Data analysis strategies for the characterization of chemical contaminant mixtures. Fish as a case study

Caroline Simonnet-Laprade, Stéphane Bayen, Bruno Le Bizec, Gaud Dervilly

To cite this version:

Caroline Simonnet-Laprade, Stéphane Bayen, Bruno Le Bizec, Gaud Dervilly. Data analysis strategies for the characterization of chemical contaminant mixtures. Fish as a case study. Environment International, 2021, 155, pp.106610. ⟨10.1016/j.envint.2021.106610⟩. ⟨hal-03257263⟩

HAL Id: hal-03257263
https://hal.inrae.fr/hal-03257263v1

Submitted on 9 May 2023

Data analysis strategies for the characterization of chemical contaminant mixtures. Fish as a case study.

Caroline Simonnet-Laprade[a], Stéphane Bayen[b], Bruno Le Bizec[a], Gaud Dervilly[a*]

[a] Laboratoire d'Étude des Résidus et Contaminants dans les Aliments (LABERCA), Oniris, INRAE, F-44307, Nantes, France

[b] Department of Food Science and Agricultural Chemistry McGill University, 21111 Lakeshore, Ste-Anne-de-Bellevue Quebec, Canada, H9X 3V9

*Corresponding author. Contact: gaud.dervilly@oniris-nantes.fr, laberca@oniris-nantes.fr

Abstract

Thousands of chemicals are potentially contaminating the environment and food resources, covering a wide spectrum of molecular structures, physico-chemical properties, sources, environmental behavior and toxic profiles. Beyond the description of the individual chemicals, characterizing contaminant mixtures in related matrices has become a major challenge in ecological and human health risk assessments. Continuous analytical developments, in the fields of targeted (TA) and non-targeted analysis (NTA), have resulted in ever larger sets of data on associated chemical profiles. More than ever, the implementation of advanced data analysis strategies is essential to elucidate profiles and extract new knowledge from these large data sets. Specifically focusing on the data analysis step, this review summarizes the recent progress in integrating data analysis tools into TA and NTA workflows to address the challenging characterization of chemical mixtures in environmental and food matrices. As fish matrices are relevant in both aquatic pollution and consumer exposure perspectives, fish was chosen as the main theme to illustrate this review, although the present document is equally relevant to other food and environmental matrices.

The key features of TA and NTA data sets were reviewed to illustrate the challenges associated with their analysis. Advanced filtering strategies to mine NTA data sets are presented, with a particular focus on chemical filters and discriminant analysis. Further, the applications of supervised and unsupervised multivariate analysis methods to characterize exposure to chemical mixtures, and their associated challenges, is discussed.

Keywords

chemical mixtures; mass spectrometry; non-targeted analysis; suspect screening;multivariate analysis; emerging contaminants

Contents

1 Characterizing contaminant mixtures in fish: a complex issue

The current inventories under the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) legislation in European Union or under the Toxic Substances Control Act (TSCA) of the United-States Environmental Protection Agency (US-EPA) indicate that over one hundred thousand chemicals, covering a wide spectrum of molecular structures and physical chemical properties, are produced globally. These chemicals may enter the environment as a consequence of their use in materials, consumer products, agriculture and industry, and the sound management of chemicals has been highlighted as one of the 17 Goals of the 2030 Agenda for Sustainable Development (United Nations, 2015). A growing evidence indicates that plants, animals and humans are continuously exposed to a multitude of chemicals over their lifetime, through various routes such as water or air (Hernández and Tsatsakis, 2017). Many chemicals are harmless or even beneficial while some others are a threat to human health and to the environment (European Chemical Agency, 2021). Some individual substances for example, such persistent organic pollutants (POPs), have been identified as a threat due to their persistence, bioaccumulation, toxic (PBT) potential, and long-term exposure to these substances, even at low-levels may be harmful (Dórea, 2008). In addition, the simultaneous exposure to multiple chemical substances may lead to additive, synergic or antagonist toxic effects ("cocktail effects") and the characterization of mixtures is now recognized as key for both environmental and human health risk assessments (Pose-Juan et al., 2016). In this line, the European Food Safety Agency (EFSA) has initiated activities to study such combined exposures through the development of harmonized methodologies for combined exposure to multiple chemicals and recently published a guidance document (EFSA, 2019). The problem associated with exposure to chemical mixtures is global and is part of an environment-food-health continuum. In this context, sentinel species are commonly used since their observations may provide information about the presence, amount, type, and effect of environmental contaminants. Fish has been recognized a relevant sentinel to monitor environmental contamination as well as suitable indicator of early contamination of the food chain (Sedeño-Díaz and López-López, 2012).

The detection, identification and quantification of a wide range of contaminants in matrices such as fish remain challenging as (i) contaminants are mostly present at trace levels, (ii) they cover a wide range of physico-chemical properties, and (iii) environmental, food, and biological samples are relatively complex matrices to analyze. Many targeted analysis (TA) methods have been developed for over half a century to detect and quantify known contaminants (metals, pesticides, POPs, etc.) in abiotic and biological matrices. While some contaminants of emerging concern (CECs) have been identified, the current surveillance framework based on TA often fails in efficiently detecting new chemical hazards, since it does not involve the treatment of unknown/unexpected signals. This is particularly alarming considering the increasing number of anthropogenic chemicals potentially reaching the environment, and a possibly even greater number of their derivatives (e.g. metabolites and degradation products), which remain to be described. To address such a challenge, methods relying on non-targeted analysis (NTA) provide a complementary and more comprehensive assessment of chemical contamination, and allow for the identification of emerging and new chemical hazards (Altenburger et al., 2019; Sobus et al., 2018).

In this context, continuous analytical developments have resulted in ever larger sets of data acquired to characterize chemical mixtures in food and environmental matrices. Depending on the initial goal of the analysis, the number of contaminants considered, the experimental design (e.g. the number of samples) and the analytical strategy (TA or NTA), gigabits or even terabits of data may now be generated within a single study. The exploration and interpretation of these large and complex data sets has thus emerged as another challenging task, and the use of advanced data processing methods has become essential for extracting the relevant information and knowledge associated to these markers of chemical exposure. Key challenges associated with data processing strategies for NTA of foods were reviewed recently in the literature (Fischer et al., 2021). With regards to data analysis tools, several methods have been developed on the basis of statistics and algorithms to describe cluster samples (e.g. according to contamination pattern) or interpret trends among variables and/or sample series. The selection of appropriate statistical tools and their use is therefore key to properly interpret the data.

This document reviews the main data analysis tools reported for the characterization of contaminant mixtures from large and complex data sets in fish samples. The first section focuses on the current challenges associated to the analysis of data resulting from the integration of TA and NTA strategies to address chemical mixtures characterization.. The second section reviews some data filtering strategies to highlight chemical mixtures and new contaminants in upon NTA. Finally, key applications of multivariate analysis methods (MAM) are presented for the exploration of large sets of data of chemicals' occurrence and the interpretation of contamination profiles. The present review focuses on methods based on LC or GC-MS, as their potential for NTA is now well established for trace contaminants. The authors nonetheless acknowledge that a range of analytical tools (e.g. FTIR, NMR, CE-MS) could be applied to NTA, with some emerging techniques (e.g. ion mobility) already anticipated to provide an additional characterization capability for the complex matrices (Mullin et al., 2020, Hernandez-Mesa et al., 2017).

While large data sets have been obtained using TA and NTA strategies for a range of environmental and food matrices, a relatively large number of studies is available on the chemical contamination of fish for both approaches. Fish are studied in the context of both aquatic pollution and consumer exposure to chemicals. Some fish species are known to accumulate relatively high concentrations of various chemicals (e.g. organic halogenated contaminants) due to their position in trophic webs (Pérez et al., 2014; Törnkvist et al., 2011). Since they are an increasingly important part of the human diet, fish have been consequently identified as a major dietary source of contaminants for consumers (Rodríguez-Hernández et al., 2016). Therefore, studies on fish contamination were primarily selected to illustrate the present review.

2 Integrating targeted and non-targeted analyses of contaminants

Current monitoring programs and studies are acquiring a continuously increasing amount of data related to chemical contaminations in environmental and food matrices. Acquired with TA or NTA methods, these data sets are often partially explored using common basic data analysis tools and critical information may be lost (Cariou et al., 2016). An in-depth interpretation of these data sets is nonetheless a challenging task and requires effective data analysis strategies. In order to better understand the associated issues, the present section introduces targeted and non-targeted analysis workflows.

2.1 Terminology

In an attempt to facilitate the discussion within the present article, a general workflow integrating various TA and NTA strategies is described in Figure 1. Both approaches may be generally described as a sequence of steps including sample preparation, acquisition of the raw data (e.g. LC or GC-MS), data processing, data analysis and interpretation. Filters are applied at various stages of the data processing and analysis to obtain a list of key compounds for interpretation. The terminology in the field is not yet standardized (Hollender et al., 2019), and some terms may be defined differently in the current literature. In the present review, the following terminology will be used:

- **Data processing** is used here as the generic term to designate all the post-acquisition steps from the transformation of raw data to extraction of relevant signal to be further analyzed (see data analysis step) in light of the research question (Pourchet et al., 2020).

- **Feature detection** is a key step of the data processing which aims at converting raw data (e.g. LC/GC-MS data) into usable data and includes tasks such as denoising, peak picking, integration and alignment . The output of this step is a list of molecular features (retention time, *m/z*), identified or not, with varying signal intensities across the samples.
- **Data analysis** is used to refer to transformation of usable and formatted data into added value and new knowledge, aiming at describe and interpret the ultimate data set. The strategy and the tools of data analysis depend on the dataset and the expected outcome. This step often involves methods based on statistics and algorithms.
- **Filtering** consists of removing signals/data corresponding to compounds which are not expected to contribute to the interpretation. It may be applied at different stages of the data processing/data analysis.
- **Data fusion:** Various analytical instrumental platforms (e.g. LC or GC-HRMS, ICPMS…) may be applied to the analysis of chemical contamination. Data fusion, sometimes called data concatenation, is an approach combining data coming from different high-throughput platforms (Smolinska et al., 2014). Data fusion may be performed at different stages of the data processing/data analysis.

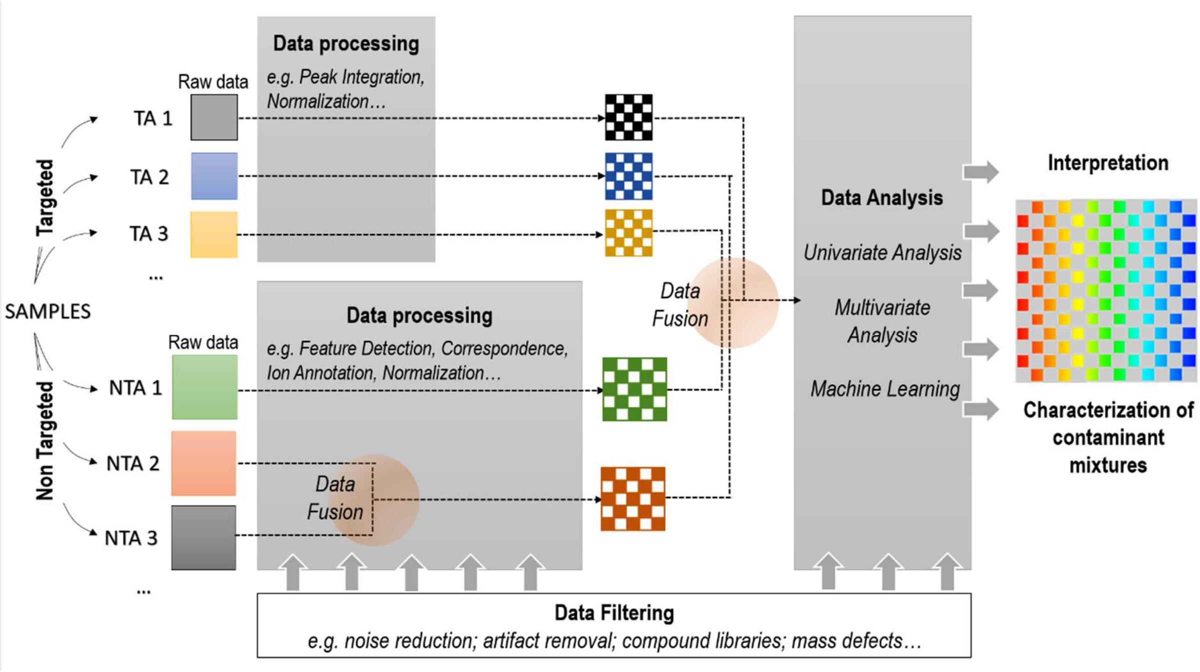


Figure 1: Integrating TA and NTA strategies to characterize contaminant mixtures.

2.2 Data acquisition and resulting data set

Taking fish as an example, a description of TA and NTA acquisition techniques and of resulting data sets is discussed in this section to understand their associated challenges in the context of data analysis.

