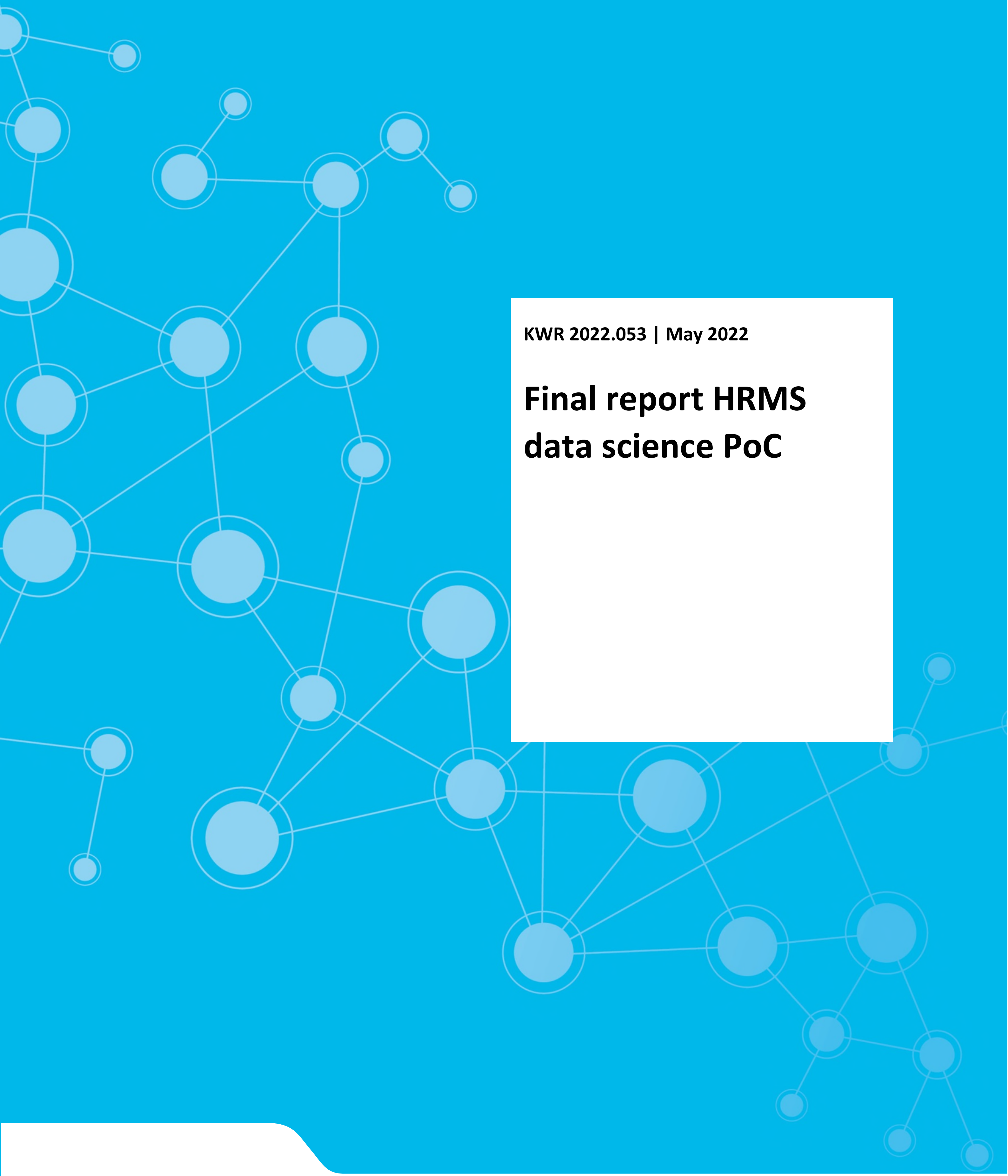

KWR 2022.053 | May 2022

# Final report HRMS data science PoC

Bridging Science to Practice

# Report

## Final report HRMS data science PoC

## KWR 2022.053 | May 2022

**Project number**
403817/001

**Project manager**
ing. Ton van Leerdam

**Client**
Rijkswaterstaat

**Author(s)**
Dr. Frederic Béen, Nienke Meekel MSc

**Quality Assurance**
dr. Peter van Thienen, Erik Emke BSc

**Sent to**
Marcel Kotte

This report is not a public document and is only provided to the client. KWR will refrain from distributing this report outside the client organisation and will therefore not provide the report to third parties, unless KWR and the client agree otherwise. The client is entitled to distribute the report subject to KWR's prior consent. KWR may attach conditions to consent to the dissemination of (parts of) the report.
Procedures, calculation models, techniques, designs of trial installations, prototypes and proposals and ideas put forward by KWR, as well as instruments, including software, that are included in research results are and remain the property of KWR.

**Keywords**

mass spectrometry, trend analysis

Year of publishing
2022

More information
Dr. Frederic Béen
T +31 30 606 9748
E frederic.been@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

# Managementsamenvatting

## *Proof of Concept voor analyse bestaande data uit verschillende water monitoring programma's – praktijkgebruik kan starten*

**Auteur(s)** Dr. Frederic Béen, Nienke Meekel MSc.

In opdracht van Rijkswaterstaat heeft KWR onderzoek gedaan naar de analyse van databestanden uit verschillende meetinstrumenten en – programma's voor onderzoek van watermonsters. Daarvoor is een eenduidige, schaalbare en overdraagbare workflow ontwikkeld. Er zijn nu twee proofs of concept of PoC's: een voor hoge-resolutie-data en een voor lage-resolutie-data, waarmee data-analisten (na een uitgebreide demonstratie) nu al bestaande gegevens kunnen analyseren om nieuwe opkomende verontreinigingen op te sporen en te identificeren.

### Belang: meer inzicht in de aanwezigheid van onbekende stoffen

Steeds vaker en op steeds grotere schaal worden monsters uit het aquatische milieu gescreend, vooral om probleemstoffen op te sporen, maar ook om de aanwezigheid van nog onbekende stoffen te detecteren. Rijkswaterstaat en andere partijen hebben daarvoor specifieke meetprogramma's opgezet. De gegevens worden bewaard en er bestaat een behoefte om deze verder te analyseren. De data kunnen bijvoorbeeld meer inzicht geven in de aanwezigheid van onbekende stoffen in (oppervlakte)water dan uit de individuele metingen voortkomt.

### Doel: Proof of Concept (PoC) voor uitgebreide data-analyse

In opdracht van Rijkswaterstaat heeft KWR onderzoek gedaan naar de analyse van databestanden uit verschillende meetinstrumenten en – programma's, met het doel een proof of concept te leveren voor de analyse van chromatografische en massaspectrometrische gegevens. De nadruk moest daarbij liggen op het ontwikkelen van een strategie die het mogelijk maakt chemische signalen (features) voorlopig te identificeren en relevante (voorheen onbekende) features te detecteren in tijd en plaats. Dit draagt bij aan de detectie van calamiteiten en van opkomende stoffen die geen onderdeel zijn van reguliere monitoring.

### Aanpak: benaderingen voor hoge- en lage-resolutie-data

Er zijn twee benaderingen toegepast en dus twee PoC's opgesteld: een voor zogenaamde lage-resolutie massaspectrometriegegevens (SPE-GC-MS, een veelgebruikte bewakingstechniek) en een tweede voor hoge-resolutie massaspectrometriegegevens (LC-HRMS, een opkomende techniek die SPE-GC-MS potentieel kan (deels) gaan vervangen). Dit was nodig wegens de verschillen tussen lage- en hoge-resolutiegegevens. Zo ontstonden twee PoC's, die een algemene gemeenschappelijke strategie delen, beide ontwikkeld volgens een stapsgewijze aanpak:

(i) invoer en filteren van de gegevens,
(ii) normalisatie,
(iii) verkennende analyse,
(iv) identificatie,
(v) validatie en
(vi) rapportage.

**Resultaten: twee PoC’s voor verkennende analyse en trendanalyse van onbekende stoffen in water**
Met de PoC’s voor respectievelijk hoge- en lage-resolutie-data kunnen gebruikers gegevens verwerken volgens een eenduidige, schaalbare en overdraagbare workflow. Zij krijgen hulp bij de selectie (prioritering) en identificatie van relevante features. Deze twee PoC’s vormen een eerste platform dat kan worden aangepast en uitgebreid op basis van feedback van gebruikers en hun specifieke behoeften/vragen. Inzet in de praktijk kan worden gebruikt om de praktische haalbaarheid van dit data analyseconcept te bepalen.

Beide ontwikkelde methodes zorgen ervoor dat meerdere metingen met elkaar vergeleken kunnen worden. Met verschillende statistische technieken (o.a. PCA, HCA) worden metingen gegroepeerd en kunnen afwijkende monsters worden gedetecteerd. Met behulp van trendanalyses kunnen onbekende stoffen die over de tijd toe- of afnemen worden geprioriteerd. Vervolgens kunnen de geprioriteerde stoffen voorlopig geïdentificeerd worden door het massaspectrum te screenen met verschillende databases. Zo is het mogelijk incidentele calamiteiten en trends te traceren.

**Toepassing: in de praktijk testen; aanbevelingen**
De ontwikkelde PoC’s vormen een eerste stap in de richting van de toepassing van de analyse van data uit verschillende meetprogramma’s of -systemen. Beide open source methoden zijn gepresenteerd tijdens een afsluitende demo. Na een uitgebreide demonstratie kunnen data-analisten de PoC’s in hun huidige staat van ontwikkeling gebruiken om bestaande gegevens te analyseren om nieuwe opkomende verontreinigingen op te sporen en te identificeren. In een volgende stap kunnen gebruikers de ontwikkelde strategieën testen, ervaring opdoen en feedback geven, zodat de werkwijze verder kan worden verbeterd.

Tijdens de ontwikkeling van de PoC’s zijn diverse aanbevelingen geformuleerd die de implementatie van deze aanpak kunnen bevorderen. Voor SPE-GC/MS data betreft dit voornamelijk het uitvoeren van blanco metingen en metingen in triplo. Voor LC-HRMS data betreft dit voornamelijk een hogere frequentie (meer monsters in de tijd), zodat eventuele trends eenvoudiger te detecteren zijn: een hogere meetfrequentie maakt de gegevensanalyse over het algemeen robuuster.

**Rapport**
Dit onderzoek is beschreven in het rapport *Final report HRMS data science PoC* (KWR 2022.053).

Jaar van publicatie
2022

Meer informatie
Dr. Frederic Béen
T +31 (0)30 606 97048
E Frederic.Been@kwrwater.nl

Keywords
mass spectrometry, trend analysis

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

# Management summary

## *Proof of Concept for analysis of existing data from different water monitoring programs – practical use can start*

**Author(s)** Dr. Frederic Béen, Nienke Meekel MSc.

KWR was commissioned by Rijkswaterstaat to investigate possibilities to analyse data generated from different analytical instruments and monitoring programs used to monitor water quality. For this purpose, a uniform, scalable and transferable workflow was developed. Two poofs of concept or PoCs were developed: one for high-resolution data and one for low-resolution data, which allow data analysts (after an extensive demonstration) to analyse existing data to detect and identify new emerging contaminants.

**Importance: more insight into the presence of unknown substances**

Samples from the aquatic environment are being screened more often and on a larger scale, mainly to detect problematic substances, but also to detect the presence of previously unknown substances. Rijkswaterstaat and other parties have set up specific monitoring programs for this purpose. The generated data are stored and there is a need to further analyse it. This data can, for example, provide more insight into the presence of unknown substances in (surface) water than can be obtained from individual measurements.

**Goal: Proof of Concept (PoC) for comprehensive data analysis**

KWR was commissioned by Rijkswaterstaat to conduct research into the analysis of data from various instruments and monitoring programs, with the aim of delivering a proof of concept for the analysis of chromatographic and mass spectrometric data. The focus was to develop a strategy that makes it possible to tentatively identify chemical signals (features) and to detect relevant (previously unknown) features based on their spatial and temporal occurrence. This contributes to the detection of calamities and of emerging substances that are not part of regular monitoring.

**Approach: workflows for high- and low-resolution data**

Two approaches have been used and thus two PoCs were developed: one for so-called low-resolution mass spectrometry data (SPE-GC-MS, a widely used monitoring technique) and a second one for high-resolution mass spectrometry data (LC-HRMS, an emerging technique that could potentially (partly) replace SPE-GC-MS). This was necessary because of the differences between low and high-resolution data. Thus, two PoCs were developed, sharing a general common strategy, both based on a stepwise approach:

(i) input and filtering of the data,
(ii) normalisation,
(iii) exploratory analysis,
(iv) identification,
(v) validation and
(vi) reporting.

### Results: two PoCs for exploratory analysis and trend analysis of unknown compounds in water

With the PoCs for high- and low-resolution data respectively, users can process combined data according to a unified, scalable and transferable workflow. They receive assistance in selecting (prioritizing) and identifying relevant features. These two PoCs form an initial platform that can be adapted and extended based on users' feedback and their specific needs/questions. Implementation in practice can be used to determine the practical feasibility of this data analysis concept.

Both developed methods ensure that analysis results can be compared. With various statistical techniques (e.g. PCA, HCA), measurements are grouped and deviating samples can be detected. Trend analyses can be used to prioritise unknown substances that increase or decrease over time. Subsequently, the prioritised substances can be tentatively identified by screening the mass spectrum with different databases. This makes it possible to trace incidental calamities and trends.

### Application: testing in practice; recommendations

The developed PoCs are a first step towards the application of the analysis of data from different monitoring programs or instruments. Both open source methods were presented during a closing demo. After an extensive demonstration, data analysts can use the PoCs in their current state of development to analyse existing data to detect and identify new emerging contaminants. In a next step, users can test the developed strategies, gain experience and provide feedback, so that the method can be further improved.

During the development of the PoCs, several recommendations were formulated that could promote the implementation of this approach. For SPE-GC/MS data, this mainly concerns performing blank measurements and measurements in triplicate. For LC-HRMS data, this mainly concerns a higher frequency (more samples in time), so that potential trends are easier to detect: a higher measurement frequency generally makes the data analysis more robust.

### Report

This research is described in the report *Final report HRMS data science PoC* (KWR 2022.053).

Year of publication
2022

More information
Dr. Frederic Béen
T +31 (0)30 606 97048
E Frederic.Been@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

**Keywords**
mass spectrometry, trend analysis

# Contents

# List of abbreviations & terminology

| | |
|---|---|
| BPC | Base Peak Chromatogram |
| EI | Electron impact ionization |
| ESI | Electrospray ionization |
| Feature | Refers to each component or peak (i.e., ion) detected during the analysis and generally consists of three parameters, namely its retention time (in minutes or seconds), its intensity (i.e., the height or area of the chromatographic peak) and, in the case of high-resolution mass spectrometry, of its accurate mass. |
| GC-MS | Gas Chromatography-Mass Spectrometry |
| HCA | Hierarchical Clustering Analysis |
| HRMS | High Resolution Mass Spectrometry |
| HWL | Het Waterlaboratorium N.V. |
| IS | Internal standard |
| KWR | KWR Water Research Institute |
| LC-MS | Liquid Chromatography-Mass Spectrometry |
| MinIenW | Dutch Ministry of Infrastructure and Water Management |
| MSC | Multiplicative Scatter Correction |
| m/z | Mass-to-charge ratio |
| PCA | Principal Component Analysis |
| PoC | Proof of Concept |
| Rt | Retention time |
| Rt-index | Retention time normalised using IS |
| RWS | Rijkswaterstaat |
| SNV | Standard Normal Variate |
| SPE | Solid Phase Extraction |
| Spectral Similarity | A function, generally involving the calculation of the scalar product of two vectors, used to determine the similarity between two mass spectra. |
| TIC | Total Ion Chromatogram |

# 1 Introduction

This project focused on developing a data analysis approach for Rijkswaterstaat (RWS) which can be applied to low- and high-resolution mass spectrometry data collected using gas and liquid chromatography (GC-MS and LC-HRMS). This project had two major goals with respect to the analysis of data:

(i) development of a strategy that allows to **tentatively identify features** (i.e., compounds detected during the analysis, but which have not been formally identified yet using reference standards);

(ii) develop a data analysis strategy that allows to detect relevant (and previously unknown) features based on relevant **temporal and/or geographical patterns** (e.g., increasing trends) and tentatively identify them.

