UNTARGETED TIME-PATTERN ANALYSIS OF LC-HRMS DATA TO DETECT SPILLS AND COMPOUNDS WITH HIGH FLUCTUATION IN INFLUENT WASTEWATER

Nikiforos A. Alygizakis[1], Pablo Gago-Ferrero[1], Juliane Hollender[2,3], Nikolaos S. Thomaidis[1,]*

[1] Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

[2] Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland

[3]Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092, Zürich, Switzerland

***Corresponding author:**

Tel: +30 210 7274317

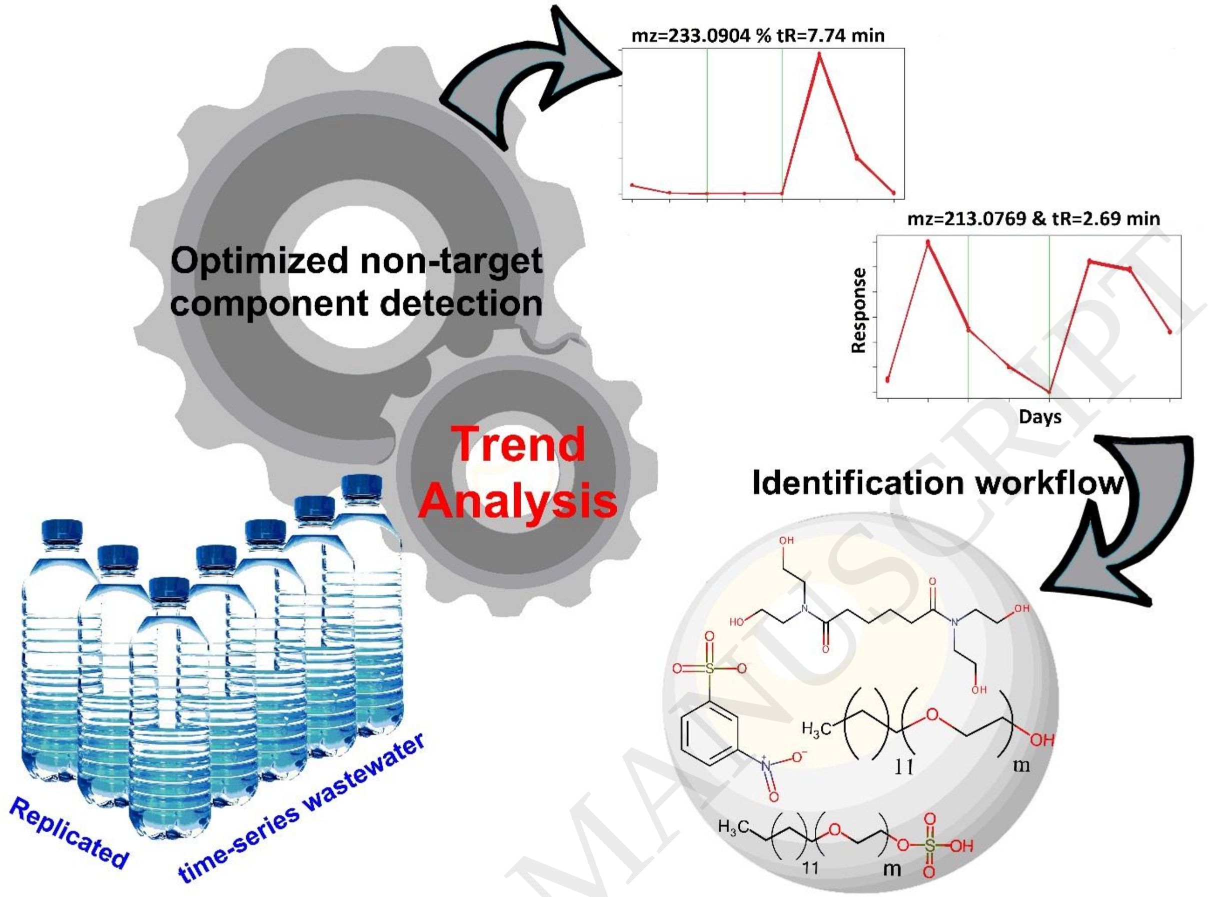
Fax: +30 210 7274750

E-mail: ntho@chem.uoa.gr

For Submission to: **Journal of Hazardous Materials**

- *Contains Supporting Information*

Graphical Abstract


mz=233.0904 % tR=7.74 min
mz=213.0769 & tR=2.69 min
Response
Days
Optimized non-target component detection
Trend Analysis
Identification workflow
Replicated
time-series wastewater

HighlightsNovel prioritization capable of detecting compounds with high fluctuation over time

- Application to LC-HRMS data of daily influent wastewater samples
- 30% of the prioritized compounds were tentatively identified
- Two compounds were reported in wastewater for the first time
- Four novel surfactant series were tentatively identified

Abstract

Peak prioritization plays a key role in non-target analysis of complex samples in order to focus the elucidation efforts on potentially relevant substances. The present work shows the development of a computational workflow capable of detecting compounds that exhibit large variation in intensity over time. The developed approach is based on three open-source R packages (xcms, CAMERA and TIMECOURSE) and includes the use of the statistical test Multivariate Empirical Bayes Approach to rank the compounds based on the Hotelling T2 coefficient, which is an indicator of large concentration variations of unknown components. The approach was applied to replicate series of 24-hour composite flow-proportional influent wastewater samples collected during 8 consecutive days. 60 events involving unknown substances with high fluctuation over time were successfully prioritized. 14 of those compounds were tentatively identified using HRMS/MS libraries, chemical databases, in-silico fragmentation tools, and retention time prediction models. Four compounds were confirmed with standards from which two never reported before in wastewater.

Keywords: Peak prioritization workflow; Trend analysis; Multivariate Empirical Bayes Approach; Non-target screening; Wastewater.

1. Introduction

Advances in high resolution mass spectrometry coupled to liquid chromatography (LC-HRMS) offer to environmental analytical chemists the opportunity to identify a continuously increasingly number of trace organic pollutants, even in highly complex environmental samples **[1]**. Target screening is insufficient to assess the quality of environmental waters as only a small portion of organic contaminants can be captured, while other relevant and potentially harmful substances cannot be detected **[2-4]**. Although still most investigations focus on target screening (where reference standards are available), there is an increasing number of studies dealing with both suspect screening (prior structural information of the suspects available, but no reference standards are available) and non-target screening (no prior information and no reference standards are available).

However, environmental samples are complex chemical mixtures containing tens of thousands of individual substances that produce a high number of peaks in LC-HRMS analysis. Their complete elucidation through the use of non-target strategies is not feasible, since it would require extensive time and effort. Thus, it is clear that the selection of the peaks of interest (*peak prioritization*) is a key step in any investigation involving non-target analysis. Depending on the goals of the study, different prioritization strategies should be applied to the set of obtained chromatographic peaks [5].

So far, most of the prioritization strategies followed intensity-based criteria in combination with the prioritization of substances with a distinctive isotopic pattern (e.g. halogenated compounds) **[4, 6-8]**, as these can be considered as relevant substances with reasonable identification changes. Other approaches used mass defect to focus identification efforts on molecular formulas outside the matrix domain in complex ~~sediment~~ samples **[9, 10].** It has also been proved useful when the objective is to find molecules with specific characteristics. An example was the detection of perfluoro-alkyl ether carboxylic acids and sulfonic acids in natural waters due to the negative mass defect of the multiple fluorine and oxygen atoms **[11].** Few studies conducted peak prioritization prior non-target analysis based on effect-directed analysis (EDA), a useful tool for identifying predominant toxicants in complex environmental mixtures combining effect testing and fractionation **[12-14]**. Other strategies include time series prioritization (prioritizing features whose intensities varied substantially over the time course of a sampling campaign in one sampling site) **[15, 16]**, are based on spatial variation **[17]** or use metabolic logic combined with multivariate statistics in order to find unknown metabolites of certain substances **[18]**. In the field of trend analysis, Schlüsener et al. **[15]**, used vendor software from SCIEX (MarkerView) to analyze long-time series LC-HRMS data coming from a sampling station of Rhine river which was affected by effluent wastewater. Afterwards, they used open-source scripts to visualize the patterns and to perform autocorrelation to search and prioritize the features with high periodic variations**.** Plassmann et al. used trend analysis to detect continuously increasing peak intensities and filter out peak signals from naturally-occurring substances in whole blood samples **[16].** Moreover, trend analysis has been used for assessing the quality of the chromatographic stability in LC-HRMS data using von Neumann trend test **[19, 20]**.

