water quality. Therefore, it is important to protect the surface water against pollution and make sure that it meets strict standards and guidelines to be able to produce high-quality drinking water. The most extensive guidelines are drawn up in the European Water Framework Directive (WFD, 2000/60/EG) (EC, 2000) and the Priority Substance Directive 2013/39/EU, amending Directives 2000/60/EC and 2008/105/EC. These directives define "good chemical status" for surface waters. A list of priority chemicals and priority hazardous substances (industrial chemicals, pesticides, and heavy metals) has been drawn up for the Netherlands. In this list, threshold levels are set for several chemicals.

Due to improved legislation and regulations, water quality has attracted much attention over the past decades, with respect to these priority compounds (Hering et al., 2010). Nowadays, dozens of anthropogenic compounds are being monitored daily by the Dutch water authority, Rijkswaterstaat (RWS). However, not all compounds can be taken into consideration for intensive monitoring since each one costs time and resources. Industry continuously develops novel compounds that end up in the environment, potentially in degraded form. Therefore, non-target screening for environmental monitoring becomes crucial to help the government/society make effective use of all available information from modern analytical platforms (Hollender et al., 2017). This allows regulatory bodies to prioritize compounds for further analysis and quantification.

Many sources of pollution exist. While factories may receive licenses to discharge certain chemicals under strict requirements, illegal discharges still happen and medicine, drugs, waste, and other toxic items still find their way into the river daily. Surface water and, in the end, drinking water are at risk by these pollutions. Regulatory bodies and drinking water producers such as the Association of River Water Supply Companies (RIWA) are therefore interested in the detection and identification of various chemicals which occur in the river. A significant benefit of correct identification is that it may help regulatory bodies to find the pollution source and take necessary steps to prevent further pollution. One way to find the source of contamination is by analyzing trends observed in the measurements of certain chemicals.

Many xC-MS techniques are widely used for the detection of chemicals (e.g., Campo et al., 2006). Pena-Abaurrea et al. (2014) described an approach using data acquired by GC × GC-TOF MS to identify potentially novel chemicals. Consequently, it has the potential to discover and analyze novel chemicals. In this paper, comprehensive analysis was done by combining GC–MS measurements and chemometric methods to analyze chemicals in the surface water of the Rhine.

Using statistical techniques like PARAllel FACtor analysis 2 (PARAFAC2), it is possible to compare the presence of unidentified chemicals across multiple measurements in the form of (mathematical) components related to the chemicals present in the samples. In this paper, we present a methodology to use the extracted components in a time series analysis and, using non-parametric tests, determine whether a component relates to an anthropogenic or naturally occurring chemical. The underlying assumption is that natural chemicals behave with a relatively predictable pattern over time. This proposed approach was validated both on thirty anthropogenic chemicals which are continuously monitored due to regulatory demands and on simulated data. By adding the suspicious untargeted chemicals to the regulatory list of targeted chemicals, it is possible to quickly detect and identify these chemicals in the future. Furthermore, illegal discharges may be tracked down to stop related pollution events (Hollender et al., 2017; Schlüsener et al., 2015).

2 Materials and Methods

2.1 Sampling

The dataset consists of water samples collected by Rijkswaterstaat from the river Rhine at Bimmen. Hourly sample collection is automated, but only four samples *per* day were chemically analyzed, unless a relevant event was detected in which case the analysis frequency was increased. The data set is composed of one GC–MS measurement per day from January 1 through December 31, 2014.

2.2 Analytical Methods

This study focuses on purge and trap GC–MS analysis, for which the water sample is spiked with a mixture of deuterated internal standards (deuterochloroform, toluene, chlorobenzene, 1,4-dichlorobenzene, and naphthalene) to a final concentration of 1.0 µg/L. GC–MS spectra were obtained using a Varian Saturn ion trap instrument (in single MS mode) with electrospray ionization (ESI) and an electron multiplier detector (EMT). Volatile components of the sample were extracted by purging with an inert gas. Interaction with the stationary face of the GC column separates chemicals based on chemical and physical properties with a retention time range of 0–22.5 min. For further identification, fractions from the GC column were injected into the mass spectrometer where they are separated based on mass-to-charge-ratio (m/z). These mass scans were acquired with 0.1 m/z resolution. Chemicals in the sample can be identified by the combination of retention time and relative intensities of signals at specific m/z values. Target chemicals can be quantified by comparing the integrated peak area to a calibration curve (see Appendix 1 Table 5 for the list of target chemicals). Unknown chemicals are challenging to identify because many potential chemicals can have similar mass spectra and retention times.

2.3 PARADISe

For the whole GC–MS spectrum, it is difficult to extract all present chemicals and it is laborious to compare each MS spectrum with chemical libraries (like NIST or ChemSpider). Therefore, it is necessary to use an automatic program to help us identify unknown chemicals. To separate the chemicals peaks, the PARADISe toolbox was used, which is based on PARAllel FACtor analysis 2 (PARAFAC2), to resolve the untargeted GC–MS data (Bro et al., 1999; Johnsen et al., 2017; Kiers et al., 1999).

PARAFAC2 analyzes three-way chemical data and has been applied in multiple research areas (Kamstrup-Nielsen et al., 2013) It allows for variable elution profiles of each present chemical over multiple GC–MS measurements. A PARAFAC2 model outputs components which are mathematical representations of chemicals based on similarities in retention time and m/z profile. We will use the term "component" in the remainder of this paper to indicate the outputs of PARADISe and "chemical" for molecules which can be found in the river.

The PARAFAC2 model can be written as

$$X_k = AD_kB_k^T, k = 1, \ldots, K \quad (1)$$

where X_k is an $I \times J$ matrix with $k = 1, \ldots, K$ as the k th data point of an $I \times J \times K$ three-way data X. Among these parameters, A is the score matrix, and B is the loading matrix, while D is diagonal $(R \times R)$ matrix containing the weights for the kth slab of X. Figure 1 demonstrates a profile separation result where panel a is the total ion current (TIC, on the vertical axis) for each retention time (in minutes, on the horizontal axis) and panel b is the weighted elution profile. Figure 1b shows the elution profiles of different components. According to the separated result, it is possible to identify whether a particular (unknown) component is present in multiple samples.