Two "proof of concept" (PoCs)[1] workflows which achieve these goals were developed, one for low-resolution (referred to as instruments which have a mass resolution between 0.5 and 1 Da, in this case GC-MS) and one for high-resolution (referred to as instruments with a higher mass resolution, in the mDa range, in this case LC-HRMS) mass spectrometry data. In fact, as will be discussed in detail below, although similar approaches can be implemented, the nature of low- and high-resolution data requires the development of distinct solutions. According to the tender call and project proposal, the project consisted of 5 sequential steps, referred to as "iterations", which are summarised in Table 1.

*Table 1: Overview of project phases as described in the initial project proposal.*

| Project phase/Part | Main activities/outcomes | Start date | End date |
|---|---|---|---|
| 1. Up to the first iteration | • Inventory of data<br>• Overview of already available strategies/methods/algorithms<br>• Define strategy for the workflow | Q2 2021 | Q3 2021 |
| 2. During the first iteration | • Workflow execution<br>• First iteration of data analysis | Q3 2021 | Q4 2021 |
| 3. During the second iteration | • Extend data analysis<br>• Include temporal and spatial component<br>• Data & requirements outside RWS | Q3 2021 | Q4 2021 |
| *Go / No Go moment* | | | |
| 4. During the third iteration | • Further extend the data analysis<br>• Detection/identification and (ii) prioritisation<br>• First description of the PoC | Q4 2021 | Q4 2021 |
| 5. After the third iteration | • First complete and working PoC | Q4 2021 | Q4 2021 |

It should be noted that due to the circumstances, in particular the fact that high-resolution data was not initially available and that it was decided to work on low-resolution data first, deviations were made from the iterations-based approach detailed in Table 1. Despite these changes, the overall strategy applied, and outputs remained in line with the goals described above.

[1] It should be noted that in the context of this work, terms such as **proof of concept (PoC), workflow or data analysis strategy** are used as synonyms, and all refer to the development of a data analysis approach allowing to process and evaluate low- and high-resolution mass spectrometry data. Proof of concept is being used as a term as this was also mentioned in the initial tender published by Rijkswaterstaat.

# 2 Low-resolution GC-MS analysis

## 2.1 Objective

The PoC which was specifically developed for GC-MS data focused on the development of a strategy to import, process and analyse data with the goal of detecting samples which deviate from normal patterns (e.g., calamities) and the features which might be causing the observed deviation, as well as to (tentatively) identify them using existing databases.

## 2.2 GC-MS data

The developed PoC for low-resolution GC-MS data relied on data collected from the monitoring stations of Bimmen and Lobith (located along the river Rhine). Both locations are equipped with an online solid-phase extraction (SPE-)GC-MS instrument. Water samples are collected over 12 hours (composite) and then an aliquot is extracted with SPE and analysed. Prior to analysis, samples are spiked with 8 stable isotopically labelled internal standards (IS), which are used for quantification and quality control. These play an important role in the development of the PoC (see Annex III for a list of the used internal standards). It should be noted that, for unknown reasons, not all internal standards could always be found in the chromatograms. Consequently, there are instances in which the PoC was implemented using a selection of IS. Furthermore, solvent or procedural blanks (i.e., including of the SPE system) were not available. As will be discussed further, this has important implications for the developed strategies and should be included in future.

## 2.3 Data analysis strategy

A multi-step strategy was developed for the processing and analysis of the data. This approach is based on commonly implemented strategies used to process and analyse large amounts of mass spectrometry data (1,2). An overview of the approach is illustrated in Figure 1. The various steps are described in more detail in the following paragraphs. The entire data analysis strategy described in this report was developed using R 4.0.2 (R Core Team, 2020) and RStudio (3) and several packages, which are described in Annex V.

Data analysis in mass spectrometry commonly uses the term “feature” which is used to refer to ions (i.e., molecules) detected during the analysis. A feature generally consists of three elements, namely its chromatographic retention time (Rt), its intensity (either peak area or height) and, in the case of HRMS instruments, its accurate mass.

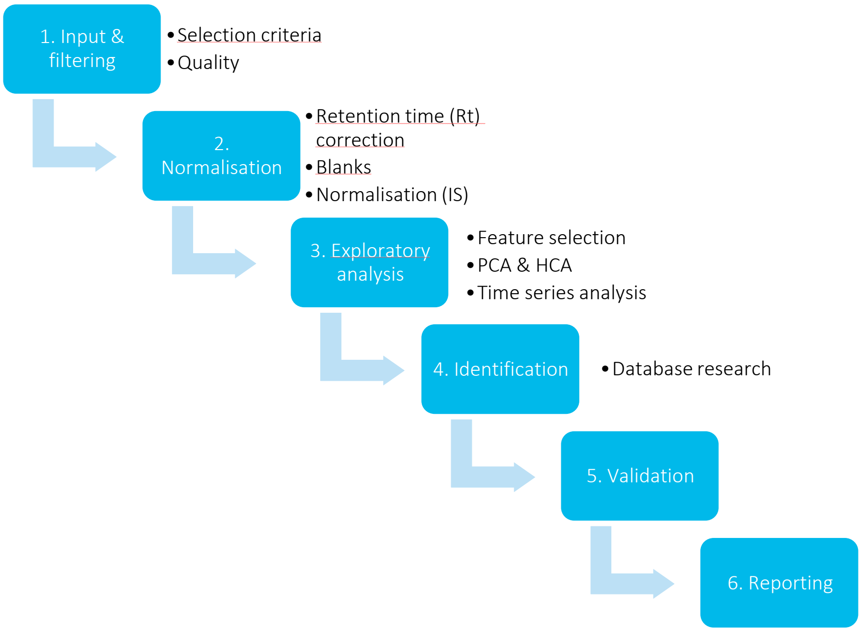

*Figure 1: Overview of the data analysis strategy. PCA = principal component analysis; HCA = hierarchical cluster analysis*

### 2.3.1 Input and filtering

The first step consisted of selecting relevant chromatograms from the large amount of data available. In fact, given that data is collected with an online system, not all recorded chromatograms are relevant (i.e., some of the data refers to calibration standards, controls, etc. which are not relevant for the development of the PoC). For filtering purposes, KWR uses the "Qxmax file", an inventory of all analyses done by the instrument and which allows to distinguish samples from other (not relevant) data files.

Data was imported using the R-package *Rawrr (4)*, which has been developed to import and read *.raw* files created by instruments of the brand *Thermo Fisher Scientific* (brand of the online GC-MS system used by RWS). The package is open source (as are all packages developed for RStudio). Alternative packages, such as *XCMS (5)* and *enviGCMS (6)*, are available for the import and (pre-)processing of data generated with instruments from other manufacturers. The import and filtering steps were implemented in a fully functional R script that allows for raw online SPE-GC-MS data to be imported and converted into a format that can be used in RStudio. Below, an example of a chromatogram imported with the *Rawrr* package is illustrated (Total Ion Chromatogram (TIC; left) and base peak plot (BPC, right)), which were visualized using the developed script in RStudio (see Figure 2).

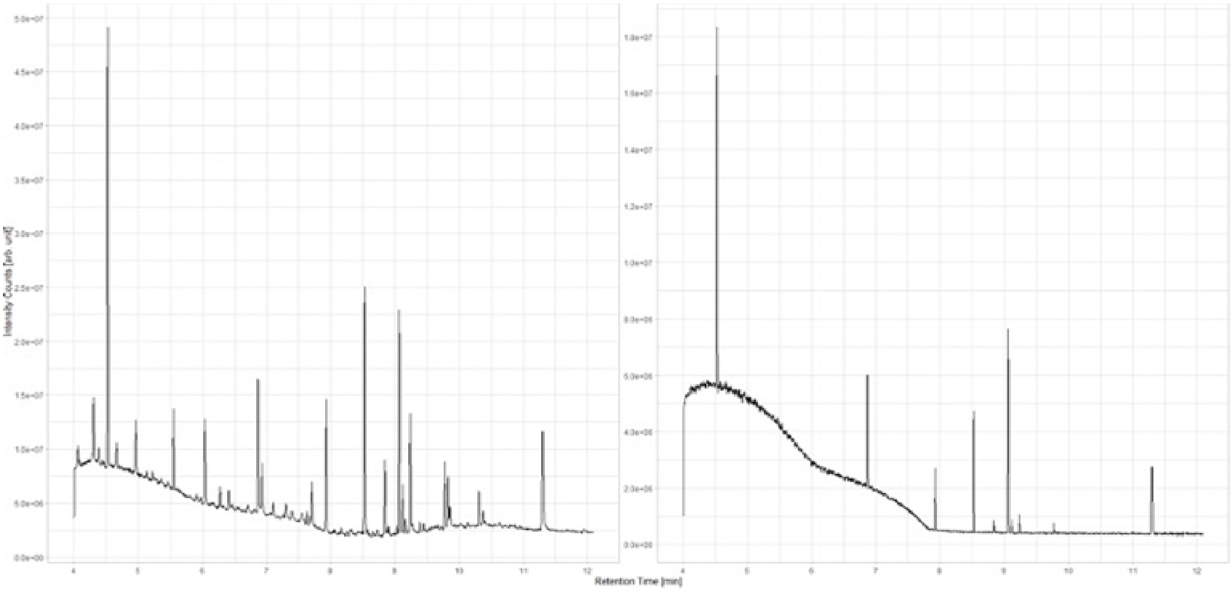

*Figure 2: Illustration of an imported chromatogram (TIC, left) and a base peak plot (right). The latter corresponds to the most intense ion (i.e., m/z) detected at each point of the analysis. Both TIC and BPC have been made with the above-mentioned R-script and the Rawrr package.*

### 2.3.2 Normalisation

Prior to data analysis, in particular to identify temporal and/or geographical patterns, it is crucial to ensure that variability in data is solely due to actual differences in sample composition and are not due to instrumental variability (e.g., natural fluctuations in detector response, ageing of the chromatographic columns, maintenance). Although the influence of external factors cannot be completely excluded, it is important to minimize it so that true patterns can be detected. In the specific case of MS data (both low and high-resolution), two main factors caused by instrumental variability can affect the results, namely changes in retention time (Rt) and intensity (or peak area). In the case of high-resolution data, deviations in the measurement of the accurate mass of detected features (also referred to as "mass drift") is a third important instrumental factor which needs to be accounted for. In the specific case of low-resolution data, this is considered less of an issue as potential errors in measured masses are assumed to fall within the accuracy of the instrument. Compared to high-resolution mass spectrometry, normalisation is more complex when working with low-resolution data, in particular, because it lacks information about the accurate mass of the detected features, which can be used to align chromatographic signals. In the case of low-resolution data, alignment has to be applied to the whole signal (or Total Ion Chromatogram, TIC) before features can be selected for further analysis. This is an important difference which substantially complicates the alignment process of low-resolution data.

#### 2.3.2.1 Retention time correction

To correct for shifts in retention time (Rt), which can occur due to random variations caused by the chromatographic separation, and ensure that only true differences in data are detected, an approach based on the calculation of retention time indexes was implemented (7). For this purpose, the retention times of internal standards (present in each chromatogram) were retrieved and used to calculate Rt-indexes for all detected features. This is done using the following formula (derived from the original approach developed by Kováts (7)):

$$Rt_y = Rt_A + \frac{(Rt_B - Rt_A)}{(Rt'_B - Rt'_A)}(Rt_x - Rt'_A)$$

where $Rt_y$ corresponds to the corrected retention time, which is calculated from the original retention time ($Rt_x$), fixed retention times for the internal standards $Rt_A$ and $Rt_B$ (derived from a user defined reference chromatogram) and the measured (actual) retention times from the internal standards $Rt'_A$ and $Rt'_B$. This non-linear approach was implemented per window, meaning that the Rt-index of each feature in the chromatogram was calculated using the

closest two internal standards (eluting left and right of the feature in question). Figure 3 illustrates the impact that retention time correction has on a chromatogram.
Because this approach relies on the Rts of internal standards, these have to be obtained for each chromatogram. However, due to the large amount of data available, this cannot be done manually. Hence, a dedicated algorithm was developed, allowing to automatically retrieve the Rts of all internal standards in each chromatogram. The algorithm relies on the specific MS-spectrum of the internal standard. In particular, a reference MS-spectrum, the expected Rt of the internal standard and a Rt search window are defined by the operator. Based on the parameters provided by the user, the algorithm uses a spectral similarity function (i.e., generally the scalar product, which allows to calculate the similarity between two mass spectra) to find which of the recorded spectra in the window matches the one defined by the user.

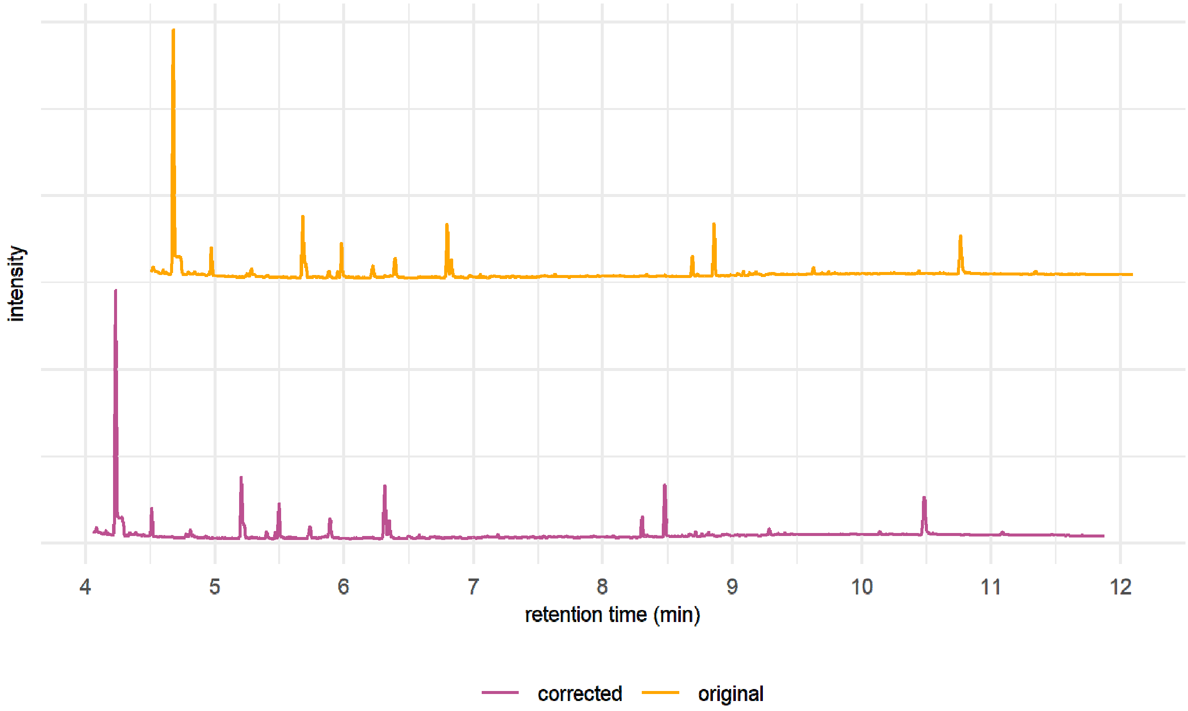

*Figure 3: Illustration of an original TIC (top) and a retention time corrected TIC (bottom) using the retention times of the internal standards.*

In addition to this ad-hoc approach, KWR investigated the possibility of implementing alternative approaches to correct for retention time shifts, namely "step-wise alignment" and "dynamic time warping". However, when compared to the IS-based approach, it becomes clear that the latter two approaches are less adequate. The reason for this is that these approaches use a reference signal to correct for retention time shifts and are less capable of accounting for large shifts as the ones observed in the available data. These two alternative techniques are more useful when chromatograms to be aligned have been measured in the same run and are hence affected only by limited Rt shifts, or in cases where no internal standards are available. However, the data which is being considered here covers multiple years and observed Rt shifts can be substantial (> 2 minutes for the same IS) and there is no adequate reference signal to be used.