The main objective of the present study was the development of an automated prioritization workflow based on open-source tools that is capable of detecting automatically compounds that exhibit large variation in their intensity over time (trend-analysis). This new prioritization approach was realized by combining the different open-source R packages xcms, CAMERA and TIMECOURSE as well as the statistical test *Multivariate Empirical Bayes Approach* (MEBA) **[21].** The statistically obtained Hotelling T2 coefficient was used as an indicator of large intensity variations to rank the compounds. MEBA seemed to be the most suitable trend test for the generated dataset, because (i) it assesses longitudinal developmental time-series, (ii) it considers the repeatability of the replicates and (iii) it is not affected by progressive variations since data is not examined sequentially. Moreover, (iv) it accounts cumulatively for large variations among the different time points and (v) it is not affected by seasonality.

The developed workflow was applied to the evaluation of influent wastewater samples in order to detect events of direct disposal (e.g. due to illegal discharges) or sudden changes in the use of any substance. Replicated time-series of 24-hour flow proportional composite influent wastewater samples were taken during 8 consecutive days from a large wastewater treatment plant (WWTP) in Athens, Greece, which receives both urban and industrial wastewater. The compounds were ranked according to the developed procedure and elucidation efforts focused on the top-prioritized ones through the application of non-target identification strategies previously developed **[7]**.

2. Materials and Methods

2.1. Chemicals and reagents

All solvents used in the present work were UPLC-MS grade. Acetonitrile (ACN) and methanol (MeOH) were purchased from Merck (Darmstadt, Germany), whereas 2-propanol of LC-MS grade was obtained from Fisher Scientific (Geel, Belgium). Distilled water was provided by a Milli-Q purification apparatus (Millipore Direct-Q UV, Bedford, MA, USA). Sodium hydroxide monohydrate (NaOH) for trace analysis ≥99.9995% and formic acid 99% were purchased from Fluka (Buchs, Switzerland). Details on the used chemicals and reagents for sample preparation and standard compounds purchased for confirmation purposes are provided in the **Supporting Information (SI, SI-1).**

2.2. Sampling and storage

24-hour composite flow proportional influent wastewater samples were collected from the WWTP of Athens (Greece) during 8 consecutive days in March 2015. The location of the WWTP of Athens can be found on section **SI-2.** The WWTP is designed with primary sedimentation, activated sludge process with biological nitrogen and phosphorus removal and secondary sedimentation. The residential population connected to the WWTP based on official census, excluding commuters, is 3,700,000 and the number of people estimated based on the number of house connections is 4,562,500. The WWTP is designed to serve a population equivalent of 5,200,000 and thus is by far the largest in Greece and one of the largest in the world. The estimated sewage flow for the collected samples was 720,000 m^3 day^{-1}.

Raw influent wastewater was collected in pre-cleaned high-density polyethylene (HDPE) bottles. The samples were filtered with glass fiber filters (pore size 0.7 μm) immediately after arrival at the laboratory. They were stored in the dark at 4 °C until analysis, which happened directly after the end of the sampling campaign.

2.3. Sample preparation and instrumental analysis

Sample extraction was carried out using a slightly modified protocol developed by Kern et al.**[22]** In-house four sorbent SPE cartridges (200 mg Strata-X, 150 mg Isolute ENV+, 100 mg Strata-X-AW and 100 mg Strata-X-CW) were preconditioned with 6 mL with MeOH and 6 mL water. Cartridges were loaded with aliquots of 100 mL (preadjusted to pH 6.5), were dried under vacuum for 1 hour and were eluted with 4 mL of 50:50 MeOH:ethyl acetate containing 2% of ammonia, followed by 2 mL of 50:50 MeOH:ethyl acetate containing 1.7% of formic acid. Extracts were evaporated under a gentle nitrogen stream to a volume of 100 µL, reconstituted to 0.5 mL with a final proportion of 50:50 MeOH:water and filtered through a 0.2 µm RC syringe filter (Phenomenex, USA) .

Analyses were carried out using an UHPLC/QTOF-MS system, equipped with a UHPLC apparatus (Dionex UltiMate 3000 RSLC, Thermo Fisher Scientific, Germany), consisting of a solvent rack degasser, auto-sampler, a binary pump with solvent selection valve and a column oven coupled to the QTOF-MS/MS analyzer (Maxis Impact, Bruker Daltonics, Bremen, Germany). An Acclaim RSLC C18 column (2.1 × 100 mm, 2.2 µm) from Thermo Fisher Scientific (Dreieich, Germany), preceded by an ACQUITY UPLC BEH C18 1.7 µm, VanGuard Pre-Column from Waters (Dublin, Ireland), and thermostated at 30 °C, was used for separation.

All the samples were first analyzed in full scan mode. The QTOF-MS system was operating in broadband collision-induced dissociation (bbCID, data-independent) acquisition mode and recorded spectra over the range of *m/z* 50−1000 with a scan rate of 2 Hz. This mode provides MS and MS/MS spectra at the same time working at two different collision energies (4 and 25 eV). A second data-dependent MS/MS acquisition was conducted using a preselected inclusion mass list containing the exact masses of the precursor ion of selected compounds. The collision energy applied was set to predefined values, according to the mass and the charge state of every ion. Detailed information on the UPLC-MS/MS performance is provided in section **SI-3**.

2.4. Computational workflow

Raw files acquired from the LC-HRMS analysis were converted to mzML file format by using Proteowizard software **[23]** with the following conversion parameters: *Peak Picking,* true 1-; *MsLevel,* 1-1 and *Threshold peak filter,* absolute 300-most intense. The computational workflow and the prioritization methodology here-in proposed is based on functions available in three R-packages. In brief, functions for peak detection, matching peaks across the samples and OBI-Warp retention time alignment are included in the XCMS R package, while functions for componentization based on retention time and peak shape and functions for annotation of adducts and isotopic peaks are included in the CAMERA R-package. TIMECOURSE package was used for prioritization using the *one sample multivariate empirical Bayes statistic* developed by Tai and Speed **[21]**. A step-wise illustration of the computational workflow can be found at **Figure 1.**

(Figure 1)

Sample feature detection was the first step and it was carried out using the function xcmsSet() with optimized parameters for QTOF MS data (CentWave parameters can be found in Table 1). After that, features representing the same analyte across samples were placed into groups using the group() function. Retention time alignment was performed using retcor() function (based on the Kernel density estimator **[24]**). Since there were feature groups with missing features from some of the samples (e.g., because an analyte is not present in a particular sample), these missing features were filled with a low intensity value with fillPeaks() function [25]. This is important in order to avoid errors due to missing values of non-detected peaks in some samples, when performing statistical analysis. Then, features were clustered according to retention time (using groupFWHM() function) and further according to the peak shape correlation coefficient (using groupCorr () function). For this purpose xcmsSet objects were converted to CAMERA objects by using xsAnnotate() function. Finally, isotopic peaks and adducts were annotated using the functions findIsotopes() and findAdducts(), respectively **[26]**. Peaks detected in the blank samples (with an intensity ratio below one order of magnitude) were removed. Target compounds were excluded based on accurate mass (±mass accuracy window of 3 mDa) and retention time (±retention time window of 0.50 min). Discussion of target screening results is out of the scope of the present manuscript. All remaining components were normalized by log2 transformation. After that, the statistical test (Multivariate Empirical Bayes Approach **[21]**) was applied and compounds were ranked based on the Hotelling T2 coefficient, which can be used as an indicator of large concentration variations among daily composite samples.