2.2.1. Targeted analysis (TA) strategies

,Many TA methods have been designed for the analysis of fish contaminants such as trace metals (Kelly et al., 2018), organochlorine pesticides (OCPs), polychlorinated biphenyls (PCBs) and other POPs (Bayen et al., 2005, Halloum et al., 2017, Abdel Malak et al., 2018), antibiotic residues (Dinh et al., 2020), synthetic musks (Zhang et al., 2015). TAs are deployed in monitoring programs (e.g. European Union Marine Strategy Framework Directive, Great Lakes Fish Contaminants Surveillance Program), generating data sets, whose size is increasing as analytical methods improve in terms of analytical performances, throughput and multi-residue capacity (McGoldrick et al., 2010). For trace organic contaminants, sample preparation usually consists of several extraction and purification steps designed to remove interfering matrix compounds and/or to concentrate the target contaminants (Ingenbleek et al., 2021). The resulting extracts are analyzed by LC or GC-MS, e.g. using single or triple quadrupoles (selected or multiple reaction monitoring modes specific to the targets) or even high-resolution mass spectrometry (HRMS). High-purity analytical standards are commonly used as reference (chromatographic retention times, quantifier/qualifier ion ratios) and the addition of isotopic labeled compounds has become a standard practice for a confident quantification. For each compound and sample, signal intensities are commonly compared to the noise, corrected using procedural blanks or normalized to the original sample weight. Additional steps may also be carried out to improve the subsequent use of statistical tools for data analysis (e.g. conversion of non-detect values, log transformation, mean centering, variance scaling, etc).

2.2.2. Non-targeted analysis (NTA) strategies

NTA may be used to screen for the presence of new contaminants or to record a broad chemical fingerprint for fish species such as salmon, cod, pike (Tian et al., 2020, 2019). NTA does not imply the pre-selection of analytes nor the systematic analysis of their pure corresponding analytical standards (Ballin and Laursen, 2018, Schulze et al., 2020). NTA relies on sample preparation steps often compromising between an exhaustive extraction of the contaminants and the removal of interfering matrix endogenous molecules, e.g. lipids (Munaretto et al., 2016). Analytical techniques coupling LC and GC systems with HRMS are used to ensure the simultaneous detection of a large range of mass in a single scan (full-scan) with high mass accuracy (± 0.001 Da) and high resolution of mass (≥ 20 000) providing excellent specificity and selectivity, but compromising the sensitivity performance somewhat (Krauss et al., 2010; Lorenzo et al., 2018).

The resulting raw data sets contain many signals, some corresponding to possible molecules of interest (e.g. contaminants), whereas others are not relevant and sometimes undesired (e.g. interfering endogenous molecules). For each of these compounds, isotopologues, multi-chargers, adducts, neutral loss and fragment ions may be recorded. As a result, several thousands of molecular features can be detected for each individual environmental (Hollender et al., 2017, Schulze et al., 2020) or food (Fisher et al, 2021) sample (). Most critically, signals corresponding to trace contaminants of interest can be tiny compared to the bulk signal of the sample. As an example, the peak height for LC-QTOF signals corresponding to bisphenols was as low as 10^3 in pike tissue extracts where the total intensities in the Total Ion Chromatogram reached about 10^8 (Tian et al., 2019). Considering the above challenges, data processing workflows need to be optimized to effectively pick up trace contaminants (Tian et al., 2019). Additional filtering and data analysis tools for the detection and identification of contaminants in NTA data are presented in Section 3 of this paper.

2.3 Integrating TA and NTA strategies through data analysis

As discussed above, up to several hundreds of chemicals are now included in environmental or food surveillance programs (Kantiani et al., 2010). While the number of monitored contaminants has gone up in the last decades, occurrence data are still often interpreted separately, following a traditional chemical class-by-class data analysis strategy. Interpretations are generally limited to relatively simple descriptive statistics such as mean, median, standard deviation (or variance) values, each variable being interpreted independently of the others. Such an approach provides little information on the exposure to chemical mixtures, or on the interactions and relationships between contaminants.

Instead, multivariate analysis should be applied more broadly to contaminant monitoring to explore more than two variables (i.e. more than two contaminants per sample sets) simultaneously and taking into account the effects of all variables on the response of interest

(Olivieri, 2008). Such approaches allow for a scientifically sound dimensionality reduction without relevant information loss. Besides, data visualization based on multivariate analysis tools often provides a simplified representation of contamination and facilitates the interpretation. Thus, such data mining approaches are interesting approach to solve multi-variate and multi-response problems as expected when studying fish contamination.

In the end, monitoring studies should aim at integrating data from both TA and NTA strategies. Indeed, the detection of an increasing number of chemicals in matrices such as fish has illustrated that contaminants cover an ever-increasing chemical space. Analytical workflows integrating both TA and NTA data appear as promising for a more comprehensive assessment of chemical mixtures. This can be achieved using data fusion at different stages of the analytical workflows (Figure 1). Finally, the integration of metadata (biological, environmental or physical-chemical parameters, spatial and temporal information) can lead to some investigation of the target systems as described for some applications below.

3 Data mining strategies to highlight contaminants in NTA workflows

As described above, NTA produces large complex sets of raw data. A key task for chemical hazard surveillance is to detect and identify contaminants, which is particularly challenging when it comes to new or emerging contaminants. Several strategies have been reported in the literature, that may be used individually or in combination to refine a list of key contaminants of interest. Some tools can be used to screen for the presence of unexpected contaminants, while others are effective at identifying new contaminants (**Table 1**). This section describes these various strategies, and includes a discussion on the importance of selecting the right approach to limit the number of false positives and false negatives.

3.1 Suspect screening using library database searching

A common approach is the screening of unexpected contaminants using libraries of compounds which is part of the more global strategies known as suspect screening.. It is carried out against a database such as MassBank (Horai et al., 2010), GNPS (Wang, 2016), Metlin (Guijas et al., 2018), MS suppliers' commercial databases, etc… that contains, at least, information on empirical formula and accurate mass of a more or less long list of compounds and additionally, can also contain information on their retention time in a defined LC system and the "*in silico*" or experimental MS/MS fragmentation compiled in libraries. Turnipseed et al. (2018) reported the use of a high-resolution mass spectrometry screening method for veterinary drug residues in incurred fish and imported aquaculture samples. On top of detecting and identifying veterinary drugs including quinolones, fluoroquinolones, avermectins, dyes, and aminopenicillins at residue levels in fish, the approach allowed for the discovery of unexpected residues and drug metabolites in various fish samples. This approach was also reported to support the identification of previously unreported contaminants in pike fish muscles (Tian et al. 2019) or to successfully extend targeted approach, revealing additional chemicals (i.e, plastic related products, pharmaceutical products, pesticides) in several samples of fish species intended for consumption (i.e, *Merluccius australis*, *Sparus aurata*, *Dicentrarchus labrax*) (Musatadi et al., 2020).

Table 1: Examples of filtering and data analysis strategies to detect and identify new contaminants in fish and other matrices via NTA.

Expected outcome	Matrix	Analytical technique	Data processing and mining/Software	Reference
Food safety assessment				
Identify unknown toxins, illegal additives or toxicants in food poisoning from fish	Mussels and oysters	C_{18} HSS T3 column HPLC-ESI-QTOF	***Case control study:*** pairwise comparison (T-test) and multivariate analysis (PCA and PCA-DA)/ MarkerView^{TM} software 1.2.1	(Dom *et al.*, 2018)
	4 fish samples including 1 control	BEH C_{18} column UHPLC-Q-Orbitrap	***Case control study:*** differential analysis combining PLS-DA and t-test/ SIMCA-P 11.0	(Fu *et al.*, 2016; 2017)
	Eel, yellow croaker, and tilapia	Supelco Ascentis Express C_{18} UHPLC-Q-Orbitrap	***Suspect-screening:*** screening of veterinary drug residues in incurred fish and imported aquaculture samples.	(Turnipseed et al., 2018)
Identify degradation products and metabolites in food	Food matrices	Zorbax Eclipse XDB-C_8 HPLC-CID-TOF	***MS fragmentation of homologues:*** identification of pesticide transformation products via "fragmentation-degradation" relationships.	(Garcıa-Reyes *et al*., 2007)
Environmental risk assessment and management				
Identify emerging bioaccumulative contaminants in biota	Lake Ontario trout	DB-5HT column GC-TQFT	***Mass defect filtering:*** screening halogenated environmental contaminants	(Jobst et al., 2013)
	European eel (*Anguilla Anguilla*) muscle	Hypersil Gold analytical column UHPLC-Q-Orbitrap	***Mass defect filtering:*** screening halogenated environmental contaminants	(Cariou *et al.*, 2016)
	Pike (*Esox lucius*) muscle	Poroshell Phenyl-Hexyl HPLC-ESI-QTOF	***Suspect-screening:*** screening plastic-related chemicals and other contaminants in samples from the St. Lawrence River, Canada	(Tian et al., 2019)
	Freshwater organisms (*Lumbriculus variegatus, Hexaenia spp, Pimephales promelas*)	DB-5HT GC column GC-FTICR	***Mass defect filtering:*** mass defect filtering on an H/Cl mass scale, H/Cl mass defect plot	(Myers *et al.*, 2014a)
	Fish livers (23 freshwater fish species)	Poroshell Phenyl-Hexyl HPLC-ESI-QTOF	***Suspect screening + Differential analysis:*** Comparison of benthic and water-column foraging strategies group. Comparison upstream and downstream of wastewater treatment plants.	(Baesu et al., 2021)
	Human blood as example of biological samples	Acquity UPLC HSS C_{18} SB column UPLC-Q-ToF or UHPLC-Orbitrap	***Time-trend screening:*** to flag reoccurring peaks in a time series. Selection of peaks displaying an increasing trend using time trend ratios and Spearman's rank correlation coefficient/ MATLAB and Microsoft Excel	(Plassmann *et al.*, 2016; 2018)
	Lake trout and walleye bream bile from Great Lakes	GC×GC-TOF HRT	***Mass defect filtering:*** mass defect filtering on an H/Cl mass scale, H/Cl mass defect plotting/ Leco, ChromaTOF v1.90.60 and Microsoft Excel	(Fernando *et al.*, 2018)
	Lake Michigan trout	UPLC-QToF	***MS fragmentation of homologues:*** screening algorithm initialized using a candidate formula matrix based on mass spectral profiles and likely fragmentation pathway/ MATLAB	(Baygi *et al.*, 2016)
Identify degradation products, metabolites, precursors in biota	*Chelonia mydas* green sea turtles	UHPLC-ESI-QTOF	***Case control study.*** multivariate analysis (PCA) to simultaneously detect biomarkers of exposure (xenobiotics) and biomarkers of effect (endogenous compounds)	(Heffernan *et al.*, 2017) and companion paper (Gaus *et al.*, 2019)
Identification of toxic compounds	Bream bile from Lake Bergumermeer, River Dommel, Amsterdam North Sea Canal (Netherlands)	GC-MSD	***Effect-directed analysis:*** identification of endocrine disruptors (ER-CALUX-assay + HPLC fractionation + GCMS analysis)	(Houtman et al., 2004)
	Liver and blubber of high-trophic-level animals	GC-MSD	***Effect-directed analysis:*** identification of dioxin-like and androgen receptor antagonist	(Suzuki et al., 2011)

3.2 Chemical filters

Many chemicals share the same fate in the environment because of similarities in terms of composition or physicochemical properties. Using the knowledge built in the fields of environmental and food sciences in the last decades, strategies have been designed to identify contaminants which may be part of homologue series or who share some composition or structural similarities.

3.2.1 Mass defect filters and isotopic profiles

The majority of the PBT substances, notably covered by the Stockholm convention, are polyhalogenated (Scheringer et al., 2012), recent studies have thus focused on identifying halogenated compounds as a screening approach to detect new contaminants. Halogenated atoms, especially chlorine and bromine, exhibit a relatively higher mass-defect (MD) (difference between the exact mass and the nominal mass of an element) as compared to other common elements (C, H, O, N), and atypical MS isotopic profiles. These two distinct attributes make halogens relatively straightforward to highlight in a mass spectrum, especially when accurate mass measurement are obtained using HRMS instruments (Kaufmann, 2012). As a result, feature filtering methods based on MD have been developed for the screening of halogenated contaminants (Sleno, 2012, Jobst et al., 2013). The principle of MD filtering is to remove all data outside a pre-defined and limited MD range. A relatively simple way to visualize and distinguish ions with a particular MD from other ions is to plot the fractional part of the *m/z* (i.e. MD) against the *m/z*. Originally based on an exact mass reference of 12.0000 for ^{12}C (International Union of Pure Applied Chemistry) or of 14.0000 for $^{12}CH_2$ (Kendrick, 1963), a modification of MD plot scale has been proposed for halogenated compounds based on the substitution of chlorine for hydrogen, thus using H/Cl mass scale of 34.0000 Da (-H/+Cl). In the corresponding H/Cl-scale MD plots, chlorinated homologue series plot on horizontal lines (see example from Cariou et al., 2016 in **Figure 2**). H/Cl and H/Br conversion factors being almost equal (1.001148 versus 1.001149, respectively), MD plots can be also effective at visualizing clusters of brominated compounds. The use of MD between the two natural and stable isotopes separated by 2 nominal atomic mass units, for both Cl and Br atoms (1.9971 for Cl and 1.9980 for Br) and ion ratio criteria is good combination to effectively identify chlorinated and brominated ion clusters. Filtration algorithms based on MD and isotopic profiles have been successfully applied to Fourier transform mass spectrometry for the screening of halogenated bioaccumulative compounds in freshwater organisms (*Lumbriculus variegatus, Hexagenia spp.,* and *Pimephales promelas*) exposed to contaminated soil from a recycling plant fire site (Myers et al., 2014b). Various bioaccumulative contaminants were identified including polychlorinated naphthalenes (PCNs), polychlorinated dibenzofurans (PCDFs), or chlorinated and mixed brominated/chlorinated anthracenes/phenanthrenes, and pyrenes/fluoranthenes. The same approach allowed the identification of 60 non-targeted halogenated species in lake trout from the Great Lakes (Fernando et al., 2018) or hexabromocyclododecane and chlorinated paraffins in muscles of the European eel (*Anguilla anguilla*) from the Loire river in France (Cariou et al., 2016). In each of these studies, the resulting thorough data filtering (from 9789 initial obtained features to 589 clusters for instance in Cariou et al., 2016) allowed for the optimization of the molecular formula assignment. In order to facilitate the wider application of this approach and accelerate the overall data processing, Léon et al. (2019) proposed a user-friendly software named HaloSeeker. The software consists in an ergonomic web user interface facilitating peak picking, deconvolution, halogenated feature filtering, MD plot and chemical formula assignment.