The use of Kováts retention time indexes is also an interesting alternative (7). This approach relies on the same equation presented above but uses pre-defined alkanes as reference standards to calculate corrected Rt (instead of mass labelled reference standards as implemented here). However, these reference alkanes need to be analysed at regular intervals to determine their Rt, which is not part of the method currently used by RWS to monitor water quality in Bimmen and Lobith. For future analyses and optimisation of the SPE-GC-MS method, KWR strongly advises RWS to include the measurement of reference alkanes at regular intervals, as the obtained Rt-indexes can also be used for identification purposes. For the sole purpose of alignment, however, the use of alkanes has no added value compared to the IS-based approach implemented here.

### 2.3.2.2 Intensity normalisation and baseline correction

The second step in the process comprises intensity normalisation and baseline correction. This step is important because baseline shifts (i.e., changes in the overall intensity of a chromatogram) can occur due to the status of the instrument, in particular column ageing and detector sensitivity, or the composition of the matrix. For this purpose, various approaches have been implemented, such as *standard normal variate* (SNV) and *multiplicative scatter correction* (MSC). The former is a straightforward “centre and scale” approach, while the latter uses a reference chromatogram to correct for intensity and baseline shifts. However, as mentioned previously, the use of reference signals in the context of this data might be problematic. These approaches were compared but none of the two seemed to outperform the other. Therefore, both approaches were implemented in the PoC and it is advised to select the best approach case by case, depending on the dataset and by testing the grouping/outputs obtained with each approach. An example of an intensity normalised chromatogram is shown in Figure 4.

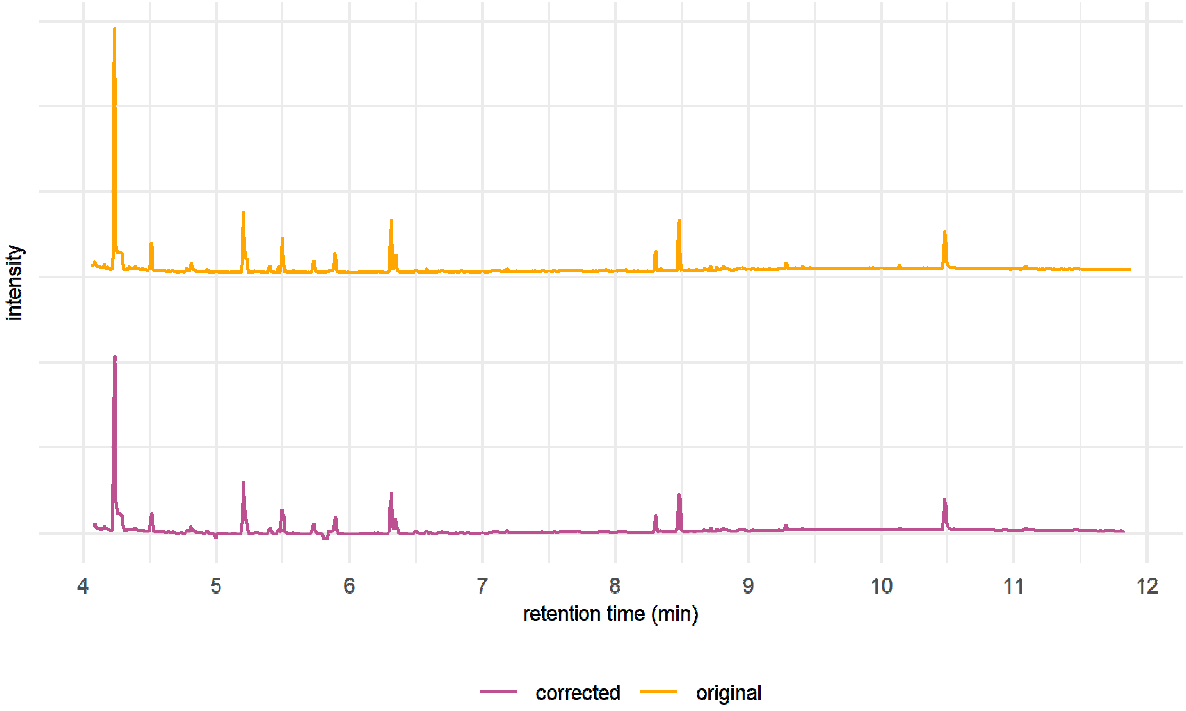

*Figure 4: Illustration of a TIC (top) and a binned and MSC corrected TIC (bottom).*

In addition to intensity normalisation, it is also important to take into account potential baseline shifts. For this, various algorithms have been developed which rely on polynomial fitting or weighted local smoothers (8). These were tested, combined with the different intensity corrections. Results showed that these did not have large effects on the datasets used here. However, all algorithms mentioned have been included in the PoC in order to let the user decide whether to apply them or not, depending on the nature of the data and the outcomes of the exploratory analysis (see 2.3.3). In particular, their effect on the grouping and calamity detection should be used to select the most appropriate technique.

### 2.3.2.3 Flow normalisation

Intensities of detected features are also dependent on the water flow rate (volume per unit of time) at the sampling points, which can cause decreases (dilution during wet seasons) or increases (concentration during dry seasons) in the concentration of chemicals in the collected samples. In the context of this work, it was not possible to include flow data due to time restrictions. However, these should be included in future updates of the developed approach as they can improve the interpretation of the data and, in particular, help detect the occurrence of calamities or interpret temporal changes. Ideally, mean 12h flows should be included as this corresponds to the sampling frequency used in this particular case. Finally, the collected samples should be 12h flow- or volume-proportional composites, to account for changes in flows during the sampling period. Nevertheless, in an ideal case, the operator would evaluate raw concentrations, mass loads and flows to determine whether the observed trends are due to an actual increase in a given feature, for instance, or whether their due to changes in flow rates and/or seasonal effects (e.g., increased used of certain compounds during a specific season).

#### 2.3.2.4 Data pre-processing and transformation

Depending on the type of data and application, additional pre-processing might be necessary to further optimise the detection of specific patterns during the exploratory data analysis (point 3 of the established strategy shown in Figure 1). For this purpose, various steps were considered, such as smoothing using a Savitzky–Golay filter, which is commonly used in signal processing to smoothen data, followed by data transformation to 1$^{st}$ or higher-order derivatives (9). The selection of the most appropriate pre-processing technique is generally done using a cross-validation approach, in particular if the goal is to classify samples based on similarities in the chemical profile (e.g., chromatographic and mass spectrometric data). However, in this case, no prior knowledge about groups is available and the selection of the most appropriate pre-processing technique was done based on the detection of known calamities (see 2.3.5 for more details about this point). Based on the outcomes of the validation step, it was decided to implement all pre-processing techniques in the PoC in order to let the user decide what pre-processing techniques suit the dataset best. For example, if the main interest goes out to compounds at low concentrations, or low ionisable compounds, smoothing might not be suitable since this will smoothen out these peaks.

### 2.3.3 Exploratory analysis

#### 2.3.3.1 Principal component and hierarchical cluster analysis

Exploratory data analysis is focused on the application of various pattern recognition algorithms to identify features of interest. Principal component analysis (PCA) and hierarchical cluster analysis (HCA) are two commonly used approaches to reduce data dimensionality and facilitate the detection of groups based on underlying patterns in the chemical signal (in this case their chromatographic and mass spectrometric profile). These visualization steps help to interpret the data. A more detailed explanation about PCA and HCA is given in box below. In this context, the idea of using these algorithms is that they create groups/clusters among a large set of samples, highlighting in particular outliers (e.g., calamities), and they help determine which feature(s) is/are causing the differentiation. The latter features are then selected for further analysis and eventual identification. The advantage of these two methods, and of other unsupervised algorithms, is that they do not require any prior information about the data and are hence ideal to detect unknown or unexpected patterns in datasets.

*Principal Component Analysis (PCA).* PCA is data a technique to reduce data dimensionality and can be used to visualize large datasets and their patterns. In fact, analytical instruments (e.g., mass spectrometry, infrared spectroscopy) often measure a large number of variables (e.g., the intensity of masses ranging from 30 to 330 m/z as is the case here), which are difficult to visualise in a 2 or 3 dimensional space. PCA is used to do just that, namely to reduce the number of dimensions and be able to visualise the data using only few dimensions (referred to as Principal Components (PCs)) and determine whether groups (i.e., samples characterised by similar chemical signals) are present in the data. PCs are calculated by linear combinations of the original data and are supposed to capture as much as possible of the variance in the original dataset. A PCA biplot (see the example figure below) shows clusters of samples based on their similarity, so if a sample is far away from the others, it has distinct properties (i.e., chemical signal, such as a mass spectrum or a chromatographic profile). If samples are close to each other, they have similar properties. Samples with deviating peaks like calamities or emerging substances are expected to form separate groups. Features/peaks that are responsible for these grouping can be examined further, using loadings. The higher the loading, the more important this variable (in this case a feature or a peak) is for the group. A biplot is a way of visualizing PCA results, as shown in the figure below. The x- and y-axis represent the first two PCs, i.e. the two directions that cover the largest amount of information of the data. Samples are represented by dots and the different colours indicate the group to which each sample belongs. The blue lines represent the loadings of the data set, so the different variables and what effect they have on the distribution of the data.

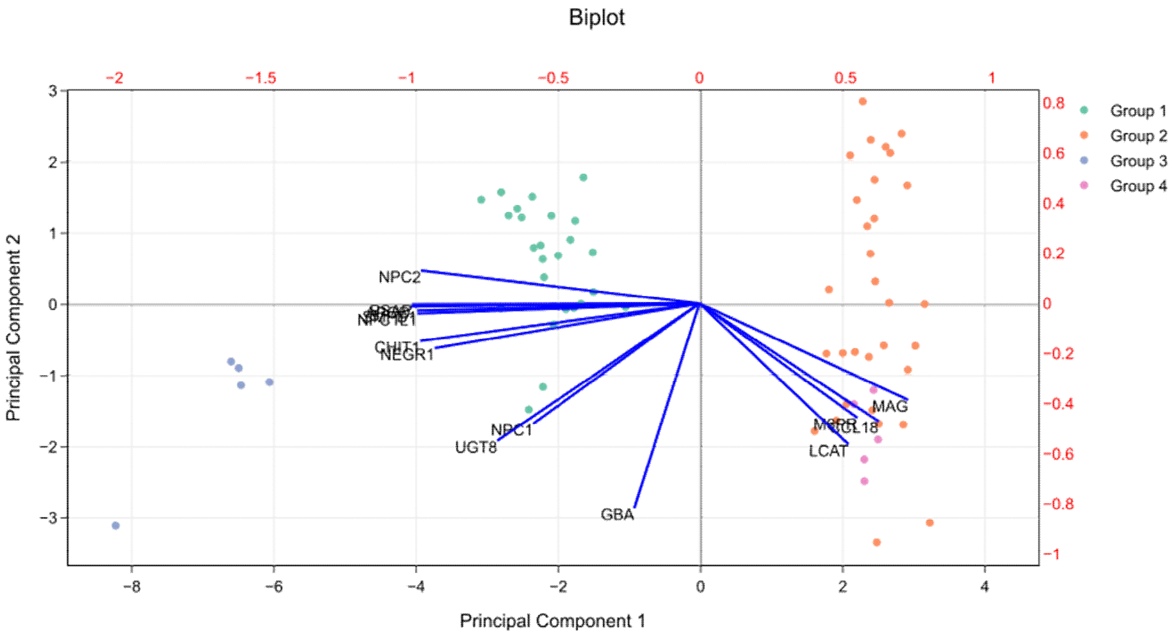

*Figure 5: Biplot figure in the box obtained from (10)*

*Hierarchical Cluster Analysis (HCA).* HCA can be used in addition or as alternative to PCA. It is a different technique but the goal remains the same: identifying groups and subsequently identification of the variables (in this case features or peaks) that cause the grouping. Samples are grouped based on similarities among them and no dimensionality reduction takes place, in contrast to PCA. Samples with deviating peaks will be grouped separately, like in PCA. The results can be visualized using a heatmap, as shown in the figure below. In this example, the x axis represents the different variables (in the context if this work that would correspond to the features and the colouring would reflect their intensity), the y-axis represents the different samples. Both dendrograms on top and at the left visualize the results of the cluster analysis, namely which samples belong to the same group (which can be determined when looking at the left dendrogram) and which features belong to the same group (which can be determined by observing the dendrogram on the top). Each time there is a split in the dendrogram, two separate groups are formed. The higher the splitting, the larger the differences between groups. In the example below, for instance, the first two samples (marked in green on the left side of the heatmap) are separated at a high level from the rest of the samples, meaning that these are the most different samples compared to the rest. The next group which can be separated are the samples marked in red, and so forth. This allows to determine which groups are present in the data set. Then, by looking at which features (vertical lines) are abundant (or vice versa) in a specific group, it is possible to determine which features are responsible for the separation. Going back to the example of the samples marked in green, one can clearly observe that features on the right side of the heatmap are particularly abundant in this group. In fact, the colours represent the intensity (red = high, blue/green = low).

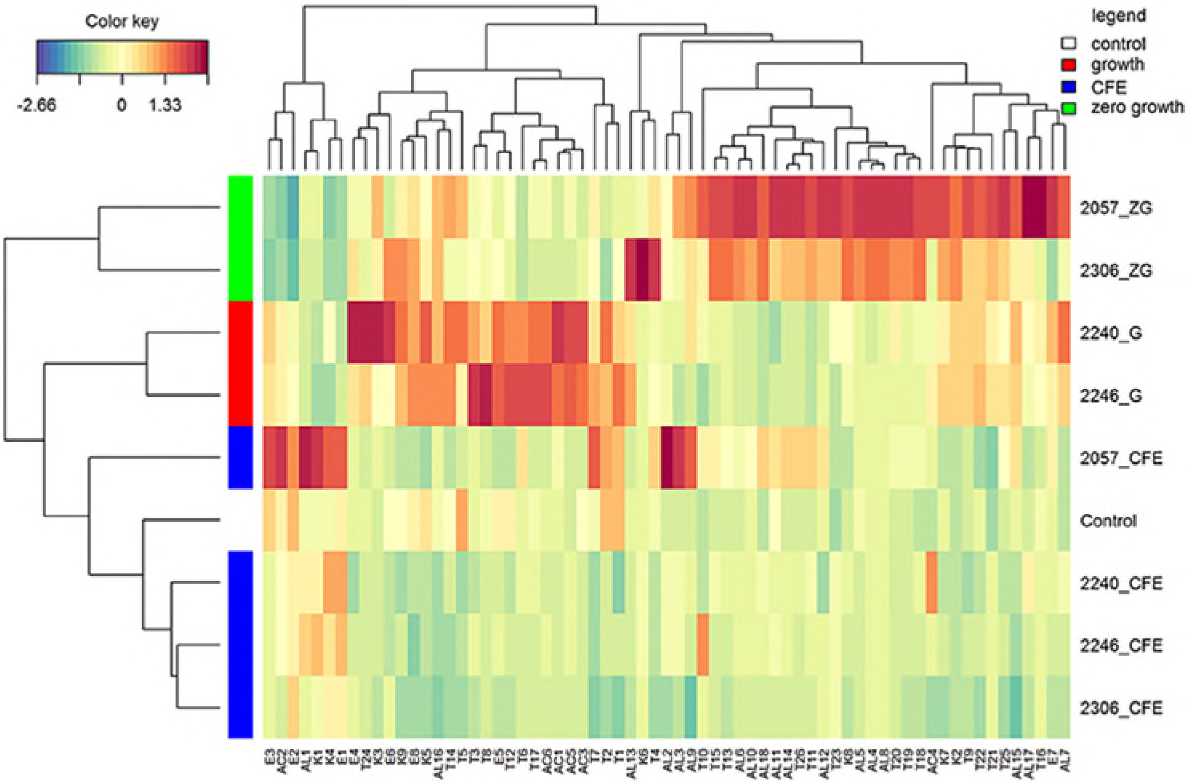

*Figure 6: Heat map taken from (11)*

An example of the application of PCA to the data used in this project is shown in Figure 7. As can be seen, samples (i.e., chromatograms) are plotted against the calculated value of the first two principal components (PCs). What is immediately visible is that there are at least three samples (i.e., 180902_LOB_06, 180902_LOB18 and 180912_LOB_12) which are clearly separated from the rest. In addition to plotting the two PCs, it is possible to include information about which features influence the observed separation the most. These are represented by the arrows shown in Figure 7 (the number indicating the Rt of the feature in question). This plot hence allows the operator to quickly identify if there are samples which strongly deviate from the rest and to get an idea of which features are involved. In Section 2.3.5, a detailed explanation of how information can be derived from heatmaps/HCA is provided.