2.5. Identification of unknown compounds

Identification of top prioritized components was based on the non-target approach established by Gago-Ferrero et al **[7]**. Possible molecular formulas were assigned by applying thresholds of mass accuracy (≤2 mDa) and isotope pattern (mSigma≤50 [27]). If elucidation of the molecular formula was not unequivocal based on mass accuracy and isotope pattern, MS/MS was also considered using Molgen-MS/MS software **[28]**. Molgen-MS/MS was used with the parameter following settings: Elements - C, H, N, O, P, S (unless there was evidence of halogens), existence filter "exist", odd electron ions (oei), ppm = 5 and acc = 15 (MS and MS/MS accuracy settings in ppm). Once determined the molecular formula, candidates were obtained through the evaluation of the MS/MS spectra, including the use of in silico fragmentation platforms (Metfrag **[29]** via Metfusion **[30]**) and the MassBank library **[31]**. Commercial importance criteria was also used through the evaluation of the number of references and data sources in Chemspider **[32]** and the number of patents in Pubchem **[33]**. The chromatographic retention time plausibility of the candidates was evaluated, using an in-house QSRR retention time prediction model [34].

In four cases, the identity of the unknowns was confirmed by purchasing the corresponding standard and comparison of the t_R and MS/MS spectrum. Spectral similarity values were calculated with the OrgMassSpecR package in R **[35, 36]**. Confirmation was considered successful only when t_R deviation was below 0.2 min and MS/MS spectrum similarity was higher than 70%. The level of confidence for the identification of the detected compounds was used according to Schymanski et al. **[37]**, where Level 1 corresponds to confirmed structures (reference standard is available), level 2 to probable structures, level 3 for tentative candidate(s), Level 4 to unequivocal molecular formulas, and level 5 to exact mass(es) of interest.

3. Results and Discussion

3.1. Optimization of the computational workflow to obtain component lists

The computational workflow established in order to obtain the compound list consists of three basic steps: peak picking, matching peaks across the samples and chromatographic t_R alignment. Different input parameters in the aforementioned steps (e.g., mass accuracy or peak width in centWave peak picking algorithm) may lead to different compound lists **[38, 39]**. Therefore, parameters were optimized by Box-Behnken fractional factorial design (IPO R-package) **[38]**. Optimized values for each parameter are summarized in Table **1** and are discussed below.

(**Table 1**)

IPO optimization is based on natural stable ^{13}C isotopic peaks. It calculates a peak picking score based on reliable peaks, meaning peaks for which their corresponding isotopes have been detected. This score combined with the total number of detected peaks and the number of low intensity peaks (isotopes may remain undetectable) is used as response variable. Peak picking parameters are tuned, so that the response variable is maximized following design of experiments method **[38]**. Optimum values for mass accuracy and peak width were almost the same in positive and negative ionization mode (17.6 ppm, ~15 sec (*minimum peakwidth*) and 50 sec (*maximum peakwidth))*. The similarity in ESI(+) and ESI(-) mode was expected since the same separation method and instrument was used for analysis of the extracts in both polarities. The obtained mass accuracy threshold is lower than those used in most of the predefined methods in R-based online platforms (Scripps center for metabolomics (xcmsonline) **[40]**), where normally ~ 30 ppm is applied for QTOF data, and therefore decreasing the number of false positives. This example shows that optimizing input parameters prior to data treatment is important for proper dataset generation and therefore prioritization. Other additional filters included such as *prefilter*, which is used in order to avoid peaks with very low intensity. When applying this filter, a given mass should be present at least in three consecutive scans with an intensity threshold (≥3000, ESI(+) and ≥1000, ESI(-)). Another filter was *scanrange*, which helps to avoid calibrant peaks by restricting peak picking to specific time intervals. In our case, calibrant substance was injected in the beginning of each chromatographic run using a 6-port valve and calibrant peaks appear for 12 consecutive scans, which were excluded by using the *scanrange* filter.

Only features existing in 3 out of the 5 replicates were kept by setting the parameter *minfrac* (minimum fraction of samples in a subgroup) to 0.6. After that, the kernel density estimator method was used for matching peaks across the samples (grouping together peaks representing the same analyte in different samples). In this regard, the parameters *bw* (bandwidth of kernels) and *mzwid* (width of overlapping *m/z* slices), which indicate time tolerance and mass accuracy, respectively, were optimized (**Table 1, part grouping of features based on kernel density estimator**).

The next step consisted of retention time alignment. It was performed by using the *ordered bijective interpolated warping (OBI-Warp)* algorithm **[24]**. Two penalty parameters, *gapInit* and *gapExtend*, which prevent the over-alignment of the chromatograms, were optimized. Optimization of grouping and retention time alignment takes place at the same time and is based on peaks appearing in all samples. Response variable is a linear combination of grouping response variable and retention time alignment response variable **[38]**. The obtained values were very similar to those obtained by Prince and Marcotte **[24] (Table 1)**. Moreover, it was observed that the maximum chromatographic drift during the analysis was 20 and 10 s (ESI(+) and ESI(-), respectively), showing the robustness of the chromatographic system.

Since ESI is a soft ionization technique, several ion species can be observed for the same compounds (e.g., adducts or isotopes). In order to obtain the final compound list, the peaks belonging to the same compound were grouped. This was conducted using the CAMERA R-package **[26]**. This package can group the peaks based on retention time and peak shape and annotates isotopic and adduct peaks. Finally, to avoid prioritizing known substances, 207 and 32 target components in positive and negative ionization respectively were excluded (target list of University of Athens consisted of 2249 compounds and is available at NORMAN Suspect list exchange http://www.norman-network.com/?q=node/236).

3.2. Prioritization methodology

To find the compounds exhibiting high fluctuation among the daily samples, the *one-sample Multivariate Empirical Bayes Approach (MEBA)* statistical test was applied. This test is suitable for longitudinal replicated developmental time-course data. Originally, this statistical test was designed to solve the problem of ranking genes in microarray experiments **[21]**. MEBA has advantages compared to other F-statistic approaches (i.e. ANOVA) since it incorporates replicate variances and the correlations among responses of time–series samples from longitudinal data.

Intensity normalization is a mandatory step for statistical hypothesis testing. Therefore, as a first step Log2 transformation was performed in the dataset (compound list), since it is the most appropriate transformation for the applied statistical test **[21]**. Then, the statistical test was applied to every compound and a score (Hotelling T2) was assigned based on the peak area values observed in the time-series samples. This score is a positive number without an upper limit, which takes into account the repeatability of the intensity among replicates representing one time point and the magnitude of change of intensity between time points. A high value indicates high fluctuation among the time series samples. Compounds were ranked according to the score and the results for the first top 30 prioritized substances in each ionization mode are summarized in **table 2** and in SI (section **SI-4, tables S4A and S4B**).

Through the evaluation of the graphs several compounds with a pollution spill trend could be observed. The graphics for the compounds with this behavior are summarized in the SI (**SI-5a**) and in **Figure 2~~1~~** (selected cases).

(**Figure 2**)

Spill trend cases were compounds detected in specific samples (normally at high intensities), while remain undetectable in the other samples. This becomes obvious for the top-ranked compounds and especially for the cases **#P1, #P2** (**Figure 2**) as well as for the others depicted in **Figure S5b**, which exhibit extreme changes in intensities and were mainly found in one daily sample. Cases of pollution spills can also include compounds that can be detected in most of the samples at low intensity but the signal increase disproportionately in specific samples. The most obvious cases are **#P12** and **#P10**, where the signal increased more than 5 and 18 times, respectively, compared to the average intensity. Compounds belonging to this pollution spill category are of crucial environmental importance, since they can reach high concentration levels and become potentially toxic for the ecosystem. The detection and identification of these substances may allow the authorities to trace the pollution source and adopt appropriate measures.

Apart from the cases of pollution spills, also compounds with dropping signal intensities during specific days were determined through the application of the developed prioritization methodology. Examples of this behavior for the compounds **#N18** and **#N16** are depicted in **Figure 2b.** The signal decreased very significantly during the weekend period indicating an industrial origin.

Several of the prioritized compounds corresponded to substances exhibiting the same time pattern. In almost all cases of successful identifications these substances were identified as surfactants belonging to different homologue series. **Figure 2c** shows an example with three different surfactants sharing the same time pattern. More examples of this behavior are depicted in Figure S6B (SI). These compounds obviously share a common origin and even might coexist in products. However, in order to draw sound conclusions with other groups of substances more successful identifications would be required.

Evaluating the ranking, substances that remained undetectable in at least one sample (very low number or zero assigned by fill gaps function) also received a relatively high score, since the statistical approach is highly affected by this fact. This is the reason why some compounds (e.g. #N15 or #P7) were prioritized even though their pattern in the rest of the time-points seems almost steady. Also, the score of a compound decreases when the repeatability within replicates of a sample time point is low, since MEBA takes into account all the replicates. This is the main reason why compounds with very similar trends are ranked slightly different (e.g. #N9-#N12, SI). Despite of the aforementioned disadvantages the prioritization approach provided good results and it was proved capable of detecting pollution spills and compounds exhibiting high fluctuation over time.