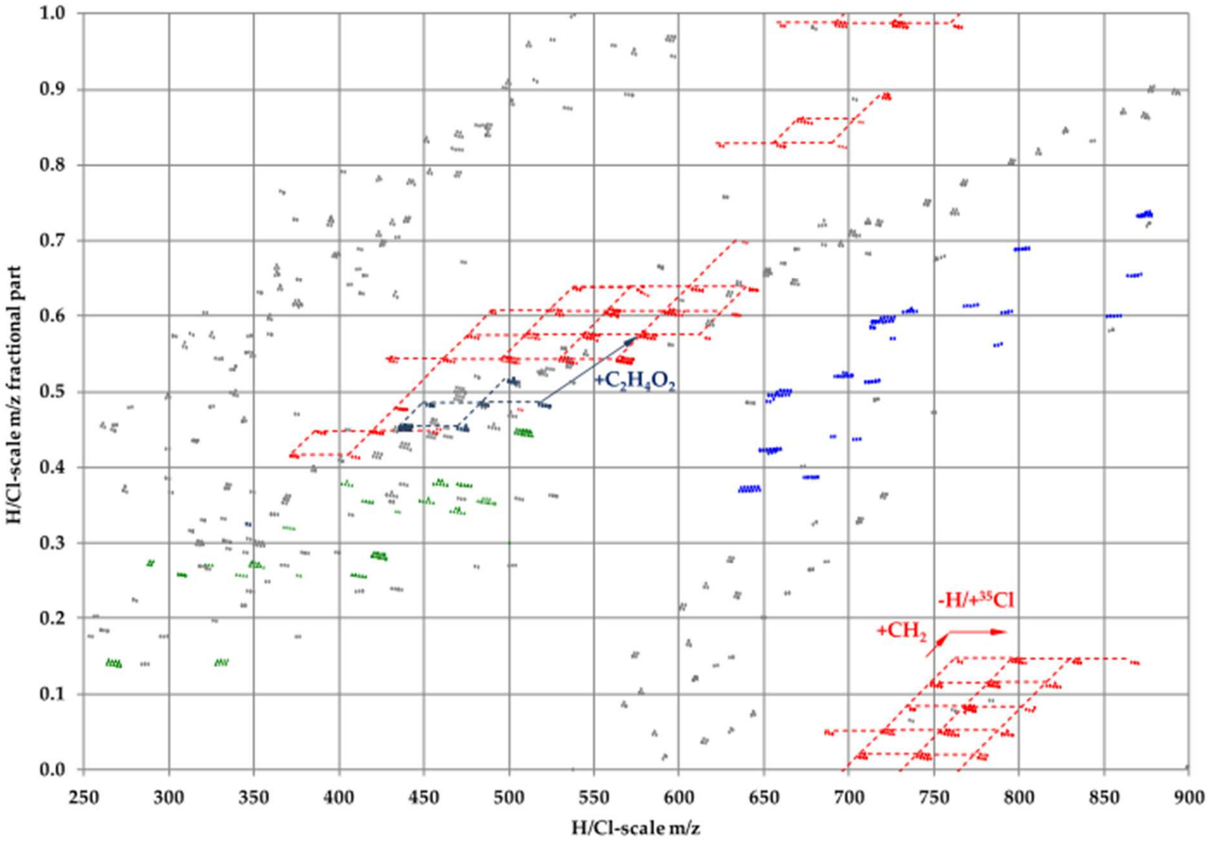


Figure 2. Example of H/Cl-scale MD plot obtained for a muscle eel sample extract reproduced with permission from Cariou et al., 2016.

Mass defect filtering was also reported for the screening of bioaccumulative fluorinated contaminants in aquatic biota, including fish (Myers et al., 2014a). The mass defect and isotopic profiles of fluorine atoms are however less specific than for Cl and Br, and their use may lead to a relatively high rate of false positives (Liu et al., 2019). A combination of CF_2-scale MD plot and homologous series searching has been proposed to flag poly- and perfluoroalkyl contaminants in full-scan data sets using mass differences of 49.997 for CF_2 units, 99.994 for CF_2CF_2 units, 64.012 for CH_2CF_2 units = 64.012 or 65.991 for CF_2O units (Liu et al., 2019). This approach can be therefore extended to other large classes of homologues which could be manufactured or used as chemical mixtures.

3.2.2 Other approaches for the identification of homologue series

In addition, compounds part of a homologue series may share similarities in terms of chromatographic or mass spectrometry behavior. Non-commercial software workflows, such as enviHomolog web (Loos and Singer, 2017), have been developed for the extraction of homologue series based on the identification of repeating patterns in the hyphenated HRMS data. Neutral loss, i.e. fragments lost as neutral molecules, has also been proposed as a feature filtering tool to screen for the presence of series of homologue compounds. Baygi *et al.* (2016) developed a candidate list screening algorithm on the basis of: (1) a molecular formula matrix for the possible ions for fluorinated homologues ($C_cO_oF_fCl_{cl}H_hS_s$, with c = 4-10, o = 2 for carboxylic forms, = 3 for carboxylic ether and sulfonate forms, = 4 for ether sulfonate form, and the summation of f, cl and h set so that all carbon atoms were fully saturated and the compound was deprotonated) previously discovered from fluoropolymer discharged impacted compartments; and (2) a candidate compound spectra matrix developed using a statistical approach developed by Yergey (1983) (see details in Baygi et al., 2016) to calculate theoretical isotopic distribution of each candidate. This algorithm allowed to reference 3570 possible compounds in Lake Michigan trout data files, highlighting the presence of 30 polyfluorinated chemical formulas reported for the first time in environmental matrices.

3.3 Differential analysis

The differential analysis approach investigates NTA data profiles among groups of samples to isolate features of interest. This strategy, similar to that implemented in metabolomics - to the nuance that it is in this case to detect markers of exposure and not effect (Hernandez-Mesa et al., 2021) - consists in the comparison of signals between two or more groups of samples of interest. It is often guided by the experimental design and relies on the application of discriminant analysis (univariate or multivariate) tools to reveal the molecular features or the compounds of interest.

3.3.1 Non-target time trend screening

Non-target time trend screening consists in comparing MS profiles of samples collected over several periods. Using time-series data sets from samples analyzed at different time points, compounds that show a meaningful trend are studied (Peters et al., 2010). The principle of this filtering strategy relies on peak occurrence and intensity assuming that reoccurring peaks with increasing (or decreasing) intensity in the time series correspond to contaminants of interest, while reoccurring peaks with constant intensity more likely refer to endogenous substances. Peaks displaying an interesting trend may be filtered from randomly fluctuating peaks using time trend ratios and Spearman's rank correlation coefficients. This strategy allows for considerable reduction of the size of datasets (Plassmann et al., 2016); it was successfully applied in environmental matrices to highlight biooaccumulative contaminants such as POPs exhibiting increasing intensity in the time series (Miller et al., 2014, Nyberg et al., 2015), while it was also reported a successful approach to investigate time series of polar contaminants in abiotic matrices (Albergamo et al., 2019). Such long-term data is also key for assessing the efficiency of measures taken to reduce contamination (Ek et al., 2021).

3.3.2 Comparison of samples of different origin.

Differential analysis can also be applied by comparing samples considered "contaminated" versus control samples. Fu et al. (2016) developed for example a data reduction strategy based on differential analysis to screen illegal additives in fish. An unsupervised partial-least square discriminant analysis (PLS-DA) was applied on UHPLC-HRMS features (*m/z*, t_R and peak response (> 1000 ions), after extraction solvent blanks, internal standard calibration and ion fusion filtration, for comparing suspected fish samples versus a control fish sample. Ions with variable importance in the PLS-DA projection (values >1.0) were selected for t-test analysis (required p-value < 0.01). Then, the retained ions were analysed by calculating the peak intensity ratio between the suspected sample and the control sample. Ions with a fold change of 10 were considered to be high risk compounds. With such approach, 69 ions were retained for database searching. Other possible questions could be addressed in applying the same strategy. For instance, the differential analysis of HRMS profiling of packaged fish fillet sample vs. unpackaged fish fillet sample could be useful to assess the impact of food packaging on chemical contamination of edible fish (provided that the fish have the same origin) and possibly identify non-intentionally added substances (Sanchis et al., 2017). The comparison of fish samples from industrial zones and unexposed area would help for discover new bioaccumulative contaminants. This approach was recently reported for the comparison of contaminant profiles in fish sampled upstream and downstream of wastewater treatment plants (Baesu et al., 2021). Through the application of differential analysis and data visualization tools such as volcano plots, erythrohydrobupropion was identified for the first time in fish livers, and was also found at higher concentrations in fish livers sampled downstream vs. upstream.

Similarly, a methodology combining a non-target HRMS analysis with multivariate statistical analysis has been proposed to simultaneously detect biomarkers of exposure (i.e. xenobiotics) and endogenous metabolites in blood of green sea turtles (*Chelonia mydas*) on the Great Barrier Reef (Heffernan et al., 2017). The simultaneous detection of exogeneous and endogenous compounds through full-scan mode may be used to identify cause-effect relationships and thus indirectly highlight toxic contaminants (Hernandez-Mesa et al., 2021). In order to investigate the potential influence of xenobiotics, HRMS profiling of case 'samples' corresponding to turtles from two coastal sites impacted by urban/industrial or agricultural activities were compared with those of 'control' sample corresponding to turtle from a remote offshore site. Prior to multivariate analysis, the number of spectral features was reduced from 4761 to less than 100 by two-to-two comparison of sites, in using several criteria: significance (p-value < 0.05), effect size (log fold-change > 0.05), monoisotopic mass (ignoring isotopes, adducts and ion products generated during the ionization process) and retention time (> 1 min). This step wise data reduction strategy allowed to focus on the most significant spectral features for subsequent identification. Then PCA established on selected features enabled the discrimination of samples according to the three sites despite inter-individual variability. The spatial difference of xenobiotic profiling was key to validate the selection of features of concern.

3.4 False positives and negatives issues

Filtering methods are critical in the identification of new contaminants in complex environmental and food matrices, such as fish tissues. However, several considerations need to be included when selecting and deploying data filtering. Inappropriate filtering parameters may be ineffective in eliminating irrelevant compounds (increasing the likelihood of false positives) or may be too stringent (false negatives) (Schulze et al., 2020).

The impact of sample preparation on the false discovery rate of contaminants is obvious, and experimental conditions are often optimized to limit the number of false negatives in complex matrices such as fish (Du et al., 2017). Instrumental conditions, for example selecting data-independent or data-dependent acquisition in HRMS, can influence the success of library searching to identify non-targeted compounds or metabolites (Wu et al. 2020). However, the choice and the parametrization of a filtering step should be also aligned with the experimental conditions (e.g. types of extraction, chromatography or ionization) and performances (e.g. mass measurement errors, retention time shifts). For example, homologue series searching and formula searching should be guided by a knowledge of chemical space covered by a specific type of sample preparation or mass spectrometry ionization mode. The parametrization of the data processing pipeline should also be considered, as each step may impact the success rate of the identification of contaminants. As an example, the type of imputation method for missing values can have major effect on the results of subsequent statistical data mining (comparison performed in Hrydziuszko and Viant, 2012; Wei et al., 2018). In that way, the selected NTA pipeline strategy should be assessed using spiked matrices or reference material on the model of what is being done in other fields of metabolomics (Ribbenstedt et al., 2018). Spiking model contaminants at trace level has been reported for eel (Wu et al., 2020), pike fish (Tian et al., 2019), but reference materials are still lacking to assess NTA workflows.