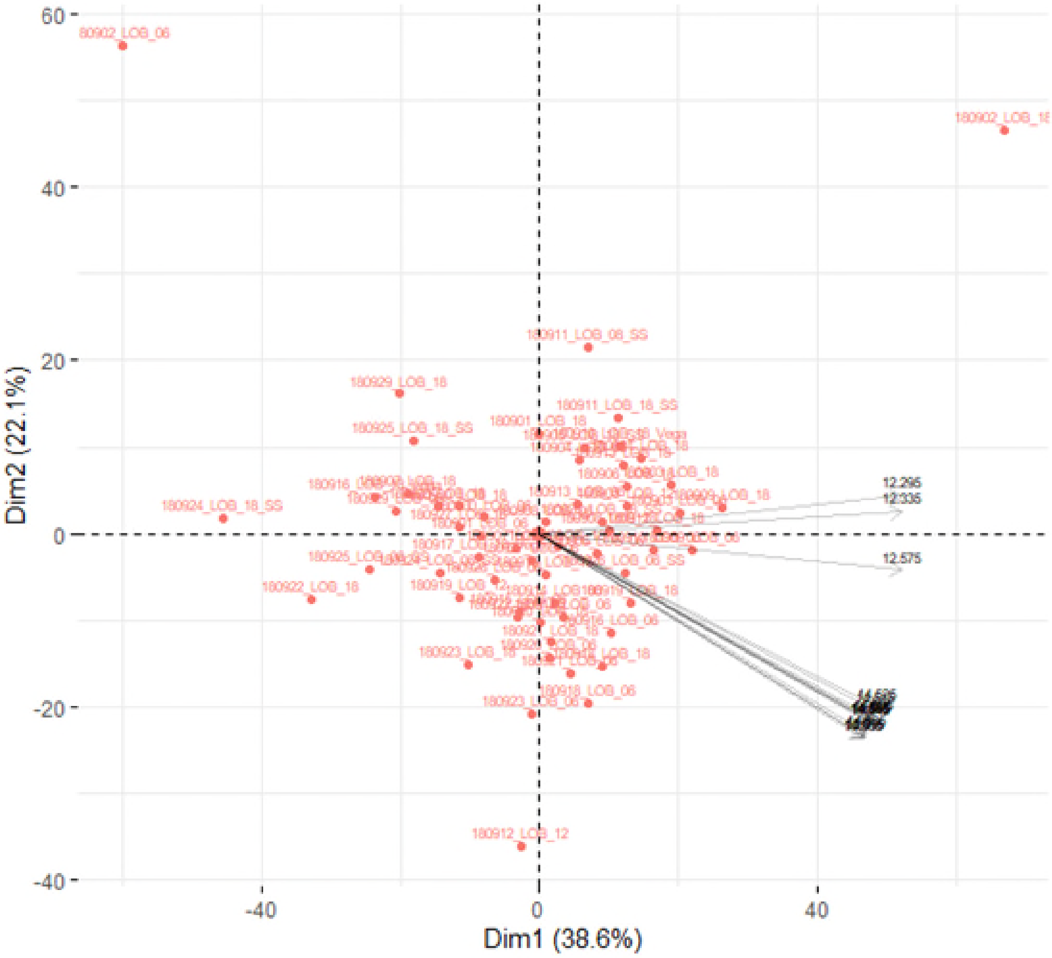

*Figure 7: Example of a Principal Component Analysis biplot (combination of PCA score plot and loading plot) of 60 chromatograms (September 2019) processed as described above. Chromatograms (red dots) are plotted against their PC1 and PC2 values and the black arrows represent the most significant features which separate the data across the two PCs. The number plotted next to each arrow is the Rt of the feature in question. As can be seen, there are three samples which can be clearly distinguished from the rest of the dataset. The purpose of the PCA is exactly to determine whether such "very different" samples are present in the data.*

### 2.3.3.2 Unsupervised machine learning algorithms

The use of additional unsupervised classification algorithms, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) or Self-Organising Maps (SOMs), were initially contemplated as a complement to PCA and HCA. However, based on finding from the validation step (see 2.3.5), it was decided that these were currently not necessary as satisfying results were obtained with PCA and, in particular, HCA.

### 2.3.4 Identification

After potentially relevant features have been detected through the exploratory data analysis, they need to be (tentatively) identified by comparison with existing spectral libraries. For this purpose, it is first necessary to retrieve the mass spectrum of the selected features. An algorithm was written which allows the user to select a feature of interest and, by back-calculating its original retention time (given that all features were aligned using Rt-indexes (see 2.3.2.1)), obtain its mass spectrum at the specific retention time specified by the user. The latter is then exported in text format (i.e., .txt) and can be used to search in spectral libraries such as MetFrag (12), NIST (13) and MassBank (14).

### 2.3.5 Validation

As detailed in the previous paragraphs, the developed PoC involves various steps and data processing approaches whose goal is to improve the detection and identification of relevant features. However, the selection of the most effective set of pre-processing strategies and guaranteeing that no unwanted artifact is being introduced in the data is crucial. For this purpose, developing a validation strategy is essential. In a conventional classification or pattern recognition approach, validation would be performed using a test set, which consists of a subset of the data not used for training purposes and solely serves to test the developed model/algorithm.

In this case, however, given that the goal is to detect samples which might contain features indicative of a pollution event, the optimization of data pre-processing and pattern recognition was done using known historic data of calamities. More specifically, chromatograms of samples which are known to contain specific compounds, linked to known pollution events, were randomly introduced into a subset of the whole dataset. These were then used to determine whether the developed PoC could highlight their presence (i.e., whether they could be labelled as specific groups within PCA and/or HCA) and whether a positive identification could be achieved based on the comparison of their MS-spectra with existing databases.

#### 2.3.5.1 Aniline

The first calamity selected for validation was an increased concentration (up to 6 µg/L) of aniline in the river Rhine. Specifically, 67 chromatograms from samples collected in Bimmen and Lobith in November 2019 were used. These chromatograms were processed as described above and an HCA was performed and the results were plotted in a heatmap (Figure 8). As described above, each horizontal line represents one chromatogram, while each vertical line represents a feature (the color gives an indication of the intensity of the feature). The left-hand side of the heatmap also shows a dendrogram, which simply illustrates the groups that were made (i.e., samples are grouped together when they are characterized by a similar chemical signal or, in other words, when they all have a similar set of features at similar intensities). A closer look at the groups of samples that we made shows that a set of six chromatograms (shown at the top of the heatmap) is clearly separated from the rest of the dataset and that this group of samples is characterized by a particularly intense feature (marked by the black circle in Figure 8). By retrieving the mass spectrum of this feature and by comparing it with the MassBankEU database, one obtains a match with aniline with a similarity score of 0.9817 (as calculated by MassBankEU). This can be also confirmed by visually inspecting the two mass spectra, as shown in Figure 9.

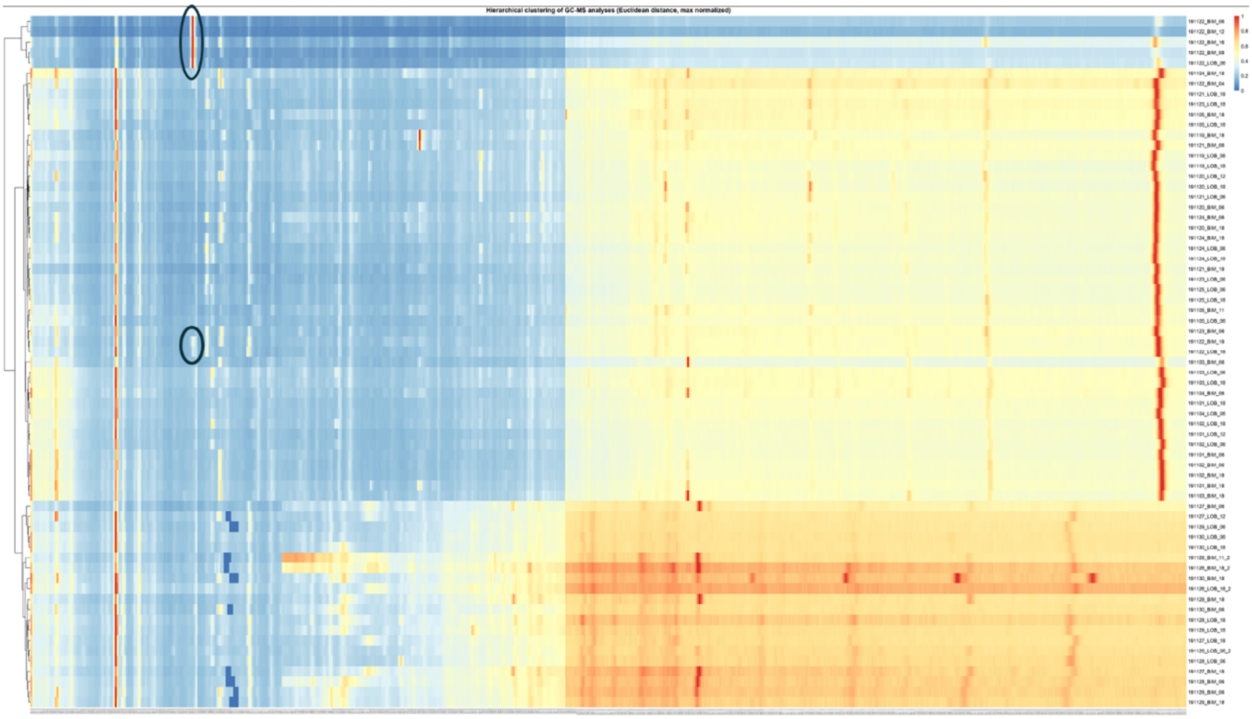

*Figure 8: Heatmap of a Hierarchical Cluster Analysis (Euclidean distance, max normalised) of Bimmen and Lobith data from November 2019, after alignment based on IS, binning and normalisation using MSC, the solvent peak (t0) and internal standard peaks were excluded from the TIC's before conducting HCA. The highlighted features in the figure correspond to aniline. Each line represents a sample (i.e., chromatogram), where the x-axis represents the retention time (i.e., feature). The colour scale indicates signal intensity. The goal of this heatmap is to illustrate the grouping of samples made based on similarities in their chromatographic profile and illustrate (through the colour gradient) which features differ between the groups.*

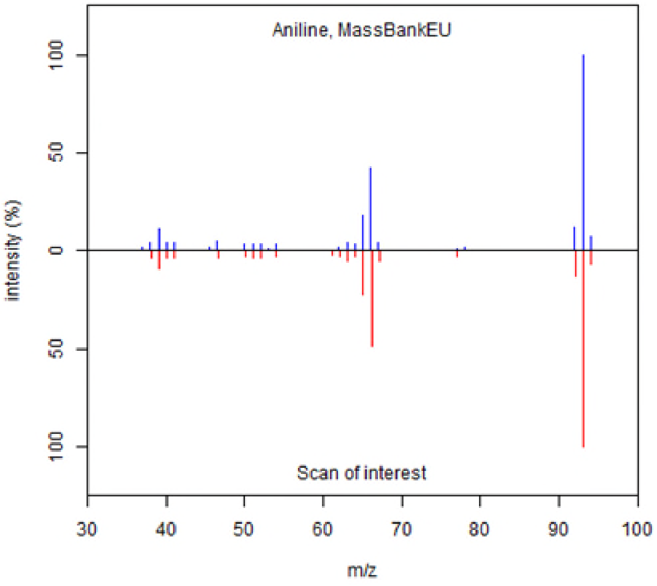

*Figure 9: MS spectra of Aniline as found in MassBankEU (top, SPLASH: splash10-00kf-9000000000-9a543eee5081cc927e82) and the measured MS spectrum (bottom).*

### 2.3.5.2 Phenol

The second calamity selected for validation was an increased concentration (1.4 µg/L) of phenol in the river Rhine. The analysed dataset consisted of all 60 measurements in Bimmen and Lobith of that month, September 2018. Similarly, to the previous example of aniline, HCA was performed and the results were plotted as a heatmap which is shown in Figure 10. Also in this case, a specific group containing only one sample is made. Contrarily to the previous example, however, the separation of this group from the rest of the dataset occurs at a lower level, which makes it less obvious to detect this calamity compared to the case of aniline in which 6 samples were immediately separated by the rest of the group. Nevertheless, one can clearly see that this sample is one of the few in the entire dataset which contains a very intense feature in the first part of the chromatogram (marked in blank in Figure 10). By retrieving the mass spectrum of this feature and by comparing it to MassBankEU, as done previously for aniline, a match with phenol at a similarity score 0.9776 (as computed by MassBankEU) is obtained. This can be also confirmed by visually inspecting the two mass spectra, as shown in Figure 11.

In this specific dataset, additional high-level groups can be observed (in particular at the top of the heatmap). When analyzing this kind of data, it would be very interesting to focus on these groups and try to and identify the features which are responsible for this separation. For instance, a group characterized by intense features which repeat at regular intervals is clearly visible in the heatmap. Given their repetitive nature, these features could be a part of homologues series of chemicals which only differ in size (e.g., length of an alkyl chain).

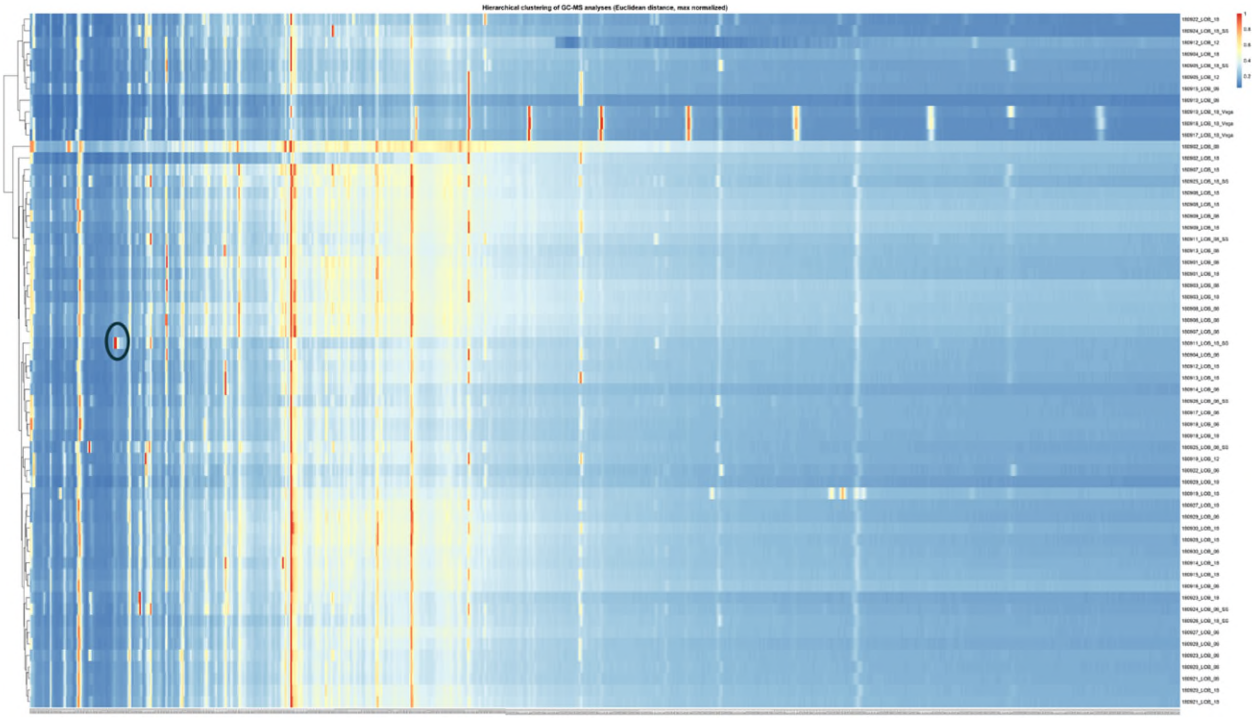

*Figure 10: Heatmap of an Hierarchical Cluster Analysis (Euclidean distance, max normalised) of Bimmen and Lobith data from September 2018, after alignment based on IS, binning and normalisation using MSC, the solvent peak (t0) and internal standard peaks were excluded from the TIC's before conducting HCA. The highlighted features in the figure correspond to phenol. Each line represents a TIC, where the x-axis represents the retention time. The color scale indicates signal intensity. The goal of this heatmap is to illustrate the grouping of samples made based on similarities in their chromatographic profile and illustrate (through the colour gradient) which features differ between the groups.*