The Varian Saturn GC–MS instrument outputs GC–MS spectra in Varian's proprietary.SMS format. This was converted to.CDF format with Open Chrom (OpenChrom® 1.2.0 "Alder," https://www.openchrom.net) for use with PARADISe (Version 2.3, http://www.models.life.ku.dk/paradise). The.CDF files were imported into the PARADISe toolbox and then intervals were selected in regions where there are possible existing peaks. Intervals were chosen without overlap to avoid duplicate peak detection. Considering the deconvolution capability and accuracy, a maximum of 8 components per interval was set. The maximum number of iterations was set to 2000, which was enough to reach model convergence in a reasonable computation time. The model was finalized by selecting the right number of components, according to the core consistency and model fit percentage, such that the components correspond to the number of chemicals present in the sample. Thereafter, a table of samples with components and total ion current (TIC) was created, which was used for trends analysis.

2.4 Preprocessing

To accurately identify pollution events, component concentrations should be transformed into loads. That is, concentrations are affected by natural phenomenon like rainfall (which increases the amount of water in the river), making variations in natural and anthropogenic components harder

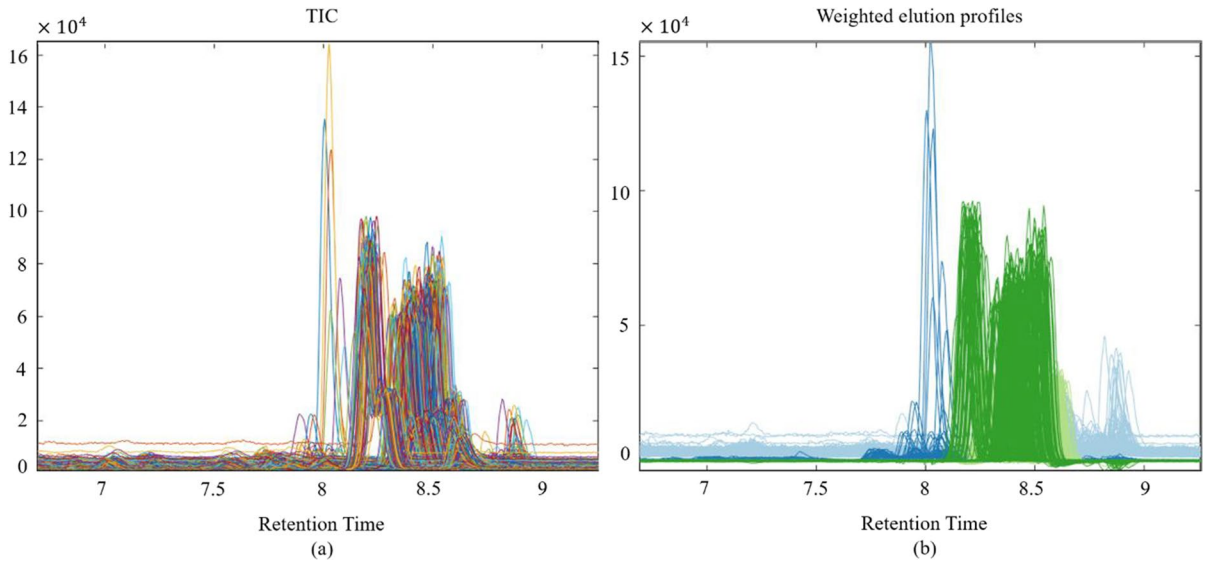

Fig. 1 Taken from the PARADISe software output, each line in panel **a** represents the elution profiles of a sample in the retention time interval given on the x-axis. By applying PARAFAC2 to the data, the mass spectra associated to each data point can separate co-eluted chemicals. Panel **b** shows the same samples as in **a** but then colored according to the similarities in mass spectra present in the sample

to distinguish. While concentrations are important criteria for the toxicity and regulatory monitoring, loads are related to the amount of component that entered the river and are independent of how much water there is in the river at a current time. Therefore, all concentrations obtained by the GC–MS measurements are divided by the water flow (in m^3/s). Furthermore, to compensate for changes in ionization efficiency between GC–MS runs, all TICs are divided by the average TIC of the 5 internal standards (Appendix 2), which is standard practice in analyzing GC–MS data. Figure 2 shows an example of how the patterns change due to this preprocessing. Additionally, we corrected for the water flow, resulting in the data being represented as loads instead of concentrations.

2.5 Tests for time trends

For the current application, non-parametric tests (García et al., 2009) were devised to classify components as natural or anthropogenic based on the variability in their concentration and specifically based on patterns in their concentration over time. In total, five tests on the variability of each component can distinguish anthropogenic components from natural components. If a component passes all five tests, it is defined as natural; otherwise, it is anthropogenic.

Overall variations in river flow and the abundance of natural components in the water do not vary extremely. As a basic test of the proposed methodology, random normally distributed data was generated and analyzed with each test. For these tests, the simulated data was generated with a mean of 3 and a standard deviation of 1 to simulate variation in naturally occurring components over a year (not considering seasonal trends).

2.5.1 Period Test

The *period test* is designed to identify production breaks over time: factories might have production breaks due to maintenance. Observing periods with concentrations of zero are unlikely to occur for natural chemicals (except for seasonal variations) and hence may indicate anthropogenic origin. A period of 7 days with a concentration of zero was considered a possible production break. If a production break was found, the component was classified as anthropogenic. If no production break was found during the entire period, the component was not classified as anthropogenic according to this test. Experience has shown that production breaks, like planned