Hollender *et al.* (2017) pointed out the limitations related to suppression of signals in matrix-rich samples and the biases that can generate samples comparison. For differential analysis, the definition of the control or reference group of samples is critical to dissociate contaminants from endogenous compounds. Homogeneity among the sample populations in terms of age, gender, species is often key to limit inter-individual and interspecies variability and better highlight, using discriminant analysis, the variability related to the “treatment” only (exposition to additives, exposition to industrial sources).

4 Multivariate analysis to characterize contaminant mixtures

The chemical contamination profile of fish may be impacted by several factors including contamination sources, physical and chemical environmental parameters and uptake of pollutants by fish, itself influenced by a variety of factors such as exposure pathways (e.g. through water or diet), elimination processes, growth rate, age, lipid contents, etc. (Wenning and Erickson, 1994). Besides, the environmental fate of chemicals and their trophic transfer obviously depend also on their own physico-chemical properties. Because of the multitude of possible combinations of influencing factors, the description and interpretation of fish contamination profiles can be intricate task. As reviewed by Mas et al. (2010) and Wenning and Erickson (1994) for instance, various types of multivariate methods, or "chemometric multivariate methods", have been developed and are now available in common statistical software packages (See examples in Table 2). However, the selection of efficient data analysis methods is not always straightforward since it is dependent on the goal of the study and key properties of the datasets. The present section provides a brief description of some multivariate analysis tools, their applications to contaminant mixtures in matrices such as fish, and some considerations to properly interpret their results.

4.1 Categories of multivariate analysis methods (MAMs)

Multivariate analysis methods have been applied for several decades in environmental studies to reduce dimensions, to classify variables or samples, to select variables or to predict phenomenon in order to simplify interpretation of environmental systems. MAMs may be categorized under two main categories: unsupervised multivariate analysis methods (UMAMs) and supervised multivariate analysis methods (SMAMs). The selection of a MAM is critical to provide an appropriate interpretation. Gibert et al., (2018) recently reviewed the differences between UMAMs and SMAMs, and proposed guidelines to select the appropriate methods according to the scientific question and the structure of data sets. Briefly, the main goal of UMAMs is to provide an in-depth understanding of the system and a general description of the global interactions. SMAMs aim to explain the specific behavior of a response variable (defined as variable of interest to be explained) by explanatory or independent variables. In the first case, all the variables are processed equivalently without *a priori*. In the latter case, a prediction is assumed for the response variable and predictor variables are used to explain it.

There are two groups of UMAM techniques (Gibert et al., 2018): (i) associative methods which help to identifying relationships among variables (e.g. contaminant concentrations) and include for instance principal component analysis (PCA) and correspondence analysis (CA); and (ii) descriptive methods which are used to assess relationships among objects (e.g. samples, sampling locations, fish species, fish tissues, etc.) and include self-organizing maps, statistical clustering, etc. SMAMs are seldom applied to only describe the system but may be used to build predictive methods (e.g. multiple linear regressions, analysis of variance such as ANOVA) or classifier/discriminant methods (e.g. decision-trees, discriminant analysis). Table 2 summarizes key applications of MAM to data sets in the context of contaminant mixtures in fish and their interest in environmental and health risk assessment.

Table 2: Applications of MAMs for the assessment of contaminant mixtures in fish.

Types of contaminants	Matrix	MAM	Interpretation	Software	Reference
23 trace metals, 80 PCBs, chlorinated hydrocarbons, OCPs, BFRs	3 species of Eurasian caviar	HC Squared Euclidean distance	Identify groups of caviar samples Determine within-group linkages	Excel and SPSS, version 4	(Wang et al., 2008)
23 OCPs, 18 PCBs	Ten common aquatic product species from Northeast China	PCA	Assess species-specific bioaccumulation Identify groups of species according to contaminants concentrations	not specified	(Fu et al., 2018)
7 OCPs, 19 PCBs	Muscle samples of 3 *Cyprinidae* species from Vransko Lake (Croatia)	SOM DT	Identify patterns among OCP and PCB congeners in freshwater fish searching for clustering based on different fish species and sampling months. Classify samples according to fish species or seasons	MATLAB STATISTICA,	(Romanić et al., 2018)
PCBs, OCPs, PBDEs	Whole fish and fillet of 5 species from Charleston Harbor and tributaries (South Carolina, USA)	Heat map + complete linkage clustering	Identify patterns of contaminant loads by fish species and location	not specified	(Fair et al., 2018)
PCDDs, PCDFs, PCBs	Liver of coalfish and cod, eel, pike perch, farmed salmon	PCA	Investigate differences in congener profiles of marine fish, shellfish and farmed fish (salmon)	not specified	(Van Leeuwen et al., 2007)
7 OCPs, 17 PCBs	Fillet of edible marine fish species from Adriatic Sea	SOM DT	Identify OCP and PCB pattern in marine fish according to species, years and fishing zone Classify samples according to fish species and sampling seasons	MATLAB STATISTICA	(Vuković et al., 2018)
PBDEs, PCBs, OCPs	The patagonian silverside (*O. hatcheri*) collected along the Negro River	PCA	Reveal the relationship among sampling sites and the accumulation of contaminants in each fish tissues	InfoStat 2008	(Ondarza et al., 2014)
18 PCBs, 7 PBDEs, 17 PFASs, BPA, 5 OH-PAHs, 4 Aps	Muscle and bile of European eel *Anguilla anguilla*	PCA	Discriminate contaminant levels in the muscle and bile of eels from different sites and life stage, as well as their biometric parameters	STATISCA, version 7	(Couderc et al., 2015)
58 PCBs, 6 PBDEs	Whole fish and eggs of fish (Chinook and salmon, brook trout, mottled sculpin)	PERMANOVA, NMDS	Compare and assess relationships between POP pattern of resident fish species of Great Lakes and with migratory salmon	R version 3.0.3	(Gerig et al., 2015)
19 contaminants (OCPs, PCBs)	Salmonids and cyprinids fish	PCA	Discriminate fish species according to organochlorine contaminant profiles and identify variables responsible of the variance.	PLS Toolbox v3.5	(Peré-Trepat et al., 2006)
7 PCBs, 18 OCPs, 16 PAHs	Eel muscle tissues	PCA DA	Characterize the correlations between PCB, OCP, PAH concentrations and biological responses Classify the different sampling sites	ADE	(van der Oost et al., 1997)
168 organic chemicals	Fish tissues	SOM, canonical correlation analysis	Investigate deviations from linear relationships between log BMF and log K_{ow} calculated from concentrations of contaminants in fish tissue and identify structure-related bioaccumulation patterns	MATLAB 2014	(Grisoni et al., 2018)
OCPs, PCBs	Muscle and liver of fish from European mountain lakes	PCA, PLS	Assess the dependence of compounds on geographical and temperature and physiological parameters	MATLAB 6.5, PLS 3.5 Toolbox	(Felipe-Sotelo et al., 2008)
PCBs, α-HCH, HCB and trace metals	Liver and muscle of Canadian Atlantic Cod	PCA with ANCOVA and MANCOVA	Investigate time trends of contaminant levels in fish tissue	SYSTAT v 5.0	(Misra et al., 1993)
16 PAHs, 29 PCBs	Liver and muscle of sharks from Galveston Bay	PERMANOVA SIMPER analysis PCA	Compare liver and muscle congener profiles among the three species Determine the congeners contributing to the greatest differences between species Investigate and visualize correlation between contaminant concentrations in fish and biomarker activity	R version 3.3.3	(Cullen et al., 2019)

		partial redundancy analysis (pRDA),	Determine which congeners were correlated with EROD and GST activity		
21 PCBs, 28 OCPs	Muscle tissues of fish from the Yadkin Pee Dee River (Caroline, USA)	PCA, Linear mixed effect model	Identify relationships between environmental contaminants and intersex occurrence and severity Predict intersex potential	JMP Pro 12	(Grieshaber et al., 2018)
28 PCBs, 5 OCPs, 2 PBDEs, 4 trace metals	Liver of flounder from two estuarine areas in the Netherlands	PCA	Visualize correlations between contaminant concentrations and biomarker responses	not specified	(Schipper et al., 2009)

HC: hierarchical cluster analysis; PCA: Principal Component Analysis; SOM: self-organizing maps; DT: Decision Tree; PERMANOVA: Permutational multivariate analysis of variance; NMDS: non-metric multidimensional scaling; PLS: Partial least-square regression; (M)ANCOVA: (multivariate) analysis of covariance.

4.2 Applications of unsupervised multivariate analysis methods (UMAMs)

Unsupervised descriptive and associative multivariate methods are commonly reported to explore data sets associated to the study of multi-contamination of fish since they do not require prior assumptions on the target system. The application of UMAMs allows reducing the complexity of a system by grouping homogeneous objects (e.g. fish samples having similar contamination profiles) or associated variables (e.g. identify relationships among contaminants or with environmental and biological parameters).

4.2.1. Descriptive UMAMs

The application of descriptive UMAMs to environmental/food samples such as fish allows for the description and the categorization of sample groups according to homologous contamination patterns. Cluster analysis is a widely used method to partition a set of objects into two or more clusters based on their similarities (Johnson and Wichern, 2002). Hierarchical cluster analysis indicates sample groupings by ranking inter-sample similarities (linkage clustering) and the resulting output data are represented on a dendrogram, i.e. a tree on which the more the link height between nodes (samples) decreases, the more the similarity between nodes is high. For instance, Wang et al. (2008) performed a hierarchical cluster analysis (HCA) to conduct a preliminary assessment of health risks associated with the consumption of caviar, and identified different groups of caviar samples according to the concentrations of a hundred contaminants including PCBs, chlorinated hydrocarbons, OCPs, BFRs and trace metals (reproduced in **Figure 3A**). Using HCA, several groups were distinguished, first by species, and then origin, supporting a discussion based on trophic levels and/or contamination sources. A similar approach, using the combination of heat map and complete linkage clustering, allowed for the simultaneous visualization of the patterns of PCBs, OCPs and PBDEs across various fish species from multiple locations (Fair et al., 2018). Heat map colors allow for the visualization of the relative contaminant levels in each samples in comparison to the average in all the samples.

Romanić et al. (2018) reported the application of Kohonen self-organizing maps (SOM) to identify pattern of OCP and PCB congeners in 3 freshwater *Cyprinidae* species collected at three different sampling periods in Vransko Lake (Croatia) (**Figure 3B**). The SOM consists in a regular neuron network (usually a two-dimensional grid), where input data are distributed using a finite set of models with the following principle: more similar models become automatically associated with nodes that are adjacent in the grid, whereas less similar models are situated farther away from each other in the grid (Kohonen, 2013). Such an approach has proved particularly interesting for describing the contamination patterns of the three fish species and for identifying the main variables that explained the observed differences (Romanić et al., 2018).

4.2.2. Associative UMAMs

Another common approach for data reduction is to identify and combine correlated variables. Principal components analysis (PCA) is probably one of the most commonly used MAM (Table 2). PCA is of particular interest to highlight correlations between different variables and to visually discriminate groups of samples. PCA consists of a projection of variables as points in bi or tri-dimensional space in preserving most of the existing relations among samples and variables (Abdi and Williams, 2010). Dimensions of the new space are created by the associations of correlated variables and are called principal components (PCs). PCA is often combined with clustering analysis to distinguish sample groups in a 2D new space. One of the first studies attesting the power of PCA modelling of multivariate data such as those encountered in complex chemical mixtures study in aquatic biota (Stalling et al., 1985) was performed using poorly performing computer processes compared to those available today.

Benefiting from computer advances, applications of PCA has generalized. Van der Oost *et al.*, (1997) demonstrated for instance the importance of monitoring biota such as fish for the assessment of freshwater pollution since no clear discrimination between moderately and heavily polluted sites could be made using PCA on sediments only. In their study, the joint application of univariate analysis methods, PCA and discriminant analysis on a data set including PCBs, OCPs and PAHs concentrations in eels (*Anguilla anguilla)* from six Amsterdam freshwater sites, allowed for: (i) the classification of the environmental quality of the sites resulting from sample discrimination, (ii) the identification of contaminants responsible to this ranking, (iii) the examination of relationships between exposure to organic trace pollutants and biochemical responses in eel. The combination of univariate analysis and PCA has been also successfully applied to discriminate muscle and bile samples of European eel *Anguilla anguilla collected* along the Loire Estuary in France according to the pattern of an extended number of class of contaminants (PCBs, PBDEs, PFASs, BPA, OH-PAHs, APs) and biometric parameters (Couderc et al., 2015), reproduced in **Figure 3C**). The variability among eels was mainly explained by the trophic level, body weight, lipid weight, and PBDE contents on the first component and PFAS and gonadosomatic index on the second component. Correlations between biometric parameters (body weight and trophic level) and concentrations of PCBs and PFAS were also identified through this MAM approach. This method allowed for the distinction between eel individuals from two sites, Bellevue and Haute Indre, the former presenting the highest PFAS and PCB levels. The additional consideration of biomarkers of effects (e.g. oxidative stress, biotransformation enzyme, genotoxic parameters) in PCA may provide insights on the possible cause-effect relationships as illustrated by Schipper et al. (2009) for instance. It should be noted though, as pointed by Bellavia et al. (2019), that PCA allows the identification of individual contribution to the mixture, but PCA is not a quantification method of the contribution of each component of the mixture on observed effects.