3.3. Identification of top-ranked prioritized compounds

Identification efforts were focused on the first 30 prioritized components in each ionization mode (60 potential compounds in total) and the results are summarized in **Tables 2** (tentatively identified compounds) and **S4a**, **S4b** for prioritized but not tentatively identified compounds). In ESI (+), two substances, 3,6,9,12-tetraoxatetracosan-1-ol and N,N,N`,N`-tetrakis(2-hydroxyethyl)hexanediamide, were confirmed with the corresponding commercial standard, reaching the confidence level 1. Eight additional substances were tentatively identified; all of them as tentative candidates (level 3). For eight compounds it was not possible to go beyond the determination of the unequivocal molecular formula (level 4) and the remaining twelve compounds remained as exact mass of interest (level 5). In ESI (-), two substances, 3-nitrobenzenesulfonic acid and lauryl sulfate, were confirmed, one reached confidence level 2 and five compounds reached level 3. For twelve additional compounds an unequivocal molecular formula was assigned (level 4) and ten peaks remained at level 5.

An interesting case was the identification of the compound N,N,N`,N`-tetrakis(2-hydroxyethyl)hexanediamide (CAS: 6334-25-4) (case **#P2** in **table 2**). A peak corresponding to *m/z* 321.2033 (t_R 2.70 min) was prioritized and the unequivocal molecular formula $C_{14}H_{28}N_2O_6$ was assigned based on the mass accuracy, the isotope pattern and the annotation of the fragments. There were 38 compounds with this formula in the ChemSpider database. The MS/MS spectra indicated a neutral loss of 105.078 (corresponding to $C_4H_{11}NO_2$) and the loss of a H_2O molecule. The structure corresponding to the confirmed compound received the highest MetFrag score and was within the top 4 MetFusion candidates. Moreover, this compound was the one with the highest commercial importance (38 data sources, 41 references and 7 patents in Chemspider and PubChem, respectively) in comparison with the other candidates. In addition, the confirmed compound received the closest predicted retention time, indicating that models for prediction of chromatographic behavior can be useful for helping in revealing the identity of unknown compounds. Finally, the identity of the substance was confirmed with a commercial standard. This substance was present in 3 out of the 8 evaluated days, two of them at an almost negligible intensity (~ 3 until 9×10^4) and at very high intensity on the other day (Wednesday 11th March 2015 (3.6×10^7). Therefore, this is a characteristic example of a pollution spill-trend. This chemical is mainly used in the fabrication of adhesives, where it is added in order to enhance their performance by acting as cross-linker **[41]**. An intensive use of this substance during the specific day of 11th March 2015 by some adhesive industry with resulting high concentrations in the discharged wastewater or an event of direct disposal of this chemical into the sewage system are plausible hypothesis to explain the observed behavior. Another interesting example of a compound with a pollution spill trend can be found in the case **#N1** (**Table 2** and **Figure S6b**). This compound was tentatively identified as hydroxybenzenesulfonic acid (level 3). On a specific day (Wednesday, 4th of March 2015), it was determined at a concentration 5 times higher than the average of the remaining days of the sampling campaign.

A compound showing lower concentration levels during the weekend was 3-nitrobenzenesulfonic acid (CAS 98-47-5, case **#N16**, **table 2**), which was confirmed with a standard, reaching level 1. This compound is used in electrical/electronics, photographic, and textile processing industries **[42]**. The behavior of this compound can be explained due to the fact that these industries do not operate (at least at the same level) during weekends leading to decreasing concentrations. Another substance which also showed lower levels during the weekend days was tentatively identified (level 3) as the glucuronated derivative of 3-methylcyclopent-2-enone (CAS: 251914-61-1) (case **#P9**). This chemical is used in the food industry as a color additive **[43]**. For other compounds following exactly the same trend (summarized in **Figure S6B**), it was not possible to go beyond level 4, mostly due to the high number of potential candidates.

Several surfactants belonging to the homologous series $CH_3(CH_2)_{11}(CH_2CH_2O)_XSO_4$ (X=1…12), which have not been previously reported in wastewater, were detected and identified as it is shown in **Figure 3**. These compounds corresponded to the cases #P20 m/z 437.1973 (t_R: 12.66 min), #P22 m/z 481.2233 (t_R: 12.62 min), #P19 m/z 525.2544 (t_R: 12.81 min) and #P21 m/z 569.2756 (t_R: 12.89 min) (**Table S5a**). In both ionization modes, consistent peak shapes and constant increase of t_R were observed when increasing the chain length. The proposed structures can explain all the fragments obtained in the ESI(+)-QTOFMS (**Figure 3**). All the spectra corresponding to the homologous series were very similar, showing in all cases characteristic fragments at *m/z* 45.0334, 89.0597, 133.0859 and 177.1121, corresponding to the group $(CH_2CH_2O)_x$ (x=1-4). The protonated adduct could not be detected in ESI(+)-QTOFMS. However, the adducts $[M+NH_4]^+$ and $[M+K]^+$ showed high intensity, in agreement with other studies dealing with identification of surfactants **[7, 44]**. Substances with the same molecular formulas and time trend were also detected in ESI(-)-QTOFMS. The presence in the MS/MS spectra of the characteristic fragments with m/z=79.7574 (SO_3^-), m/z=96.9601 (HSO_4^-) and 122.9758 ($C_2H_3SO_4^-$) supports the proposed structures. Although ChemSpider and PubChem databases only provided linear chain candidates, ramified compounds may exist (and similar MS/MS spectra are expected). Therefore, a level of confidence 3 was assigned to these substances.

(**Figure 3**)

Other identified surfactants included the substances $CH_3(CH_2)_{12}(CH_2CH_2O)_4OSO_3H$ (level 3), $CH_3(CH_2)_{11}(CH_2CH_2O)_4OH$ (level 1) and $CH_3CH{=}CH(CH_2)_{15}(OCH_2CH_2)_{10}OH$ (level 3) and the additional compounds belong to the respective homologue series detected through retrospective analysis (**Figure S7A, S7B and S7C, respectively**). In all these cases consistent t_R shifts, peak shapes and MS/MS spectra were observed. All these identifications indicate that several surfactants and their corresponding transformation products remain unreported in wastewater yet.

The spectra of the successfully confirmed substances were uploaded in MassBank database (AU4064, AU4065, AU4066, AU4067) in order to assure their easy accessibility for the community of analytical environmental chemists.

4. Conclusions

The developed computational workflow was successfully optimized for critical parameters as demonstrated for influent samples from the WWTP of Athens. The statistical test MEBA, which ~~have~~ had not been used before in such identification workflows, was successfully used to prioritize compounds with large concentration variations among the samples. This success of the workflow was demonstrated by tentative identification of 14 compounds wherefrom two compounds detected for the first time in raw wastewater.

The development of new prioritization methods capable to prioritize and identify unknown compounds in environmental samples is important as non-target screening becomes wide-spread. Smart prioritization strategies combining the power of LC-HRMS with advanced statistics can lead to a much better understanding of the environment from a chemical point of view. However, the current lack of an interface to host the developed prioritization approaches prevents transparent comparison of the different approaches and standardization of the methods. It also complicates the application of multiple methods to the same set of samples which may lead to the identification of an increasing number of unknown compounds. The development of unified interfaces that solve the aforementioned limitations in combination with platforms for the storage of large mass spectrometric data would provide important advances to better understand the presence and fate of micropollutants in the environment.

Acknowledgements

This project was implemented under the Operational Program «Education and Lifelong Learning» and funded by the European Union (European Social Fund) and Greek National Resources – ARISTEIA 624 (TREMEPOL project). Nikiforos A. Alygizakis would like to thank John S. Latsis Foundation (http://www.latsis-foundation.org/eng) for the financial support. Authors acknowledge Emma Schymanski for providing the R-scripts for plotting homologue series.

References

[1] B. Petrie, R. Barden, B. Kasprzyk-Hordern, A review on emerging contaminants in wastewaters and the environment: current knowledge, understudied areas and recommendations for future monitoring, Water research, 72 (2015) 3-27.