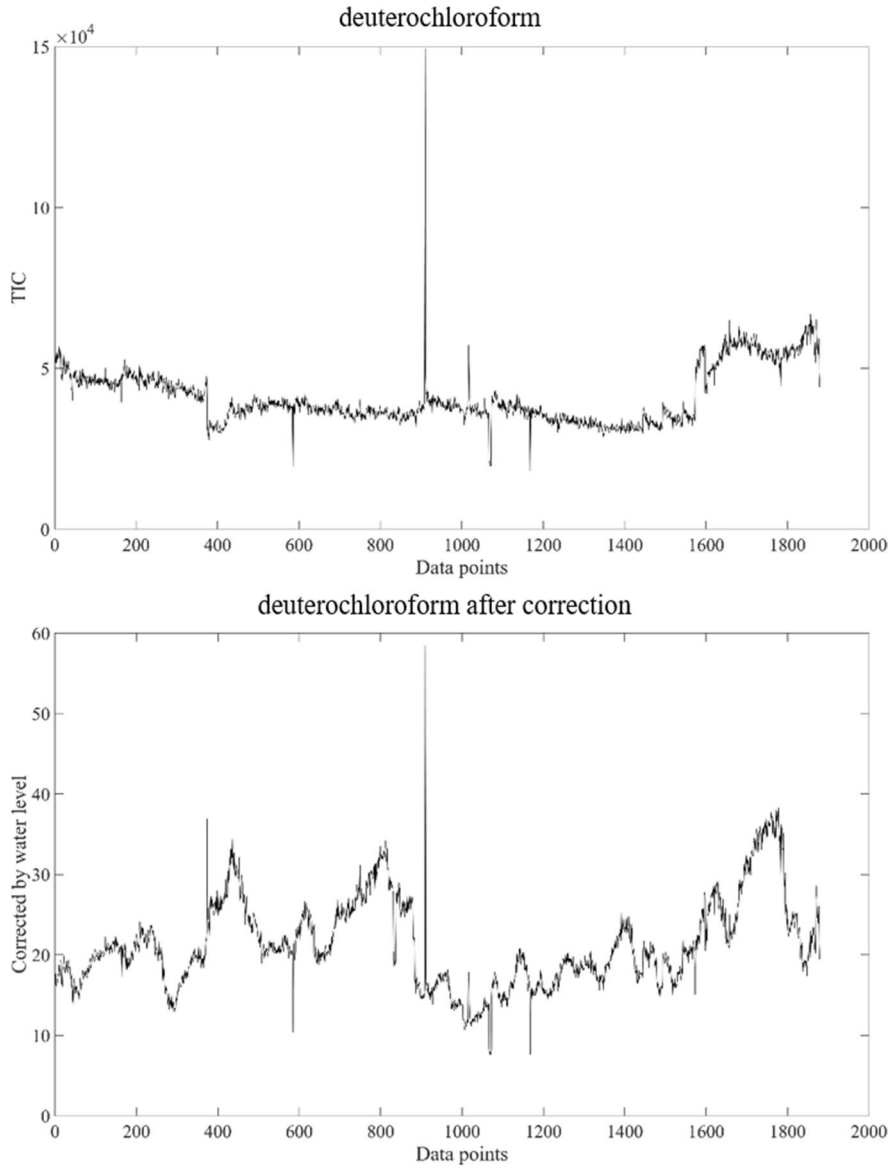

Fig. 2 Observed total ion currents (TIC) (**a**) and corrected TIC (**b**) of the internal standard deuterochloroform over time, corrected with respect to river flow. Variability in the corrected TIC relates to the natural variations in the river water flow

maintenance, take 7 days or more. Smaller breaks will not be detected by this test.

2.5.2 Peak Test

Theoretically, chemicals in nature will generally not fluctuate in extreme manners. Quickly emerging signals may therefore indicate anthropogenic origin. We assume that the concentration of a natural chemical changes gradually over time and that components with patterns that are strongly spiked may be anthropogenic. In this test, we search for these spikes. All the component's concentrations were sorted in order from small to large. Then, the first quartile (Q1) and the third quartile (Q3) of the data were calculated from this data.

To define peaks, the limit value was calculated with the following equation:

$$Limit\ value = Q3 + (c(Q3 - Q1)) \quad (2)$$

If one or more values above the limit value were found, the component was classified as anthropogenic. The variable c in Eq. (2) is used to specify

the cutoff for what is considered a peak. Higher values will lead to stricter classification. For this application, c was rather arbitrarily set to 3.

2.5.3 Extreme Test

In the extreme test, it was determined whether an unknown component contained outliers. By setting a limit value higher than the peak test, the data could be evaluated for extreme values.

Assigning extreme values was done like the peak test, but here ten times the interquartile range was used. This value was added to Q3 to reach the limit value. This results in the following equation:

$$Limit\ value = Q3 + (10(Q3 - Q1)) \quad (3)$$

Components with values over this limit value are classified as anthropogenic. This test may reduce the set of results from the peak test.

2.5.4 Day Test

The *day test* is based on a periodic difference in concentrations for days of the week. Production cycles will likely follow weekly patterns, so a deviation on a certain day, with respect to the other days, could indicate an anthropogenic origin.

The standard deviation was calculated from the data corresponding to the same day of the week. Ultimately, seven standard deviations per component were calculated. The highest standard deviation was compared with the lowest one. If the highest standard deviation was more than 1.6 times the lowest standard deviation, the null hypothesis was rejected. The value of 1.6 was chosen ad hoc, specifically for this dataset. Any value can be used, but it was determined that a value of 1.6 classified unknown components correctly in this research, according to the optimal results evaluated based on the RIWA dataset.

2.5.5 Visual Inspection Check

Visual inspection of the time trend of the unknown component was performed when all other tests were negative. For patterns over time which stand out, the null hypothesis was rejected, and the alternative hypothesis was accepted. The visual inspection was added for components whose outcome was "natural" with all other tests. With an extra visual check, it was possible to determine if a natural component indeed looks like a natural occurring component, or whether there was an abnormal pattern that the tests did not recognize. Note that in this study, the visual check did not influence whether a component was classified as anthropogenic. Classification was only based on the statistical tests described before.

3 Results

3.1 Assessment of Tests

3.1.1 Application to Simulated Data

The tests were applied to simulated datasets. A thousand normally distributed data points were generated with an average of three and a standard deviation of one, as natural components to validate the tests in this study. In statistics, a 95% confidence interval could represent the reliability of the data, so the error smaller than 5% could be acceptable in this dataset. For repeatability, five simulated data sets of 1000 components were assessed. The results are given in Table 1. The table shows that the type I errors are all smaller than 5%, which indicates that the tests work well on the simulated data (e.g., the extreme test does not identify more false positives than can be expected from the random data).

