Multivariate Analysis of Large µ-FTIR Data Sets in Search of Microplastics

Wander L.[1,2,*], Vianello A.[3], Vollertsen J.[3], Braun U.[1], Paul A.[1]

[1]Bundesanstalt für Materialforschung und -prüfung, Richard-Willstätter-Straße 11, 12489 Berlin, Germany
[2]Humboldt-Universität zu Berlin, Department of Chemistry, Brook-Taylor-Str. 2, 12489 Berlin, Germany
[3]Aalborg University, Department of Civil Engineering, Thomas Manns Vej 23, 9220 Aalborg, Denmark

*corresponding author: e-mail: lukas.wander@bam.de

Abstract
µ-FTIR spectroscopy is a widely used technique in microplastics research. It allows to simultaneously characterize the material of the small particles, fibers or fragments, and to specify their size distribution and shape. Modern detectors offer the possibility to perform two-dimensional imaging of the sample providing detailed information. However, data sets are often too large for manual evaluation calling for automated microplastic identification. Library search based on the comparison with known reference spectra has been proposed to solve this problem.
To supplement this 'targeted analysis', an exploratory approach was tested. Principal component analysis (PCA) was used to drastically reduce the size of the data set while maintaining the significant information. Groups of similar spectra in the prepared data set were identified with cluster analysis. Members of different clusters could be assigned to different polymer types whereas the variation observed within a cluster gives a hint on the chemical variability of microplastics of the same type. Spectra labeled according to the respective cluster can be used for supervised learning. The obtained classification was tested on an independent data set and results were compared to the spectral library search approach.
Keywords: Microplastics, µ-FTIR, Multivariate Data Analysis

1. Introduction

Numerous studies on the occurrence of microplastics in the environment have been published in recent years. In most cases it is the number of synthetic polymer particles, their chemical identity (i.e. the polymer type), size distribution and morphology which are reported. The analysis focused on individual particles mainly relies on microscopic techniques such as µ-FTIR spectroscopy. Modern FTIR microscopes can be equipped with focal plane array detectors. This grants access to hyperspectral imaging of the sample. Up to several thousand spectra are recorded at a time and large areas can be mapped yielding up to several million spectra in total. The detailed spatio-chemical information available enables a profound investigation of the samples but at the same time renders the analysis more challenging.

Multivariate data analysis offers a set of tools to cope with vast and complex datasets. It can improve the speed of analysis significantly and help to interpret the data. For example, the use of dimensionality reduction techniques such as PCA can aid to visualize structures within the data but have received little attention in the field of microplastics so far.

2. Results

FTIR imaging of marine sediment samples was performed in transmission mode in the spectral range 3752-848 cm^{-1}. Approximately 3.2 million spectra were recorded covering a sample area of about 1x1 cm. Performing µ-FTIR mapping of such large areas, the tradeoff for fast scanning is usually reduced spectral quality. In addition, environmental samples contain highly varying particles. Natural organic residue can be found next to inorganic particles. Particles vary in size and morphology which can have significant impacts on the spectra. Absorbance spectra in this case are influenced by physical effects such as light scattering. Overall, relatively poor spectra with low signal to noise ratio and distorted baselines are obtained.

Mathematical preprocessing of the data prior to multivariate analysis is an essential step to obtain good results. In the example presented in figure 1, techniques such as derivatives, smoothing and normalization were used to prepare the spectra for further analysis. Spectra with a very low signal to noise ratio were completely removed prior to analysis. PCA is a well-known efficient tool to reduce the dimensionality of datasets, in ideal cases separating information and noise. Here, starting from 754 spectral variables (i.e. wavenumbers at which the spectra were recorded), a reduction to 15 latent variables was feasible.

The distribution of spectra can be visualized in this new PCA scores space. A high density of spectra, roughly 95 % of all datapoints, was observed in one region of the plot of the first two principal components (red frame in figure 1A). Spectra located in this area were identified as background originating from areas of the sample slide which are not covered by particles. Spectra with defined absorption features carrying the actual chemical information spread outwards from this region. Depending on the direction in space, different substances can be identified when comparing the spectra to a set of library spectra.