[2] M. Krauss, H. Singer, J. Hollender, LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns, Analytical and bioanalytical chemistry, 397 (2010) 943-951.

[3] M. Ibáñez, J.V. Sancho, F. Hernández, D. McMillan, R. Rao, Rapid non-target screening of organic pollutants in water by ultraperformance liquid chromatography coupled to time-of-light mass spectrometry, TrAC Trends in Analytical Chemistry, 27 (2008) 481-489.

[4] E.L. Schymanski, H.P. Singer, P. Longree, M. Loos, M. Ruff, M.A. Stravs, C. Ripolles Vidal, J. Hollender, Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry, Environmental science & technology, 48 (2014) 1811-1818.

[5] J. Hollender, E.L. Schymanski, H.P. Singer, P.L. Ferguson, Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go?, Environmental science & technology, 51 (2017) 11505-11512.

[6] C. Hug, N. Ulrich, T. Schulze, W. Brack, M. Krauss, Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening, Environmental pollution, 184 (2014) 25-32.

[7] P. Gago-Ferrero, E.L. Schymanski, A.A. Bletsou, R. Aalizadeh, J. Hollender, N.S. Thomaidis, Extended suspect and non-target strategies to characterize emerging polar organic contaminants in raw wastewaters with LC-HRMS/MS, Environmental science & technology, 49 (2015) 12333-12341.

[8] P. Gago-Ferrero, E.L. Schymanski, J. Hollender, N.S. Thomaidis, Chapter 13 - Nontarget Analysis of Environmental Samples Based on Liquid Chromatography Coupled to High Resolution Mass Spectrometry (LC-HRMS), in: S. Perez, P. Eichhorn, D. Barcelo (Eds.) Applications of Time-of-Flight and Orbitrap Mass Spectrometry in Environmental, Food, Doping, and Forensic Analysis, Elsevier, 2016, pp. 381-403.

[9] A.C. Chiaia-Hernandez, E.L. Schymanski, P. Kumar, H.P. Singer, J. Hollender, Suspect and nontarget screening approaches to identify organic contaminant records in lake sediments, Analytical and bioanalytical chemistry, 406 (2014) 7323-7335.

[10] Y. Verkh, M. Rozman, M. Petrovic, A non-targeted high-resolution mass spectrometry data analysis of dissolved organic matter in wastewater treatment, Chemosphere, 200 (2018) 397-404.

[11] M. Strynar, S. Dagnino, R. McMahen, S. Liang, A. Lindstrom, E. Andersen, L. McMillan, M. Thurman, I. Ferrer, C. Ball, Identification of Novel Perfluoroalkyl Ether Carboxylic Acids (PFECAs) and Sulfonic Acids (PFESAs) in Natural Waters Using Accurate Mass Time-of-Flight Mass Spectrometry (TOFMS), Environmental science & technology, 49 (2015) 11622-11630.

[12] W. Brack, Effect-directed analysis: a promising tool for the identification of organic toxicants in complex mixtures?, Analytical and bioanalytical chemistry, 377 (2003) 397-407.

[13] C.M. Gallampois, E.L. Schymanski, M. Krauss, N. Ulrich, M. Bataineh, W. Brack, Multicriteria approach to select polyaromatic river mutagen candidates, Environmental science & technology, 49 (2015) 2959-2968.

[14] R. Zaja, S. Terzic, I. Senta, J. Loncar, M. Popovic, M. Ahel, T. Smital, Identification of P-glycoprotein inhibitors in contaminated freshwater sediments, Environmental science & technology, 47 (2013) 4813-4821.

[15] M.P. Schlusener, U. Kunkel, T.A. Ternes, Quaternary Triphenylphosphonium Compounds: A New Class of Environmental Pollutants, Environmental science & technology, 49 (2015) 14282-14291.

[16] M.M. Plassmann, E. Tengstrand, K.M. Aberg, J.P. Benskin, Non-target time trend screening: a data reduction strategy for detecting emerging contaminants in biological samples, Analytical and bioanalytical chemistry, 408 (2016) 4203-4208.

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

[18] J.E. Schollee, E.L. Schymanski, S.E. Avak, M. Loos, J. Hollender, Prioritizing Unknown Transformation Products from Biologically-Treated Wastewater Using High-Resolution Mass Spectrometry, Multivariate Statistics, and Metabolic Logic, Analytical chemistry, 87 (2015) 12121-12129.

[19] J. von Neumann, Distribution of the Ratio of the Mean Square Succesive Difference to the Variance, The Annals of Mathematical Statistics, 4 (1941) 367-395.

[20] T. Bader, W. Schulz, K. Kummerer, R. Winzenbacher, General strategies to increase the repeatability in non-target screening by liquid chromatography-high resolution mass spectrometry, Analytica chimica acta, 935 (2016) 173-186.

[21] Y.C. Tai, T.P. Speed, A multivariate empirical Bayes statistic for replicated microarray time course data, The Annals of Statistics, 34 (2006) 2387-2412.

[22] S. Kern, K. Fenner, H. Singer, R.P. Schwarzenbach, J. Hollender, Identification of transformation products of organic contaminants in natural waters by computer-aided prediction and high-resolution mass spectrometry, Environmental science & technology, 43 (2009) 7039-7046.

[23] M.C. Chambers, B. Maclean, R. Burke, D. Amodei, D.L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T.A. Baker, M.Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S.L. Seymour, L.M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E.W. Deutsch, R.L. Moritz, J.E. Katz, D.B. Agus, M. MacCoss, D.L. Tabb, P. Mallick, A cross-platform toolkit for mass spectrometry and proteomics, Nat Biotechnol, 30 (2012) 918-920.

[24] J.T. Prince, E.M. Marcotte, Chromatographic alignment of ESI-LC-MS Proteomics data sets by ordered bijective interpolated warping, Analytical chemistry, 45 (2006) 6140-6157.

[25] C.A. Smith, E.J. Want, G. O'Malle, R. Abagyan, G. Sluzdak, XCMS: Processing Mass Spectometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching and Identification., Analytical chemistry, 78 (2006) 779-787.

[26] C. Kuhl, R. Tautenhahn, C. Bottcher, T.R. Larson, S. Neumann, CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets, Analytical chemistry, 84 (2012) 283-289.

[27] B. Daltonics, Challenges in Metabolomics addressed by targeted and untargeted UHR-Q-TOF analysis, in, 2010.

[28] M. Meringer, E.L. Schymanski, Small Molecule Identification with MOLGEN and Mass Spectrometry, Metabolites, 3 (2013) 440-462.

[29] S. Wolf, S. Schmidt, M. Muller-Hannemann, S. Neumann, In silico fragmentation for computer assisted identification of metabolite mass spectra, BMC bioinformatics, 11 (2010) 148.

[30] M. Gerlich, S. Neumann, MetFusion: integration of compound identification strategies, Journal of mass spectrometry : JMS, 48 (2013) 291-298.

[31] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, MassBank: a public repository for sharing mass spectral data for life sciences, Journal of mass spectrometry : JMS, 45 (2010) 703-714.

[32] RSC ChemSpider, ChemSpider (www.chemspider.com), Royal Society of Chemistry, (2018) Accessed 01 Aug 2018.

[33] NCBI PubChem, PubChem (https://pubchem.ncbi.nlm.nih.gov), National Center for Biotechnology Information, (2018) Accessed 01 Aug 2018.

[34] R. Aalizadeh, N.S. Thomaidis, A.A. Bletsou, P. Gago-Ferrero, Quantitative Structure–Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples, J Chem Inf Model, 56 (2016) 1384-11398.

[35] OrgMassSpecR: Organic Mass Spectrometry. R package version 0.4-4., http://CRAN.R-project.org/package=OrgMassSpecR.

[36] E.S. Stein, Optimization and Testing of Mass Spectral Libray Search Algorithms for Compound Identification., J. Am. Soc. Mass Spectrom., 5 (1994) 859-866.

[37] E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, Identifying small molecules via high resolution mass spectrometry: communicating confidence, Environmental science & technology, 48 (2014) 2097-2098.

[38] G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, S. Neumann, G. Trausinger, F. Sinner, T. Pieber, C. Magnes, IPO: a tool for automated optimization of XCMS parameters, BMC bioinformatics, 16 (2015) 118.