3.1.2 Application to Target Components

The results of assessing the components in the targeted regulatory monitoring set are provided in Table 2. Indeed, every component was correctly

Table 1 Results of applying the tests on simulated data with five repetitions. False positives are situations in which the tests indicate that a component is anthropogenic

Repetition	False positives	Error rate (%)
1	37	3.7
2	30	3
3	37	3.7
4	27	2.7
5	36	3.6

Table 2 Results of the non-parametric tests for the thirty anthropogenic chemicals of the Rijkswaterstaat dataset. The data was corrected for internal standards and water flow; 0 indicates that the chemical was classified as natural, 1 indicates that it was classified as anthropogenic

Component	Test 1	Test 2	Test 3	Test 4	Visual check	Result
Methyl tertiary-butyl ether (MTBE)	1	1	1	1	1	Anthropogenic
Diisopropyl ether	1	1	1	1	1	Anthropogenic
Cis-1,2-Dichloroethene	1	1	1	0	1	Anthropogenic
Ethyl tertiary-butyl ether (ETBE)	1	1	1	1	1	Anthropogenic
Chloroform	0	1	0	0	1	Anthropogenic
Ethyl sec-butyl ether (ESBE)	1	1	1	0	1	Anthropogenic
1,1,1-Trichloroethane	1	1	1	1	1	Anthropogenic
Cyclohexane	1	1	1	1	1	Anthropogenic
1,2-Dichloroethane	1	1	1	1	1	Anthropogenic
Benzene	0	1	1	1	1	Anthropogenic
Tertiary amyl methyl ether (TAME)	1	1	1	1	1	Anthropogenic
Trichloroethylene	1	1	1	0	1	Anthropogenic
Tertiary amyl ethyl ether (TAEE)	1	1	1	0	1	Anthropogenic
Methylisothiocyanate	1	1	1	0	1	Anthropogenic
Toluene	0	1	1	1	1	Anthropogenic
1,1,2-Trichloroethane	1	1	1	1	1	Anthropogenic
Tetrachloroethylene	1	1	0	0	1	Anthropogenic
Chlorobenzene	1	1	1	1	1	Anthropogenic
Ethylbenzene	1	1	1	1	1	Anthropogenic
m/p-Xylene	1	1	1	1	1	Anthropogenic
o-Xylene	1	1	1	1	1	Anthropogenic
Styrene	1	1	1	1	1	Anthropogenic
Cumene	1	1	1	1	1	Anthropogenic
n-Propylbenzene	1	1	1	0	1	Anthropogenic
2-Chlorotoluene	1	1	1	0	1	Anthropogenic
t-Butylbenzene	1	1	1	1	1	Anthropogenic
1,2,4-Trimethylbenzene	1	1	1	0	1	Anthropogenic
1,2-Dichlorobenzene	1	1	1	0	1	Anthropogenic
Hexachlorobutadiene	1	1	1	0	1	Anthropogenic
Naphthalene	1	1	1	0	1	Anthropogenic

classified as being anthropogenic. Because the proposed methodology had low type I error rates in the generated data of non-anthropogenic components, and high power in this test set of known anthropogenic components, the methodology is valuable for analysis of unknown components.

Especially for the visual check, we can take Fig. 2 of deuterochloroform as an example. It passed three tests but failed the peak test (test 2). Besides, the high TIC after correcting for water flow also indicates an anthropogenic origin with the visual check.

3.2 Extracting Untargeted Components with PARADISe

In PARADISe, 12 retention time intervals that do not include peaks for the internal standards were selected from the GC–MS data file (details are given in Table 3).

Take interval 3 for example; the TICs of all samples in the selected retention time interval are shown in Fig. 3. The main goal is to decompose the data and find all of the components inside. Therefore, the number of components was evaluated, using fit

Table 3 Details of intervals definition and optimal number of components as determined by core consistency. In total, 12 intervals are selected with different retention time

Interval	Start (minutes)	End (minutes)	Optimal number of components
1	2	2.6674	1
2	2.6674	3.5305	2
3	3.5354	4.8185	5
4	4.8234	6.1583	2
5	6.1583	7.4659	3
6	7.4708	9.0515	6
7	9.0633	11.4151	2
8	11.4268	12.3896	2
9	13.5208	14.6768	2
10	15.3008	17.2995	1
11	19.0695	19.9409	1
12	21.6190	23.4635	3

percentage and core consistency. When the number of components was set to 6, the core consistency reached almost 100 (Fig. 4), and the fit percentage was high (Fig. 4). When the number of components increased more, the core consistency dropped, suggesting that there are indeed 6 components in the data. We analyzed the component mass spectra, weighted elution profiles, and residuals to check the result; see Fig. 5.

There is no pattern in the residuals, which indicates that spectra are well decomposed, and all components are extracted. Components 5 and 6 show clear peaks in the elution profiles, which indicates they relate to actual chemicals in the sample. Components 1, 2, 3, and 4, on the other hand, show quite low intensity which can be judged as background. Therefore, components 5 and 6 represent the components in interval 3.

3.3 Application of Trend Analysis to Untargeted Pseudo-components

The same procedure as described above was applied to all intervals. That resulted in 12 extracted components for use in further trend analysis. Then, we can output a result table with retention times of all extracted components and their corresponding TICs per sample. Like the targeted data, these untargeted

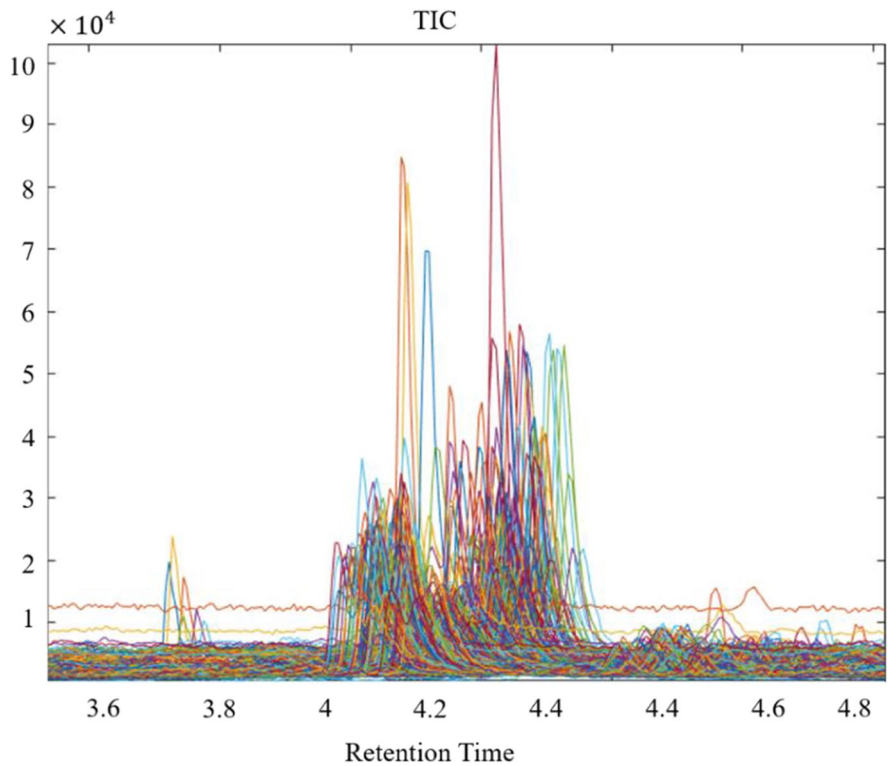

Fig. 3 Observed chromatograms over interval 3, used to illustrate how to determine the number of components. The retention time interval is given on the *x*-axis and total ion current on the *y*-axis. Each color line represents one sample