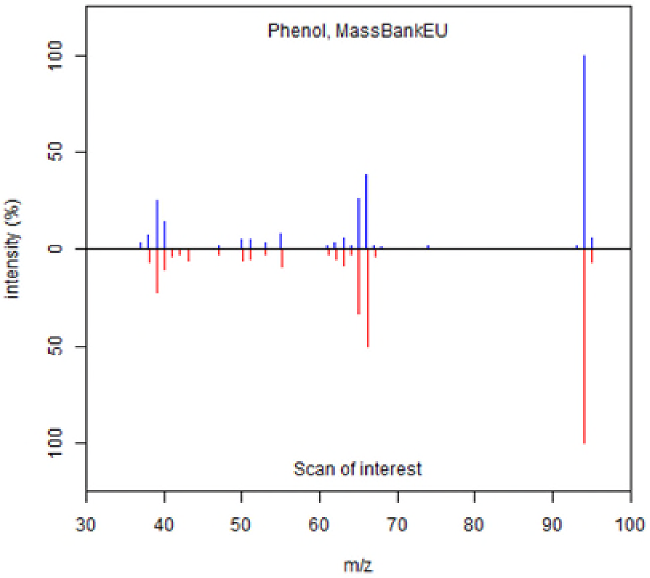

*Figure 11: MS spectra of Phenol as found in MassBankEU (top, SPLASH: splash10-00kf-9000000000-6fb456992902a13931f9) and the measured MS spectrum (bottom).*

## 2.4 Outcomes of the validation and recommendations for using the PoC with low-resolution data

The validation showed that using the developed approach, it is possible to highlight the presence of specific compounds (i.e., calamities) and to identify them using existing spectral libraries. However, in both cases, it is clear that the chances of detecting a new compound/calamity strongly depend on the number of samples that are being processed simultaneously and the amount of background noise present in the chromatograms. In the first case, findings from tests run by KWR suggest that the analysis should be performed using one month's worth of samples (i.e., approximately 60 chromatograms) at most. Ideally, one would want to perform the analysis with fewer samples, for instance one or two weeks. This would avoid having too much information being represented in a single heatmap/PCA plot. In fact, it is clear from the outcomes of the validation that, although the features related to the calamities were grouped separately from the rest of the samples, there is still a large amount of information present which could complexify the interpretation. This is particularly true for features which appear in few or just one chromatogram (as is the case for phenol). On the other hand, however, it is possible that large discharges might be visible in samples collected over longer periods (days or weeks). In these cases, one might need broader time windows to observe the increase and then the decrease of one or multiple previously unknown features. It is hence important to analyse the data using different time windows to make sure that both short- (i.e., acute) and long-lasting (i.e., large disposal) events can be detected.

Regarding the background noise, inclusion of blanks which can be used to remove parasite peaks is highly desirable. Furthermore, noise removal would most likely also benefit the data analysis as a larger number of samples can be processed and visualized as only relevant features would be included. In fact, based on experience from other screening methods indicates that the number of relevant features can be reduced substantially by removing signals present in the blanks (e.g., 50% or more in the case of LC-HRMS analysis).

## 2.5 Recommendations for the improvement of the SPE-GC-MS method

Following the analysis of the SPE-GC-MS data, it was possible to identify three points in the current method(s) used by RWS that can be improved in future. Firstly, the measurement of blanks (either solvent or, ideally, procedural (SPE) blanks) at regular intervals needs to be included in the monitoring scheme. Blanks are essential to ensure that no contamination or carry-over is taking place and play an important role in the development of any automated data analysis strategy. In addition, they would allow improving the data analysis approach as irrelevant features would be filtered out and more samples could be analysed simultaneously. Currently, the data analysis can still be performed, however the absence of blanks serious complexifies the detection of calamities or new contaminants, in particular, if these are present at low concentrations. In fact, the operator will need to interpret the outcomes of the exploratory analysis very carefully as potentially relevant features might be present at intensities comparable to the background (which would have been removed if blanks were available). Secondly, alkanes should be measured at regular intervals to allow the calculation of Kováts retention indexes for normalisation and identification purposes. Finally, the analysis of replicates could also help improve the data analysis as artefacts and/or non-reproducible features could be excluded (as is done for the LC-HRMS data, see 3.4).

## 2.6 Perspectives for future applications

A data analysis strategy was developed allowing to import, process, evaluate online SPE-GC-MS data and identify features of interest. The approach was developed using data generated by a Thermo Fisher Scientific instrument, however, the goal of the PoC is that it can be used with data from any manufacturer. Although the developed algorithms can be used with data from any manufacturer, there are a number of aspects that will always need to be customized given that different laboratories and/or instruments will provide slightly different data. In particular, the following aspects will need to be customized prior to implementation of the PoC in another laboratory:

1. Import and conversion of data to a usable format. This is rather straightforward and there are numerous packages, depending on the manufacturer, which can be used to convert proprietary raw files to open-source formats such as mzML.

2. Criteria to select relevant chromatograms. As described previously, this is currently being done using the Qxmax file inventory. Another laboratory might have a different approach to record which data represents a sample and which not. However, assuming all samples are measured using the same method, minimum information requirements are: sample name, file location, sampling date and time, analysis date and time, corresponding blank and the applied ionization mode (i.e. positive or negative).

3. Information about internal standards. Laboratories are likely to use different internal standards and/or methods, hence the retention time of internal standards needs to be retrieved (using the developed algorithm) and the normalisation step needs to be updated (Rt-indexes of features need to be computed using retention time windows specific to the method in question). Furthermore, mass-labelled internal standards should always be preferred where possible.

4. Information about methods used. Different method parameters (e.g., gradient, oven program, run time, injection volume, chromatographic column and maintenance) will lead to different results, which might hinder the comparison of data from different methods. The current GC-MS PoC is not yet cross-compatible between different instruments. This issue can partly be solved by using Rt-indexes (e.g., Kováts retention indexes), however differences will most likely always exist between different methods/laboratories. Detailed records of maintenance should also be kept to help interpret findings.

5. Availability and usability of blank samples. Subtraction of features that occur in blanks is currently not included in the developed PoC as these were not available. In future versions of the PoC, this can easily be implemented.

Despite the abovementioned points, which will need to be adapted/optimized for each new method/laboratory, there are a number of elements which are universal and independent of the method used, namely:

- Type of samples: the developed approach can be implemented regardless of the type of samples and sample processing technique which is being used. In this specific case, samples were processed with SPE prior to analysis. This is however not a requirement as the sample type or preparation does not affect the data analysis used here.

- Sampling frequency: the data used to develop the PoC was collected at very high frequency (i.e., one sample every 12h), however, this is once more not a requirement as the approach can be implemented with much smaller datasets without any change needed. However, should the goal be to identify specific spatial or temporal trends, then it is important to have sufficient data to guarantee that external factors (e.g., seasonal effects, noise) can be accounted for.

- Real-time applications: currently, the PoC was developed using historical data (2016-2021) provided by RWS. However, in its current design, the PoC can be implemented also for real-time applications, in the sense that newly analysed samples can be immediately processed and compared to previous samples.

# 3 High-resolution LC-HRMS analysis

## 3.1 Objective

The PoC which was specifically developed for LC-HRMS data focused on the development of a strategy to import, process and analyse data with the goal of detecting samples which deviate from normal patterns (e.g., calamities) and the features which might be causing the observed deviation, as well as to (tentatively) identify them using existing databases. However, in this case, specific attention was given to developing an approach which allows to analyse time trends as this was not implemented for GC-MS data, due to the absence of LC-HRMS data at the beginning of this project and consequent time constraints.

## 3.2 LC-HRMS data

Although significantly different from HRMS data, the work that was performed on low-resolution data has allowed us to define an overall strategy which can be implemented also for HRMS data. The major difference between low- and high-resolution data is that the latter is substantially easier to pre-process. As discussed above, accounting for retention time shifts in high-resolution data is less complex because of the increased specificity that accurate mass information offers. In fact, this can be used to easily detect and group features across numerous chromatograms without the need to first align the whole signal. Furthermore, a larger and more comprehensive set of data analysis packages exists for HRMS compared to low-resolution data. One point, however, where HRMS-data is more complex is the identification step, in particular if data has been acquired using liquid chromatography coupled to HRMS (as is the case here). In fact, the ionization in such instruments, which mainly occurs using electrospray ionization (ESI), is substantially less reproducible and standardized compared to electron impact (EI), which is used in GC-MS instruments. Hence, the comparison with existing databases is less obvious and an expert operator is still required to evaluate whether there is a true match between the experimental MS and the one from the database. Finally, databases are far from being comprehensive, thus the risk that a detected feature might not be identifiable through available databases is non-negligible.

High-resolution data was not available at RWS and hence had to be retrieved from other laboratories. Het Waterlaboratorium (HWL), which is a member of the supervisory board of this project, indicated that it had sufficient HRMS data to perform the tasks as described in the project plan. An official request from RWS/MinIenW was made to HWL to provide HRMS data. For this purpose, KWR established a detailed list of requirements that HRMS data should fulfill in order to be able to develop the described PoC also for this kind of high-resolution data (memo of Frederic Béen, 4th October 2021). On 6th January 2022, HWL provided data which fulfilled the requirements. For the data analysis of LC-HRMS data, it was decided to develop the PoC using the existing, internationally widely used package *'patRoon'* (15) developed by the University of Amsterdam. One of the advantages of *patRoon*, besides the fact that it has been implemented by numerous research groups across the world, is that it is a platform which makes use of various validated and commonly used algorithms and packages used to process HRMS data. In addition, KWR has extensive knowledge in the use of this package (and has been directly involved in its development).

The data consisted of 68 measurements of surface water samples from 3 locations (Lekkanaal/Rhine, Meuse and IJsselmeer) over a 2-year period. All samples, including procedural blanks, were measured in triplicate in both positive and negative ionization mode. Samples and blanks were all spiked with internal standards, as given in Annex IV. Data was acquired with a Bruker Daltonics maXis series Impact II micrOTOF mass spectrometer, using a quadrupole mass analyser and electrospray ionization. The total runtime of the chromatogram was around 20 minutes. Mobile phase A consisted of 95% 5mM ammonium formate in 5% MeOH and mobile phase B consisted of 5 mM ammonium formate in MeOH. An example of two chromatograms (TIC) is shown in Figure 12. In this proof of principle, only data acquired in the positive ionization mode was considered.

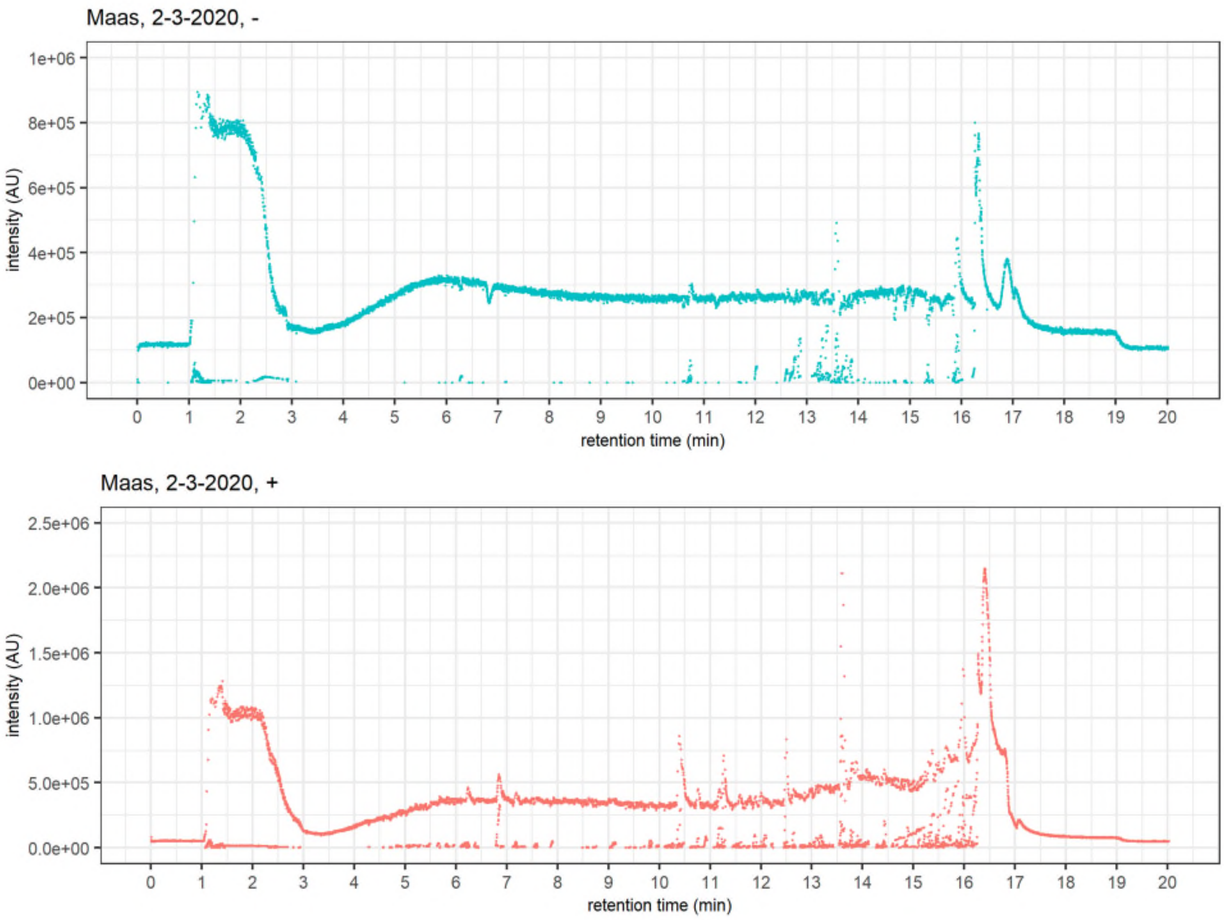

*Figure 12: Two TICs of a sample from the river Meuse, in negative mode (top) and positive ionization mode (bottom).*

## 3.3 Data analysis strategy

Globally, the strategy implemented for high-resolution data follows similar steps as those illustrated in Figure 1, and described in Section 2, except for some of the details related to importing (step 1) and normalisation (step 2) which, as will be described in detail in the following sections, differs due from low-resolution data due to the fact that information about accurate mass is available. With respect to the actual data analysis steps, PCA and HCA were implemented for LC-HRMS as previously done for GC-MS data, however, additional emphasis was put here on the development of the trend analysis which was not included in the GC-MS data due to time constraints and also because LC-HRMS data was not initially available.

## 3.4 (Pre)processing

### 3.4.1 Re-calibration

Given that data was acquired using a Bruker instrument, re-calibration was necessary prior to exporting the raw data as this instrument uses an external calibration. In agreement with HWL, this step was kindly performed by the University of Amsterdam as proprietary Bruker software is required and KWR does not have access to such software. Not having worked directly with raw Bruker data before, KWR was not aware of this step. In future, if data from Bruker has to be processed, then re-calibration needs to be carried out before exporting the raw data.