[39] R. Tautenhahn, C. Bottcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC bioinformatics, 9 (2008) 504.

[40] R. Tautenhahn, G.J. Patti, D. Rinehart, G. Siuzdak, XCMS Online: a web-based platform to process untargeted metabolomic data, Analytical chemistry, 84 (2012) 5035-5039.

[41] M. Puig, L. Cabedo, J.J. Gracenea, A. Jiménez-Morales, J. Gámez-Pérez, J.J. Suay, Adhesion enhancement of powder coatings on galvanised steel by addition of organo-modified silica particles, Progress in Organic Coatings, 77 (2014) 1309-1315.

[42] J.A. Brown, M.D. M.P.H., Haz-Map®: Information on Hazardous Chemicals and Occupational Diseases, in, 2015.

[43] G.A. Burdock, Encyclopedia of food and color additives, in: G. Jaffe (Ed.) Encyclopedia of food and color additives, CRC Press, U.S.A., 1997, pp. 1751.

[44] E.L. Schymanski, H.P. Singer, J. Slobodnik, I.M. Ipolyi, P. Oswald, M. Krauss, T. Schulze, P. Haglund, T. Letzel, S. Grosse, N.S. Thomaidis, A. Bletsou, C. Zwiener, M. Ibanez, T. Portoles, R. de Boer, M.J. Reid, M. Onghena, U. Kunkel, W. Schulz, A. Guillon, N. Noyon, G. Leroy, P. Bados, S. Bogialli, D. Stipanicev, P. Rostkowski, J. Hollender, Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis, Analytical and bioanalytical chemistry, 407 (2015) 6237-6255.

Table 1. Parameters used for the computational analysis

Input Parameter	POSITIVE ESI	negative ESI
CentWave parameters		
ppm	17.6	17.6
Minimum peak width	14.34	15.5
Maximum peak width	50	50
prefilter	3, 3000	3, 1000
scanrange	20 until 1840	20 until 1840
fitgauss	TRUE	TRUE
integrate	TRUE	TRUE
Retention Time alignment based on OBI-Warp algorithm		
Distance function	cor_opt	cor_opt
gapInit	0.3	0.27
gapExtend	2.4	2.36
Grouping of features based on kernel density estimator		
bw	5	5
mzwid	0.032	0.0305
minfrac	0.6	0.6
minsamp	2	2
max	50	50

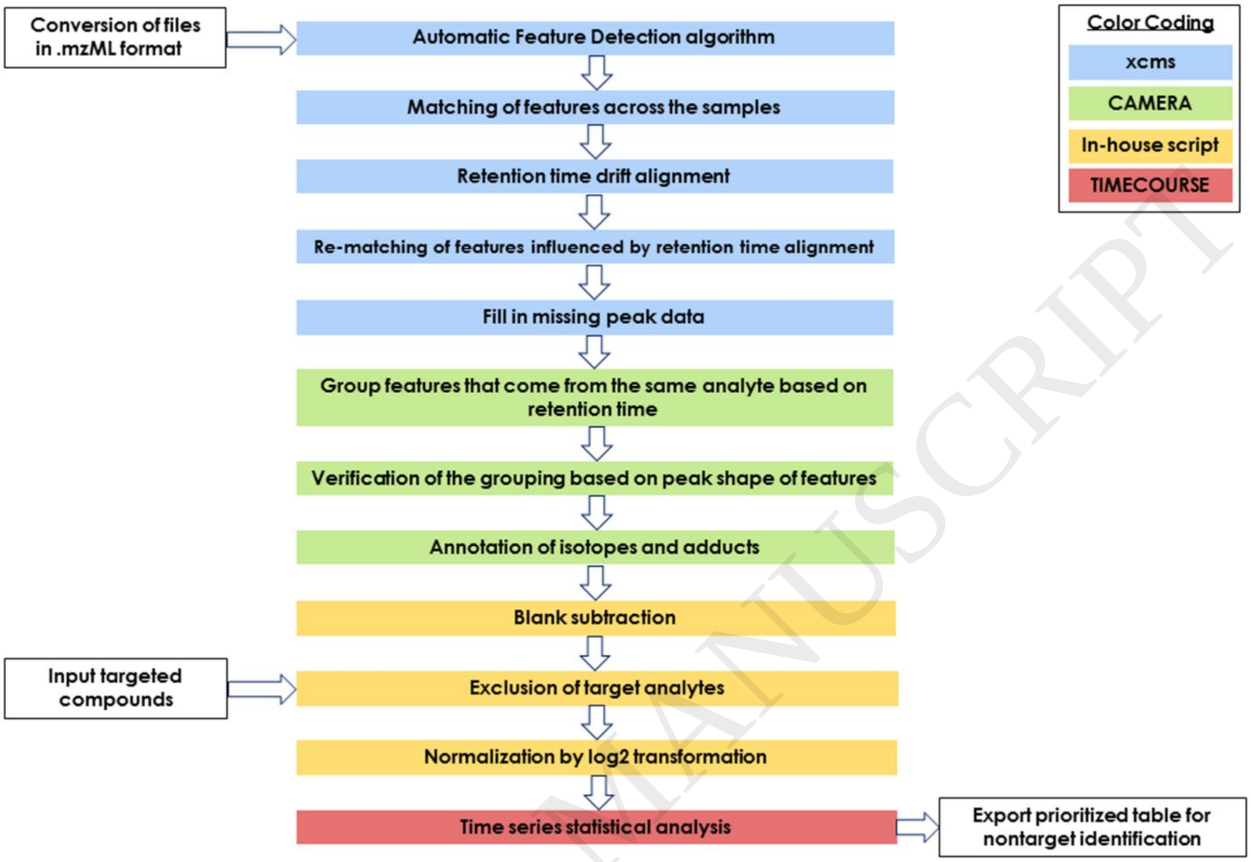

Conversion of files in .mzML format
Automatic Feature Detection algorithm
Matching of features across the samples
Retention time drift alignment
Re-matching of features influenced by retention time alignment
Fill in missing peak data
Group features that come from the same analyte based on retention time
Verification of the grouping based on peak shape of features
Annotation of isotopes and adducts
Blank subtraction
Input targeted compounds
Exclusion of target analytes
Normalization by log2 transformation
Time series statistical analysis
Export prioritized table for nontarget identification
Color Coding
xcms
CAMERA
In-house script
TIMECOURSE

Figure 1. Compiled and optimized workflow for detecting compounds with a characteristic intensity fluctuation over time.