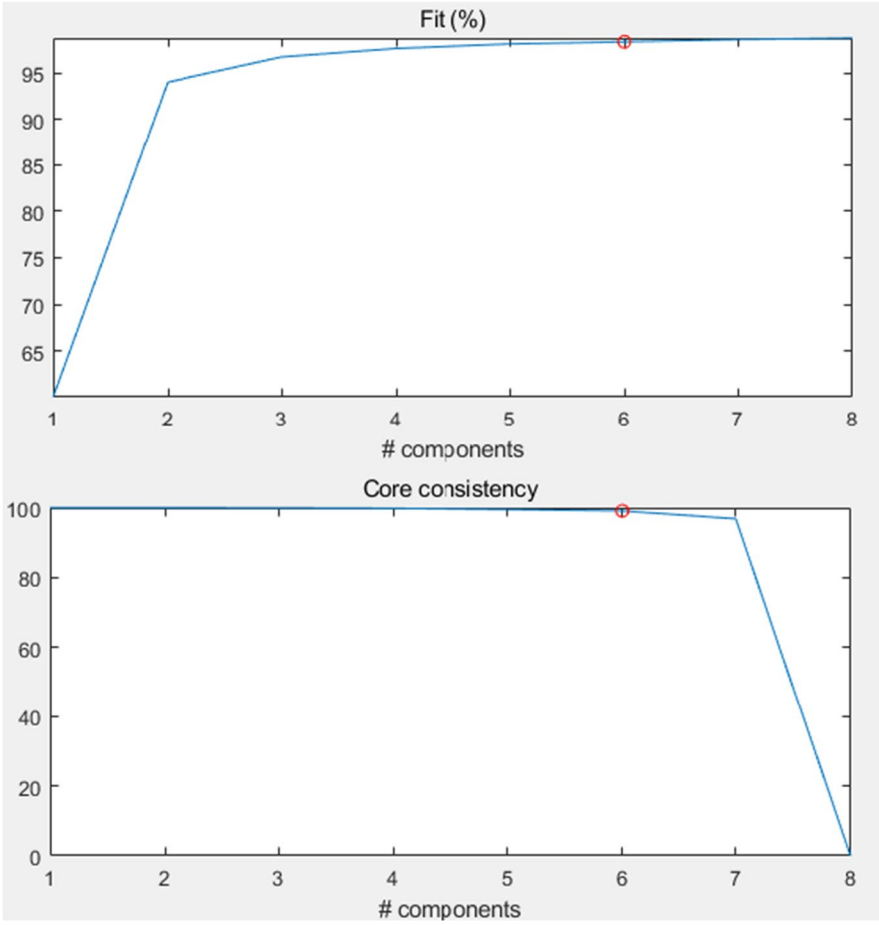

Fig. 4 PARADISe provides measures of core consistency and fit percentage as standard output for each specified interval 3. This information is used to estimate how many components should be extracted

components were corrected for river flow and intensity of internal standards. The 12 extracted components were subjected to trends analysis.

In Fig. 6, the twelve components with their time trends are extracted, which represent unknown chemicals present in the Rhine. A first interesting result is that components 11 and 12 follow similar patterns as the one observed for the internal standard deuterochloroform (cf. Fig. 2). Such similarity indicates a constant concentration in the river water. For a naturally occurring component, this is unlikely as the load may be constant, but not its concentration. Component 11 is even more suspicious as it has concentration measurement of (close to) zero, which is highly unlikely for a natural component. It has been observed that polluters control their waste streams to match the river water flow to avoid detection. Component 11 may be one such pollutant where the zero concentration was related to a break in production.

To investigate whether these 12 components are anthropogenic, the hypothesis tests described in Section 2.5 were applied to these components, and the results can be found in Table 4.

4 Discussion

In the presented work, some assumptions were made regarding patterns in naturally occurring chemicals. When implementing the tests presented in this work, seasonal trends will have to be included if one wants to rely on some of the tests. However, visual inspection of the patterns of suspicious components may give clear indications of seasonal trends. Variations outside of the expectations may still relate to influences like waste streams as the composition of the polluted water can change unpredictably, even when considering seasonal variations. In the future work, simulations could