### 3.4.2 Feature detection and grouping

Next, the re-calibrated data was converted to the open format mzML, using the ProteoWizard msconvert tool(16), embedded in patRoon (15). Subsequently, features were generated using the OpenMS (17) software, also embedded in patRoon. The resulting features were filtered to exclude all features eluting before or during the solvent peak. In the specific case of the data provided by HWL, a threshold was set at 180 seconds. The remaining features were grouped and aligned using the *OpenMS* algorithm. More specifically, features in different samples having accurate mass, isotopic pattern and retention time within a defined threshold were considered to be the same feature. Thresholds used for grouping are reported in the PoC script and users can modify them based to the instruments used and the type of data being analysed. Features which were not present in all replicates and/or features whose relative standard deviation of the intensities within triplicates was above 75% were removed. Once more, these criteria can be modified by the users in the PoC script. This is a standard procedure done to remove features which are likely artefacts, not reproducible or related to background contaminations. A downside of these steps is that the risk of losing relevant features is posed, such as features eluting prior to or within the solvent peak

The steps described above were performed using patRoon’s default settings. It should be noted that these parameters can be further optimized if deemed necessary using a design of experiments approach embedded in patRoon. In fact, m/z and retention time thresholds have a significant impact on the outcomes of feature grouping. This is shown in Figure 13, where in the first 180 analyses, the internal standards atrazine-d5 is labelled as being feature group ‘M221_R547_5652’, while this same group does not contain any feature in the subsequent 72 analyses. In these subsequent 72 analyses, an additional group, labelled ‘M221_R560_5653’ was present having the same m/z but with a retention time shift of on average 13 seconds. Tweaking the retention time threshold allowed for all features related to atrazine-d5 to be grouped together (i.e., ‘M221_R550_5692’). The fact that these groups all contained atrazine-d5 was confirmed by inspecting the MS2 spectra of the feature (see Figure 14).

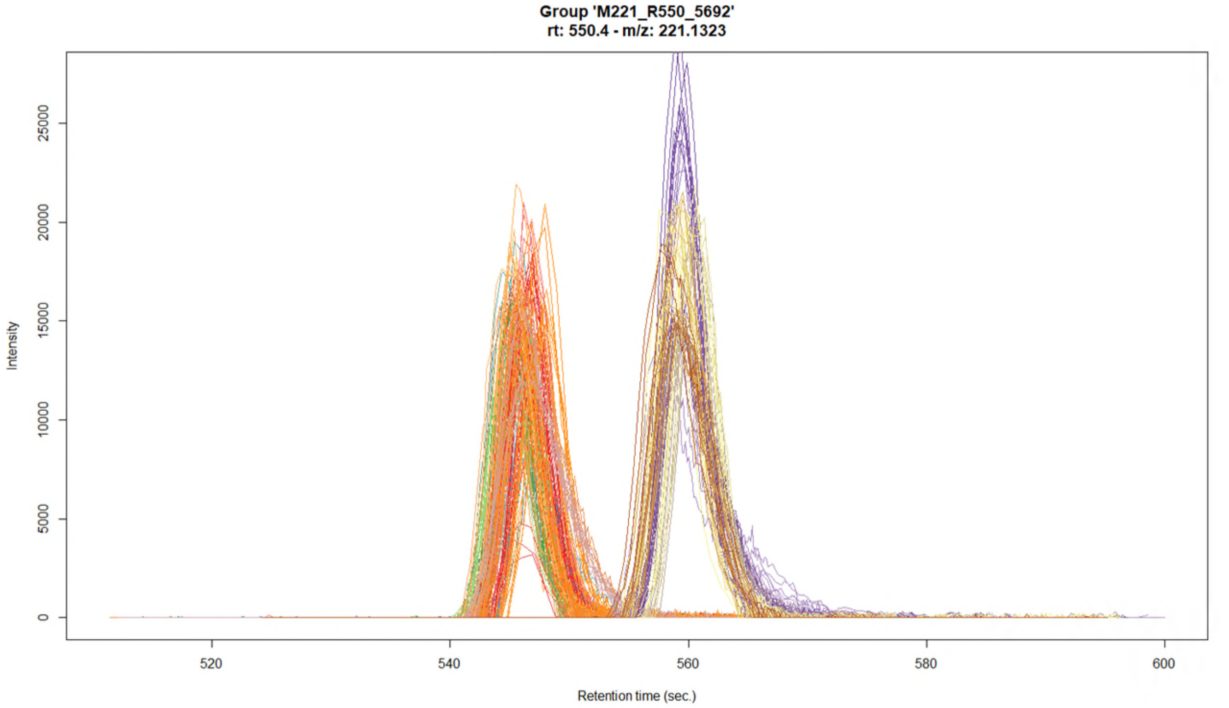

*Figure 13: Chromatographic data of feature group ‘M221_R550_6592’, every line is a measurement. The left peak is generated by the data from the first 180 analyses, whereas the right peak is from the other 72 analyses.*

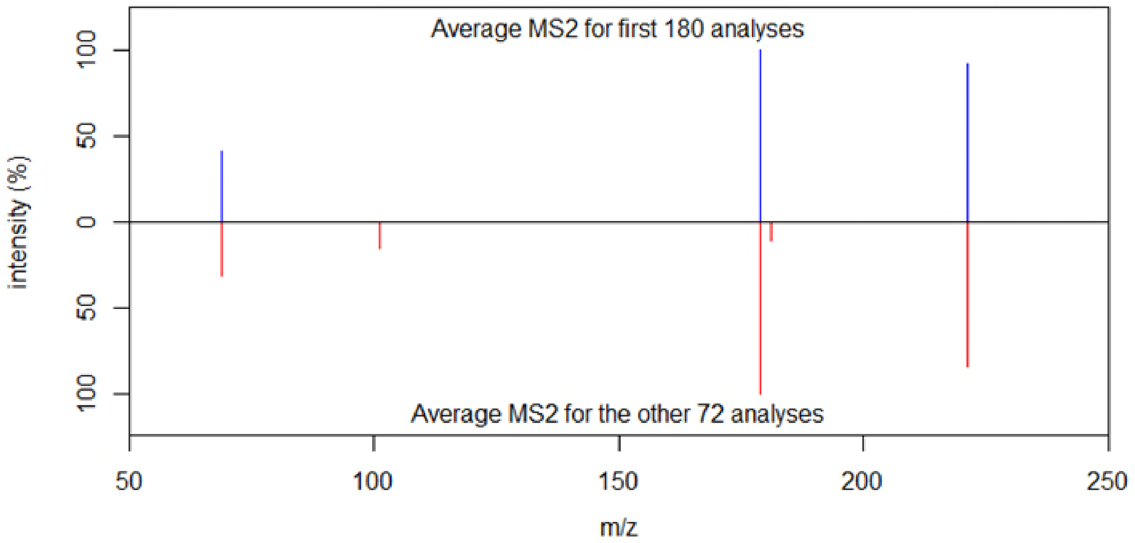

*Figure 14: Average MS2 peaklist obtained using the mzR algorithm, for the feature group 'M221_R550_5692'*

A similar situation was observed for other IS, for example neburon, which was attributed to two feature groups with a retention time difference of ± 11.9 seconds. The shift occurred between the same sample sets as for atrazine-d5, attributed to a different analysis batch. These examples show that setting grouping parameters is important and that the use of internal standards is highly valuable also for HRMS data to ensure that features are being grouped (i.e., aligned) correctly. This is particularly important when comparing data from samples that have been analysed at different times. In fact, as for GC-MS data, retention time shifts can occur due to random variability but also due to column ageing or slight differences in eluents or ageing of the HPLC seals or other factors.

### 3.4.3 Noise removal and blank filtration

Grouped features (also referred to as feature list) were converted into a data frame and feature intensities were normalised (i.e., divided) by the intensity of atrazine-d5. This should in the future be extended to all available internal standards and feature intensities should be normalised per retention time windows, as done for the SPE-GC-MS data. Due to time constraints, this could not be implemented at this stage. Nevertheless, the decision of whether or not to normalize the data using internal standards will remain to the description of the operator, based on the effect that this step has on the outcomes. In fact, it cannot be excluded that this approach might introduce artefacts due to local ion suppression or enhancement. However, this issue is present even without IS-normalisation as features across different samples might have different intensities at equivalent concentrations due to different matrix effects. For this reason, one would ideally use a sufficient number of IS covering the whole chromatogram.

Subsequently, an intensity filter was applied. Namely, all features whose intensity was <10% of that of the internal standards were removed. Generally, in HRMS data analysis workflows, features whose intensity falls below a certain threshold (e.g., 50k to 500k counts depending on the type of sample being analysed) are filtered out to remove noise/background peaks. However, this approach is generally used when samples were analysed in one sequence (i.e., all at the same time). In this case, however, analyses were carried out at different times and it cannot be assumed that the background will be the same across all samples. Hence, it was decided to remove features based on their intensity relative to the internal standards (as the latter was analysed in all samples and should allow taking instrumental variability into account). Subsequently, the remaining features were averaged for all triplicates in order to have one intensity per feature group per sample. Afterwards, blank correction was performed by removing all feature groups with intensities < 5x intensity in the blank samples. Similarly to the previous step, this is a common step that allows to reduce the number of features to be analysed by removing features which are present in blanks in substantial levels. Finally, the intensity of the blank was removed from features which were also present in blanks, although with an intensity of at least 5 times compared to the blank. This was done in order to correct for the variation in the background between batches and measurement series. Similarly to the IS-normalisation step, the decision of whether or not to apply this background removal is at the discretion of the operator who will chose based on the effect that this step has on the outputs. In the specific case of the data used here, the difference in results with and without background subtraction was minimal (i.e., overall the same features were prioritised based on their increasing trend over time).

During these preprocessing steps, it appeared that some measurements needed to be excluded from the data set as they caused fatal errors, internal standards could not be found or did not cover a complete set. These were the 12 measurements done on 17 June 2019, the 12 measurements done on 7 October 2019 and an extra sample from the IJsselmeer from June 2020. Currently, no information is available about why these samples could not be processed. More detailed discussions with HWL will be necessary to identify the source of these issues.

Overall, 240 measurements were considered in the trend analysis, covering 60 samples. Below, an overview is given about how many features were detected in the different datasets.

(Pre)processing statistics:

1. Initial dataset consisted of 252 analyses. After removal of data which caused fatal errors (see above), 240 analyses (i.e., chromatograms) were used.
2. A total of 336,103 features were found.
3. After filtering all features prior to 180 seconds, 5.28% was removed. Remaining: 318,009 features.
4. After feature grouping, 22,547 groups were formed with an average of 12.4 features per group.
5. After filtering all feature groups not abundant in all replicates: 242,541 features remained in 9398 groups.
6. For all 60 samples, feature intensities were averaged and corrected for presence in the blank.
7. Feature groups that were not present in any of the measurements after blank correction were removed, giving 6858 feature groups.
   - IJsselmeer: 5,482 feature groups
   - Lekkanaal-Rijn: 4,230 feature groups
   - Maas: 4,910 feature groups

## 3.5 Data analysis

### 3.5.1 Principal component and hierarchical clustering analysis

After pre-processing, in particular removal of features which occur also in blanks, there is still an important number of feature groups which makes the interpretation of the data very complex. Hence, more appropriate approaches need to be implemented to select features characterized by specific (temporal) patterns. Similarly to low-resolution data, HCA and PCA were used to visualize and detect features of interest.

With respect to PCA, results obtained for high-resolution data were analogous to what was previously observed with GC-MS data, namely that it can be used to highlight the presence of groups of samples or, as shown in Figure 15, detect the presence of “outliers”. However, the immediate identification of which features are responsible for the separation might be less straight forward compared to the more visual approach offered by HCA and heatmaps.

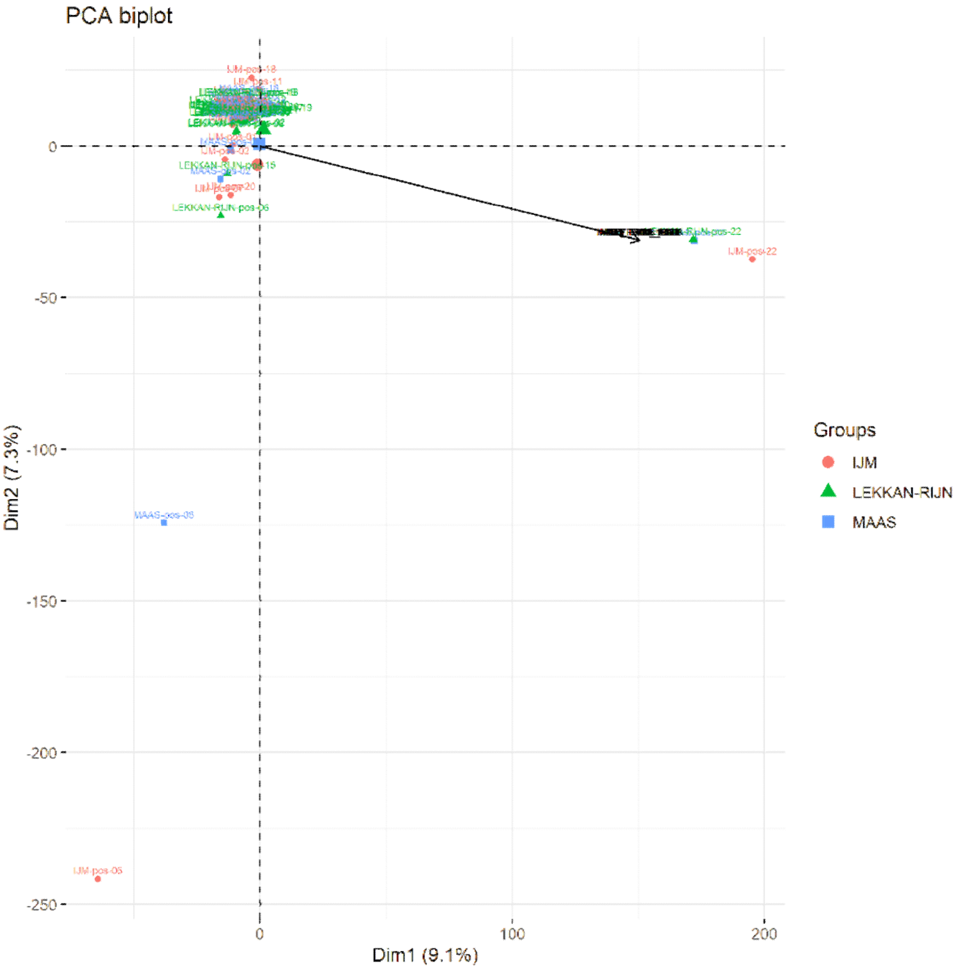

*Figure 15: Example of PCA of the different samples, groups are generated based on sample location (as shown by the different colors). The x- and y-axis represent the first two dimensions and the black arrow represents the features which are responsible for the separation of samples along the first PC. In fact, as can be seen, there a few samples which can be differentiated from the rest of the dataset. In particular, there is a group of samples which shows substantially different values along with the first PC (Dim1). In this particular PCA plot, the amount of variance explained by the first principal components is very low (< 10%) due to the large number of features still present in the dataset and their variability.*

Similarly to what was done for GC-MS data, sample groups can be made and visualized using HCA and heatmaps, as shown in Figure 16. Here a cluster analysis using Pearson’s correlation and Ward.D2’s method for clustering were used to compute the illustrated heatmap. Because of the large variability in feature intensities, the latter were log-transformed prior to HCA. However, compared to GC-MS data, there is an extremely large amount of features still present in the data, which makes it difficult to distinguish specific features which might be of interest in a certain group. To facilitate the analysis of this large amount of data, it is possible to perform the grouping not between samples but between features, as is shown in the following Figure 17. In this case, the operator can more easily observe group of features which occur in specific samples and can more easily select them for further investigation. Finally, it is also possible to combine both sample and feature grouping in a single heatmap, however this substantially complicates the interpretation of the results. In the PoC, the operator can toggle between the two types of heatmaps and can decide which one is more appropriate based on the data set under investigation.

It should be noted, however, that due to time constraints, a dedicated algorithm to select features from the obtained heatmaps (as was done for GC-MS data) could not be implemented. As will be discussed below, it was decided to focus here on the implementation of an approach to select features characterized by increasing temporal trends. Nevertheless, the information about which features belong to which clusters is stored in the features list after HCA. Operators can hence filter relevant features based on the cluster they belong to and their intensity. In future developments of the PoC, a more automated selection algorithm can be implemented. Furthermore, HCA in combination with correlation testing (as described in the next section) could be implemented in future as an additional or alternative approach to detect and select features showing (increasing) temporal trends (18).