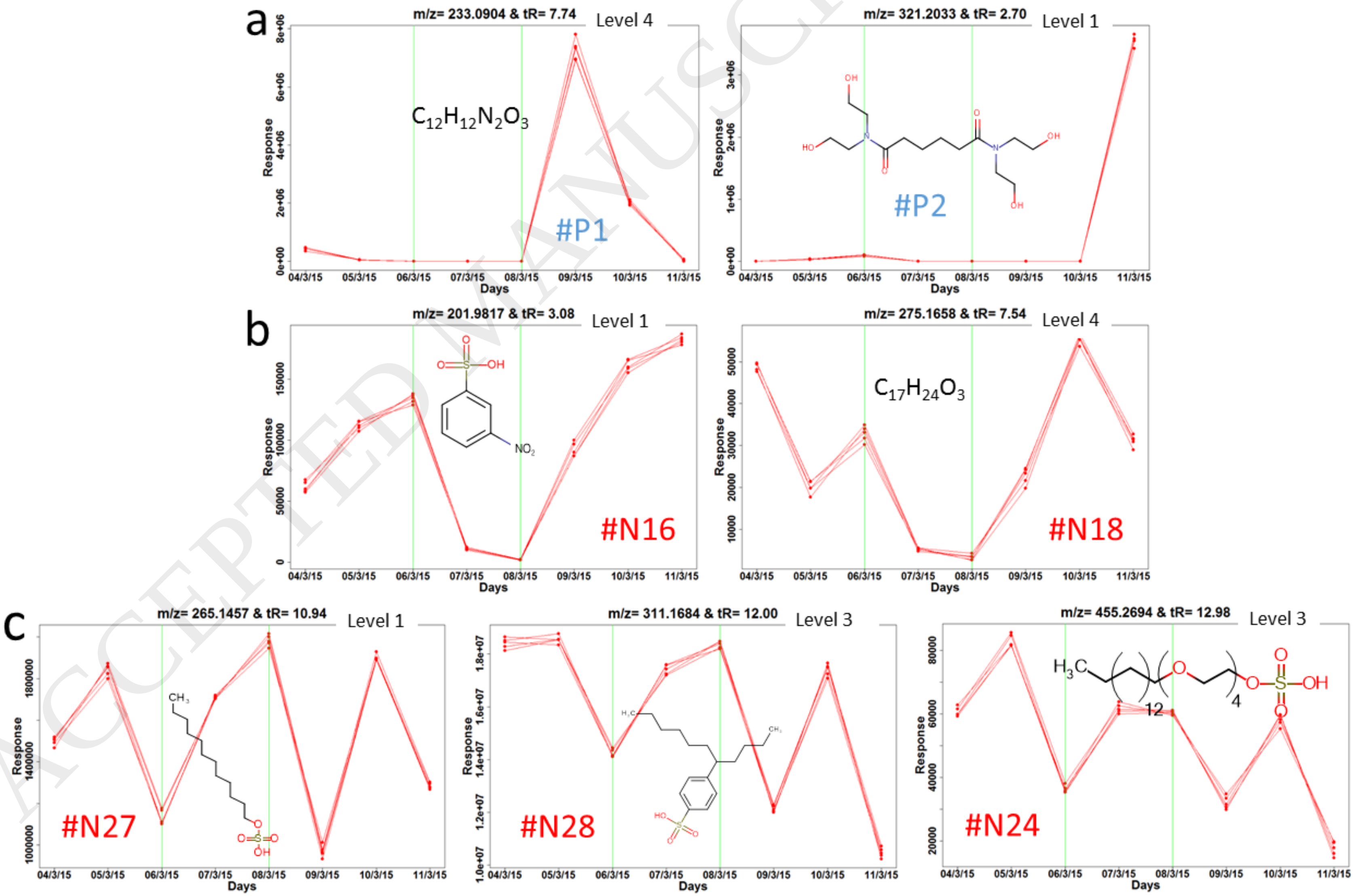


Figure 2. (a) Examples of pollution spills (events of direct disposal of chemicals into the sewage system); (b) Examples of compounds with dropping response during the weekend (The space between the green lines correspond to the weekend); (c) Surfactant compounds sharing a common time trend.

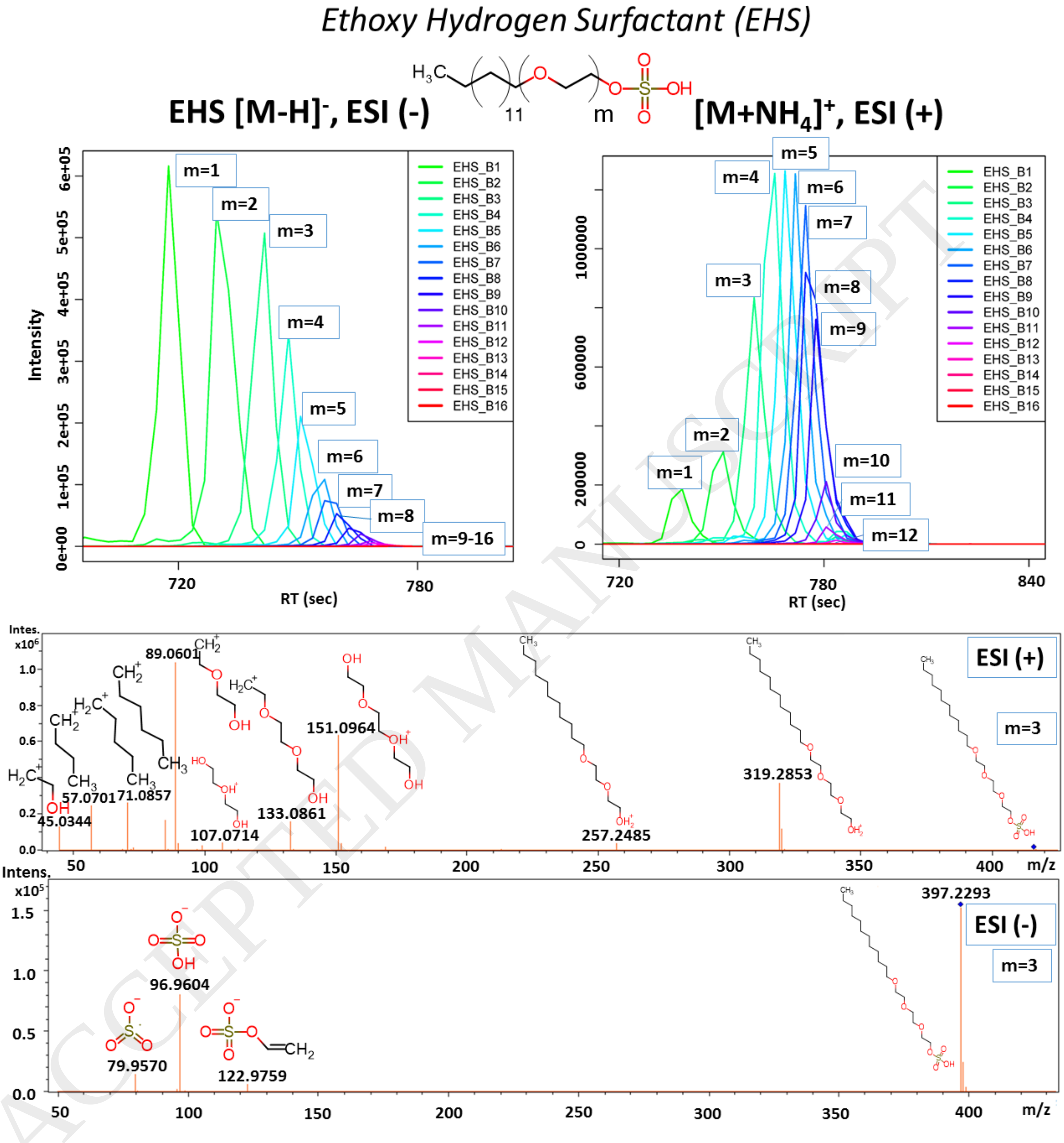
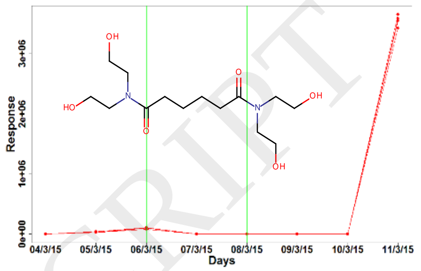
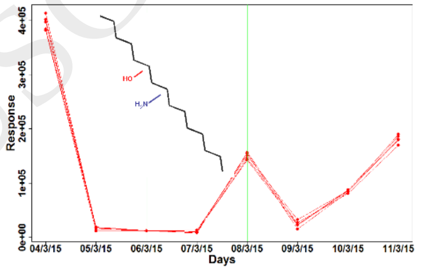
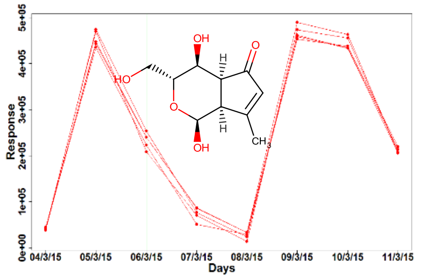
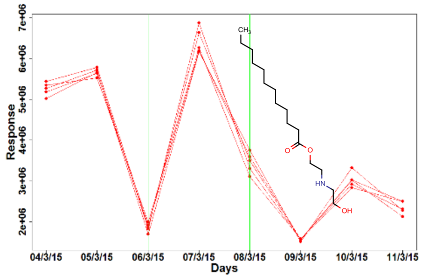
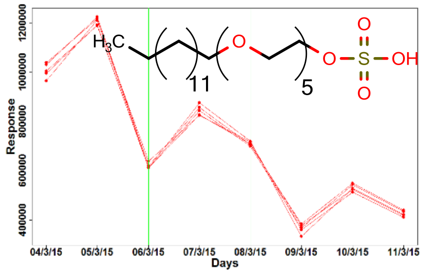
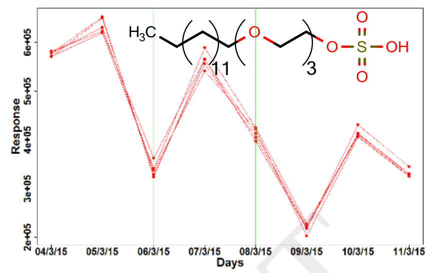
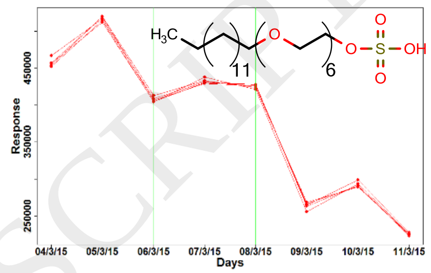
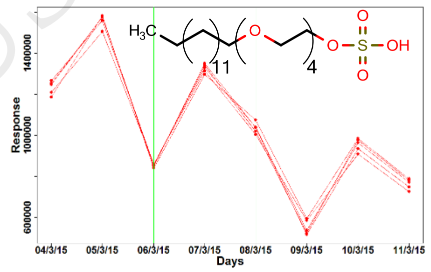
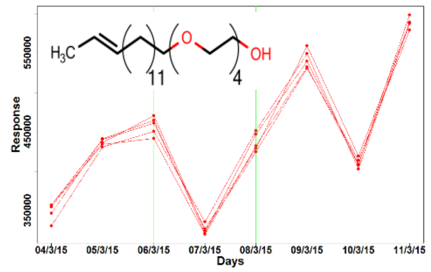
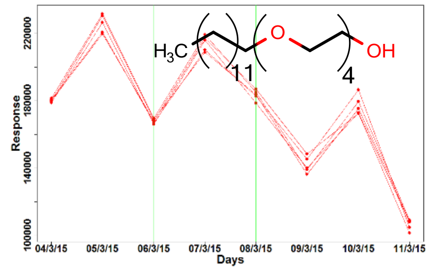
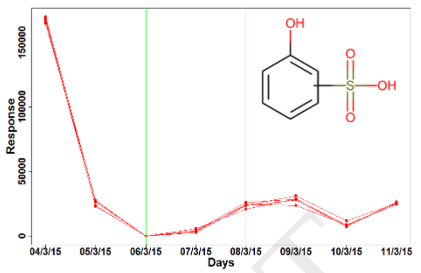
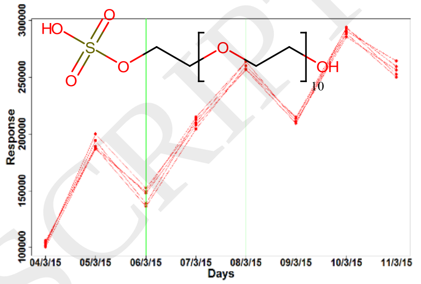
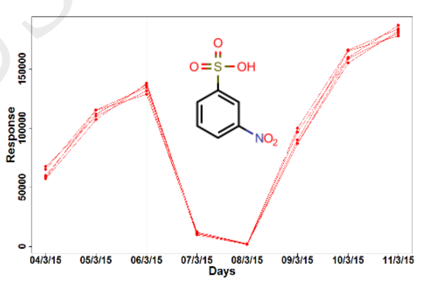
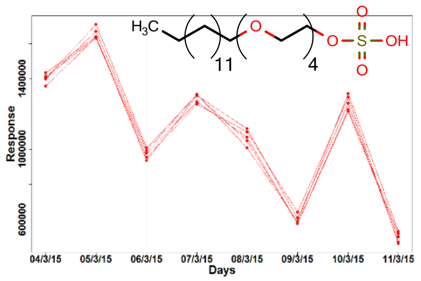
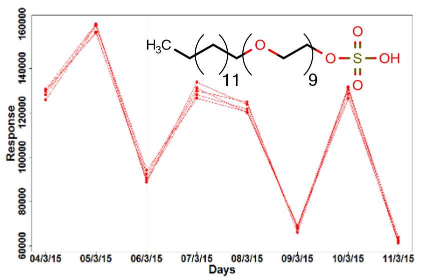