4.3 Applications of supervised multivariate analysis methods (SMAMs)

The choice of a SMAM rather than an UMAM depends on the possibility to perform an assumption on the target system (i.e., contamination profiles of two groups of fish samples are differentiated by the concentration of one chemical substance and the question is what are the variables that may explain this difference). SMAMs allow for the statistical test of assumption using the entire dataset, and may be used to build predictive models.

4.3.1. Discriminant SMAMs

Fish contamination can be explored through supervised discriminant methods (Table 2). Among these approaches, decision tree (DT) analysis was recently reported to assess fish multi-contamination (Romanic et al., 2018; Vukovic et al., 2018). DT analysis is a supervised learning algorithm that can be used in both regression and classification problems (Debska and Guzowska-Swider, 2011). DT consists in a tree-shaped graphical representation of every possible outcome of a decision. Tree starts with a root node which represents all the samples and is further divided in homogeneous sub-nodes according to successive decision rules (values of single variables that best divide the data into two or more groups as homogeneous as possible). Romanic et al. (2018) applied DT models, in combination with SOM analysis (see section SDAM), to discriminate freshwater fish samples according to species and sampling seasons (2014 and 2016). Vukovic et al. (2018) reported the same approach (SOM combined with DT) to investigate POPs in edible fish species from different fishing zones of Croatian Adriatic. Results from DT (**Figure 3D**) indicated that fish collected on two sampling dates (2014 and 2016) could distinguished from each other based on PCB-74 levels (threshold at 0.066 ng.g $^{-1}$). In both these studies, DT models provided complementary results to the

SOM approach, pointing at the levels of a specific variable that may discriminate fish samples.

Discriminant SMAM may be also combined to UMAM. In a recent study, Cullen et al., (2019) combined PCA and a partial redundancy analysis (pRDA) to study POP contamination in shark species from the northwestern Gulf of Mexico. pRDA aims to summarize linear relationship between components of response variables and explanatory variables in removing the effect of one or more explanatory variables with strong effect (Anderson, 2017). Cullen et al. (2019) evaluated, through pRDA, correlations between POP congeners and biomarker responses (ethoxyresorufin-O-deethlyase, EROD and glutathione S-transferase, GST) while limiting the effect of interspecific variability of POP concentrations between the 3 studied shark species (*Carcharhinus leucas*, *Carcharhinus limbatus*, *Sphyrna tiburo*). This method may be particularly useful to highlight weakly pronounced relationships, especially when the sample sets are heterogeneous.

4.3.2. Predictive SMAMs

Predictive SMAMs often involve establishing a regression model to explain a variable with others. The analysis of variance (ANOVA) is probably the most common statistical method for hypothesis testing on fish multi-contamination (Table 2). ANOVA is a type of general linear model which aims at testing if the means of two or more populations are equal, and assesses the effect of (and interactions between) various factors (dependent variable) on some variable response (Henson, 2015). The multivariate extension of ANOVA, MANOVA (for multivariate analysis of variance), simultaneously takes into account multiple response variables (Henson, 2015). Thus, MANOVA may be used to assess similarities/differences in contaminant patterns among different fish species and location for instance (e.g. Faira et al., 2019).

Predictive SMAMs have also been recently applied to elucidate contaminant transport. For example, Gerig et al. (2015) applied a combination of Permutational multivariate analysis of variance (PERMANOVA) and non-metric multidimensional scaling (NMDS) to determine if the migratory Pacific salmon (*Oncorhynchus tshawytscha*, *O. kisutch*) could be a source of POP contaminants to stream-resident fish in Great Lakes tributaries. PERMANOVA is the non-parametric (based on permutation tests) version of MANOVA (based on sums of squared distances) that partitions variance in a distance matrix by calculating a distance based F-statistic (Anderson, 2017, 2001). As with PCA, NMDS aims at projecting input data of a target system into a new space with a reduced number of dimensions (example from Gerig et al., 2015 in **Figure 3E**) in order to create a straightforward representation of relationships between objects and descriptors (Agarwal et al., 2007). However, unlike PCA, NMDS relies on rank orders (distances) for ordination and does not require normal distribution of data (often the case when studying ecological systems) (Agarwal et al., 2007). In Gerig et al. (2015), the joint use of these both methods, less stringent than parametric methods, allowed the verification of hypothesis that (1) salmon PCB and PBDE congener patterns differed among Great Lakes basins and (2) resident consumer fish species from reaches with salmon have more similar POP patterns with salmon than resident consumer fish species from reaches without salmon.

Partial least square (PLS) regression is another approach to assess simultaneously the effects of various factors on fish contamination. PLS regression is an extension of the multiple linear regression model that assess relationship between response variable and a set of predictor variables. PLS is relatively less reported, but was successfully applied to assess the relative importance of the geographical, temperature and physiological variables (predictor variables) affecting the accumulation of OCPs in different fish samples from European mountain and to find potential systematic patterns in these dependencies (Felipe-Sotelo et al., 2008). In this study, PLS was deemed complementary to PCA, because PLS is not affected by correlation among predictor variables. This can be useful when dataset including geographical and physical-chemical variables for example, may be correlated.

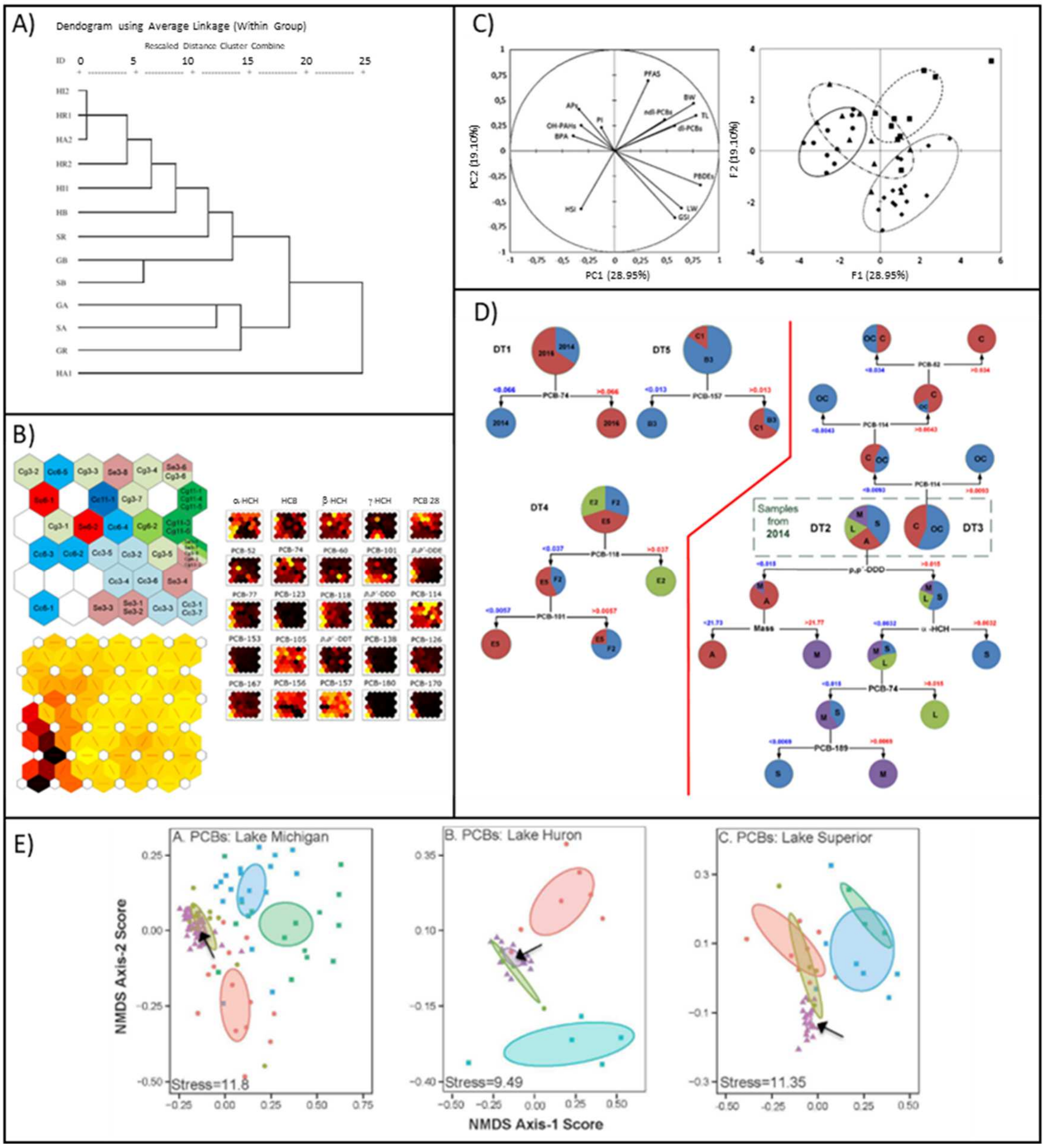


Figure 3: Examples of result representations from unsupervised and supervised data analysis methods: (A) dendogram from cluster analysis of Eurasian caviar samples according to organic (PCBs, OCPs, BFRs, OCs) and inorganic compounds (from Wang et al., 2008); (B) the Kohonen self-organizing maps (SOM) of OCP and PCB patterns in freshwater fish from Vransko Lake (from Romanić et al., 2018); (C) Principal Component Analysis (correlation loading on the left and sample representation on the right) of biometric parameters and contaminants in the European eel tissues from the Loire Estuary (from Couderc et al., 2015); (D) Decision Tree classification of the organochlorine compounds found in edible fish species from different zones of Croatian Adriatic, according to sampling year (DT1) and coastal (DT4) and off coast fisheries zone (DT5), fish species sampled in 2014 (DT2) and fisheries zones (DT3) (from Vuković et al., 2018); (E) non-metric multidimensional scaling (NMDS) plots of PCB pattern for salmon spawners and resident fish in stream reaches with and without salmon from lakes Michigan, Huron and Superior (from Gerig et al., 2015).

4.4 Considerations when applying MAMs

MAMs generally facilitate the interpretation of complex systems, such as contaminant mixtures in fish, and provide simplified visualization of the results. Interpretations of contamination profiles, relationships between environmental variables and occurrence of contaminants, based on MAMs often provide a strong rationale for the implementation of a customized management approach of the food or environmental system. However, based on the present review, the applications of MAMs are still limited, and were mostly applied to the levels of regulated contaminants (e.g. PCBs, dioxins, PBDEs) determined through targeted analysis. The limited number of MAM applications may be explained by the complexity of the data sets, and a lack of guidelines to select and apply appropriate MAM. But a deeper root for this issue remains the relatively poor understanding of the impact of data processing, data fusion and data filtering on the outcome of data analysis, particularly for NTA data.

As introduced in section 2.2, data sets obtained using both TA and NTA approaches are often complex. First, unbalanced experimental design is common in food or environmental surveillance, as it is often difficult to obtain an equal number of samples for all tested groups (e.g. sites, species, age, etc.). The data may contain both quantitative and qualitative variables (e.g. metadata). Non-normal or multimodal data distributions are often encountered among fish contamination levels, environmental parameters (e.g. temperature, pH, turbidity) or biological parameters (e.g., gender, age, lipid contents, biomarkers). Contaminant concentrations in fish can be extremely variable, even within the same study, because the fate of contaminants is multi-factor dependent. As an example, the sum of 25 PCBs in marine benthic fish from the Belgian North Sea and the Western Scheldt Estuary ranged 20-3200 ng g^{-1} ww (Voorspoels et al., 2004). Finally, missing values (e.g. non acquired data or non-detected value) are very common, especially for emerging contaminants.

The selection of an appropriate MAM starts with the clear formulation of the expected scientific outcome. Table 2 provides some clear examples of applications for each tool. Still, more systematic guidelines are needed for the selection and the parametrization of MAMs for specific food safety and environmental management applications. To achieve standardization in the field, software, scripts, and parameters should be first more systematically reported in the literature. The comparison of various tools should also be more frequently tested to explore the potential advantages and bias of different methods. In the end, and as noted by Gibert et al. (2018), statistical software could provide a greater intelligent assistance to support the selection or the parametrization of data analysis steps, which is currently uncommon.

Finally, the impact of data processing, data fusion and filtering on the output of data analysis is still poorly understood. Hohrenk et al. (2020) recently compared the list of molecular features obtained from four data processing tools applied to the same initial raw data set (river water samples). Only about 10% overlap were observed among the features between all four programs, and between 40-55% of features for each software did not match with any other program. Tian et al. (2019) also described the influence of data processing on the detection and identification of model contaminants in pike muscle tissues using NTA, and parameters related to peak height showed a significant influence on the number of model compound identified. As concluded by Fischer et al. (2021) in a recent review on data processing, poor or unreliable results can be obtained if data processing parameters are not optimized for the dataset/application. Similarly, different strategies have been developed for the fusion of data from different instruments. The type of data fusion is known to impact data analysis in the field of metabolomics (Hendriks et al., 2011). Finally, as described in section 4, data filtering influences the data input for analysis.