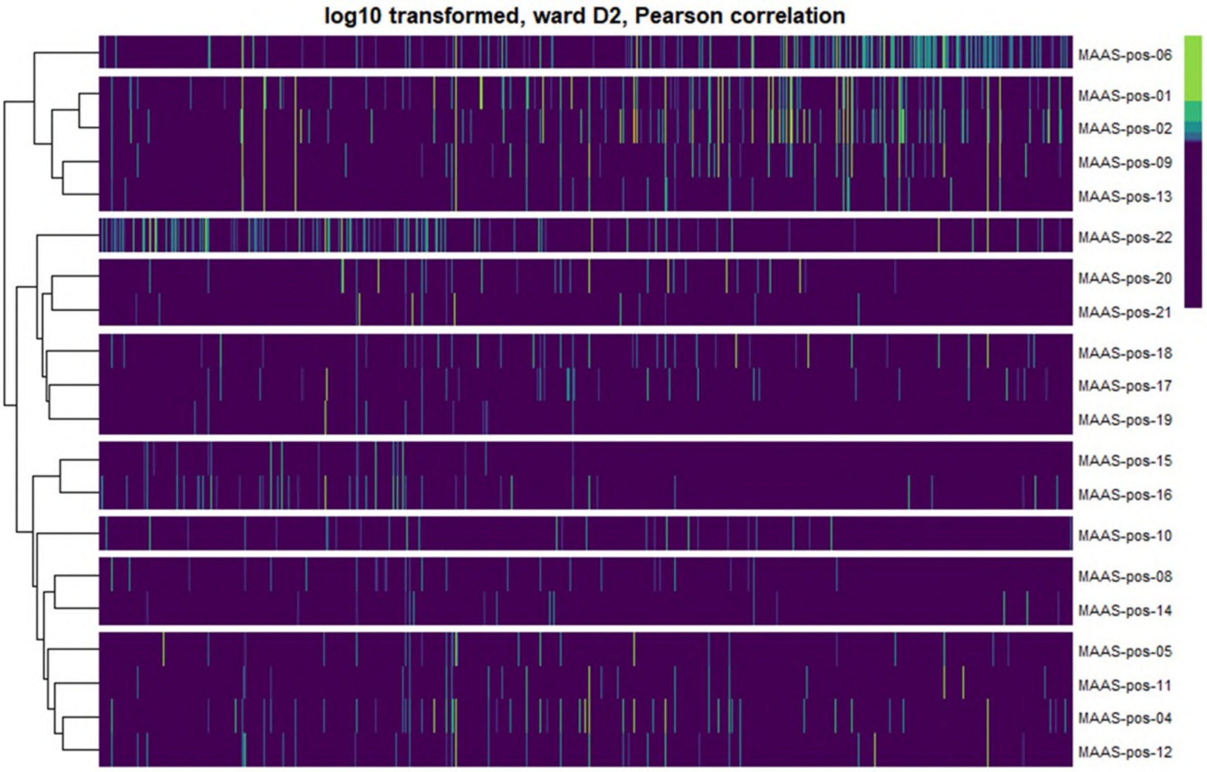

*Figure 16: Heatmap of hierarchical cluster analysis performed on samples from the river Meuse. The x-axis represents the different features and the colour illustrates their intensity (after log-transformation) while on the y-axis samples are being grouped based on similarities in their features (occurrence and intensity).*

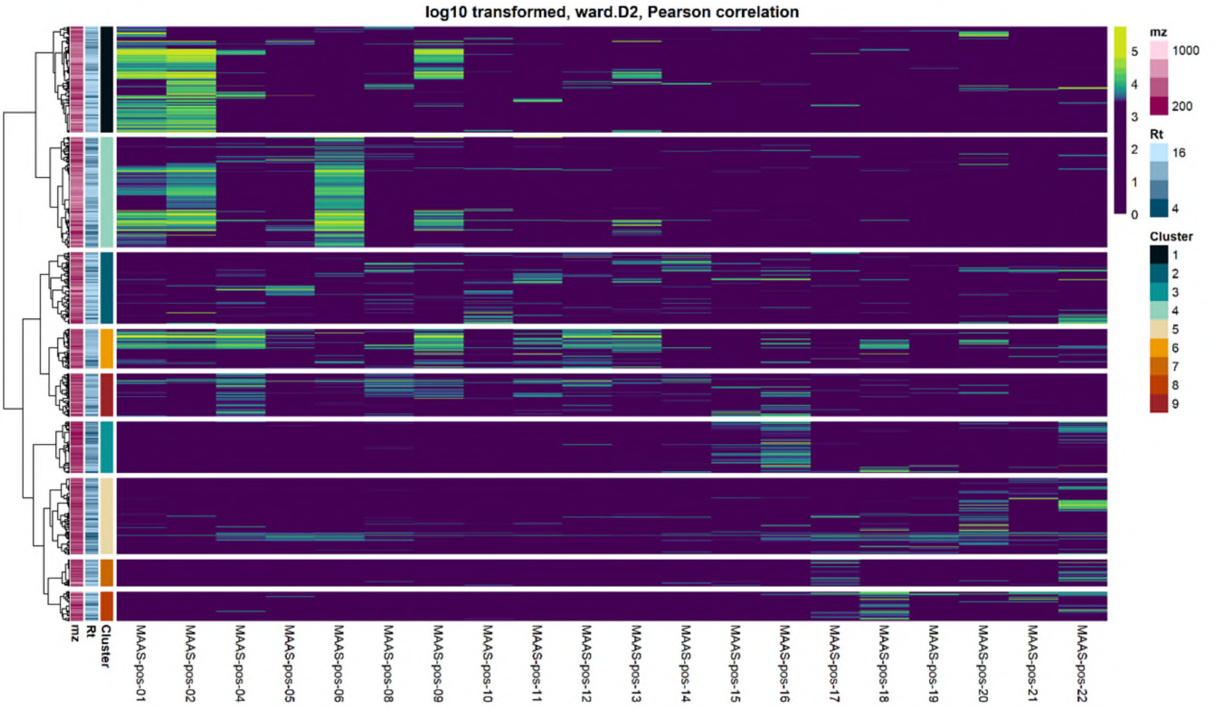

*Figure 17: Heatmap of hierarchical cluster analysis performed on samples from the river Meuse. Contrarily to the previous plot, the x-axis here represents samples, while the y-axis illustrates groups of features. This allows to group features based on their occurrence (or absence) in certain samples and allows to analyse multiple features at the same time instead of having to look at individual ones as is the case in Figure 16. For example, it appears clearly that samples 01 and 02 are both characterized by the presence of two groups of features, and it is more straightforward for the operator to determine which these are and to investigate them further. Additional information such as the m/z, retention time and cluster number of the features was added to this plot.*

### 3.5.2 Outcomes and perspectives

The above examples show how HCA and PCA can be implemented in the LC-HRMS data analysis workflow. Based on the obtained results, the operator can select the relevant feature and extract further information like in the approach for the GC-MS dataset. As discussed above, PCA appears to be more appropriate to rapidly detect outliers while HCA can be more easily used to detect groups of features differentiating sample groups. Nevertheless, it remains important that operators understand how the PoC is built and get accustomed to working with it (as well as getting used to the data itself). Only by gaining experience with this kind of tool will it become possible to easily detect anomalies and identify the features which are causing them.

## 3.6 Trend analysis

Various studies reported in the literature have applied trend analyses to NTS data using various techniques. These techniques range from relatively simple (linear regression analysis (19), Spearman's rank correlation coefficient (20), Mann-Kendall correlation coefficients (21), time-trend ratio's (20) and rarity scores (22)) to more complex approaches like hierarchical clustering (23) and multivariate empirical Bayes approach combined with Hotelling T2 (24). The most suitable approach depends on the shape of the trend, namely whether it is monotonically increasing or decreasing, or nonmonotonic. Regression analysis can be applied if the trend appears to be linear and a trend line can easily be fitted (19). Otherwise, non-parametric tests can be applied, such as Mann-Kendall or Spearman's rank correlation tests. The latter are more suitable to test for the presence of monotonic trends which may not be linear. Both tests are similar, however Mann-Kendall correlation test is generally considered more robust. The case of nonmonotonic trends (i.e., that both increase and decrease over time) was not covered in the current workflow.

In the context of this work, a multistep approach was developed using four different techniques to detect the presence of trends. This consisted of linear regression analysis using a linear model and a log-linear model, and two similar non-parametric tests, i.e. Mann-Kendall correlation coefficient and Spearman correlation coefficient. Regression analyses are used to determine if the estimated slope of the regression line is different from zero and its sign will provide an indication about whether the observed trend is increasing or decreasing. Hence, the obtained p-values and slopes can be used to determine if the data exhibits a linear trend. In the case of the non-parametric tests, a correlation (ρ), which ranges between -1 and 1, is calculated (together with a p-value to assess the significance of the observed correlation) to determine if a monotonic trend is present in the data.

Initially, all four statistical analyses were applied to all feature groups that were present in at least three samples. This was chosen to ensure that enough data points were available to adequately determine the presence of a trend. Next, feature groups with a p-value < 0.05 in in the Mann-Kendall test were selected. These were then further filtered according to the calculated correlation (ρ), which should be a positive value, indicating an increasing trend. For the river Meuse samples, this resulted in 101 feature groups with increasing trends, as shown in Figure 18.

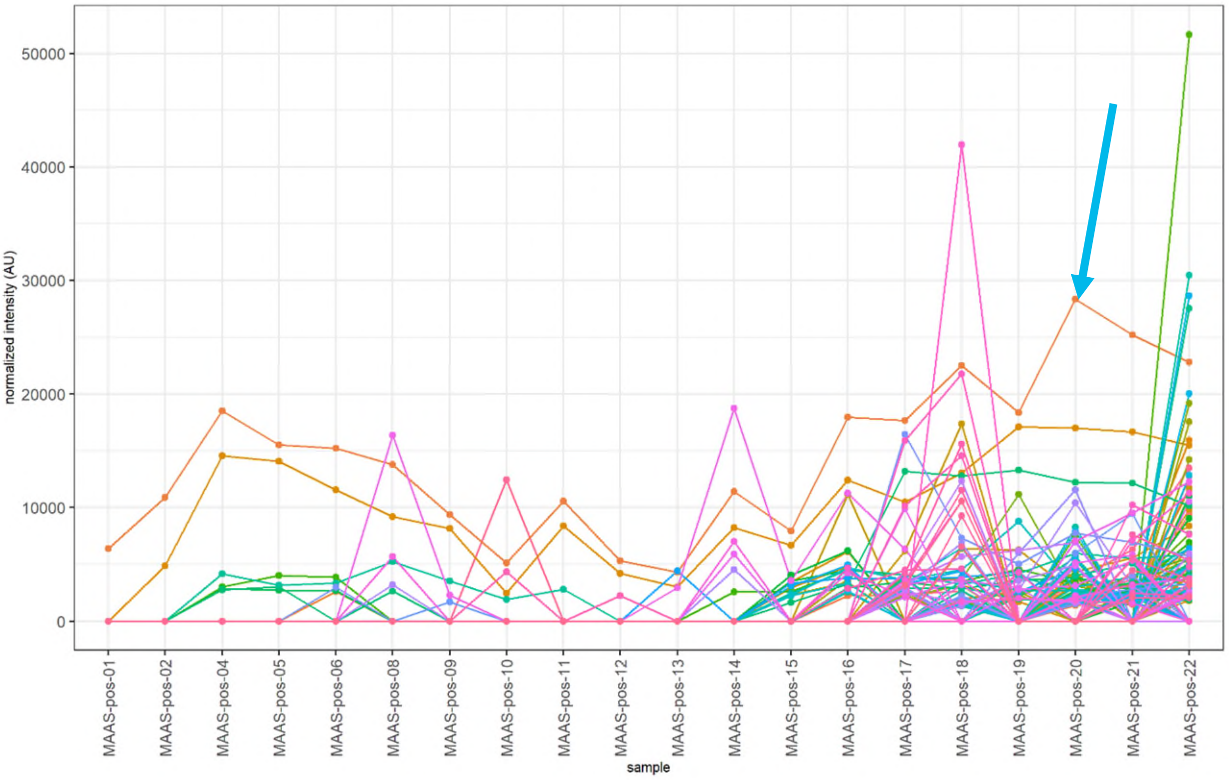

*Figure 18: Feature groups having a p-value below 0.05 for Mann-Kendall correlation coefficient and a positive ρ (n = 101). Only samples from the river Meuse were considered here. The blue arrow indicates one of the features which were selected for subsequent analysis (see below). AU = arbitrary units*

The feature group 'M120_R355_1744', (marked by the blue arrow in Figure 18) was subjected to a tentative identification attempt. The SIRIUS algorithm (embedded in *patRoon*) generated three possible formulas, whereas the GenForm algorithm (embedded in *patRoon*) generated one possible formula, namely $C_6H_6N_3$, which was also present in the SIRIUS results. The most likely candidate (out of a total of 26) formed using MetFrag was 2H-benzotriazole, as shown in the left pane of Figure 19. A further comparison was made with the MassBankEU spectral library and the result of the comparison between the measured spectra and the one from the database is shown in the right pane of Figure 19. As can be seen, an almost perfect match between the two MS2 spectra was obtained, which is a strong indication that the selected feature is likely a benzotriazole-like compound. This example illustrates that the developed approach can be used to highlight features with increasing trends and that these can potentially be tentatively identified based on their MS2 spectra using existing libraries.

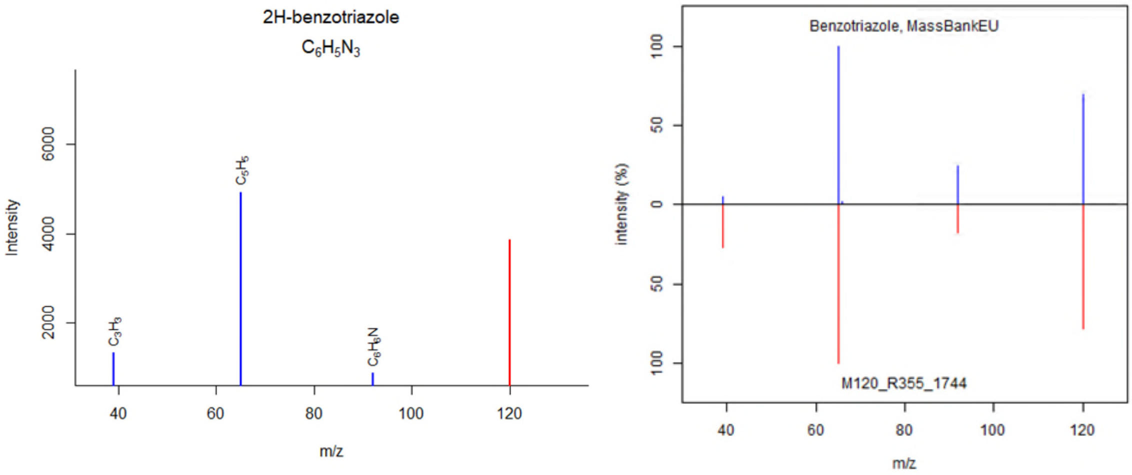

*Figure 19: MS2 of feature group 'M120_R355_1744', the MetFrag annotation is given at the left (where the red peak represents the precursor and the blue peaks are the assigned peaks), and the comparison with a benzotriazole MS2 from MassBankEU (spectrum in blue, SPLASH splash10-01b9-9400000000-79f1def14297918ce9e5) is given at the right.*

However, it should be noted that the trend analysis approach implemented here does not allow to take into account for seasonal effects, that might be misinterpreted as increasing (or decreasing) trends although they are not. In the future, more advanced time series analysis approaches and larger data sets covering longer time periods should be considered, in particular if long time series are to be included and analysed, as the impact of seasonal effects in these datasets might be substantial. In fact, considering the amount of time needed to formally identify a detected feature (i.e., purchase and analysis of reference material or, if the latter is not available, analysis using orthogonal methods such as NMR), one needs to be sure that the selected feature is really showing an increasing trend. However, except for trends, features being constantly present at the same concentrations, or features undergoing seasonal effects might be of interest as well.

## 3.7 Outcomes of the validation and recommendations for using the PoC with high-resolution data

In the case of HRMS data provided by HWL, information about calamities that can be used for validation purposes, as was done for low-resolution data, was not available at the time of the development of the PoC. However, findings such as the one illustrated above (i.e., 2H-Benzotriazole) could be verified by comparing with results from targeted analysis (provided that the compound is included in routine monitoring conducted by HWL) and/or by comparing the retention time and MS2 spectra of the selected feature with that of a reference standard analysed by HWL with the same instrument. If possible, this exercise should be carried out for a selection of features of interest to confirm the accuracy of the obtained results.