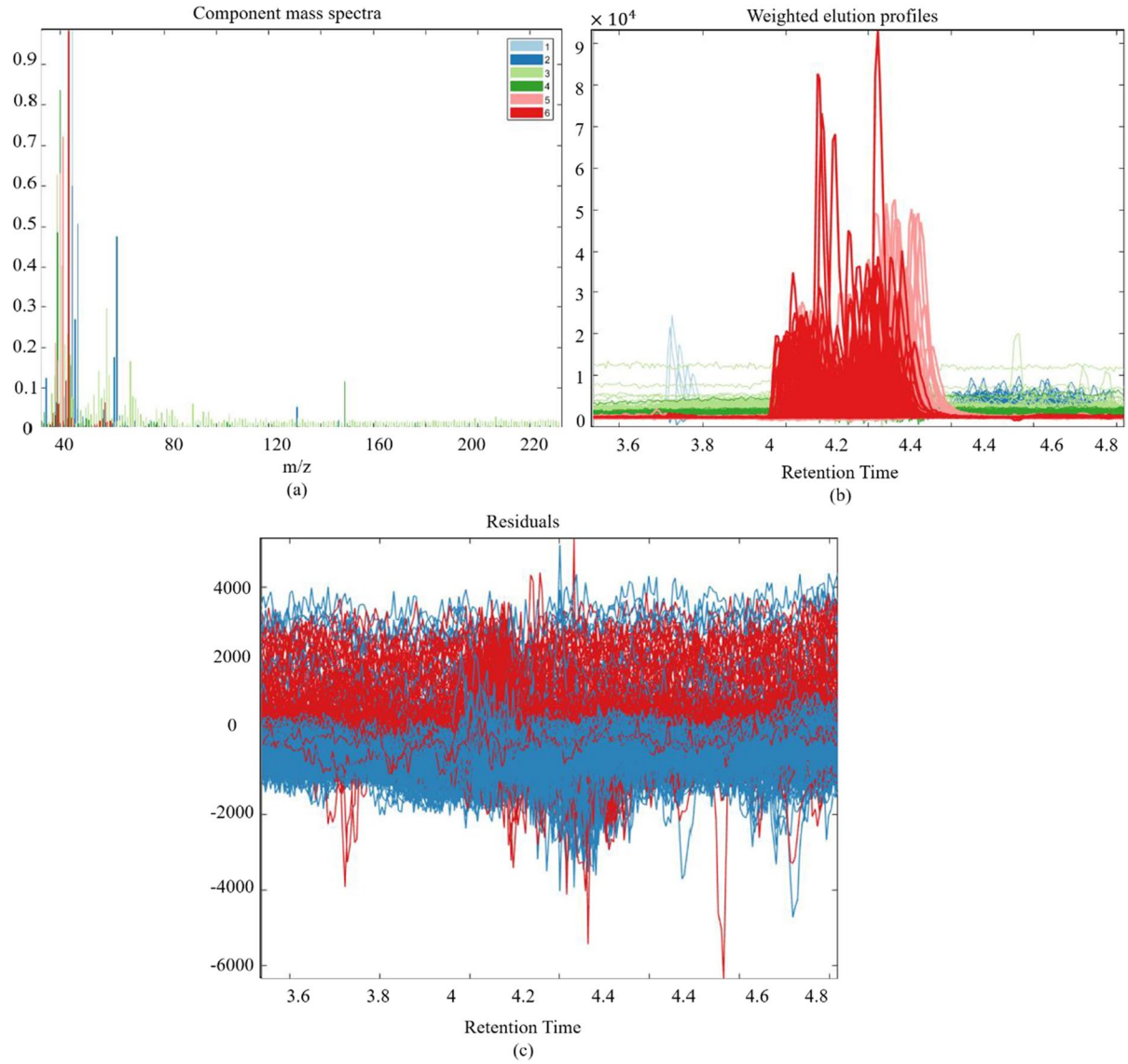

Fig. 5 Detailed information about the component extraction in a particular time interval can be obtained by looking at **a** the mass spectra of the extracted components; **b** the grouping of the chromatograms, now called "weighted elution profiles"; and importantly to (**c**) the residuals after component extraction. Patterns in the residuals are an indication that more components are present in the data

incorporate seasonal trends and other environmental characteristics to make the simulations more realistic.

Selection of retention time windows and number of components is done manually, and these choices will influence the results. Too narrow time windows will reduce the accuracy of the extracted components and increase the computational load, while too wide time window will include too many components, which makes it difficult to separate them. Usually, one peak per retention is ideal. For example, when shifted peaks appear partly in a different retention time window, not all ions belonging to the peak are counted in its TIC. In that case, a slight change in the position of the window would lead to a different TIC. Similarly, the number of components is generally ambiguous, and sometimes it is not clear how many components there should be. Always choosing more components, however, would result in overfitting. An automated procedure (Risum &

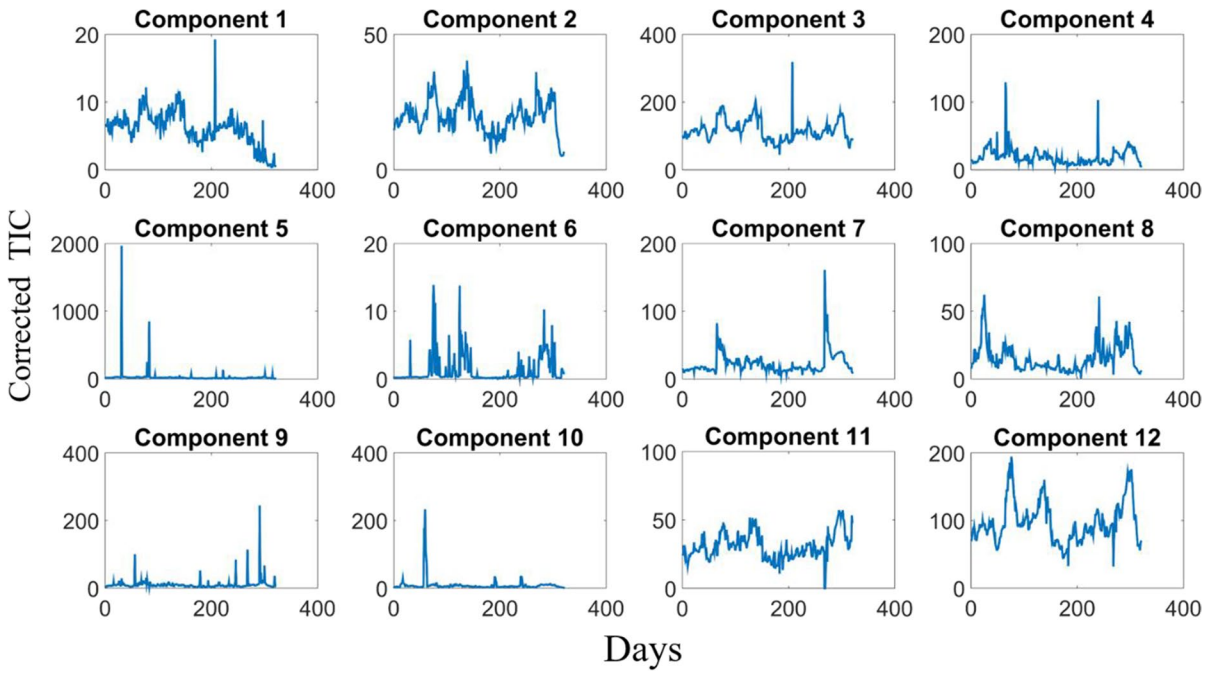

Fig. 6 The trends of twelve components obtained from PARADISe analysis. All the data was corrected for river flow and internal standards. For each subplot, *x*-axis is the data points of each day, while *y*-axis is the corrected TIC

Table 4 Results of the four statistical hypothesis tests, to judge whether the component is anthropogenic. In the test columns, 1 means it failed to pass the test while 0 means pass. In label column, 1 represents anthropogenic and 0 is natural. The last column indicates which component the component is most likely to be according to the NIST look-up. The top-1 percentage is given in parentheses