Based on the above considerations, several but non-exhaustive recommendations can be made when selecting and applying MAM to study chemical mixtures:

- ***Check the compatibility between the type of variables of the data set*** (categorical, discrete, continuous) and the statistical principles on which MAM are based.
- ***Assess the normality of the data distribution***. Skewed data distributions are common, and 100-base normalization or log-transformation may be applied where necessary (Morris et al., 2019). When data normality cannot be verified, non-parametric methods should be selected rather than parametric ones (Mas et al., 2010).
- ***Check the comparability of data***. The interpretation of MAM results has to consider possible bias obtained from heterogeneous datasets (i.e., including both single and average values).
- ***Describe the approach for missing values***. Multivariate methods rely on a sample covariance matrix of which estimators require complete data vectors on all subjects (Pesonen et al., 2015) and this requirement is often not met in context of contaminant monitoring as some chemicals may be present at too low levels in fish to be detected (< LOD). The question of non-detected data is key as it will also impact any reported means of the concentrations and standard deviations (Pesonen et al., 2015). While the general consensus is that statistical methods (e.g., maximum likelihood estimation (MLE), non-parametric Kaplan-Meier method, regression order statistics (ROS)

approaches (Helsel, 2012) cause less bias than common and/or recommended substitution methods (typically “zero”, LOD, half of LOD, upper, lower and middle bound) (EFSA, 2010; Arcella and Gómez Ruiz., 2018), none of them has been selected as the most suitable approach. Conclusions may vary according to the dataset, and the degree of censoring can have a large effect (EFSA, 2010; Helsel, 2010; Leith et al., 2010).

- Similarly to what is commonly done for sample preparation and instrument analysis, ***assess the impact of data processing, data fusion and filtering steps and report experimental conditions*** (algorithms, scripts, parameters). Although standards are still lacking in the field, current best practices consist in testing the impact of data processing using procedural blanks, pooled samples and pooled QC samples, reference samples, replicates, or spiked samples (Gika et al., 2014). Tian et al. (2019) for example optimize the selection of the data processing parameters using spiked model contaminants in fish tissues.

5 Conclusion

Progresses in the analytical characterization of environmental contamination has resulted in the production of large datasets and consequently to the development of efficient data analysis strategies favored by machine learning advances. Chemical or statistical filtering of NTA datasets are effective, almost fundamental, strategies for identifying new chemicals in complex matrices, while keeping the number of false-positives and –negatives low. MAMs are an essential tool for describing and interpreting big data sets to extract unique insights on chemical mixtures in fish. These strategies can also be advantageously coupled with biological approaches, such as EDA, to characterize the effects associated with the exposure to chemical pollutants, in particular by considering the effects of mixtures (Houtman, 2004, Suzuki, 2011, Simon, 2015). Knowledge on sample or compound discriminations, as well as the identification of factors that may influence the environmental behavior or the toxic potential of chemicals, are essential for risk assessment and the implementation of preventive or remedial measures. However, to date, the application of these tools is still limited, particularly for biological matrices. Addressing the knowledge gaps summarized in this paper may influence a more widespread implementation of data analysis strategies to interpret contaminant mixtures in food and environmental matrices.

6 Acknowledgements

Caroline Simonnet-Laprade has received funding from the French Région Pays de la Loire through the "Recherche-Formation-Innovation: Cap-Aliment Food 4 Tomorrow (RFI-Food4.2)" program (Grant FISHCONTAM).

7 REFERENCES

Abdel Malak, I., Cariou, R., Vénisseau, A., Dervilly-Pinel, G., Jaber, F., Babut, M., Le Bizec, B., 2018. Occurrence of Dechlorane Plus and related compounds in catfish (Silurus spp.) from rivers in France. Chemosphere 207:413-420. http://dx.doi.org/10.1016/j.chemosphere.2018.05.101

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2, 433–459.

Agarwal, S., Lanckriet, G., Wills, J., Kriegman, D., Cayton, L., Belongie, S., 2007. Generalized non-metric multidimensional scaling. J. Mach. Learn. Res.

Albergamo V., Schollée JE., Schymanski EL., Helmus R., Timmer H., Hollender J,, de Voogt P. Nontarget Screening Reveals Time Trends of Polar Micropollutants in a Riverbank Filtration System. Environmental Science & Technology 2019 53 (13), 7584-7594

Altenburger, R., Brack, W., Burgess, R.M., Busch, W., Escher, B.I., Focks, A., Hewitt, L.M., Jacobsen, B.N., López de Alda, M., Ait-Aissa, S., Backhaus, T., Ginebreda, A., Hilscherová, K., Hollender, J., Neale, P.A., Schulze, T., Schymanski, E.L., Teodorovic, I., Tindall, A.J., de Aragao Umbuzeiro, G., Vrana, B., Zonja, B., Krauss, M., 2019. Future water quality monitoring: improving the balance between exposure and toxicity assessments of real-world pollutant mixtures. Environ. Sci. Eur. 31:12.

Anderson, M.J., 2017. Permutational multivariate analysis of variance (PERMANOVA). Wiley StatsRef Stat. Ref. Online 1–15. https://doi.org/10.1002/9781118445112.stat07841

Anderson, M.J., 2001. A new method for non parametric multivariate analysis of variance. Austral Ecol. 26, 32–46.

Arcella, D., Gómez Ruiz, J.A., 2018. Use of cut-off values on the limits of quantification reported in datasets used to estimate dietary exposure to chemical contaminants. EFSA Support. Publ. 15. https://doi.org/10.2903/sp.efsa.2018.en-1452

Baesu, A., Ballash, G., Mollenkopf, D., Mažeika, P., Sullivan, S., Wittum, T., Bayen, S. 2021. Suspect screening of pharmaceuticals in fish livers based on QuEChERS extraction coupled with high resolution mass spectrometry. Science of the Total Environment. Accepted for publication.

Ballin, N.Z., Laursen, K.H., 2018. To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. Trends Food Sci. Technol. https://doi.org/10.1016/j.tifs.2018.09.025

Bayen, S., Koroleva, E., Lee, H.K., Obbard, J.P., 2005. Persistent organic pollutants and heavy metals in typical seafoods consumed in Singapore. J. Toxicol. Environ. Heal. - Part A 68, 151–166. https://doi.org/10.1080/15287390590890437

Baygi, S.F., Crimmins, B.S., Hopke, P.K., Holsen, T.M., 2016. Comprehensive Emerging Chemical Discovery: Novel Polyfluorinated Compounds in Lake Michigan Trout. Environ. Sci. Technol. 50, 9460–9468.

Bellavia, A., James-Todd, T., Williams, P.L., 2019. Approaches for incorporating environmental mixtures as mediators in mediation analysis. Environ. Int. 123, 368–374.

Cariou, R., Omer, E., Léon, A., Dervilly-Pinel, G., Le Bizec, B., 2016. Screening halogenated environmental contaminants in biota based on isotopic pattern and mass defect provided by high resolution mass spectrometry profiling. Anal. Chim. Acta 936, 130–138. https://doi.org/10.1016/j.aca.2016.06.053

Couderc, M., Poirier, L., Zlouk-Vergnoux, A., Kamari, A., Blanchet-Letrouvé, I., Marchand, P., Vénisseau, A., Veyrand, B., Mouneyrac, C., Le Bizec, B., 2015. Occurence of POPs and other persistent organic contaminants in the European eel (Anguilla anguilla) from the Loire estuary, France. Sci. Total Environ. 505, 199–215.

Cullen, J.A., Marshall, C.D., Hala, D., 2019. Integration ofmulti-tissue PAH and PCB burdens with biomarker activity in three coastal shark species from the northwestern Gulf of Mexico. Sci. Total Environ. 650, 1158–1172.

Debska, B., Guzowska-Swider, B., 2011. Decision trees in selection of featured determined food quality. Anal. Chim. Acta 705, 261–271.

Dinh, Q.T., Munoz, G., Vo Duy, S., Tien Do, D., Bayen, S., Sauvé, S., 2020. Analysis of sulfonamides, fluoroquinolones, tetracyclines, triphenylmethane dyes and other veterinary drug residues in cultured and wild seafood sold in Montreal, Canada. J. Food Compos. Anal. 94. https://doi.org/10.1016/j.jfca.2020.103630

Dom, I., Biré, R., Hort, V., Lavison-Bompard, G., Nicolas, M., Guérin, T., 2018. Extended Targeted and Non-Targeted Strategies for the Analysis of Marine Toxins in Mussels and Oysters by (LC-HRMS). Toxins (Basel). https://doi.org/10.3390/toxins10090375

Dórea, J.G., 2008. Persistent, bioaccumulative and toxic substances in fish: Human health considerations. Sci. Total Environ. 400, 93–114. https://doi.org/10.1016/j.scitotenv.2008.06.017

Du, B., Lofton, J.M., Peter, K.T., Gipe, A.D., James, C.A., McIntyre, J.K., Scholz, N.L., Baker, J.E., Kolodziej, E.P., 2017. Development of Suspect and Non-Target Screening Methods for Detection of Organic Contaminants in Highway Runoff and Fish Tissue with High-Resolution Time-of-Flight Mass Spectrometry Bowen. Environ. Sci. Process. Impacts 1185–1196.

Dürig, W., Tröger, R., Andersson, P.L., Rybacka, A., Fischer, S., Wiberg, K., Ahrens, L., 2019. Development of a suspect screening prioritization tool for organic compounds in water and biota. Chemosphere 222, 904–912. https://doi.org/10.1016/j.chemosphere.2019.02.021

EFSA, 2010. Management of left-censored data in dietary exposure assessment of chemical substances. EFSA J. 8, 1–96. https://doi.org/10.2903/j.efsa.2010.1557

EFSA, 2019. Guidance on harmonized methodologies for human health, animal health and ecological risk assessment of combined exposure to multiple chemicals. EFSA J. 17, https://doi.org/10.2903/j.efsa.2019.5634.

Ek, C., Faxneld, S., Nyberg, E., Rolff, C., Karlson, AML. 2021. The importance of adjusting contaminant concentrations using environmental data: A retrospective study of 25 years data in Baltic blue mussels. Science of The Total Environment, 762. https://doi.org/10.1016/j.scitotenv.2020.143913.

European Chemicals Agency (ECHA), 2021, https://echa.europa.eu/information-on-chemicals

Fair, P.A., White, N.D., Wolf, B., Arnott, S.A., Kannan, K., Karthikraj, R., Vena, J.E., 2018. Persistent organic pollutants in fish from Charleston Harbor and tributaries, South Carolina, United States: A risk assessment. Environ. Res. 167, 598–613. https://doi.org/10.1016/j.envres.2018.08.001

Fair, P.A., Wolf, B., White, N.D., Arnott, S.A., Kannan, K., Karthikraj, R., Vena, J.E., 2019. Perfluoroalkyl substances (PFASs) in edible fish species from Charlestion Harbor and tributaries, South Caroline, United States: Exposure and risk assessment. Environ. Res. 171, 266–277.

Felipe-Sotelo, M., Tauler, R., Vives, I., Grimalt, J.O., 2008. Assessment of the environmental and physiological processes determining the accumulation of organochlorine compounds in European mountain lake fish through multivariate analysis (PCA and PLS). Sci. Total Environ. 404, 148–161.

Fernando, S., Renaguli, A., Milligan, M.S., Pagano, J.J., Hopke, P.K., Holsen, T.M., Crimmins, B.S., 2018. Comprehensive analysis of the Great Lakes top predator fish for novel halogenated organic contaminants by GCxGC-HR-ToF mass spectrometry. Environ. Sci. Technol. 52, 2909–2917.

Fisher, C.M., Croley, T.R., Knolhoff, A.M. 2021. Data Processing Strategies for Non-Targeted Analysis of Foods using Liquid Chromatography/High-Resolution Mass Spectrometry. TrAC Trends in Analytical Chemistry, 116188

Fu, L., Lu, X., Tan, J., Zhang, H., Zhang, Y., Wang, S., Chen, J., 2018. Bioaccumulation and human health risks of OCPs and PCBs in freshwater products of Northeast China. Environ. Pollut. 242, 1527–1534.

Fu, Y., Zhao, C., Lu, X., Xu, G., 2017. Nontargeted screening of chemical contaminants and illegal additives in food based on liquid chromatography–high resolution mass spectrometry. TrAC - Trends Anal. Chem. 96, 89–98. https://doi.org/10.1016/j.trac.2017.07.014

Fu, Y., Zhou, Z., Kong, H., Lu, X., Zhao, X., Chen, Y., Chen, J., Wu, Z., Xu, Z., Zhao, C., Xu, G., 2016. Non-targeted Screening Method for Illegal Additives based on Ultra High Performance Liquid Chromatography-High Resolution Mass Spectrometry. Anal. Chem. 88, 8870–8877.

Garcıa-Reyes, J.F., Molina-Diaz, A., Fernandez-Alba, A.R., 2007. Identification of Pesticide Transformation Products in Food by Liquid Chromatography/ Time-of-Flight Mass Spectrometry via “Fragmentation-Degradation” Relationships Juan. Anal. Chem. 79, 307–321.