After removing the background data, spectra of particles were treated with a slightly changed preprocessing strategy. In a consecutive approach clustering was then used to extract similar spectra. Substances were assigned to clusters by comparing the spectra to reference spectra published by Primpke et al. (2018) which are commonly encountered in environmental samples (figure 1B)). Four clusters were initially found: 1) containing plant residue, 2) polyamides (including natural polyamides such as proteins) 3) microplastics; this cluster was later subdivided into different polymer types, and 4) spectra which could not clearly be identified based on the limited library. The spatial distribution is shown in figure 1C. Depending on the initial threshold values used to separate spectra from the background, size and number of particles found changed. Good agreement on the entire dataset was observed between the exploratory results and an automated library search approach.

3. Conclusion and Outlook

In the presented exploratory approach, no prior assumptions are used when searching for microplastics. The distribution of datapoints offers the possibility to study variation within substance classes and eventually discover substances potentially overlooked in a targeted analysis. Using cluster analysis, labeled data as input to supervised learning methods can be obtained. Once a classification model has successfully been trained, spectra from new samples can be analyzed very quickly.

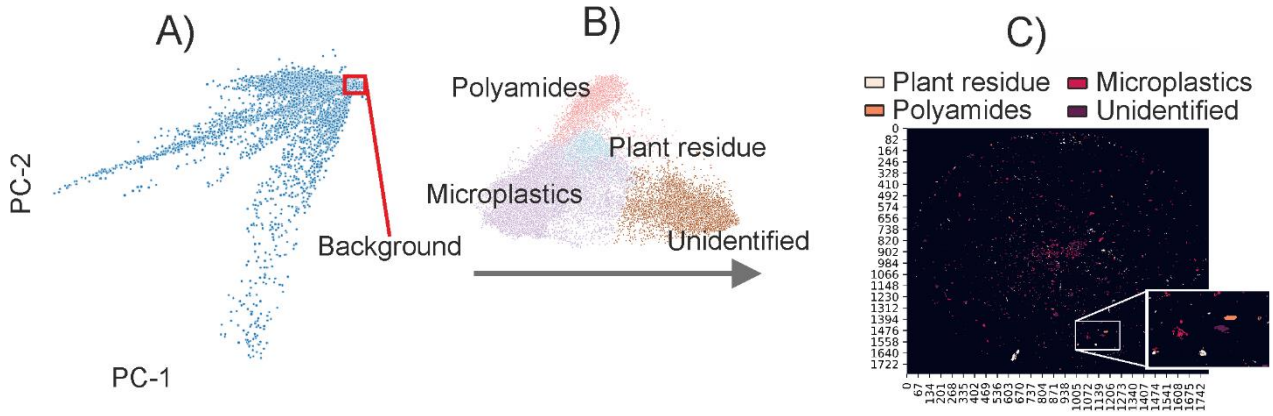

Figure 1. Exploratory analysis of a large µ-FTIR dataset: A) Separation of particle from background spectra by PCA, B) clustering of the remaining data and identification of substances, and C) 2D visualization of the results.

References

Hidalgo-Ruz, V., Gutow L., Thompson R. C. and Thiel M. (2012), Microplastics in the Marine Environment: A Review of the Methods Used for Identification and Quantification, *Environmental Science & Technology*, **46**, 3060-3075.

Hufnagl B., Steiner D., Renner E., Löder M. G. J., Laforsch C. and Lohninger H. (2019), A method for the fast identification and monitoring of microplastics in environmental samples using random decision forest classifiers, *Analytical Methods*, **11**, 2277-2285.

Huppertsberg S. and Knepper T. P. (2018), Instrumental analysis of microplstics – benefits and challenges, *Analytical and Bioanalytical Chemistry*, **410**, 25, 6343–6352.

Primpke S., Wirth M., Lorenz C. and Gerdts G. (2018), Reference database design for the automated analysis of microplastic samples based on Fourier transform infrared (FTIR) spectroscopy, *Analytical and Bioanalytical Chemistry*, **410**, 21, 5131-5141.

Primpke S., Lorenz C., Rascher-Friesenhausen R. and Gerdts G. (2017), An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and image analysis, *Analytical Methods*, **9**, 1499-1511.

Renner G., Nellessen A., Schwiers A., Wenzel M., Schmidt T. C., Schram J. (2018), Data Preprocessing & Evaluation used in the Microplastics Identification Process: A Critical Review & Practical Guide, *Trends in Analytical Chemistry*, **111**, 229-238.