Similarly to what was recommended for the GC-MS data, a good balance has to be struck between the amount of data being analysed. Too large data sets might make the interpretation of the outcomes of the data analysis very complex due to the large number of features and/or samples. On the other hand, a too small data set might hinder the detection of calamities if these last multiple days or even weeks (in which case enough data before and after the calamity is required in order to detect the change in feature composition). Similarly, for the detection of trends, it might be useful to actually increase the number of samples measured and analysed, so that changing trends can be detected earlier and stronger inferences can be made about their significance. Hence, it remains the task of the operator to perform the analysis with different time windows, get acquainted with the data and ensure that features affecting water quality both on the long- and short-term can be detected using the developed PoC.

## 3.8 Perspectives for future applications

A data analysis strategy was developed allowing to import, process, select and identify features detected by LC-HRMS. The approach was developed using data generated by a Bruker Daltonics instrument, however, it can be adapted for use with raw data from any manufacturer. Nevertheless, like for GC-MS data analysis, there are a number of aspects that will always need to be customized given that different laboratories or instruments will provide slightly different data. In particular, the following aspects will need to be customized prior to implementation of the PoC in another laboratory/on another instrument:

1. Information about internal standards. Laboratories are likely to use different internal standards and/or methods, hence the retention time of internal standards needs to be retrieved and the normalisation step needs to be updated. Mass labelled reference standards are always to be preferred where possible.

2. Information about methods used. Different method parameters (e.g., gradient, oven program, run time, injection volume and chromatographic column) will lead to different results, which might hinder the comparison of data from different methods. This issue can partly be solved by using retention time alignment, however, differences will most likely always exist between different methods/laboratories.

3. Availability and usability of blank samples. This PoC involves the subtraction of features that occur in blanks. This is a requirement for the method to work. However, in the case that all samples have been analysed simultaneously (i.e., same sequence), blank subtraction might be less relevant as the instrumental background is assumed to affect all samples equally.

# 4 Conclusion

In the context of this project, two approaches were developed for the analysis of low- (GC-MS) and high-resolution (LC-HRMS) data. Whilst sharing an overall similar processing strategy, two distinct workflows were developed due to the intrinsic differences in the types and frequency of data used.

From a technical perspective, the two approaches represent a proof of concept of what can be achieved with this kind of data and can serve as platforms that can be further expanded and improved as experience is gained by users/laboratories. In fact, further developments and improvements are possible and, in some cases, even needed. For instance, the method used for GC-MS analyses should be improved by adding blank samples which can then be used to reduce noise in the data. Concrete tests with data from other manufacturers should also be carried out to determine if any specific steps need to be included or modified. Furthermore, more advanced time series analysis approaches might be necessary for the detection of trends in larger data sets (e.g., multiple years' worth of data). Finally, the development of a graphical user interface (GUI) should be also contemplated, although this often comes at the loss of flexibility in tweaking specific aspects of the data analysis, which are possible when working with a script-based approach. Despite the fact that analysing and interpreting mass spectrometric data remains a challenging task that requires an experienced operator, in particular when trying to formally identify selected features, the developed approach can help to substantially narrow down the number of features to identify by focusing only on those showing relevant patterns.

From a practical perspective, the PoCs in their current state can be used by analysts/operators to analyse both historic and contemporary data (including newly analysed samples almost in real-time) to (i) rapidly detect the presence of samples deviating from normal patterns (e.g., due to a discharge or a calamity), (ii) detect the presence of new and potentially relevant features (i.e., chemicals), (iii) highlight differences across sampling locations and determine which features are responsible for these differences, (iv) detect features characterized by increasing (or decreasing if necessary) trends over time and, last but not least, (v) tentatively identify features of interest by comparison with existing databases. As mentioned previously, to be able to efficiently carry out all the above mentioned data-analysis activities, users will need to acquire experience with the developed PoCs and these will have to be constantly updated and improved based on users' feedback.

# 5 Acknowledgements

The authors would like to thank Rick Helmus from the University of Amsterdam for the support in re-calibrating the Bruker data, as well as Het Waterlaboratorium (HWL), in particular Eelco Pieke, for providing the data used in the development of the approach.

# 6 References

1. Watson NE, VanWingerden MM, Pierce KM, Wright BW, Synovec RE. Classification of high-speed gas chromatography–mass spectrometry data by principal component analysis coupled with piecewise alignment and feature selection. J Chromatogr A. 2006 Sep;1129(1):111–8.

2. Domingo-Almenara X, Brezmes J, Vinaixa M, Samino S, Ramirez N, Ramon-Krauel M, et al. eRah: A Computational Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC/MS-Based Metabolomics. Anal Chem. 2016 Oct 4;88(19):9821–9.

3. RStudio Team. RStudio: Integrated Development for R [Internet]. Boston: RStudio Inc.; 2020. Available from: https://rstudio.com/

4. Kockmann T, Panse C. rawR - Direct access to raw mass spectrometry data in R. bioRxiv. 2020 Jan 1;2020.10.30.362533.

5. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Anal Chem. 2006 Feb 1;78(3):779–87.

6. Yu M, Wang T. enviGCMS [Internet]. 2020. Available from: https://cran.r-project.org/web/packages/enviGCMS/enviGCMS.pdf

7. Kováts E. Gas-chromatographische Charakterisierung organischer Verbindungen - Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. Helv Chim Acta. 1958;41(7):1915–32.

8. Wang Z, Zhang M, Harrington P de B. Comparison of three algorithms for the baseline correction of hyphenated data objects. Anal Chem. 2014 Sep 16;86(18):9050–7.

9. Rinnan Å, Berg F van den, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal Chem. 2009 Nov 1;28(10):1201–22.

10. Ngo L. How to read PCA biplots and scree plots [Internet]. BioTuring's Blog. 2018 [cited 2022 May 6]. Available from: https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/

11. Ricci A, Levante A, Cirlini M, Calani L, Bernini V, Del Rio D, et al. The Influence of Viable Cells and Cell-Free Extracts of Lactobacillus casei on Volatile Compounds and Polyphenolic Profile of Elderberry Juice. Front Microbiol [Internet]. 2018 [cited 2022 May 6];9. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2018.02784

12. MetFrag - project [Internet]. [cited 2022 Mar 25]. Available from: https://ipb-halle.github.io/MetFrag/projects/metfragweb/

13. NIST. NIST Standard Reference Database 1A [Internet]. NIST. 2014 [cited 2022 Mar 25]. Available from: https://www.nist.gov/srd/nist-standard-reference-database-1a

14. MassBank | Database | Search [Internet]. [cited 2022 Mar 25]. Available from: https://massbank.eu/MassBank/Search

15. Helmus R, ter Laak TL, van Wezel AP, de Voogt P, Schymanski EL. patRoon: open source software platform for environmental mass spectrometry based non-target screening. J Cheminformatics. 2021 Dec;13(1):1.

16. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012 Oct;30(10):918–20.

17. Röst HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods. 2016 Sep;13(9):741–8.

18. Chiaia-Hernández AC, Günthardt BF, Frey MP, Hollender J. Unravelling Contaminants in the Anthropocene Using Statistical Analysis of Liquid Chromatography–High-Resolution Mass Spectrometry Nontarget Screening Data Recorded in Lake Sediments. Environ Sci Technol. 2017 Nov 7;51(21):12547–56.

19. Veenaas C, Bignert A, Liljelind P, Haglund P. Nontarget Screening and Time-Trend Analysis of Sewage Sludge Contaminants via Two-Dimensional Gas Chromatography–High Resolution Mass Spectrometry. Environ Sci Technol. 2018 Jul 17;52(14):7813–22.

20. Plassmann MM, Tengstrand E, Åberg KM, Benskin JP. Non-target time trend screening: a data reduction strategy for detecting emerging contaminants in biological samples. Anal Bioanal Chem. 2016 Jun;408(16):4203–8.

21. Dürig W, Alygizakis NA, Menger F, Golovko O, Wiberg K, Ahrens L. Novel prioritisation strategies for evaluation of temporal trends in archived white-tailed sea eagle muscle tissue in non-target screening. J Hazard Mater. 2022 Feb 15;424:127331.

22. Krauss M, Hug C, Bloch R, Schulze T, Brack W. Prioritising site-specific micropollutants in surface water from LC-HRMS non-target screening data using a rarity score. Environ Sci Eur. 2019 Jul 22;31(1):45.

23. Albergamo V, Schollée JE, Schymanski EL, Helmus R, Timmer H, Hollender J, et al. Nontarget Screening Reveals Time Trends of Polar Micropollutants in a Riverbank Filtration System. Environ Sci Technol [Internet]. 2019 Jun 18 [cited 2019 Jul 1]; Available from: http://pubs.acs.org/doi/10.1021/acs.est.9b01750

24. Alygizakis NA, Gago-Ferrero P, Hollender J, Thomaidis NS. Untargeted time-pattern analysis of LC-HRMS data to detect spills and compounds with high fluctuation in influent wastewater. J Hazard Mater. 2019 Jan 5;361:19–29.

25. Bengtsson H, Ahlmann-Eltze C, Bravo HC, Gentleman R, Gleixner J, Hickey P, et al. matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors) [Internet]. 2021 [cited 2021 Nov 9]. Available from: https://CRAN.R-project.org/package=matrixStats

26. Roudier P, Kuhn M, Liland KH, Mevik BH, Wickham H, Viscarra Rossel R. spectacles [Internet]. 2021. Available from: https://cran.r-project.org/web/packages/spectacles/spectacles.pdf

27. Mevik BH, Wehrens R. pls [Internet]. 2021. Available from: https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf

28. signal: Signal processing [Internet]. [cited 2022 Apr 1]. Available from: https://r-forge.r-project.org/projects/signal/

29. Liland KH, Almøy T, Mevik BH. Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. Appl Spectrosc. 2010 Sep 1;64(9):1007–16.

30. Kassambara A, Mundt F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses [Internet]. 2020 [cited 2021 Nov 9]. Available from: https://rpkgs.datanovia.com/factoextra/index.html

31. Kolde R. pheatmap [Internet]. 2018. Available from: https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf

32. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.

33. Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, et al. GGally: Extension to “ggplot2” [Internet]. 2021 [cited 2022 Apr 1]. Available from: https://CRAN.R-project.org/package=GGally

34. Dodder N, Mullen K. OrgMassSpecR: Organic Mass Spectrometry [Internet]. 2017 [cited 2020 Sep 4]. Available from: https://CRAN.R-project.org/package=OrgMassSpecR

# I Process of PoC development

The development of the PoC involved a close collaboration between KWR and the laboratory of RWS and HWL, which provided the used data but also relevant contextual/methodological information. Furthermore, RWS organized regular meetings with the supervisory commission, during which KWR presented the progress and very detailed and constructive discussions took place. The feedback provided by the commission were integrated in the PoC, which were adapted according to the suggestions received. Finally, on the 31st of March 2022, a demonstration of the PoCs was given to laboratory personnel of RWS and Aquon (HWL could unfortunately not attend). This exchange was very interesting and allowed us to in particular identify future requirements for a smooth implementation of the developed PoCs.

# II Feedback from the Demo session

On the 31st of March 2022, a demo of the PoCs took place at KWR. Frederic Béen (KWR), Nienke Meekel (KWR), Chris Lukken (RWS) and René Lindenburg (Aquon) participated. The demonstration focused on both developed approaches (i.e., GC-MS and LC-HRMS). Globally the exchange was very constructive and it appeared that the use of the PoCs for RWS could occur on the very short term (despite the lack of long term HRMS data). On the other hand, Aquon indicated that the lack if this type of instrumentation currently limits the amount of data available and hence the implementation of the developed PoC. Furthermore, some technical/detailed feedback about the PoC were given and these will be included in the final versions. Finally, the discussion was brought on the topic of the strategic/operational choices which need to be made to allow/facilitate the implementation of the developed approaches. These topics will need to be discussed with the supervisory commission as, together with technical aspect, are crucial for an efficient implementation of data analysis platforms in laboratories.

# III Internal standards GC-MS analyses

List of internal standards which are used for quantification and quality control of the GC-MS samples. Three internal standards were not used for alignment since they could not be detected in multiple chromatograms.

| Internal standard | Monoisotopic mass (Da) | Bruto formula | Used for Rt alignment? |
|---|---:|---|---|
| Toluene-d8 | 100.1128142 | $C_6D_5CD_3$ | No |
| Chlorobenzene-d5 | 117.0393616 | $C_6D_5Cl$ | Yes |
| 1,4-Dichlorobenzene-d4 | 149.9941125 | $C_6D_4Cl_2$ | Yes |
| Naphthalene-d8 | 136.1128142 | $C_{10}D_8$ | Yes |
| 1,4-Dibromobenzene-d4 | 237.89308 | $C_6D_4Br_2$ | Yes |
| Terbuthylazine-d5 | 234.140807 | $C_9D_5H_{11}ClN_5$ | No |
| Phenanthrene-d10 | 188.1410178 | $C_{14}D_{10}$ | Yes |
| Chrysene-d12 | 240.1692213 | $C_{18}D_{12}$ | No |

# IV Internal standards LC-HRMS analyses

Internal standards used by HWL in their LC-HRMS method and their average retention times (list was provided by HWL).

| Compound | Bruto formula | Retention time (min) | Monoisotopic mass (Da) |
|---|---|---|---|
| Metformine-d6 | C4H6D6ClN5 | 2.64 | 171.1158 |
| Fenuron-d5 | C9H7D5N2O | 6.38 | 169.1263 |
| Bentazon-d7 | C10D7H5N2O3S | 6.47 | 247.1008 |
| Chloridazone-d5 | C10H3ClD5N3O | 6.6 | 226.067 |
| Carbetamide-d5 | C12H11D5N2O3 | 7.81 | 241.1475 |
| Monuron-d6 | C9H5D6ClN2O | 8.13 | 204.0937 |
| Metobromuron-d6 | C9H5D6BrN2O2 | 9.33 | 264.0381 |
| Atrazine-d5 | C8H9D5ClN5 | 9.38 | 220.1252 |
| Chlorbromuron | C9H10BrClN2O2 | 10.71 | 291.9614 |
| Chlooroxuron-d6 | C15D6H9ClN2O2 | 11.1 | 296.1199 |
| Diclofenac-d4 | C14H7D4Cl2NO2 | 11.46 | 299.0418 |
| Neburon | C12H16Cl2N2O | 11.81 | 274.064 |
| Diazinon-d10 | C12H11D10N2O3PS | 12.12 | 314.1638 |
| Metconazole-d6 | C17H16D6ClN3O | 12.29 | 325.1828 |
| Fenofibrate-d6 | C20H15D6ClO4 | 13.43 | 366.1505 |
| Quinoxyfen-d4 | C15H4D4Cl2FNO | 13.77 | 311.0218 |

# V Overview of used packages

## V.I Used packages for pre-processing of SPE/GC-MS data

To perform the various data processing described above, the following packages are being used:

*(i)* *rawrr*, an R package for direct access to data from Thermo Fischer Scientific, this was used to read the data from the raw files.

(ii) *matrixStats (25)*, an R package with functions to perform operations on matrices, in this case used for binning;

(iii) *spectacles (26)*, an R package developed for processing spectroscopy data; this was used to perform Standard Normal Variate normalisation;

(iv) *pls (27)*, an R package for multivariate regression methods; this has been used to perform Multiplicative Scatter Correction;

(v) *signal (28)*, an R package for signal processing, this was used to perform Savitzky-Golay smoothing;

(vi) *baseline (29)*, an R package for baseline correction of spectra, this was used to perform modified polynomial fitting and Gaussian weighting;

As for the previous packages, all packages mentioned above are open source.

## V.II Used packages for exploratory analysis of SPE/GC-MS and LC-HRMS data

To perform the exploratory data analysis, the following packages have been used:

(i) *Factoextra (30)*, an R package for the visualisation of multivariate data analysis;

(ii) *Pheatmap (31)*, an R package for the visualisation of hierarchical clustering;

(iii) *ggplot2 (32), an R package for visualization.*

## V.III Used packages for the pre-processing and analysis of HRMS data

(i) *patron (15)*, an R package for workflows for mass spectrometry based non-target analysis;

(ii) *GGally (33)*, an R package for plotting and visualization of trends, extension to ggplot2;

(iii) *OrgMassSpecR (34),* an R package for organic mass spectrometry, this has been used to calculate and visualize spectrum similarity of MS and MS2 spectra.