Gaus, C., Villa, C.A., Dogruer, G., Heffernan, A., Vijayasarathy, S., Lin, C.-Y., Flint, M., Hof, C.M., Bell, I., 2019. Evaluating internal exposure of sea turtles as model species for identifying regional chemical threats in nearshore habitats of the Great Barrier Reef. Scien 658, 732–743.

Gerig, B., Chaloner, D., Janetski, D.J., Rediske, R.R., O'Keefe, J.P., Moerke, A., Lamberti, G., 2015. Congener patterns of persistent organic pollutants establish the extent of contaminant biotransport by Pacific salmon in the Great Lakes. Environ. Sci. Technol. 50, 554–563. https://doi.org/10.1021/acs.est.5b05091

Gibert, K., Izquierdo, J., Sànchez-marrè, M., Hamilton, S.H., Rodríguez-roda, I., Holmes, G., 2018. Which method to use? An assessment of data mining methods in Environmental Data Science. Environ. Model. Softw. 1–25. https://doi.org/S1364815218308715

Gika, H.G., Theodoridis, G.A., Plumb, R.S., Wilson, I.D., 2014. Current practice of liquid chromatography-mass spectrometry in metabolomics and metabonomics. J. Pharm. Biomed. Anal. 87, 12–25. https://doi.org/10.1016/j.jpba.2013.06.032

Grieshaber, C.A., Penland, T.N., Kwak, T.J., Cope, W.G., Heise, R.J., Law, J. Mac, Shea, D., Aday, D.D., Rice, J.A., Kullman, S.W., 2018. Relation of contaminants to fish intersex in riverine sport fishes. Sci. Total Environ. 643, 73–89.

Grisoni, F., Consonni, V., Vighi, M., 2018. Detecting the bioaccumulation patterns of chemicals through data- driven approaches. Chemosphere 208, 273–284.

Guijas, C., Montenegro-Burke, JR., Domingo-Almenara, X., Palermo, A., Benedikt Warth, Hermann, G., Koellensperger, G., Huan, T., Uritboonthai, W., Aisporna, A.E., Wolan, D., Spilker, ME., Benton, HP., Siuzdak, G. 2018. METLIN: A Technology Platform for

Identifying Knowns and Unknowns. Analytical Chemistry 90 (5), 3156-3164. DOI: 10.1021/acs.analchem.7b04424

Halloum, W., Cariou, R., Dervilly-Pinel, G., Jaber, F., Le Bizec, B., 2017. A comparative study of different ionization modes (EI, CI, APCI) using GC-MS/MS for the identification of 18 organophosphorus flame retardants and plasticizers. Journal of Mass Spectrometry 52:54-61. http://dx.doi.org/10.1002/jms.3899.

Heffernan, A.L., Gómez-Ramos, M.M., Gaus, C., Vijayasarathy, S., Bell, I., Hof, C., Mueller, J.F., Gómez-Ramos, M.J., 2017. Non-targeted, high resolution mass spectrometry strategy for simultaneous monitoring of xenobiotics and endogenous compounds in green sea turtles on the Great Barrier Reef. Sci. Total Environ. 599–600, 1251–1262. https://doi.org/10.1016/j.scitotenv.2017.05.016

Helsel, D., 2010. Much ado about next to nothing: Incorporating nondetects in science. Ann. Occup. Hyg. 54, 257–262. https://doi.org/10.1093/annhyg/mep092

Helsel, D.R., 2012. Statistics for Censored Environmental Data Using Minitab and R, (2nd ed.),. ed.

Hendriks, M.M.W.B., Eeuwijk, F.A. va., Jellema, R.H., Westerhuis, J.A., Reijmers, T.H., Hoefsloot, H.C.J., Smilde, A.K., 2011. Data-processing strategies for metabolomics studies. TrAC - Trends Anal. Chem. 30, 1685–1698. https://doi.org/10.1016/j.trac.2011.04.019

Henson, R.N., 2015. Analysis of Variance (ANOVA). Brain Mapp. An Encycl. Ref. 477–481.

Hernández, A.F., Tsatsakis, A.M., 2017. Human exposure to chemical mixtures: Challenges for the integration of toxicology with epidemiology data in risk assessment. Food Chem. Toxicol. 103, 188–193. https://doi.org/10.1016/j.fct.2017.03.012

Hernández-Mesa, M., Escourrou, A., Monteau, F., Le Bizec, B., Dervilly-Pinel, G., 2017. Current applications and perspectives of ion mobility spectrometry to answer chemical food safety issues. TrAC Trends in Analytical Chemistry 94, 39-53. https://doi.org/10.1016/j.trac.2017.07.006.

Hernández-Mesa, M., Le Bizec, B., Dervilly, G., 2021. Metabolomics in chemical risk analysis – A review, Analytica Chimica Acta, 1154, https://doi.org/10.1016/j.aca.2021.338298.

.

Hohrenk, L.L., Itzel, F., Baetz, N., Tuerk, J., Vosough, M., Schmidt, T.C., 2020. Comparison of Software Tools for Liquid Chromatography-High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. Anal. Chem. 92, 1898–1907. https://doi.org/10.1021/acs.analchem.9b04095

Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ. Sci. Technol. 51, 11505–11512. https://doi.org/10.1021/acs.est.7b02184

Hollender, J., van Bavel, B., Dulio, V., Farmen, E., Furtmann, K., Koschorreck, J., Kunkel, U., Krauss, M., Munthe, J., Schlabach, M., Slobodnik, J., Stroomberg, G., Ternes, T., Thomaidis, N.S., Togola, A., Tornero, V., 2019. High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management. Environ. Sci. Eur. 31. https://doi.org/10.1186/s12302-019-0225-x

Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 45(7):703–14. doi:10.1002/jms.1777.

Houtman, C.J., Van Osstveen, A.M., Brouwer, A., Lamoree, M.H., Legler, J., 2004. Identification of estrogenic compounds in fish bile using bioassay-directed fractionation. Environ. Sci. Pollut. Res. 38, 6415–6423.

Hrydziuszko, O., Viant, M.R., 2012. Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. Metabolomics. https://doi.org/10.1007/s11306-011-0366-4

Ingelbleek, L, Lautz, L, Dervilly, G, Darney, K, Astuto, MC, Tarazona, J, Liem, D, Kass, GEN, Leblanc, JC, Verger, P, Le Bizec ,B, Dorne, JLCM. 2021. Risk assessment of chemicals in food and feed: principles, applications and future perspectives, in Environmental Pollutant Exposures and Public Health. From Book Series: Issues in Environmental Science and Technology, 50, 1-38http://dx.doi.org/10.1039/9781839160431-00001, eISBN: 978-1-83916-043-1.

Jobst, K.J., Shen, L., Reiner, E.J., Taguchi, V.Y., Helm, P.A., McCrindle, R., Backus, S., 2013. The use of mass defect plots for the identification of (novel) halogenated contaminants in the environment. Anal. Bioanal. Chem. 405, 3289–3297. https://doi.org/10.1007/s00216-013-6735-2

Johnson, R.A., Wichern, D.W., 2002. Applied Multivariate Statistical Analysis, Prentice H. ed.

Kantiani, L., Llorca, M., Sanchís, J., Farré, M., Barceló, D., 2010. Emerging food contaminants: A review. Anal. Bioanal. Chem. 398, 2413–2427. https://doi.org/10.1007/s00216-010-3944-9

Kaufmann, A., 2012. The current role of high-resolution mass spectrometry in food analysis. Anal. Bioanal. Chem. https://doi.org/10.1007/s00216-011-5629-4

Kelly, B.C., Myo, A.N., Pi, N., Bayen, S., Leakhena, P.C., Chou, M., Tan, B.H., 2018. Human exposure to trace elements in central Cambodia: Influence of seasonal hydrology and food-chain bioaccumulation behaviour. Ecotoxicol. Environ. Saf. 162, 112–120. https://doi.org/10.1016/j.ecoenv.2018.06.071

Kendrick, E., 1963. A Mass Scale Based on CH2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds. Anal. Chem. 35, 2146–2154. https://doi.org/10.1021/ac60206a048

Kohonen, T., 2013. Essentials of the self-organizing map. Neural Networks 37, 52–65.

Krauss, M., Singer, H., Hollender, J., 2010. LC-high resolution MS in environmental analysis: From target screening to the identification of unknowns. Anal. Bioanal. Chem. 397, 943–951. https://doi.org/10.1007/s00216-010-3608-9

Leith, K.F., Bowerman, W.W., Wierda, M.R., Best, D.A., Grubb, T.G., Sikarske, J.G., 2010. A comparison of techniques for assessing central tendency in left-censored data using PCB and p,p′DDE contaminant concentrations from Michigan’s Bald Eagle Biosentinel Program. Chemosphere 80, 7–12. https://doi.org/10.1016/j.chemosphere.2010.03.056

Léon, A., Cariou, R., Hutinet, S., Hurel, J., Guitton, Y., Tixier, C., Munschy, C., Antignac, J-P., Dervilly-Pinel, G., Le Bizec B., 2019. HaloSeeker 1.0, a user-friendly software to highlight halogenated chemicals in non-targeted high resolution mass spectrometry dataset. Analytical Chemistry 91:3500–3507. https://www.ncbi.nlm.nih.gov/pubmed/30758179.

Liu, Y., D’Agostino, L.A., Qu, G., Jiang, G., Martin, J.W., 2019. High-resolution mass spectrometry (HRMS) methods for nontarget discovery and characterization of poly- and per-fluoroalkyl substances (PFASs) in environmental and human samples. Trends Anal. Chem. 121. https://doi.org/10.1016/j.trac.2019.02.021

Lorenzo, M., Campo, J., Picó, Y., 2018. Analytical challenges to determine emerging persistent organic pollutants in aquatic ecosystems. TrAC - Trends Anal. Chem. 103, 137–155. https://doi.org/10.1016/j.trac.2018.04.003

Loos, M., Singer, H. 2017. Nontargeted homologue series extraction from hyphenated high resolution mass spectrometry data, Journal of Cheminformatics 9(1), 12.

Mas, S., Juan, A. de, Tauler, R., Olivieri, A.C., Escandar, G.M., 2010. Application of chemometric methods to environmental analysis of organic pollutants: A review. Talanta 80, 1052–1067.

McGoldrick, D.J., Clark, M.G., Keir, M.J., Backus, S.M., Malecki, M.M., 2010. Canada's national aquatic biological specimen bank and database. J. Great Lakes Res. 36, 393–398. https://doi.org/10.1016/j.jglr.2010.02.011

Miller, A., Nyberg, E., Danielsson, S., Faxneld, S., Haglund, P., Bignert, A. 2014. Comparing temporal trends of organochlorines in guillemot eggs and Baltic herring: advantages and disadvantage for selecting sentinel species for environmental monitoring. Mar. Environ. Res., 100, 38-47.

Misra, R.K., Uthe, J.F., Chou, C.L., Scott, D.P., Musial, C.J., 1993. Trend analysis using a multivariate procedure for data with unequal residual covariance and regression coefficient matrices: Application to Canadian Atlantic cod contaminant data. Mar. Pollut. Bull. 26, 73–77. https://doi.org/10.1016/0025-326X(93)90094-Z

Morris, A.D., Letcher, R.J., Dyck, M., Chandramouli, B., Cosgrove, J., 2019. Concentrations of legacy and new contaminants are related to metabolite profiles in Hudson Bay polar bears. Environ. Res. 168, 364–374. https://doi.org/10.1016/j.envres.2018.10.001

Mullin, L., Jobst, K., DiLorenzo, R.A., Plumb, R., Reiner, E.J., Yeung, L.W.Y., Jogsten, I.E., 2020. Liquid chromatography-ion mobility-high resolution mass spectrometry for analysis of pollutants in indoor dust: Identification and predictive capabilities. Anal. Chim. Acta 1125, 29–40. https://doi.org/10.1016/j.aca.2020.05.052

Munaretto, J.S., May, M.M., Saibt, N., Zanella, R., 2016. Liquid chromatography with high resolution mass spectrometry for identification of organic contaminants in fish fillet: screening and quantification assessment using two scan modes for data acquisition. J. Chromatogr. A. https://doi.org/10.1016/j.chroma.2016.06.018

Musatadi, M., González-Gaya, B., Irazola, M., Prieto, A., Etxebarria, N., Olivares, M., Zuloaga, O., 2020. Focused ultrasound-based extraction for target analysis and suspect screening of organic xenobiotics in fish muscle. Sci. Total Environ. 740, 139894. https://doi.org/10.1016/j.scitotenv.2020.139894

Myers, A.L., Jobst, K.J., Mabury, S.A., Reiner, E.J., 2014a. Using mass defect plots as a discovery tool to identify novel fluoropolymer thermal decomposition products. J. Mass Spectrom. https://doi.org/10.1002/jms.3340

Myers, A.L., Watson-Leung, T., Jobst, K.J., Shen, L., Besevic, S., Organtini, K., Dorman, F.L., Mabury, S.A., Reiner, E.J., 2014b. Complementary nontargeted and targeted mass spectrometry techniques to determine bioaccumulation of halogenated contaminants in freshwater species. Environ. Sci. Technol. 48, 13844–13854. https://doi.org/10.1021/es503090s

Nyberg, E., Faxneld, S., Danielsson, S., Eriksson, U., Miller, A., Bignert, A. 2015. Temporal and spatial trends of PCBs, DDTs, HCHs, and HCB in Swedish marine biota 1969–2012.