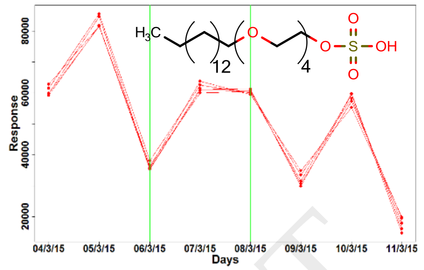
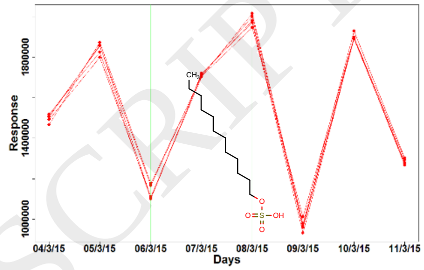
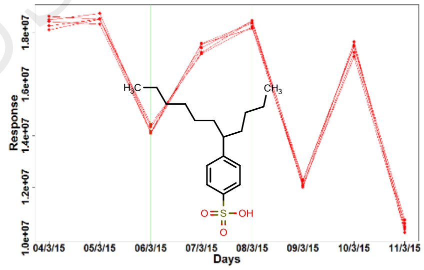
Figure 3. Tentative identification (Level 3) of a novel ethoxy hydrogen surfactant (EHS) homologue series in negative and positive ionization mode. Spectra correspond to m=3.

Table 2. Summary of the results for the prioritized and tentatively identified compounds in positive and negative ionization mode.

Rank	m/z	Molecular formula (Name if available)	t_R (min)	Pred. t_R (min)	Level of confidence	Time trend
#P2	321.2033	$C_{14}H_{28}N_2O_6$ N,N,N',N'-Tetrakis(2-hydroxyethyl) hexanediamide)	2.70	2.80	1	
#P3	259.2822	$C_{16}H_{35}NO$	12.98	-	3	
#P9	215.0916	$C_{10}H_{14}O_5$ (1S,3R,4S,4aS,7aS)-1,4-dihydroxy-3-(hydroxymethyl)-7-methyl-3,4,4a,7a-tetrahydro-1H-cyclopenta[c]pyran-5-one)	3.40	3.34	3	
#P16	288.2539	$C_{16}H_{33}NO_3$ (2-[(2-Hydroxyethyl) amino] ethyl laurate)	11.74	9.84	3	
#P19	525.2544	$C_{22}H_{46}O_9S$	12.81	-	3	

#P20	437.1973	$C_{18}H_{38}O_7S$	12.66	-	3	
#P21	569.2756	$C_{24}H_{50}O_{10}S$	12.89	-	3	
#P22	481.2233	$C_{20}H_{42}O_8S$	12.62	-	3	
#P27	726.5726	$C_{38}H_{76}NO_{11}$	14.98	-	3	
#P29	363.3105	$C_{20}H_{42}O_5$ (3,6,9,12-Tetraoxatetracosan-1-ol)	13.51	13.05	1	

#N1	172.9914	$C_6H_5O_4S$	3.27	-	3	
#N3	581.2464	$C_{22}H_{46}O_{15}S$ (10 GES)	4.30	7.90*	2B	
#N16	201.9817	$C_6H_4NO_5S$ (3-nitrobenzenesulfonic acid)	3.08	3.17	1	
#N21	441.2546	$C_{20}H_{42}O_8S$ (3,6,9,12-Tetraoxatetracos-1-yl hydrogen sulfate)	12.46	10.11*	3	
#N22	661.3863	$C_{30}H_{62}O_{13}S$ (3,6,9,12,15,18,21,24,27-nonaoxanonatriacontyl hydrogen sulfate)	12.73	11.21	3	

#N24	455.2694	$C_{21}H_{44}O_8S$ (2-[2-[2-(2-tridecoxyethoxy) ethoxy] ethoxy] ethyl hydrogen sulfate)	12.98	10.20	3	
#N27	265.1457	$C_{12}H_{26}O_4S$ (lauryl sulfate)	10.94	10.05	1	
#N28	311.1684	$C_{17}H_{28}O_3S$ (C11-LAS)	12.00	12.10	3§ MassBank record: ETS00014	

§mix of isomers (spectra present in MassBank as a mix of isomers)

*Out of the domain of the retention time prediction model