Component	Test1	Test2	Test3	Test4	Label	Top-1 hit (%)
1	0	1	0	0	1	2-Hexanamine, 4-methyl- (28.57%)
2	0	1	0	0	1	1,3-Benzenediamine, 2,4-dinitro-N3 (44.50%)
3	0	1	0	0	1	Octodrine (23.23%)
4	0	1	0	1	1	Acetic anhydride (37.88%)
5	0	1	1	1	1	Propane, 2-methoxy-2-methyl- (65.59%)
6	0	1	1	1	1	Formic acid, 1,1-dimethylethyl ester (45.65%)
7	0	1	1	1	1	Dodecanal (42.36%)
8	0	1	0	0	1	Oxirane, ethyl- (71.26%)
9	0	1	1	1	1	Benzene (80.30%)
10	0	1	1	1	1	Ethyl Acetate (73.04%)
11	0	0	0	0	0	Trichloromethane (88.71%)
12	0	1	0	0	1	Methylene Chloride (97.39%)

Bro, 2018) for setting retention time windows and the number of components is currently being developed but is, to our knowledge, not yet available publicly.

In the untargeted analysis of GC–MS data, 12 unknown components were extracted from the RWS dataset. Among the tests, the peak test is the strongest indicator of anthropogenic origin, as all the defined anthropogenic components are detected by this test. This indicates that anthropogenic components indeed show bigger fluctuations in concentration than naturally occurring components. A peak test can be the main indicator to monitor the components' concentration over time. As internal standards are relatively constant, if not controlled for river water flow, we must compare the patterns of other components to those of the internal standards. Components 11 and 12 in Fig. 6 follow similar patterns

as the one observed for the internal standard deuterochloroform. Importantly, component 11 has measurement of zero concentrations, which is highly suspected to be anthropogenic. Such pollutant may be difficult to detect if not for the kinds of test presented in this paper. The next step could be to match the extracted PARAFAC2 component to a (online) chemical database. If a match is found, regulatory bodies may choose to further investigate the source of this chemical and identify the presence of it much more rapidly in the future.

The current research was based on one analytical platform used by the responsible regulatory body. When other platforms are available, additional tests could be devised. For example, another method to determine anthropogenic origin might be to evaluate whether a component contains fluoride, an element which is hardly found in naturally occurring components. The data used in this paper did not allow for identification of fluoride in the MS spectrum, but other application may include high-resolution MS measurements if they are available.

Another promising approach is to predict the occurrence of components across multiple measuring stations. While a component may not be exactly identified, the label it gets from the PARADISe analysis may still be used to identify the same component across different measurement stations. Work has been by some of the current authors to relate untargeted GC–MS measurements of different measuring stations along the Rhine using a statistical path model called Process PLS (van Kollenburg et al., 2021). The results of this approach are forthcoming.

Various instruments are widely used to detect organic chemicals, like gas chromatography–mass spectrometry (GC–MS), and liquid chromatography–mass spectrometry (LC–MS). With high sensitivity and good separation capability, GC–MS is widely used in organic chemical detection. It is especially good at quantifying chemicals with low boiling point and good thermal stability. With the successful application of chemometrics in water diagnosis, the same methodology can also be applied to time trends extracted from other commonly used analytical platforms.

5 Conclusion

We have used a PARAFAC2-based method to extract components from water data and have proposed statistical hypothesis tests to judge whether a component in the water is anthropogenic. The method was validated with simulated data and successfully applied to real data with satisfactory results. For empirical data, we evaluated all the targeted components which support the accuracy of the hypothesis. As for the untargeted components in river Rhine, in total, twelve components were identified and only one was recognized as natural while the others were classified as anthropogenic, which provides compelling evidence that studying unknown components can be highly valuable for regulatory bodies, helping them to help focus their attention on the most suspicious pollutants.

Because quantitative chemical analysis can be quite laborious and costly, screening for anthropogenic components can be particularly useful for regulatory bodies. Also, identification of potential unknown components requires more attention to deal with, which increases treatment cost. According to the distribution of mass spectra line and intensity difference, after comparing with the data library, the possibility of certain component will be given and the one with maximum probability is selected.

Furthermore, according to the component characteristics, like m/z constitution and intensity, we could tentatively identify the chemical it represents. In the future, we could track back the location of discharge and help regulators deal with pollution with more data like factory distribution and components' concentration distribution, besides river water flow.

Funding This work was supported by NWO-TA-COAST project "Outfitting the Factory of the Future with Online analysis" (grant 053.21.114) and the National Key Research and Development Program of China 2019YFC0118202. Yangwei Ying is supported by China Scholarship Council grant 201706320265. Part of this work was done supported by the ECSEL Joint Undertaking (JU) under grant agreement No 826589.

Data Availability The data that support the findings of this study are available from RIWA-Rijn. Restrictions apply and the data is not publicly available but can be made available from the authors upon reasonable request and with permission of RIWA-Rijn.

Declarations

Conflict of Interest The authors declare no competing interests.

Appendix 1. Targeted Anthropogenic Volatile Chemicals

Table 5 Thirty anthropogenic chemicals with five internal standards, monitored multiple times a day by Rijkswaterstaat

Internal standards
Deuterochloroform
Toluene-D8
Chlorobenzene-D5
1,4-Dichlorobenzene-D4
Naphthalene-D8
Anthropogenic chemicals
Methyl tertiary-butyl ether (MTBE)
Diisopropyl ether
Cis-1,2-Dichloroethene
Ethyl tertiary-butyl ether (ETBE)
Chloroform
Ethyl sec-butyl ether (ESBE)
1,1,1-Trichloroethane
Cyclohexane
1,2-Dichloroethane
Benzene
Tertiary amyl methyl ether (TAME)
Trichloroethylene
Tertiary amyl ethyl ether (TAEE)
Methylisothiocyanate
Toluene
1,1,2-Trichloroethane
Tetrachloroethylene
Chlorobenzene
Ethylbenzene
m/p-Xylene
o-Xylene
Styrene
Cumene
n-Propylbenzene
2-Chlorotoluene
t-Butylbenzene
1,2,4-Trimethylbenzene
1,2-Dichlorobenzene
Hexachlorobutadiene
Naphthalene

Table 5

Appendix 2. Correcting for Internal Standards and Water Flow

The available data on internal standards from Bimmen comprised 5745 measurement points of five internal standards; let $C_{1,1}$ represent the first measurement of the first internal standard, $C_{1,2}$ as the first measurement of the second internal standard, etc. then the data matrix can be represented as

$$\mathrm{BIM}_{5745*5} = \begin{bmatrix} C_{1,1} & \cdots & C_{1,5} \\ \vdots & \ddots & \vdots \\ C_{5745,1} & \cdots & C_{5745,5} \end{bmatrix}_{5745*5} \tag{4}$$