Ambio, 44 (3), 484-497.

Olivieri, A.C., 2008. Analytical advantages of multivariate data processing. One, two, three, infinity? Anal. Chem. https://doi.org/10.1021/ac800692c

Ondarza, P.M., Gonzalez, M., Fillmann, G., Miglioranza, K.S.B., 2014. PBDEs, PCBs and organochlorine pesticides distribution in edible fish from Negro River basin, Argentinean Patagonia. Chemosphere 94, 135–142.

Peré-Trepat, E., Olivella, L., Ginebreda, A., Caixach, J., Tauler, R., 2006. Chemometrics modelling of organic contaminants in fish and sediment river samples. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2006.04.005

Pérez, F., Llorca, M., Köck-Schulmeyer, M., Škrbić, B., Silva, L.F.O., da Boit Martinello, K., Al-Dhabi, N.A., Antić, I., Farré, M., Barceló, D., 2014. Assessment of perfluoroalkyl substances in food items at global scale. Environ. Res. 135, 181–189. https://doi.org/10.1016/j.envres.2014.08.004

Pesonen, M., Pesonen, H., Nevalainen, J., 2015. Covariance matrix estimation for left-censored data. Comput. Stat. Data Anal. 92, 13–25. https://doi.org/10.1016/j.csda.2015.06.005

Peters S, Janssen HG, Vivó-Truyols G. 2010. Trend analysis of time-series data: A novel method for untargeted metabolite discovery. Anal Chim Acta. 17;663(1):98-104. doi: 10.1016/j.aca.2010.01.038.

Plassmann, M.M., Fischer, S., Benskin, J.P., 2018. Non-target time trend screening in human blood. Environ. Sci. Technol. Lett. 5, 335–340.

Plassmann, M.M., Tengstrand, E., Åberg, K.M., Benskin, J.P., 2016. Non-target time trend screening: a data reduction strategy for detecting emerging contaminants in biological samples. Anal. Bioanal. Chem. 408, 4203–4208. https://doi.org/10.1007/s00216-016-9563-3

Pose-Juan, E., Fernández-Cruz, T., Simal-Gándara, J., 2016. State of the art on public risk assessment of combined human exposure to multiple chemical contaminants. Trends Food Sci. Technol. 55, 11–28. https://doi.org/10.1016/j.tifs.2016.06.011

Pourchet, M., Debrauwer, L., Klanova, J., Price, E.J., Covaci, A., Caballero-Casero, N., Oberacher, H., Lamoree, M., Damont, A., Fenaille, F., Vlaanderen, J., Meijer, J., Krauss, M., Sarigiannis, D., Barouki, R., Le Bizec, B., Antignac, J.P., 2020. Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues. Environ. Int. 139, 105545. https://doi.org/10.1016/j.envint.2020.105545

Ribbenstedt, A., Ziarrusta, H., Benskin, J.P., 2018. Development, characterization and comparisons of targeted and non-targeted metabolomics methods. PLoS One. https://doi.org/10.1371/journal.pone.0207082

Rodríguez-Hernández, Á., Camacho, M., Henríquez-Hernández, L.A., Boada, L.D., Ruiz-Suárez, N., Valerón, P.F., Almeida González, M., Zaccaroni, A., Zumbado, M., Luzardo, O.P., 2016. Assessment of human health hazards associated with the dietary exposure to organic and inorganic contaminants through the consumption of fishery products in Spain. Sci. Total Environ. 557–558, 808–818. https://doi.org/10.1016/j.scitotenv.2016.03.035

Romanić, S.H., Vuković, G., Klinčić, D., Sarić, M.M., Župan, I., Antanasijević, D., Popović, A., 2018. Organochlorine pesticides (OCPs) and polychlorinated biphenyls (PCBs) in Cyprinidae fish: Towards hints of their arrangements using advanced classification methods. Environ. Res. 165, 349–357.

Sanchis, Y., Yusà, V., Coscollà, C., 2017. Analytical strategies for organic food packaging contaminants. J. Chromatogr. A 1490, 22–46. https://doi.org/10.1016/j.chroma.2017.01.076

Scheringer, M., Strempel, S., Hukari, S., Ng, C.A., Blepp, M., Hungerbuhler, K., 2012. How many persistent organic pollutants should we expect? Atmos. Pollut. Res. 3, 383–391. https://doi.org/10.5094/APR.2012.044

Schipper, C.A., Lahr, J., van den Brink, P.J., George, S.G., Hansen, P.-D., da Silva de Assis, H.C., van der Oost, R., Thain, J.E., Livingstone, D., Mitchelmore, C., van Schooten, F.-J., Ariese, F., J. Murk, A., Grinwis, G.C.M., Klamer, H., Kater, B.J., Postma, J.F., van der Werf, B., Vethaak, A.D., 2009. A retrospective analysis to explore the applicability of fish biomarkers and sediment bioassays along contaminated salinity transects. ICES J. Mar. Sci. 66, 2089–2105.

Schulze, B., Jeon, Y., Kaserzon, S., Heffernan, AL., Dewapriya, P., O'Brien, J., Gomez Ramos, MJ., Gorji, SG., Mueller, JF., Thomas, K.V., Saer Samanipour, S. 2020. An assessment of quality assurance/quality control efforts in high resolution mass spectrometry non-target workflows for analysis of environmental samples. TrAC Trends in Analytical Chemistry,133. 2020, https://doi.org/10.1016/j.trac.2020.116063.

Simon, E., Lamoree, M.H., Hamers, T., de Boer, J., 2015. Challenges in effect-directed analysis with a focus on biological samples. TrAC - Trends Anal. Chem. 67, 179–191. https://doi.org/10.1016/j.trac.2015.01.006

Sleno, L., 2012. The use of mass defect in modern mass spectrometry. J. Mass Spectromery 47, 226–236.

Smolinska, A., Hauschild, A.C., Fijten, R.R.R., Dallinga, J.W., Baumbach, J., Van Schooten, F.J., 2014. Current breathomics - A review on data pre-processing techniques and machine learning in metabolomics breath analysis. J. Breath Res. 8. https://doi.org/10.1088/1752-7155/8/2/027105

Sobus, J.R., Wambaugh, J.F., Isaacs, K.K., Williams, A.J., McEachran, A.D., Richard, A.M., Grulke, C.M., Ulric, E.M., Rager, J.E., Strynar, M.J., Newton, S.R., 2018. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. J. Expo. Sci. Environ. Epidemiol. 28, 411–426.

Sedeño-Díaz , J., López-López E., 2012. Fish as sentinel organisms: from the molecular to the population level, a review.In book: New Advances and Contributions to Fish BiologyChapter: Fish as sentinel organisms: from the molecular to the population level, a reviewPublisher: Intech OpenEditors: Hakan Turker. DOI: 10.5772/54825

Stalling, D.L., Norstrom, R.J., Smith, L.M., Simon, M., 1985. Patterns of PCDD, PCDF, and PCB contamination in Great Lakes fish and birds and their characterization by principal component analysis. Chemosphere 14, 627–643. https://doi.org/10.1111/ane.12608

Suzuki, G., Tue, N.M., Van Der Linden, S., Brouwer, A., Van Der Burg, B., Van Velzen, M., Lamoree, M., Someya, M., Takahashi, S., Isobe, T., Tajima, Y., Yamada, T.K., Takigami, H., Tanabe, S., 2011. Identification of major dioxin-like compounds and androgen receptor antagonist in acid-treated tissue extracts of high trophic-level animals. Environ. Sci. Technol. 45, 10203–10211. https://doi.org/10.1021/es2024274

Tian, L., Verreault, J., Houde, M., Bayen, S., 2019. Suspect screening of plastic-related chemicals in northern pike (Esox lucius) from the St. Lawrence River, Canada. Environ. Pollut. 255.

Tian, L., Zheng, J., Goodyer, C.G., Bayen, S., 2020. Non-targeted screening of plastic-related chemicals in food collected in Montreal, Canada. Food Chem. 326, 126942. https://doi.org/10.1016/j.foodchem.2020.126942

Törnkvist, A., Glynn, A., Aune, M., Darnerud, P.O., Ankarberg, E.H., 2011. PCDD/F, PCB, PBDE, HBCD and chlorinated pesticides in a Swedish market basket from 2005 - Levels and dietary intake estimations. Chemosphere 83, 193–199. https://doi.org/10.1016/j.chemosphere.2010.12.042

Turnipseed, S.B., Storey, J.M., Wu, I.L., Gieseker, C.M., Hasbrouck, N.R., Crosby, T.C., Andersen, W.C., Lanier, S., Casey, C.R., Burger, R., Madson, M.R., 2018. Application and evaluation of a high-resolution mass spectrometry screening method for veterinary drug residues in incurred fish and imported aquaculture samples. Anal. Bioanal. Chem. 410, 5529–5544. https://doi.org/10.1007/s00216-018-0917-x

UN, 2015. Take Action for the Sustainable Development Goals [WWW Document]. URL https://www.un.org/sustainabledevelopment/sustainable-development-goals/

van der Oost, R., Vindimian, E., van den Brink, P.J., Satumalay, K., Heida, H., Vermeulen, N.P.E., 1997. Biomonitoring aquatic pollution with feral eel (Anguilla anguilla). III. Statistical analyses of relationships between contaminant exposure and biomarkers. Aquat. Toxicol. 39, 45–75.

Van Leeuwen, S.P.J., Leonards, P.E.G., Traag, W.A., Hoogenboom, L.A.P., De Boer, J., 2007. Polychlorinated dibenzo-p-dioxins, dibenzofurans and biphenyls in fish from the Netherlands: Concentrations, profiles and comparison with DR CALUX® bioassay results. Anal. Bioanal. Chem. https://doi.org/10.1007/s00216-007-1352-6

Voorspoels, S., Covaci, A., Maervoet, J., De Meester, I., Schepens, P., 2004. Levels and profiles of PCBs and OCPs in marine benthic species from the Belgian North Sea and the Western Scheldt Estuary. Mar. Pollut. Bull. 49, 393–404.

Vuković, G., Romanić, S.H., Babić, Ž., Mustać, B., Štrbac, M., Deljanin, I., Antanasijević, D., 2018. Persistent organic pollutants (POPs) in edible fish species from different fishing zones of Croatian Adriatic. Mar. Pollut. Bull. 137, 71–80.

Wang, W., Batterman, S., Chernyak, S., Nriagu, J., 2008. Concentrations and risks of organic and metal contaminants in Eurasian caviar. Ecotoxicol. Environ. Saf. https://doi.org/10.1016/j.ecoenv.2007.06.007

Wang, M., Carver, J., Phelan, V. et al. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34, 828–837. https://doi.org/10.1038/nbt.3597

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., Ni, Y., 2018. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Sci. Rep. https://doi.org/10.1038/s41598-017-19120-0

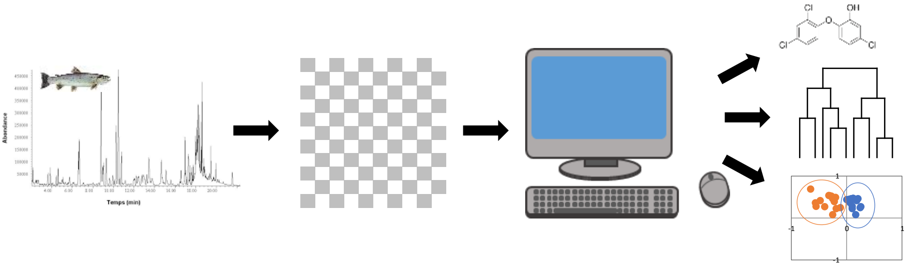
Wenning, R.J., Erickson, G.A., 1994. Interpretation and analysis of complex environmental data using chemometric methods. Trends Anal. Chem. 13, 446–457.

Wu, I.L., Turnipseed, S.B., Storey, J.M., Andersen, W.C., Madson, M.R., 2020. Comparison of data acquisition modes with Orbitrap high-resolution mass spectrometry for targeted and non-targeted residue screening in aquacultured eel. Rapid Commun. Mass Spectrom. 34, 1–15. https://doi.org/10.1002/rcm.8642

Yergey, J.A., 1983. A general approach to calculating isotopic distributions for mass spectrometry. Int. J. Mass Spectrom. Ion Phys. 52, 337–349. https://doi.org/10.1016/0020-7381(83)85053-0

Zhang, H., Bayen, S., Kelly, B.C., 2015. Co-extraction and simultaneous determination of multi-class hydrophobic organic contaminants in marine sediments and biota using GC-

EI-MS/MS and LC-ESI-MS/MS. Talanta 143, 7–18. https://doi.org/10.1016/j.talanta.2015.04.084


Abundance
450000
400000
350000
300000
250000
200000
150000
100000
50000
4.00
6.00
8.00
10.00
12.00
14.00
16.00
18.00
20.00
Temps (min)
Cl
OH
O
Cl
Cl
1
-1
0
1
-1