Then, for each column of the internal standards, the standard deviation was calculated (5).

$$std_n = std\begin{pmatrix} C_{1,n} \\ \vdots \\ C_{5745,n} \end{pmatrix} = \begin{bmatrix} \sqrt{\frac{1}{5745}\sum_{n=1}^{5745}\left(C_{n,1} - \overline{C}_1\right)^2} \\ \cdots \sqrt{\frac{1}{5745}\sum_{i=1}^{5745}\left(C_{n,5} - \overline{C}_5\right)^2} \end{bmatrix} \tag{5}$$

A total of five standard deviations were calculated (6).

$$\mathrm{STD}_{\mathrm{BIM}} = \begin{bmatrix} \mathrm{std}_1 & \cdots & \mathrm{std}_5 \end{bmatrix}_{5*1} \tag{6}$$

Next, each element of the BIM matrix (4) was divided by the standard deviation of its respective column (Equ. 6), resulting in 5745 scaled values per internal standard, resulting in (with abuse of notation):

$$\mathrm{BIM}^*_{5745\times5} = \frac{\mathrm{BIM}_{5745\times5}}{\mathrm{STD}_{\mathrm{BIM}}} = \begin{bmatrix} C^*_{1,1} & \cdots & C^*_{1,5} \\ \vdots & \ddots & \vdots \\ C^*_{5745,1} & \cdots & C^*_{5745,5} \end{bmatrix} \tag{7}$$

To get the corrected concentration, $\mathrm{Conc}^*_{n,x}$, of an observed chemical x in the river water at time-point n, each measured concentration $\mathrm{Conc}_{n,x}$ was divided by the average of the 5 corrected internal standards (see 8 and 9). For the corrected data, the last step was done by dividing the data by the water flows at time n to obtain loads instead of concentrations (10). The final corrected data consisted out of 5745 estimated loads per component. The tests were performed on the loads from (Equ. 10).

$$\bar{C}^*_n = \mathrm{mean}\left(C^*_{n,1} + C^*_{n,2} + C^*_{n,3} + C^*_{n,4} + C^*_{n,5}\right) \tag{8}$$

$$\mathrm{Conc}^*_{n,x} = \frac{\mathrm{Conc}_{n,x}}{\bar{C}^*_n} \tag{9}$$

$$\mathrm{Load}_{n,x} = \frac{\mathrm{Conc}^{*}_{n,x}}{\mathrm{Water\ flow}_{n}} \tag{10}$$

Because internal standards have a fixed concentration added into every water sample. The average of all the internal standards will the most robust way to calibrate other components.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Bro, R., Andersson, C. A., & Kiers, H. A. L. (1999). PARAFAC2-Part II. Modeling chromatographic data with retention time shifts. *Journal of Chemometrics, 13*, 295–309.

Campo, E., Ferreira, V., López, R., Escudero, A., & Cacho, J. (2006). Identification of three novel compounds in wine by means of a laboratory-constructed multidimensional gas chromatographic system. *Journal of Chromatography A, 1122*, 202–208.

Diehl P., Gerke T., Jeuken A., Lowis J., Steen R., van Steenwijk J., Stoks P. and Willemsen H. (2005). Early warning strategies and practices along the river Rhine. The Rhine. Springer, Berlin, Heidelberg, pp 99-124

EC (2000). Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for Community action in the field of water policy. OJ L327, 22.12.2000.

García, S., Molina, D., Lozano, M., & Herrera, F. (2009). A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: A case study on the CEC' 2005 Special Session on Real Parameter Optimization. *Journal of Heuristics, 15*, 617–644.

Hering, D., Borja, A., Carstensen, J., Carvalho, L., Elliott, M., Feld, C. K., Heiskanen, A., Johnson, R. K., Moe, J., & Pont, D. (2010). The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. *Science of the Total Environment, 408*, 4007–4019.

Hollender, J., Schymanski, E. L., Singer, H. P., & Ferguson, P. L. (2017). Nontarget screening with high resolution mass spectrometry in the environment: Ready to go? *Environmental Science and Technology, 51*, 11505–11512.

Johnsen, L. G., Skou, P. B., Khakimov, B., & Bro, R. (2017). Gas chromatography – mass spectrometry data processing made easy. *Journal of Chromatography A, 1503*, 57–64.

Kamstrup-Nielsen, M. H., Johnsen, L. G., & Bro, R. (2013). Core consistency diagnostic in PARAFAC2. *Journal of Chemometrics, 27*, 99–105.

Kiers, H. A. L., Berge, J. M. F. T., & Bro, R. (1999). PARAFAC2-Part I. A direct fitting algorithm for the PARAFAC2 model. *Journal of Chemometrics, 13*, 275–294.

Loos, R., Gawlik, B. M., Locoro, G., Rimaviciute, E., Contini, S., & Bidoglio, G. (2009). EU-wide survey of polar organic persistent pollutants in European river waters. *Environmental Pollution, 157*, 561–568.

Pena-Abaurrea, M., Jobst, K. J., Ruffolo, R., Shen, L., McCrindle, R., Helm, P. A., & Reiner, E. J. (2014). Identification of potential novel bioaccumulative and persistent chemicals in sediments from Ontario (Canada) using scripting approaches with GC×GC-TOF MS analysis. *Environmental Science and Technology, 48*, 9591–9599.

Risum A.B. and Bro R. (2018). Fully automated PARAFAC2 based analysis of GC-MS data. The 17th Chemometrics in Analytical Chemistry Conference, Halifax, Canada

Ruff, M., Mueller, M. S., Loos, M., & Singer, H. P. (2015). 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 Research, 87*, 145–154.

Schlüsener, M. P., Kunkel, U., & Ternes, T. A. (2015). Quaternary triphenylphosphonium compounds: A new class of environmental pollutants. *Environmental Science and Technology, 49*, 14282–14291.

van Kollenburg, G., Bouman, R., Offermans, T., Gerretzen, J., Buydens, L., van Manen, H., & Jansen, J. (2021). Process PLS: Incorporating substantive knowledge into the predictive modelling of multiblock, multistep, multidimensional and multicollinear process data. *Computers & Chemical Engineering, 154*, 107466.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.